UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas

Estudio de la sostenibilidad medioambiental y social de las empresas europeas

Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

AUTOR/A: Marqués Boigues, Enrique

Tutor/a: Peiró Signes, Ángel

CURSO ACADÉMICO: 2021/2022

A Ángel, por su tiempo y dedicación a lo largo del verano

Abstract

The main objective of the study is to analyze what business characteristics define companies that actively take actions for environmental sustainability (recycling or reusing materials, reducing the consumption of natural resources or the impact on them, saving energy or using sustainable energy sources or develop sustainable products or services) or social (improve the working conditions of your workers, promote and improve diversity and equality in the workplace, assess the impact of your company on society or involve employees in the management of the company). The study will use Eurobarometer 486 on "SMEs, start-ups, scale-ups and entrepreneurship" focusing on the barriers and challenges faced by SMEs in Europe as they grow, transition to more sustainable business models and digitalisation. The survey covers more than 16,000 companies and machine learning techniques will be used to propose a model that differentiates companies that are oriented towards environmental and social sustainability. The study will include the motivated analysis of the variables to be included in the model, the treatment of data for modeling, the evaluation of different machine learning techniques for the selection of the most appropriate technique, the adjustment and optimization of the model and its interpretation. Based on the results, the most relevant characteristics, their relative importance and the direction of their impact will be extracted. The differences between the groups will also be compared using statistical techniques. The results and conclusions of this study will be of interest for the orientation of the public financing policies proposed by public bodies and for the managers of Spanish companies that intend to move towards a more sustainable model.

Resumen

El objetivo principal del trabajo es analizar qué características empresariales definen a las empresas que toman acciones de forma activa para la sostenibilidad medioambiental (reciclar o reutilizar materiales, reducir el consumo de recursos naturales o el impacto sobre ellos, ahorrar energía o utilizar fuentes de energía sostenibles o desarrollar productos o servicios sostenibles) o social (mejorar las condiciones de trabajo de sus trabajadores, promover y mejorar la diversidad y la igualdad en el lugar de trabajo, evaluar el impacto de su empresa en la sociedad o implicar a los empleados en la gestión de la empresa). Para el estudio se utilizará el Eurobarómetro 486 sobre "PYME, start-ups, empresas en expansión y espíritu empresarial" se centra en las barreras y desafíos que enfrentan las PYME en Europa cuando crecen, hacen la transición a más modelos de negocio sostenibles y digitalización. La encuesta cubre a más de 16.000 empresas y se utilizarán técnicas de aprendizaje automático para proponer un modelo que diferencie a las empresas que se orientan a la sostenibilidad medioambiental y social. El estudio incluirá el análisis motivado de las variables a incluir el modelo, el tratamiento de los datos para la modelización, la evaluación de distintas técnicas de aprendizaje automático para la selección de la técnica más adecuada, el ajuste y optimización del modelo y su interpretación. En base a los resultados se extraerán características más relevantes, su importancia relativa y la dirección de su impacto. También se comparará mediante técnicas estadísticas las diferencias entre los grupos. Los

resultados y conclusiones de este estudio serán de interés para la orientación de las políticas públicas de financiación propuestas desde organismos públicos y para los managers de las empresas españolas que pretendan transitar hacia un modelo más sostenible.

Tabla de contenido

Tabla de imágenes

Tabla de figuras

1. INTRODUCCIÓN

El estudio planteado en este trabajo está motivado por conocer qué características definen a las empresas de la Unión europea a realizan acciones a favor del medioambiente y la sostenibilidad.

El auge internacional y la preocupación de las personas y de las empresas sobre los problemas medioambientales y sociales que acontecen hoy en día, hace de este estudio un trabajo de interés en el ámbito de la administración, tanto desde el punto de vista de los gestores empresariales como de los encargados de las políticas públicas. En la actualidad, toda empresa que no se preocupan por el medioambiente, la salud, la igualdad… no son bien vistos por la sociedad, incrementándose cada día más la demanda a las empresas para que se comprometan e integren políticas renovadoras para revolucionar la sociedad.

Estas acciones medioambientales o sociales se pueden aplicar de distinta manera. En este estudio se va a estudiar las dos aproximaciones que toman las empresas para afrontar estos retos: Las acciones externas y las acciones internas. Resulta coherente diferenciar las distintas acciones que pueda tomar la organización ya que dependen de muchos factores que pueden determinar que camino sigue cada negocio. A lo largo del trabajo se pretende estudiar muchos de los factores que pueden estar llevando a las empresas a tomar estas decisiones en materia de sostenibilidad y poder llegar así a conclusiones útiles para la sociedad.

No solo la sociedad en general reclama medidas sociales y medioambientales. “Los tiempos en los que el plan de gestión ambiental de una empresa era visto como un gasto innecesario en la cuenta de resultados de la entidad a final de su ejercicio económico se difuminan de manera progresiva, al tiempo que los grupos de interés y stakeholders exigen a

las corporaciones el cumplimiento de sus objetivos medioambientes de una manera fáctica y en el medio y largo plazo". (CTMA consultores, 2020) Además, posibles inversores y fondos solo invierten en empresas "verdes" o con buenas prácticas medioambientales y sociales. Existen índices bursátiles que solo incluyen este tipo de empresas, como podría ser el FTSE4Good del IBEX, que motiva y llevan grandes cantidades de inversión a las compañías que forman parte de este selecto grupo.

Para las empresas europeas es vital la necesidad de renovarse y afrontar el nuevo paradigma al cual se enfrentan hoy en día. Pero, sobre todo pensando el futuro incierto que se avecina, hay que estar preparado para cualquier acontecimiento y tener la capacidad de actuar y adaptarse rápidamente. Debido al constante cambio climático que dio comienzo con la revolución industrial y concretamente los últimos 50 años; buscar soluciones que incluyan un impacto positivo en el medioambiente es imperativo para que las empresas puedan adaptarse al complejo sistema que se está planteando. Prepararse para ese futuro se puede realizar de infinidad de formas distintas, tantas como a las empresas se les pueda ocurrir. Es necesario pues que las empresas cambien, se adapten, se transformen para ser mas sostenibles y, la eco-innovación es una de las herramientas más poderosas para hacerlo. Una sociedad que es capaz de adaptarse rápido a las demandas del cliente y las oportunidades e inconvenientes que surgen es la que tiene más papeletas tiene de crecer y progresar en una situación tan complicada como la actual. El paradigma actual reclama una transformación digital y verde profunda de las empresas y las PYMEs se están empezando a dar cuenta de la realidad.

Según afirma este titular de la revista digital itReseller: "8 de cada 10 pymes confían en la digitalización como vía para aumentar sus ingresos" (itReseller Tech&Consulting,

2022). Las PYMEs ya observan los beneficios de la transformación digital y afirman que gracias a ella aumentan hasta un 50% su cartera de clientes. Es esencial que las PYMEs europeas con ayuda de las instituciones a transformarse, ya que son el tejido vital de la sociedad económica europea.

Por otra parte, prácticamente todas las personas realizan prácticas medioambientales o sociales a nivel personal. Ya sea el reciclaje en el hogar, controles de gasto de agua o luz, voluntariado, etc. Para poder reconducir y tener un planeta más sostenible a largo plazo, no solo deben actuar las personas a nivel personal. Los grandes cambios lo deben hacer los estados y las sociedades que son las potencias que influyen en las economías y los impactos medioambientales. Las empresas deben verse influenciadas por aquello que hacen sus clientes y aplicar a gran escala las prácticas que estos realizan en sus hogares. Como podría ser la regla de las 3 erres: Reducir, Reutilizar y Reusar.

Existen muchas formas con las cuales las empresas pueden aplicar cambios de cara a convertir el mundo más sostenible. El desarrollo de productos sostenibles resulta una forma idónea de llevar a cabo estos objetivos, sumado a aspectos importantes a tener en cuenta como puede ser la innovación, la inversión o la creación de trabajo. Un producto nuevo, más sostenible (ya sea reciclado o con materiales menos contaminantes) no solo es bueno para el medio ambiente si no que aporta una imagen de empresa innovadora y comprometida con los problemas contemporáneos.

Existen diversas agencias agendas internacionales y programas de la ONU como son la agenda 2030 y los ODS que marcan una hoja de ruta a seguir tanto para las empresas, gobiernos y personas en busca de un desarrollo sostenible. No solo estas organizaciones claman por llevar a cabo estas políticas; expertos del IPPC (grupo intergubernamental de

expertos sobre cambio climático) afirman que existe un calentamiento global de 1,5ºC de temperatura media respecto a niveles preindustriales (Masson-Delmotte, 2019). Estos estudios demuestran la necesidad de realizar un cambio a nivel social y medioambiental en busca de ese desarrollo sostenible tan deseado e impulsado desde la ONU y sus distintas propuestas y medios.

El COVID19 ha supuesto muchos cambios. Pero no todo ha sido malo. Ha abierto muchas nuevas oportunidades y sobre todo ha motivado a las personas a cambiar de perspectivita y estar más comprometidos con el medioambiente y la salud. Además, con motivo de la pandemia, la UE *"*La UE se propone el objetivo de conseguir la neutralidad climática en 2050, transformando la UE en una economía sostenible y climáticamente neutra basándose en los siguientes pilares: Descarbonización, Eficiencia Energética, Contaminación 0, Economía Circular, Movilidad Sostenible y de la granja a la mesa" (PWC (PriceWaterhouse&Coopers), s.f.). Todos estos objetivos pretenden hacerlo apoyando proyectos de transición ecológica de empresas europeas con los fondos europeos NextGeneration. Esto supone una "obligación" para todas las compañías sea cuales sea su tamaño, a realizar cambios en busca de un mundo más sostenible y tener acceso a estos fondos para poder cambiar el mundo.

En este contexto, analizar las características que definen a las empresas europeas que activamente toman acciones en pro de la sostenibilidad medioambiental o social resulta de gran interés. Conocer las facetas que definen a una empresa decide innovar o no permite realizar campañas y políticas más efectivas, por ejemplo, focalizadas en aquellas reacias a no participar activamente o poder actuar de forma que estas compañías acaben sumándose y colaboren a hacer un mundo más sostenible.

Los resultados y conclusiones extraídos de este análisis pueden ayudar a gobiernos o a empresas a tomar decisiones, por ejemplo, en materia de ayudas y subvenciones, para promover y tener una economía más comprometida con el medioambiente. También pueden ayudar a los gestores de las empresas a evaluar qué tipo de características y comportamientos empresariales deben de promoverse para ser más sostenible.

2. DESCRIPCIÓN DEL CONTEXTO

Antes de comenzar el estudio, se debe comprender todos aquellos aspectos que están relacionados con la investigación que han sido introducidos previamente. Poniendo especial hincapié en el medioambiente, la sostenibilidad y la innovación. Así mismo se comentarán otros factores como el reciclaje, los ODS o el COVID. Se espera comprender el entorno y las características de todo aquello que hace que las empresas decidan participar activamente en acciones empresariales sobre sostenibilidad medioambiental o social.

2.1 Medioambiente y sostenibilidad

Desde la revolución industrial iniciada de mediados del siglo XVIII, los diferentes cambios tecnólogos y económicos propiciaron una transformación del mundo hacia sociedades cada vez más modernas. Se dejo de lado la artesanía y las actividades agrícolas y ganaderas trabajadas de forma manual, para pasar a un mundo movido por máquinas y procesos que requieren cantidades de energía inanimada.

Desde este proceso iniciado a en Inglaterra sobre 1750, hasta más de 200 años después, no hubo preocupación alguna por la sostenibilidad y el medioambiente. Es más, durante las dos grandes guerras, el derroche y el uso masivo de recursos naturales sin preocupación alguna por el planeta (llegando incluso a lanzar varias bombas nucleares) nos indican que a lo largo de dos siglos el ser humano no ha mirado por la naturaleza.

A finales del siglo XIX, un científico sueco llamado Svante Arrhenius aviso de los problemas de las emisiones de CO2 y un posible calentamiento global. Pero no es hasta

mediados de los años 50 donde crece la preocupación por la materia. Fueron apareciendo nuevas investigaciones que demostraban que hay poco vapor de agua en la parte más alta de la atmosfera o que el C02 producido por los combustibles no resultaba absorbido por los océanos de manera inmediata. Este último descubrimiento fue realizado por Hans E. Suess usando la prueba del carbono 14.

Desde estos descubrimientos, y las constantes evidencias que daba el planeta de lo que está sucediendo, un número mayor de científicos y personas contrastadas fueron haciéndose eco de la problemática. A lo largo de los distintos años 70 se crean ONGs ambientalistas como Greenpeace o el programa de las naciones unidas para el medio ambiente. Todos estos movimientos fueron los precursores de las políticas actuales.

A lo largo de los años, sobre todo mediante la educación en los colegios y las distintas campañas en los medios de comunicación, se ha promovido y concienciado a la población sobre el cuidado del planeta y el medio ambiente. Por ejemplo, según ANFEVI (Asociación nacional de fabricantes de envases de vidrio) en Europa se recicla el 73% de botellas de vidro, llegando a ser el 98% en países como Dinamarca. El presidente de FEVE, Vitaliano Torno comenta: "La elevada tasa del 73% sitúa al modelo de envasado de vidrio entre los mejores modelos de negocio con el objetivo de minimizar la generación de residuos y combatir el agotamiento de materias primas. Sin embargo, se debe seguir invirtiendo todos los recursos disponibles para seguir protegiendo el medioambiente" (Anfevi, s.f.).

No solo la población en general está concienciada. En su mayoría, las empresas también lo están. Las sociedades son los mayores productores y no solo los consumidores deben cuidar el medioambiente. "Lo ilustró hace dos años un informe elaborado por Carbon Majors Report, en el que ponía cifras al problema: **el 71%** de las emisiones de CO2 eran

emitidas, según sus cálculos, por tan sólo **100 empresas**, la mayor parte de ellas contadas entre las más grandes de la economía global." (Mohorte, 2020) Resulta esencial por tanto estudiar cuales son las características de las empresas que realizan prácticas por el medioambiente para así poder enfocar mejor las políticas de cara a las grandes empresas que no lo hacen. el CDP, "cada vez son más las grandes empresas que apoyan la transición a una economía libre de carbono y se han comprometido a obtener energía de origen 100 % renovable. En este grupo, lideran el cambio compañías como Apple, Facebook, Google o Ikea." Realmente, las grandes empresas contaminantes, son todas las que tienen como su principal la quema de combustibles fósiles, siendo China Coal la que más CO2 emite a la atmosfera. Resulta evidente que estás empresas deben realizar esfuerzos para reducir estas emisiones y practicar responsabilidad social corporativa.

Los objetivos de los principales gobernantes de Europa es que a **medio/largo plazo toda la energía** que alimenta a la unión sea proveniente de las energías **verdes.** Preferiblemente, energías renovables como la eólica o la solar. Desde finales de los años 70 que se consideraron opciones viables hasta hoy en día ha habido una evolución constante en busca del objetivo principal que es dejar de lado los combustibles fósiles. Hoy en día esto resulta imposible porque no se ha desarrollado aun la tecnología suficiente para almacenar energía renovable en baterías. Es decir, toda la energía generada renovablemente, debe ser consumida al instante. Es por eso por lo que en la actualidad se debe tener una energía fósil de backup. Queda patente que el futuro sostenible es de la mano de la energía renovable. En un informe elaborado por Bloomberg y Acciona durante Cumbre del Clima en Madrid, **las energías renovables** proporcionarán el **68% de la demanda energética en España allá para el año 2030**, un aumento a tener en cuenta respecto al 40% en 2021.

Al hilo de lo comentado respecto de la necesidad de tener un combustible fósil de soporte constante a las energías renovables, no hay que olvidar la crisis del gas natural a nivel mundial (debido a diferentes causas geopolíticas y de mercado). Esta crisis hace evidente la necesidad de los países europeos de apostar por las energías limpias y renovables y evitar ser tan dependientes de países productores y altamente contaminantes.

2.2 Innovación

No hay progreso medioambiental ni sostenibilidad sin innovación. Se conoce a la innovación como un proceso que modifica ideas o elementos, mejorándolos o creando nuevos con un impacto positivo en el mercado o para la sociedad. No solo resulta imprescindible la innovación, la implementación de estas novedades en los procesos o productos.

Hoy en día estamos ante la cuarta revolución industrial donde la automatización, el internet de las cosas, macrodatos y la nube forman parte de esta nueva revolución. Esta serie de cambios van de la mano de una economía en busca de la descarbonización y mucho más sostenible. La innovación también se aplica al campo del medioambiente y la búsqueda de la sostenibilidad. No es posible descarbonizar las economías y ser más sostenibles sin innovación. Ya que innovación es sinónimo de progreso. Uno de los últimos logros realizados por la ciencia ha sido sustituir los fertilizantes agrícolas de nitrógeno (provoca gases de efecto invernadero) por unos cultivos nuevos que no necesitan estos fertilizantes, ya que son capaces de producir los suyos propios (Yanes, 2022).

Empresas como Apple han reducido su huella de carbono en 4,3 millones de toneladas métricas en 2019 a través de innovaciones de diseño y estudio de materiales y contenido

reciclado en sus productos. Además, tiene planes de 0 emisiones de aquí a 10 años. (Aznar, 2020) Se observa con este ejemplo que la innovación es el camino a seguir para lograr los objetivos planteados por las autoridades. La innovación empresarial es vital, pero debe venir de la mano de políticas claras y ayudas por parte de los gobiernos con el fin de progresar lo más eficientemente posible.

2.3 Las tres erres

Es común conocer la regla de las tres erres (Reciclar, Reusar y Reutilizar) como una propuesta de hábitos de consumo de una manera responsable. Especialmente enfocada en el ciudadano y el consumidor final, su principal finalidad fue reducir el volumen de basura generada. Pese a estar enfocada al usuario de a pie, las empresas deben aplicar esta regla adaptada a sus situaciones. No se puede llegar a cumplir los objetivos medioambientales solo fomentando buenas prácticas con ciudadanos, se necesita un consenso general de todos los aspectos de la sociedad.

1. Reducir. La primera de las erres y probablemente la más importante para las empresas. "También conocida como minimización de residuos, es la acción de disminuir, simplificar o eliminar el consumo y/o uso bienes o energía." (Significados, s.f.) Existen prácticas muy simples como podría ser reducir los embalajes de un solo uso u contaminantes o complicadas como reducir la emisión de gases contaminantes sin dejar de ser productivos. Cualquier iniciativa por pequeña que sea es provechosa para el medio ambiente.
2. Reutilizar. La segunda de las erres está enfocada en reducir de manera indirecta el impacto negativo de los seres humanos en el medio ambiente. La finalidad es darle

una segunda vida a un objeto de manera que se evite producir de más. Sobre todo, es importante aplicar esta regla a aquellos productos que nos son biodegradables o contaminantes como pueden ser las bolsas de plástico. Una manera muy eficaz para las empresas es promoviendo o publicitando a los consumidores que reutilicen sus productos.

3. Reciclar. No debemos de olvidarnos del reciclaje, "que consiste en el proceso de someter los materiales a un proceso en el cual se puedan volver a utilizar, reduciendo de forma verdaderamente significativa la utilización de nuevos materiales." (rss (Responsabilidad social empresarial y sustenibilidad), 2022) Debido al aumento exponencial de población y la escasez de recursos, se ha vuelto vital la necesidad de reciclar lo máximo posible. Si la sociedad no recicla, pronto nos podríamos encontrar ante un paradigma incierto y lleno de dudas sobre cómo mantener de una manera sostenible a la población y al planeta en general.

Por esta razón, uno de los objetivos del estudio es estudiar cuales son las características que forman las empresas que aplican las tres erres como aspecto interno con la finalidad de ser más sostenibles. Sin la aplicación de esta simple regla, será muy complicado para las futuras generaciones vivir sosteniblemente.

2.4 Desarrollo sostenible, ODS y agenda 2030

"El desarrollo sostenible representa la transición de la sociedad actual a una sociedad más respetuosa con el medio ambiente. Es un modo de desarrollo cuyo objetivo es garantizar el equilibrio entre el crecimiento económico, la preservación del medio ambiente y el bienestar social." (Garrett, 2022)

Según la ex primera ministra de Noruega, Gro Harlem Brundtland, lo importante es no comprometer las capacidades de las generaciones futuras. Por lo tanto, se llega a la conclusión de que el objetivo es producir y consumir eficientemente y de manera consecuente con las capacidades del planeta tierra.

Existen distintos grupos o campos en las cuales el desarrollo sostenible puede actuar: sostenibilidad económica (reducir pobreza), sostenibilidad ambiental (reducir el impacto negativo de los humanos sobre la tierra), sostenibilidad social (un mundo más igualitario y justo) y la sostenibilidad política (buenas acciones y poder liderar para lograr los objetivos anteriores). Es crucial tomar acciones lo antes posible para poder avanzar y lograr una economía prospera y sostenible en el plazo más corto.

Un claro ejemplo de aplicación de los campos de desarrollo sostenible en todos los ámbitos podría ser la decisión del Parlamento Europeo de prohibir la venta de coches combustibles fósiles o no sostenibles en territorio comunitario a partir de 2035. Esta decisión política busca una sostenibilidad ambiental reduciendo el co2 emitido por coches tradicionales, y sostenibilidad económica y social a que impulsa las economías incentivando a un consumo ecológico. Pero, ¿es un objetivo realista? "De momento, un coche eléctrico sigue siendo una propuesta cara para un consumidor", confirma el experto en automóviles Conor Faughnan. "Pero tal y como están los combustibles actualmente, los atractivos son muy evidentes". (Euronews en español, 2022) Es necesario impulsar desde las instituciones las ayudas a las empresas para el desarrollo e innovación de los vehículos ecológicos (ya sea eléctricos o impulsados por hidrógeno) ya que, si no, resulta complicado ver cumplido el objetivo marcado.

"Los **Objetivos de Desarrollo Sostenible** (ODS) constituyen un llamamiento universal a la acción para poner fin a la pobreza, proteger el planeta y mejorar las vidas y las perspectivas de las personas en todo el mundo". (UN (Naciones Unidas), s.f.) Pero para entender los ODS, no podemos dejar de lado a la agenda 2030 y es que "Los países miembros de la ONU acordaron 17 objetivos como parte de la agenda 2030. La agenda, aprobada en 2015 por la asamblea general de las Naciones Unidas, establece una visión transformadora hacia la sostenibilidad económica, social y ambiental además de ser la guía de referencia para la comunidad internacional hasta 2030." (Cepal, s.f.) La agenda pone la dignidad y la dignidad de todas las personas en el centro, junto al desarrollo sostenible necesario para progresar de la mano del planeta. De todos los objetivos planteados para cumplir con la agenda, los relacionados con el estudio son los siguientes:

- Objetivo 7 – Energía asequible y no contaminante
- Objetivo 8 – Trabajo decente y crecimiento económico
- Objetivo 9 – Industria, innovación e infraestructura
- Objetivo 11 – Ciudades y comunidades sostenibles
- Objetivo 12 – Producción y consumo responsables
- Objetivo 13 – Acción por el clima
- Objetivo 15 – Vida de ecosistemas terrestres
- Objetivo 17 – Alianzas para lograr los objetivos

Desde las instituciones públicas europeas y el gobierno español, se hace mucho eco sobre el pacto verde y la transición ecológica, pero ¿Qué es realmente la transición ecológica? "La transición ecológica se ocupa de la propuesta y ejecución de las políticas del gobierno en materia de energía y medioambiente para la transición a un modelo productivo y social

más ecológico" (Ministerio para la transición ecologica, Gobierno de España, 2018). Tras la pandemia y los problemas geopolíticos y energéticos, se ha propulsado este proyecto con el fin de ser más sostenibles y menos dependientes de las energías no renovables.

2.5 Covid19 y Fondos europeos

A principios de marzo de 2020 se desató una catástrofe mundial sanitaria que paralizo el mundo prácticamente al completo en todos los sentidos. Esto supuso un cambio de mentalidad tanto para las personas como para las sociedades. Las personas se preocupan más por su salud y por el medio ambiente. Los millones de víctimas a lo largo de los últimos dos años, nos obliga a la sociedad a tomar parte activa para mejorar el mundo en el que vivimos.

A raíz de la pandemia y sus consecuencias, la unión europea ha planteado un paquete de fondos sin parangón. Estos fondos europeos, llamados NEXTGEN, tienen la intención de volver a impulsar las economías europeas tras el COVID, pero de una manera sostenible y renovada. Haciendo hincapié en la transición ecológica, la digitalización y la apuesta por los jóvenes. Debido a estas ayudas las empresas y los estados están ante una oportunidad inmejorable para innovar a nivel social o medioambiental ya que se eliminan grandes barreras como la financiación o el acceso a recursos.

No solo los fondos europeos, debido a las distintas crisis que han ido surgiendo y al conflicto bélico ocurrido en territorio ucraniano, estamos ante una crisis energética de calado. También existen ayudas para empresas con prácticas sostenibles como el programa LIFE de la UE o el programa de sostenibilidad para pymes de la Cámara de Comercio de España para actuaciones relativas a la eficiencia energética

Por lo tanto, nos encontramos ante un paradigma incierto, pero repleto de oportunidades a nivel de inversión y cambios que las empresas deben saber aprovechar y ser el motor de los distintos cambios en pro de un mundo más sostenible

2.6 Prácticas sociales empresariales

No solo solo la innovación proporciona cambios positivos en la sociedad. Las empresas deben también practicar cambios en pro de sus trabajadores. Existen cambios ligeros que mejoran sustancialmente las relaciones, como ofrecer café gratuito a los trabajadores o crear un buen clima de trabajo donde reine la confianza y el respeto. Otros suponen más costes, pero la rentabilidad es abrumadoramente superior. Las mejoras laborales (ya sea en el aspecto salarial o de conciliación familiar) y un mejor entorno de trabajo (acceso a psicólogos, médicos o guardería) consiguen que los empleados sean más eficientes y eso es vital para el correcto funcionamiento de las empresas.

Desde la pandemia que dio comienzo en marzo de 2020 donde todos nos vimos obligados a encerrarnos en casa y adoptar nuevas maneras de asumir las responsabilidades. Surgió el teletrabajo y se ha convertido, una vez superados aquellos terribles meses, en una forma de trabajo de lo más común. Poder realizar el trabajo desde casa, facilitando la conciliación familiar e incluso en muchos casos, siendo mucho más eficiente. Fomentar esta práctica entre los trabajadores es una acción social que puede llevar a las empresas a otro nivel.

En los últimos tiempos, la igualdad en los puestos de trabajo, sin distinguir entre raza, sexo o religión es una obligación. Pese a que aún queda camino por realizar, no se debe dejar

de insistir a las empresas para que no se relajen. Una empresa con un entorno de trabajo donde impera la igualdad y el respeto crece más y más rápido que una aferrada a valores anticuados.

No puede haber progreso medioambiental solo de la mano de la innovación y el reciclaje. Esta debe ir acompañada por unas políticas sociales claras y apostando por la igualdad y el respeto para conseguir tener éxito.

2.7 Las empresas y las acciones medioambientales

La responsabilidad social corporativa (RSC) es la estrategia diseñada por las empresas para de manera voluntaria adquieren un compromiso activo para contribuir al mejoramiento social económico o medioambiental. Esta práctica, pese a ser ética para las empresas, cada día se vuelve más necesario ya que la sociedad exige a las empresas especial atención a las practicas sociales y sostenibles.

Inversores, Instituciones y los propios clientes premian y se interesan por las compañías con un código ético y prácticas sociales y ambientales. Distintos índices miden la sostenibilidad, en el caso de España, el FTSE4Good. Este índice clasifica a las empresas según un ranking sobre las cuales realizan más prácticas sostenibles. Gracias a estos índices las empresas pueden captar inversores externos y nuevos clientes.

Desde el Sepe, se han puesto en marcha a lo largo de este año el Programa NEOTEC, que consiste en ayudas a nuevos proyectos empresariales innovadores. “Las ayudas financiarán la puesta en marcha de nuevos proyectos empresariales, que requieran el uso de tecnologías o conocimientos desarrollados a partir de la actividad investigadora y en los que

la estrategia de negocio se base en el desarrollo de tecnología." (Ministerio de trabajo y economia social, 2022) El apoyo de las instituciones es básico para hacer funcionar la cadena de la innovación y la sostenibilidad medioambiental. Estos beneficios, no solo de instituciones públicas sino también de inversores privados premian a aquellos interesados en realizar acciones con impacto positivo en la sociedad creando así un modelo más sostenible.

No solo es importante las actividades sociales o medioambientales de carácter externo. Tiene el mismo valor o incluso superior las mejoras generales en busca de un modelo de procesado interno de las operaciones más sostenible o la creación/mejora de los productos mediante materiales reciclables o menos contaminantes. Tomando conciencia de la coyuntura actual respecto al cambio climático, de nada sirve realizar acciones de cara a la galería, pero no tomar un papel determinante y claro en los aspectos internos. Empresas como Google tienen extremadamente claro estos conceptos y llevan a cabo proyectos como crear centros de datos eficientes o lugares de trabajo sostenible. Así como ser la primera gran empresa a nivel mundial en abastecerse en el 100% de consumo electrónico anual con energía verde (Google, s.f.). Liderar en los aspectos ecológicos y sociales en todos los ámbitos convierte a las empresas en los verdaderos motores en busca del equilibrio humano con el medio ambiente.

3. OBJETIVOS DEL ESTUDIO

En este estudio se pretende analizar la importancia del medioambiente para las compañías europeas respecto a dos decisiones: Reciclar y/o innovar o no hacer nada. A lo largo de la investigación han surgido distintas preguntas que han ayudado a formar unos objetivos claros y concisos. Las cuestiones son las siguientes:

- ¿Por qué las compañías deciden innovar?
- ¿Qué aspectos definen a las empresas innovadoras?
- ¿Qué diferencia a las empresas que deciden innovar y a las que deciden reciclar o reusar?
- ¿Realizan prácticas activas las empresas en pro del medio ambiente?
- ¿Es útil la inteligencia artificial para modelizar un modelo de innovación medioambiental?
- ¿Por qué las empresas hoy en día realizan tantas acciones con relación a la responsabilidad social corporativa?
- ¿Qué es la transición ecológica y en que afectan o actúan las empresas?

Una vez planteadas las preguntas se han llegado a la conclusión de la necesidad de este estudio. No solo para las propias empresas, sino también para las distintas instituciones públicas o privadas (de carácter inversor). A continuación, se detallan los objetivos claramente definidos:

- **Definir las características de las empresas que realizan prácticas medioambientales activas para el entorno**. Ya sean acciones internas (reciclar o reusar) o externas (innovación) es importante conocer cuáles son las variables que definen que empresas realizan prácticas medioambientales positivas. Gracias a este estudio usando inteligencia artificial, instituciones públicas o inversores privados pueden identificar con garantías que tipos de empresas cumplen con los compromisos medioambientales. De esta manera, pueden decidir en el tipo de empresa que invertir o detectar que tipo de sociedad no está actuando en el medio ambiente. Resulta vital para poder influir en el mercado, conocer las características de las empresas para poder actuar y fomentar las prácticas medioambientales positivas.
- **Definir las características de las empresas que realizan prácticas medioambientales activas para las propias empresas.** Resulta interesante también enfocar el estudio para las propias empresas. De esta manera las empresas pueden observar y estudiar qué características deberían adoptar para adaptarse (dentro de sus posibilidades) a los requisitos que la sociedad exige hoy en día.
- **Definir las acciones sociales sostenibles**. Las acciones sociales que puedan realizar las empresas a nivel externo y sobre todo a nivel interno también pueden ayudar a comprender como actúan las empresas que realizan acciones medioambientales. Tiene un sentido lógico que una empresa comprometida con el medio ambiente y la sociedad también busque un progreso social dentro de la misma.

- **La utilidad del *MACHINE LEARNING* para poder definir las particularidades de las empresas.** Uno de los objetivos es demostrar la utilidad de la inteligencia artificial, concretamente del *MACHINE LEARNING*, para poder establecer con precisión las variables que explican el modelo planteado. Se decide usar este tipo de tecnología para poder dejar de lado suposiciones, prejuicios… pero, sobre todo, debido a la gran cantidad de datos y la complejidad operacional del modelo planteado.

4. DATOS Y VARIABLES

4.1 Base de datos

Resulta muy común encontrar una gran variedad de información respecto al cambio climático, medioambiente, sostenibilidad… La mayoría de esta información es muy generalista o precisa en un solo tema. La obtención de información propia resulta gesta titánica debido a la magnitud de datos necesaria para plantear el modelo propuesto. Debido a estas dificultades se ha decidido extraer la información que sirve como medio para obtener las preguntas planteadas de un cuestionario realizado para la Comisión Europea.

Los datos que nutren al estudio pertenecen a las respuestas del cuestionario *FLASH EUROBAROMETER 486 (Pymes, start-ups, scale-ups y emprendimiento*). El pase de las encuestas se realizó entre febrero y mayo de 2020 y fue publicada en septiembre de 2020. La encuesta, fue requerida por la comisión europea, como insumo decisivo para la "Estrategia de las PYME para una Europa sostenible y digital" (Comisión europea, 2020). El cuestionario, centrado en las PYMES europeas más doce países externos a la unión, se centra en los retos que tienen las empresas al crecer tratando de entender la transición a modelos más sostenibles y digitales. La encuesta fue realizada vía telefónica, la mayoría de las respuestas fueron obtenidas antes del confinamiento de marzo 2020 y cubre más de 16000 empresas y 385 variables.

Del informe dado por la comisión se pueden extraer algunas ideas interesantes para la compresión del análisis planteado:

- **6 de cada 10 empresas encuestadas han realizado algún tipo de innovación** en el último periodo de 12 meses desde la realización de la encuesta. Un 21% han introducido innovaciones con beneficios medioambientales y un 17% mejoras sociales.
- **La mayoría de las empresas** (71% de las encuestadas) encuentran **barreras de entrada a la innovación**. Para un tercio de las empresas, tanto los problemas legales administrativos y medioambientales o la financiación suponen dificultades añadidas.
- Solo 1 de cada 5 PYMEs tienen planes de digitalización comparado con cerca de la mitad de las grandes multinacionales.
- Más del 90% de los encuestados realizan al menos una acción social o medioambiental. De estos, dos tercios apuestan por las mejoras laborales y el uso de materiales reciclados y reusados. La mitad apuestan por reducir el consumo o mejorar la diversidad y la igualdad en el puesto de trabajo. Solo un 30% trata de desarrollar productos o servicios sostenibles.
- El 70% de las PYMEs encuentran que al menos una barrera les impide convertirse en sostenibles, ya que estos cambios les impedirían tener éxito a largo plazo con un impacto positivo en la sociedad o el medio ambiente. También comenta un 30% que una de las barreras para los cambios sostenibles es la falta de demanda para estos cambios.

Como se puede observar, estamos ante una base de datos completa que se ajusta a la perfección al modelo planteado para el estudio sobre innovación medioambiental y social para las empresas de la unión.

El cuestionario, base de datos e informes pueden ser obtenidos de manera gratuita (previa autorización) en las siguientes webs:

- https://search.gesis.org/research_data/ZA7637
- https://europa.eu/eurobarometer/surveys/detail/2244

4.2 Limpieza de la base de datos, tuneado del modelo y procesamiento de datos

Como se ha comentado anteriormente, la base de datos obtenida del cuestionario FLASH EUROBAROMETER 486 tiene más de 16000 empresas y 385 variables distintas. Para la correcta ejecución del trabajo, se ha procedido a una limpieza y preparación de la base de datos para la aplicación de modelos de aprendizaje automatizado. Este proceso consiste en eliminar aquellas variables explicativas que no son de interés para los objetivos propuestos, modificar las variables para que puedan ser tratadas por los algoritmos, o la realización de distintos filtrados para la selección de la muestra utilizada para el estudio. Además, en este paso también se selecciona y trata la variable o variables objetivo.

Para procesar todos los datos y la modelización del estudio en general (procesado, análisis, machine learning…), se ha utilizado el lenguaje de programación PHYTON a través del entorno JUPYTER LAB del lanzador ANACONDA. PHYTON es un lenguaje de programación muy potente que es utilizado en todos los sistemas operativos y presente en prácticamente todas las aplicaciones y programas. Es el lenguaje más común en proyectos de análisis de datos e inteligencia artificial. Tiene numerosas ventajas como su sencillez, la capacidad de poder trabajar en distintas plataformas, su potencia y su acceso gratuito. Según

opina Ángel Robledano, “Python es ideal para trabajar con grandes volúmenes de datos ya que, el ser multiplataforma, favorece su extracción y procesamiento, por eso lo eligen las empresas de Big Data” (Robledano, 2019).

Una vez instaladas todas las extensiones necesarias para iniciar el proyecto, se comienza cargando la base de datos en el entorno de trabajo. Tras visualizar las 385 variables que contiene la base de datos, se procede a eliminar todas aquellas que no se consideran interesantes para explicar el modelo (por ejemplo, variables de agrupación utilizadas por Eurostat para agrupar países). Los criterios seguidos para esta limpieza son los siguientes:

- La clasificación de países europeos según su orden de entrada a la unión: ‘eu-6’, ‘eu-12’… De igual manera, para aquellos de la eurozona: ‘eurozn-13’, ‘eurozn-15’…
- Destinos aspectos descriptivos como pueden ser: ‘edition’, ‘survey’…
- La pregunta 18 y todas las variables que ella conlleva, ya que es de respuesta abierta y no ayuda a definir con precisión el modelo. La pregunta es la siguiente: “¿Podría explicar con un poco más de detalle o dar ejemplos de cuáles son exactamente los problemas principales con los que se encuentra su empresa?”
- Todas las variables que son pesos para ponderar: ‘w1’, ’w5’… Estas variables permiten ponderar casos cuando la muestra que se ha obtenido no es reflejo de las proporciones de empresas reales de la población. En este contexto, no son necesarias ya que la muestra sí que es un reflejo real de la población.
- Para finalizar se eliminan variables como: 'vq1', 'q2a', 'vq2a', 'q2b', 'q3a', 'vq3a', 'q3b','q4a', 'vq4a', 'q4b’. Estas variables tienen información repetida en otras variables que si que se usan para el estudio.

Tras completar la limpieza inicial de variables, se procede a escoger solo los países pertenecientes a la UE ya que las empresas que forman parte de unión son las protagonistas del estudio. Además, para poder trabajar de manera exacta agrupamos los países en dos grupos, aquellos que están por encima de la media en términos de eco-innovación y los que no. Para ello utilizamos el índice de eco-innovación para el año 2021 y recodificamos la variable país en una nueva variable Isocoutry de tipo dummy. Tal y como se muestra en la Imagen 1.

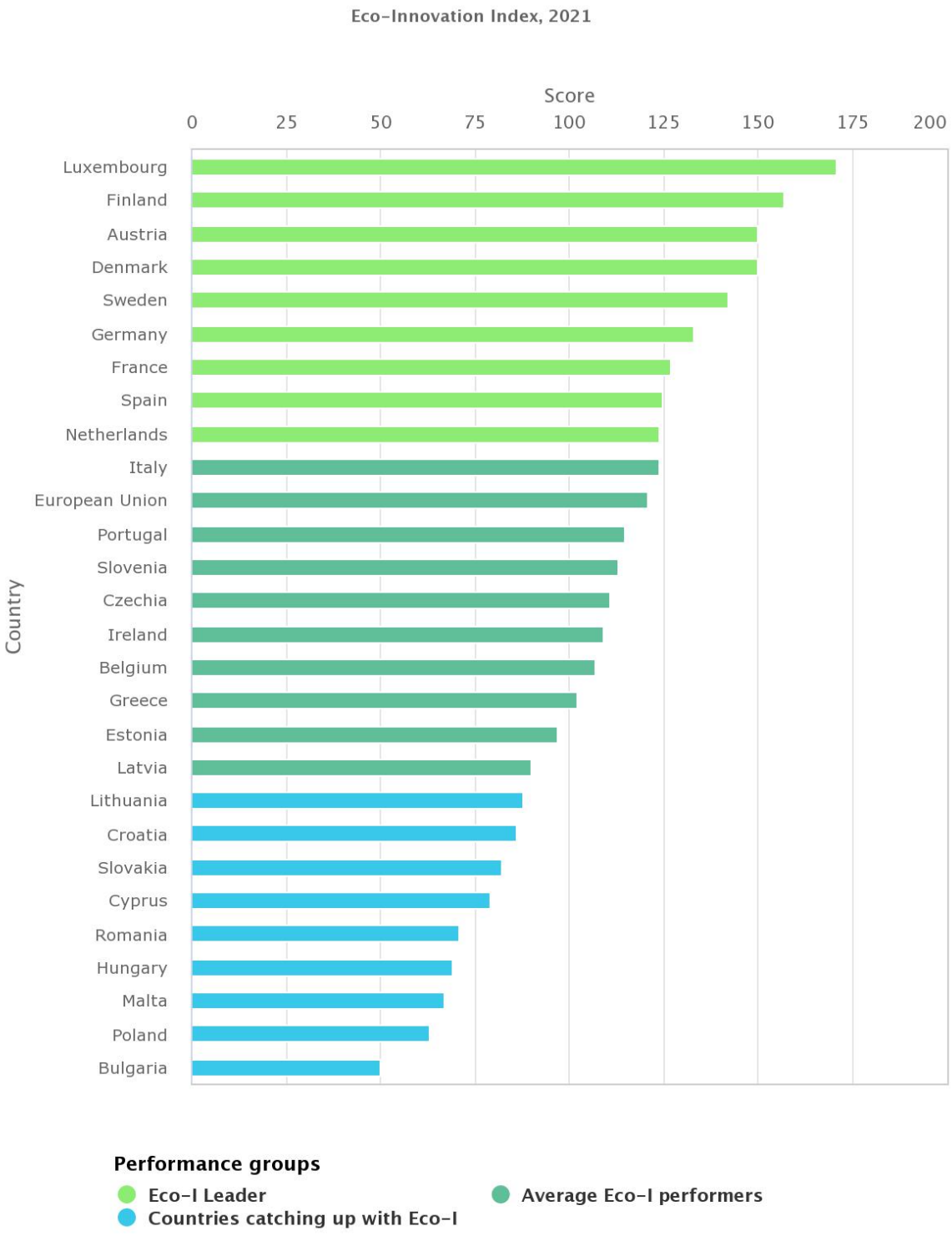

Imagen 1 THE ECO-INNOVATION SCOREBOARD AND THE ECO-INNOVATION INDEX 2021 (Unión europea, 2021)

Por lo tanto, los grupos serán:

Isocountry – 1 (eco-innovation leaders): Luxemburgo, Finlandia, Austria, Dinamarca, Suecia, Alemania, Francia, España y Países bajos.

Isocountry – 0 (resto UE-27): Italia, Portugal, Eslovenia, República checa, Irlanda, Bélgica, Grecia, Estonia, Lituania, Letonia, Croacia, Eslovaquia, Chipre, Rumania, Hungría, Malta, Polonia y Bulgaria.

Dentro de la encuesta hay preguntas con múltiples respuestas (el entrevistado puede seleccionar una o varias de las respuestas propuestas). En la base de datos de la encuesta original cada respuesta se codifica en forma una variable distinta. De esta forma, para cada una de las posibles respuestas planteadas, tenemos dos posibles valores: que se haya mencionado dicha respuesta o que no (NOT MENTIONED). Al dicotomizar las variables se generan variables asociadas a la no mención de una determinada respuesta que ya están implícitas en la variable creada para la correspondiente respuesta. Por tanto, se eliminan todas estas variables "NOT MENTIONED" asociadas a su respectiva respuesta. También se desestiman las variables generadas cuando las respuestas son del tipo "no sabe no contesta" ya que no aportan información relevante.

Teniendo en cuenta que las variables objetivo del estudio se encuentran en la pregunta 24, se procede también a eliminar de entre las variables predictoras a la variable 'q24.9_None (DO NOT READ OUT)' ya que, esta variable corresponde con la que se intenta predecir. Concretamente hace referencia a la respuesta: Ninguna; de la pregunta: En términos de sostenibilidad medioambiental y social, ¿cuáles de las siguientes acciones está llevando a cabo su empresa de forma activa? Por lo tanto, esta variable solo es un reflejo del grupo 0 (aquellas empresas que ni reciclan/reúsan y no innovan) el cual se explicará en un futuro

apartado más en detalle. Resulta contraproducente tener una variable que es la respuesta que intento predecir, ya que el objetivo del trabajo es descubrir que características hacen que las empresas decidan o no realizar innovación en términos de sostenibilidad social o medioambiental.

Para finalizar, el lenguaje y el entorno utilizado, no admite símbolos en las variables. Es por tanto que se modifican todas para evitar que surjan problemas al analizar el modelo. Se eliminan las comas, barras laterales, porcentajes paréntesis y corchetes por un espaciado.

Por lo tanto, nos encontramos ante una base limpia, clara y preparada para poder modelizar y obtener resultados.

4.3 Definición variables

Para definir la variable objetivo se ha realizado un clúster de variables, ya que nos encontramos ante una base de datos con distintas variables que podrían ser por si solas variables objetivo. De esta manera, agrupamos las variables en dos grupos que nos permitirán analizar cuáles son las características (sociales, innovación, estructura…) que tienen las empresas que realizan las acciones agrupadas en cada grupo.

La variable objetivo proviene de la pregunta número 24 de la encuesta. La pregunta de la encuesta es como sigue "En términos de sostenibilidad medioambiental y social, ¿cuáles de las siguientes acciones está llevando a cabo su empresa de forma activa?". Por tanto, la pregunta intenta desvelar el tipo de acciones tomadas por las empresas en el ámbito medioambiental y social. Puesto que nuestro objetivo es determinar las características de las empresas comprometidas con la sostenibilidad nos centraremos en las 4 primeras respuestas

ofrecidas por la encuesta. Dentro de estas respuestas encontramos actividades orientadas a la reducción, reciclado y reúso y, por otro lado, una indicado el desarrollo de productos y servicios sostenibles. En estas respuestas vemos claramente dos orientaciones. La primera, agrupa las actividades relacionadas con la gestión interna de la empresa (reducción, reciclado y reúso), mientras que la restante se refiere al impacto de la sostenibilidad de los productos y servicios. En definitiva, estos grupos, representan dos tipos de eco-innovaciones que pueden llevar las empresas, de tipo interno y externo.

Los grupos serán los siguientes:

- 'environmental_reduce_recible_reuse': 'q24.2_Reducing consumption of or impact on natural resources e.g. saving water or switching to sustainable resources' + 'q24.3_Saving energy or switching to sustainable energy sources' + 'q24.1_Recycling or reusing materials'
- 'q24.4_Developing sustainable products or services'

En estudios anteriores (Peiro-Signes & Segarra-Oña, ecommons, 2014, 2015) ya se indica que la innovación medioambiental es una parte de un concepto más extenso, la innovación social. De forma que las empresas que se preocupan por los aspectos medioambientales también se preocupan más por los aspectos sociales. Por ello, todas las respuestas (variables) que explican las acciones sociales que pueda realizar la empresa como la mejora de las condiciones de trabajo, la mejora de la diversidad o la igualdad, se incorporan al estudio como variables explicativas. De esta manera podemos entender mejor si la innovación medioambiental está relacionada con los cambios sociales positivos realizados por las empresas.

Una vez agrupadas las variables, podemos observar la cantidad de casos en los que las empresas pertenecen a cada grupo, a ambos o a ninguno. Como se ve a simple vista en la Figura 1, casi la mitad recicla o reúsa, pero no desarrolla productos sostenibles innovadores. Cerca de un **30% realiza ambas prácticas** y un **20% no realiza ninguna**. Tan solo un escaso 2,5% desarrolla nuevos productos o servicios sostenibles e innovadores y no recicla/reúsa. A los efectos de desarrollar un modelo general, no nos interesa discriminan grupos tan minoritarios. Por lo tanto, vamos a eliminar este grupo (grupo 2 imagen) tan particular del estudio debido al reducido número de empresas en comparación con los demás casos.

	environmental_reduce_recible_reuse	q24.4_Developing sustainable products or services	Count	Percent	Column Percent	Row Percent
0	0	0	2510	0.207181	0.305241	0.891969
1	1	0	5713	0.471564	0.694759	0.614235
2	0	1	304	0.025093	0.078109	0.108031
3	1	1	3588	0.296162	0.921891	0.385765

Figura 1 Porcentajes iniciales para cada grupo. Fuente: Elaboración propia

Para finalizar, se ha eliminado la variable 'q24.10_DKNA' del estudio ya que representa a las empresas que responden “No saben/No contestan (NS/NC)” a la pregunta: Existiendo una serie de respuestas claras tanto en términos de sostenibilidad medioambiental y social, y una respuesta que recoge a las empresas que no realiza ninguna de las acciones indicadas”, se considera aquellos que contestan NS/NC como casos perdidos de la variable objetivo. En este caso eliminamos los casos (las empresas) que desconocen las acciones llevadas a cabo en materia de sostenibilidad medioambiental y social.

4.4 Tipos de variable

Tras realizar la agrupación de variables, se va a hacer un breve resumen de las variables independiente y de control (variables explicativas del modelo). Figura 2. Existen variables tanto del tipo numérico o categórico. La mayoría son del tipo categórico ya que son la respuesta a una pregunta y esas respuestas están formuladas de manera que forman parte de este grupo. La mayoría de las numéricas fueron eliminadas ya que se correspondían con pesos para poder ponderar la muestra en los distintos ámbitos territoriales y no interesan para el estudio ya que contamos con una muestra fiel de la población. Por otro lado, algunas de las variables continuas, como por ejemplo el número de empleados, disponen de variables categorizadas en la base de datos. En estos casos tomamos la categoría más relevante como variable representativa de dicho concepto. Por ejemplo, para número de empleados, tomamos la variable que transformada en 4 categorías (1-9, 10-49, 50-249, 250 o más).

	isocntry	nace_a	q1	q2t	q3t	q4t	q5_1	q5_2	q6_1	q6_2	...	q26.1	q26.2	q26.3	q26
0	BE	M - Professional, scientific and technical act...	Between 2000 and 2014	1 to 9 employees	1 to 9 employees	More than 500,000 and up to 1 million euros	NaN	It has grown by at least 30%	Grow by less than 10% per year	Grow by between 10% and 20% per year	...	Not mentioned	Lack of consumer or customer demand	Not mentioned	N mention
1	BE	M - Professional, scientific and technical act...	Before 2000	1 to 9 employees	1 to 9 employees	More than 100,000 and up to 500,000 euros	NaN	It has grown by at least 30%	Grow by less than 10% per year	Grow by less than 10% per year	...	Not mentioned	Not mentioned	Not mentioned	It is n compatit with yo curre busines
2	BE	G - Wholesale and retail trade, repair of moto...	Between 2000 and 2014	1 to 9 employees	1 to 9 employees	More than 100,000 and up to 500,000 euros	NaN	It has remained stable	Grow by between 10% and 20% per year	Grow by between 10% and 20% per year	...	Lack of willingness among the management	Lack of consumer or customer demand	Not mentioned	N mention
3	BE	C - Manufacturing	Before 2000	1 to 9 employees	1 to 9 employees	More than 100,000 and up to 500,000 euros	NaN	It has grown by less than 30%	It does not plan to grow	Grow by less than 10% per year	...	Not mentioned	Not mentioned	Not mentioned	N mention

Figura 2 Ejemplo empresas y valor de sus variables. Fuente: Elaboración propia

Para el estudio, nos interesa que las variables sean del tipo Dummy, que sirven para clasificar identificar categorías. Haciendo uso de la función de pandas “get_dummies”, al

conjunto de variables categóricas, permite transformar cada variable categórica con n categorías en n variables dummy que tendrán un valor de 1 si la empresa pertenece a la categoría y cero en caso contrario. Por ejemplo, si tenemos una variable que contiene la cantidad de empleados que tiene una empresa: de uno a nueve, de 10 a 49, de 49 a 249 o más; se crearían tantas columnas como opciones de respuesta hubiera, es decir, en este caso 4 subvariables nuevas que solo harían referencia a cada una de las clases o categorías de número de empleados. Si en la columna de uno a nueve empleados hay un 1, resultara que esta empresa tiene una cantidad de trabajadores asociable a ese grupo y 0 en el caso de cualquiera de las otras posibles combinaciones para esa misma empresa. De esta manera, es como obtenemos las 281 variables que van a explicar el modelo, todas del tipo DUMMY. Figura 3

	index	isocntry_0	isocntry_1	nace_a_Arts entertainment and recreation	nace_a_B - Mining and quarrying	nace_a_C - Manufacturing	nace_a_D - Electricity gas steam and air conditioningsupply	nace_a_E - Water supplyseweragewaste managementremediation activ	nace_a_F - Construction	nace_a_G - Wholesale and retail trade repair of motor vehicles and	...	q26.1 willin amor manage
0	0	1	0	0	0	0	0	0	0	0	...	
1	1	1	0	0	0	0	0	0	0	0	...	
2	2	1	0	0	0	0	0	0	0	1	...	
3	3	1	0	0	0	1	0	0	0	0	...	
4	4	1	0	0	0	0	0	0	0	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
12110	13613	1	0	0	0	1	0	0	0	0	...	
12111	13614	1	0	0	0	1	0	0	0	0	...	
12112	13615	1	0	0	0	0	0	0	0	0	...	
12113	13616	1	0	0	0	1	0	0	0	0	...	
12114	13617	1	0	0	0	0	0	0	0	0	...	

Figura 3 Variables del tipo dummy. Fuente: Elaboración propia

5. METODOLOGÍA

Una vez estudiada y limpiada la base y con una idea clara de aquello con lo que se va a trabajar, se considera oportuno diseñar un modelo que pueda predecir los objetivos planteados. Este modelo debe predecir con precisión a qué grupo de los 3 indicados pertenece la empresa y qué características definen que las empresas sean eco-innovadoras con el fin de poder estudiar y tomar acción en pro de tener un tejido empresarial más sostenible y justo.

Un modelo del tipo predictivo es un conjunto de procesos y técnicas que mediante computación se analizan una gran cantidad de variables para inferir la probabilidad de que ocurra sucesos previos a su ejecución. Los modelos predictivos se pueden aplicar a infinidad de casos como por ejemplo reducir el riesgo de la empresa, reducir costes, aumentar beneficios o prever las características de un mercado. Son modelos muy útiles que ayudan a las empresas y ya son una realidad de aquellas empresas punteras en este mundo tan competitivo.

Dentro de los modelos de predicción, grosso modo se podrían dividir en dos grupos:

- Modelo de clasificación. Donde a la salida del modelo planteado es la probabilidad de ser parte de una determinada clase o grupo previamente definido.
- Modelo de regresión. Donde a la salida del modelo se obtiene un valor concreto.

Resulta evidente que el modelo a utilizar en este estudio es el **modelo de clasificación** ya que esperamos conocer con precisión a partir de los datos presentados con anterioridad a qué grupo pertenecen las empresas. Además, pretendemos evaluar qué variables son las que definen a las empresas más eco-innovadoras y las que menos. Dentro de los modelos de

clasificación tenemos la opción de utilizar técnicas de análisis multivariante (por ejemplo, análisis discriminante o regresión logística). Sin embargo, estos modelos no tienen una buena precisión cuando se trata de modelos con una cantidad importante de variables explicativas. En estos casos, las nuevas metodologías desarrolladas basadas en inteligencia artificial proporcionan resultados muy buenos. Concretamente, para el modelo de clasificación planteado se utilizará Machine learning, más conocido en los países hispanohablantes como Aprendizaje automático.

El Machine learning es una especialidad dentro de la rama de la inteligencia artificial que, mediante el uso de distintos algoritmos, permite que las máquinas dispongan de capacidad de identificar patrones o conductas en los datos y elaborar predicciones precisas. Gracias al aprendizaje, los ordenadores pueden realizar las distintas operaciones sin tener que estar programado, lo realiza de manera automática.

Existen tres categorías en las que se dividen este tipo de algoritmos, donde las dos primeras suelen ser las más habituales:

- Aprendizaje supervisado. Este tipo de algoritmos tienen etiquetas o información asociada a los datos que dan capacidad para hacer predicciones más precisas. Este campo engloba la mayoría de los problemas.
- Aprendizaje no supervisado. Algoritmos formados datos previos analizados. El objetivo es encontrar patrones que los organicen para sacar conclusiones
- Aprendizaje por refuerzo. Su función es que el algoritmo aprenda de sus errores. muy común en reconocimiento facial.

Para la investigación en curso, se plantea un **modelo de aprendizaje supervisado**. Para este modelo, antes se deben trabajar los datos. Este tratamiento de los datos ha sido

explicado en el apartado de datos y variables. Tras el proceso aplicado a los datos, se ha de modelar. Esto consiste en realizar un número de pruebas diferentes, hasta encontrar el modelo que más se adapte y sea más preciso. Para comprobar que prueba de las realizadas es la más optima, se debe tener una parte de la muestra que haga la parte de test para poder confrontar los datos. Para este trabajo se divide la muestra en dos partes. Una parte de prueba o entrenamiento con un 66,6% de los datos y una parte de la muestra en forma de test, con un 33,3% de los datos de la base. Tras obtener el modelo que mejor ajusta a los datos de entrenamiento, se obtiene un valor de predicción sobre la muestra de test o validación.

Cabe destacar que el modelo debe evitar el sobreajuste, porque no interesa que el modelo profundice ajustándose mucho a los datos de entrenamiento, ya que no podría generalizar correctamente y su capacidad ante datos nuevos se vería mermada. En un mundo como el actual, un modelo con sobreajuste no es productivo, ya que la cantidad de datos varia y aumenta con rapidez. Interesa un modelo rápido y con capacidad de aprender nuevos datos sin perder facultades.

Tras comprender el contexto de los modelos, hay que seleccionar el tipo de modelo que mejor se ajuste a lo requerido por la presente investigación. Existen una gran variedad de modelos diferentes que pueden adaptarse al estudio y son convenientes estudiarlos. Los modelos analizados son los siguientes:

- DummyClassifier
- LogisticRegression
- LinearDiscriminantAnalysis
- QuadraticDiscriminantAnalysis
- GaussianNB

- MultinomialNB
- SVC (SVM)
- KNeighborsClassifier
- BaggingClassifier
- RandomForestClassifier
- ExtraTreesClassifier
- XGBClassifier, más conocido como XGBoost

En la lista superior encontramos distintos tipos de modelos de aprendizaje supervisado, como podrían ser los modelos lineales (LogisticRegression o LinearDiscriminantAnalysis) o los basados en modelos de árboles (RandomForestClassifier o XGBClassifier) por mencionar solo algunos.

A continuación, se presenta los resultados obtenidos de analizar todos y cada uno de los modelos de aprendizaje supervisados mencionados anteriormente.

```
>DC 0.490 (0.000)
>LR 0.592 (0.010)
>LDA 0.596 (0.010)
>QDA 0.375 (0.020)
>GNB 0.411 (0.011)
>MNB 0.514 (0.009)


>SVM 0.583 (0.007)
>KNN 0.503 (0.007)
>BAG 0.590 (0.008)
>RF 0.595 (0.010)
>ET 0.598 (0.009)
>XGBC 0.598 (0.010)
```

Figura 4 Precisión modelos de aprendizaje. Fuente: Elaboración propia

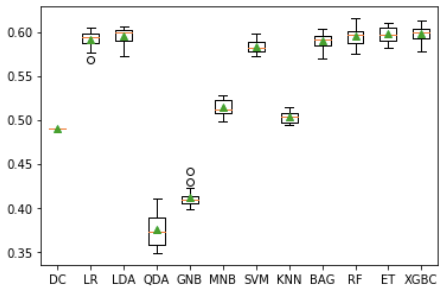

Figura 5 Precisión modelos de aprendizaje, boxplot. Fuente: Elaboración propia

Como se puede observar en la Figura 4 y la Figura 5, los modelos basados en teorías de árboles (XGBClassifier, RandomForestClassifier, y ExtraTreesClassifier) son los modelos de aprendizaje supervisados que explican con una precisión más alta el modelo presente, junto a LinearDiscriminantAnalysis (más propio de un modelo del tipo lineal).

Resulta que tanto para ExtraTreesClassifier y XGBClassifier se obtiene la misma precisión, siendo el valor final 0,598. Dicho en otras palabras, ambos presentan una precisión a la salida del modelo una precisión del 59,8%. Se aprecia como ligeramente el modelo ET obtiene una desviación típica menor, por lo que a priori resultaría más fiable. Pero este valor puede variar ligeramente según la muestra o si se realiza de nuevo los mismos pasos y podría resultar que XGBC fuera ligeramente superior. Por esto último, por la muy mínima diferencia y sobre todo por la gran ventaja computacional (mucho más eficiente) y de entrenamiento que permite XGBC, este será el modelo seleccionado.

5.1 XGBoost

Extreme Gradient Boosting, es el nombre completo del método de aprendizaje supervisado basado en árboles. Esta herramienta resulta ser la **referencia del aprendizaje**

supervisado automático para problemas de clasificación y regresión. Mediante el uso de este algoritmo se trata de buscar patrones en los datos etiquetados de manera que tras entrenar el modelo se predicen los datos de un nuevo conjunto y obtenemos la precisión a la salida del mismo.

Tianqui Chen, comenzó desarrollando el modelo como parte de un proyecto de investigación personal dentro de una comunidad de aprendizaje automático. Lo que inició en 2014 como una aplicación de terminal, fue cogiendo fama entre los usuarios y actualizándose con paquetes en Python y R. Tras el paso de los años y sus correspondientes actualizaciones, XGBoost ha acabado convirtiéndose en referencia del aprendizaje supervisado automático debido a su precisión y rendimiento. Tanto es así que en 2019 fue uno de los ganadores de los premios que otorga InfoWorld a las mejores tecnologías del año. Además, en los últimos años XGBoost ha ganado casi todas las competiciones de datos del tipo estructurados de la comunidad Kaggle.

XGBoost es el modelo basado **en arboles de decisión más optimizado** y **preciso** de la actualidad. Pero primero se va a explicar brevemente como es la evolución de este tipo de modelos hasta llegar al protagonista en cuestión.

Loa arboles de decisión simples conciben un modelo que predice la etiqueta a través de examinar un árbol y se estima que cual es la mínima cantidad de preguntas que hay que hacer para obtener la probabilidad de tomar una decisión optima. Estos modelos son ideales para predecir una categoría o la regresión para predecir un valor numérico continuo (Nvidia, s.f.). También encontramos los árboles con refuerzo de gradiente y los de bosque aleatorio que ambos combinan múltiples arboles de decisión para obtener un mejor modelo. Les diferencia como combinan, elaboran y ejecutan los distintos árboles. No se debe olvidar al

modelo Random Forest (RF) ya que XGBoost es una extensión a gran escala de este modelo. RF elabora una gran cantidad de árboles en paralelo y la predicción obtenida al final es el promedio de la predicción individual de cada árbol.

Hay que entender la diferencia entre impulsar y aumentar de gradiente. El termino impulso de gradiente mejora un modelo pobre, juntándolo con más modelos débiles para de manera conjunta, generar un modelo mejor. Aumento de gradiente es una extensión al impulso donde se trata de minimizar los errores producidos al juntar modelos débiles.

Por lo tanto, podríamos decir que XGBoost es una herramienta que combina lo mejor de todas las técnicas de toma de decisiones mediante árboles. Es escalable, forma árboles en paralelo y es preciso, contando con **aumento de gradiente** llevándolo a tener el menor error posible. Se aprovecha de su potencia y exprime a las máquinas para tener un modelo muy rápido y eficiente. La Imagen 2 muestra un breve resumen del funcionamiento de XGBoost.

Se invita a seguir usando XGBoost en un futuro ya que al ser un modelo de código abierto y muchos científicos continúan desarrollando y mejorando el modelo. Resulta super útil debido a que es una biblioteca portátil y funciona a la perfección en todos los sistemas operativos. Es, por tanto, el modelo ideal para la investigación en curso.

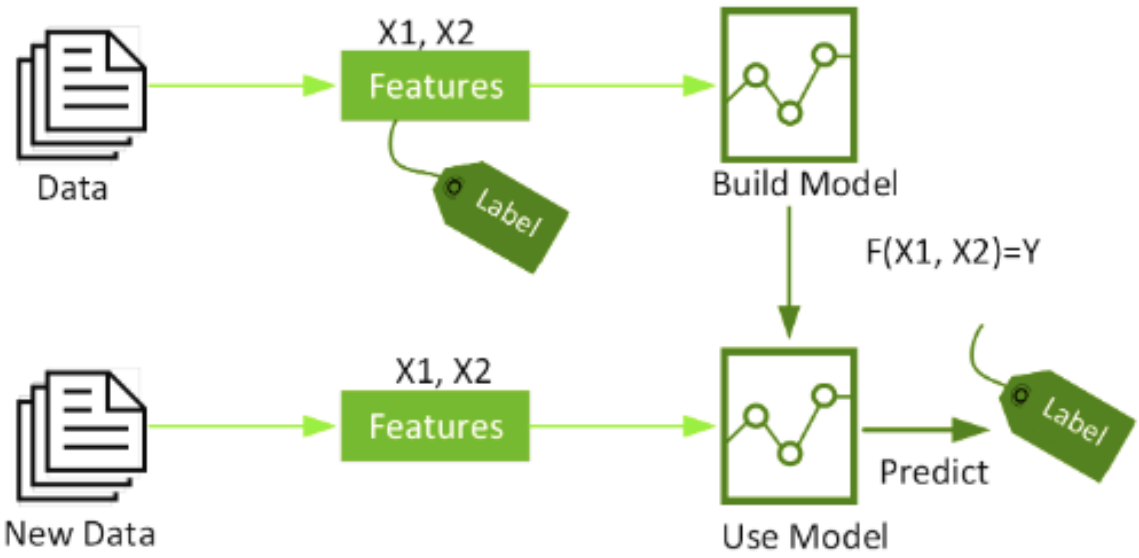

Imagen 2 Funcionamiento básico XGBoost (Nvidia, s.f.)

Para finalizar se muestra el resultado de aplicar el método XGBoost separando la muestra en test y train en la Figura 6. Se observa como aumenta la precisión del modelo, pasando de un 59,8% a un 61,58%. Hay una mejoría considerable debido a la muestra, la utilización de algunos hiperparámetros y la especificación de mlogloss.

```
# split data into train and test sets. Primera aproximación al problema
#Separamos los datos en test y train (33% y 66% respectivamente)

test_size = 0.3
X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=test_size, random_state=0,stratify=y)


# fit model on training data mlogloss porque son varias categorías
model = XGBClassifier()
eval_set = [(X_test, y_test)]
modelo=model.fit(X_train, y_train, eval_metric=["mlogloss"], eval_set=eval_set, verbose=False)

#verbose=True para que salgan todos los pasos de la cálculo de mlogloss para cada arbol
print(model)
# make predictions for test data
predictions = model.predict(X_test)
# evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
XGBClassifier(objective='multi:softprob')
Accuracy: 61.58%
```

Figura 6 Separación muestra en test/train y precisión XGBoost. Fuente: Elaboración propia

5.2 BorutaSHAP

Una vez es seleccionado el modelo XGBoost apreciamos que el número de variables que tiene el problema es considerablemente grande, unas 281 variables explicativas. Este tamaño no resulta cómodo y su manejo resulta engorroso ya que muchas variables no aportan nada y pueden influir negativamente a la salida del modelo. Por esta razón es necesario reducir esa cantidad de variables, a una cantidad manejable pero que siga explicando, con al menos, la misma precisión que la inicial.

Resulta que XGBoost de manera automática, otorga puntuaciones a las variables de entrada en el proceso de ajuste y da la información sobe la utilidad de cada una de ellas en la formación de los árboles de decisión que se plantean. Pero este resultado, el valor que se le da a las variables es buscando el mínimo óptimo usable para a predicción y varía según el modelo. Es decir, cambiaria la importancia si en vez de usar XGBoost se usara RandomForest. Los métodos basan la importancia de las variables en la precisión para aceptarlas o rechazarlas del conjunto. “Sin embargo, una disminución de la precisión al eliminar una característica es una condición suficiente para declararla como importante, pero no es suficiente para declararla como no importante” (Peiro-Signes, sciencedirect, 2022).

Usar la correlación para el filtrado de variables resultaría un error para disminuir el tamaño de la muestra. Debido a que la falta de correlación directa no quiere decir que no sea importante una variable, ya que esta puede serlo en conjunto con otras.

Es por lo cual, por lo que se necesita un método de selección que proporcione aquellas **variables relevantes** para el estudio y no solo aquellas que resulten óptimas para mejorar únicamente la precisión del modelo. Se busca que el estudio se centre en las cualidades que hacen que las empresas sean más o menos eco-innovadores y prosociales. La investigación

se centra en aumentar la precisión si no en conocer con precisión las características, comprender el fenómeno y para eso se debe evitar los factores redundantes.

Con que se decide utilizar el algoritmo de selección de características llamado BorutaShap. Este algoritmo es el resultado de la combinación de dos algoritmos independientes que, fusionados, proporcionan una **clasificación sin sesgo** y **coherente** de aquellas variables que resultan ser importantes y las que no. De esta manera, ayuda al investigador a comprender la interpretación de los resultados y obtener conclusiones coherentes.

Como se ha comentado, el algoritmo de selección escogido es la suma de dos algoritmos. El primero de estos es Boruta. Este algoritmo sigue un proceso iterativo que elimina todas las variables que, tras realizarles una prueba estadística, son irrelevantes ante las medidas aleatorias.

La segunda pieza del algoritmo de selección es Shapley Additive exPlanations. Más conocido como valores de Shapley o SHAP, estos sirven para interpretar mejor el aprendizaje automático. “Los valores SHAP son medidas de la contribución de cada característica a un modelo de aprendizaje automático. (Peiro-Signes, sciencedirect, 2022)” Es decir, los valores obtenidos indican hasta dónde puede llegar cada variable a la salida del modelo para cada una de las muestras individuales respecto del conjunto de datos de prueba. La suma de todos los valores asociados a una variable es el resultado de las predicciones y globalmente la importancia seria la media de cada contribución del tipo marginal para cada variable.

Por estas razones y combinando ambos subalgoritmos, BorutaSHAP es una técnica de selección de variables importantes para la explicación del modelo ideal. En la Figura 7 se puede observar el código característico empleado para implementar BorutaSHAP

```
#Vamos ahora con el algoritmo definitivo para elegir las variables relevant
from BorutaShap import BorutaShap
# load X and y
# NOTE BorutaPy accepts numpy arrays only, hence the .values attribute y = np.array(y_train)
#X = X_train
#y = y_train

# no model selected default is Random Forest, if classification is True it is a Classification problem
model = XGBClassifier()

# if classification is False it is a Regression problem
Feature_Selector = BorutaShap(model=model,
                              importance_measure='shap',
                              classification=True)

Feature_Selector.fit(X=X_train, y=y_train, n_trials=100, sample=False,
train_or_test = 'test', normalize=True, verbose=True)
```

Figura 7 Algoritmo BorutaSHAP. Fuente: Elaboración propia

Pero ¿Cómo es el resultado que aporta BorutaSHAP y como lo calcula? Pues el algoritmo examina las puntuaciones z (la división entre la media de perdida de precisión entre la desviación típica de la perdida) de cada variable y los compara con los atributos sombra. Los atributos sombra es una copia barajada del conjunto de datos para crear un nuevo conjunto, de esta manera se puede comprar ya que son nuevas características basadas en los mismos datos. Por lo tanto, aquellas que su valor z es mayor al valor máximo z de las variables sombra, serán catalogadas como importantes, y las que no, como no importantes. Por su parte, si supera ligeramente, con un valor marginal o la diferencia no es significativa estadísticamente, se consideran rasgos tentativos.

A continuación, se muestra la lista (Figura 8) donde aparecen los atributos que se han sido considerados como importantes. Se observa como de los 281 atributos iniciales, la lista se ha visto reducida en 37 importantes. Solo un 13,2% de los atributos son considerados por BorutaSHAP para explicar el modelo. Como se puede ver a simple vista, la mayoría de estos atributos están relacionados directamente con la eco-innovación, el progreso social y factores típicos que definen a las empresas y sus barreras. Tales como 'q25_No but it may be

considered in the future', 'q19.5_An innovation with an environmental benefit including innovations with an energy or resource efficiency benefit' o 'q24.6_Promoting and improving diversity and equality in the workplace'.

```
37 attributes confirmed important: ['q23.6_High speed infrastructure', 'q19.8_No none', 'q23.8_None of these', 'q19.6
_Social innovations such as new products services or processes that have the aim of improving society', 'q26.7_Lack o
f financial resources', 'q12a_Less than 25', 'q23.5_Big data analytics e.g. data mining and predictive analysis', 'q1
7.2_Regulatory obstacles or administrative burden', 'q24.5_Improving working conditions of its employees', 'q9.6_It h
as a patent or patent application', 'q10_Yes probably', 'q9.2_It mainly provides services', 'q16_3_Fairly good', 'q2
3.4_Smart devices e.g. smart sensors smart thermostats etc.', 'q24.8_Engaging employees in the governance of the ente
rprise', 'q24.6_Promoting and improving diversity and equality in the workplace', 'q20.3_Difficulties in predicting t
he market response', 'q26.4_It is not compatible with your current business model', 'q9.8_It has a strategy or action
plan to digitalise', 'q26.8_None of the above', 'q25_Yes and it has already been implemented', 'q19.1_A new or signif
icantly improved product or service to the market', 'q19.5_An innovation with an environmental benefit including inno
vations with an energy or resource efficiency benefit', 'q25_Yes and it is in the process of being implemented', 'q16
_6_DKNA', 'q25_No but it may be considered in the future', 'q25_Not applicable DO NOT READ OUT', 'q25_No and it will
not in the future', 'isocntry_0', 'q1_Before 2000', 'q24.7_Evaluating the impact of your enterprise on society', 'q7
a.3_Plan to grow as a result of operating in growing markets', 'q19.2_A new or significantly improved production proc
ess or method', 'q4t_100000 euros or less', 'q13.7_Predominantly family owned', 'q11.3_Other European countries outsi
de of the EU incl. Russia', 'q9.5_It is a part of a global value chain']
```

Figura 8 Variables importantes BorutaSHAP. Fuente: Elaboración propia

Ahora se observan aquellos atributos que han sido catalogados como tentativos (Figura 9), resultando ser estos, 7. Estos siete podrían ser incluidos o no, ya que su valor z supera a los de la sombra, pero no es estadísticamente significativo. De esta forma, ¿Deberíamos incluir estas variables como importantes o mantenerlas como tentativas y no incluirlas en el modelo final?

Tras estudiar el caso, se ha decido no incluir estas siete variables, pero sí que se ha considerado oportuno mencionarlas para que aquel interesado las tenga en cuenta a la hora de extraer sus propias conclusiones. Se ha llegado a esta conclusión por varios motivos. El primero es que se tienen 37 atributos que han sido considerados relevantes para la explicación del modelo y se considera un número razonable. Segundo, que, tras realizar un primer análisis con esos 37 atributos confirmados, se logran cubrir sobradamente los objetivos planteados para este trabajo. Es decir, los 37 atributos definen con precisión las características de aquellas empresas eco-innovadoras y las que no. Como consecuencia, se decide no incluirlas, pero mencionarlas ya que son atributos que podrían ayudar a complementar los resultados,

como, por ejemplo: 'q14.1_The sole founder of this enterprise' (sociedad unipersonal), 'q22_Your enterprise has adoptedis planning to adopt basic digital technologies but not advanced digital technologies ...' (innovación básica) o 'q17.8_Difficulties with digitalisation' (problemas con la digitalización).

```
7 tentative attributes remains: ['q21.7_Internal resistance to change', 'q14.1_The sole founder of this enterprise',
'q7a.2_Plan to grow as a result of introducing some kind of innovation', 'q22_Your enterprise has adoptedis planning
to adopt basic digital technologies but not advanced digital technologies ...', "q26.3_Lack of awareness about how to
integrate sustainability into the enterprise's business model", 'q17.8_Difficulties with digitalisation', 'q19.4_A ne
w way of selling your goods or services']
```

Figura 9 Variables tentativas BorutaSHAP. Fuente: Elaboración propia

Para finalizar, se vuelve a guardar el nuevo fichero con los datos finales extraídos por BorutaSHAP, es decir los atributos importantes, y se crean los ficheros test y train para su posterior entrenamiento y lectura de los resultados (Figura 10).

```
 # Returns a subset of the original data with the selected features
subset = Feature_Selector.Subset()
train_BorutaShap=pd.concat([subset, y_train], axis=1)
# guardamos el fichero con los datos subset. Estos son con las variables finales de X_train c
train_BorutaShap.to_csv ('train_BorutaShap_enterprises.csv')

#filtramos estas variables también en la matriz X_test
# filtramos solo las columnas que se encuentran en la lista de variables seleccionadas
select_X_test=X_test.filter(list(subset.columns))
test_BorutaShap=pd.concat([select_X_test, y_test], axis=1)
test_BorutaShap.to_csv ('test_BorutaShap_enterprises.csv')
```

Figura 10 Salvado muestra tras BorutaShap, train/test. Fuente: Elaboración propia

6. RESULTADOS

Una vez sabemos el método de machine learning y las variables optimizadas que vamos a utilizar a lo largo del trabajo, al siguiente paso se le podría llamar "entrenamiento del modelo".

Usando la muestra separada en "test y train" realizado para optimizar las variables usando el algoritmo de inteligencia artificial BorutaShap, se va a entrenar el modelo. ¿Qué quiere decir "entrenar el modelo"? Pues significa que, usando la muestra train y aplicándole unos parámetros distintos, se consigue mejorar la precisión del modelo final comprobando con la muestra test.

En la Figura 11, se puede observar cómo se prepara la muestra en distintas matrices. Existen dos parejas de matrices X e Y, de las cuales cada una representa la muestra para el entrenamiento y la otra para la prueba.

```
In [4]: #cogemos la última columna como variable objetivo uno Objet11 como categórica
        #Vemos el número de columnas de la matriz
        numcolumns=len(train.columns)
        numcolumns

        y_train = train.iloc[0:,numcolumns-1] #Asigna las última columna a la matriz Y
        # seleccionamos el resto como matriz de variables para la predicción y a esa matriz le quitaremos aquellas que no quer
        amos
        X_train = train.iloc[0:,0:numcolumns-1]  #Asigna las primeras  columnas a la matriz X

        y_test = test.iloc[0:,numcolumns-1] #Asigna las última columna a la matriz Y
        # seleccionamos el resto como matriz de variables para la predicción y a esa matriz le quitaremos aquellas que no quer
        amos
        X_test = test.iloc[0:,0:numcolumns-1]  #Asigna las primeras  columnas a la matriz X

        X_train.head()
```

Figura 11 Separación muestra en matrices X e Y. Fuente: Elaboración Propia

A continuación (Figura 12), se puede observar el modelo inicial sin realizar ningún tipo de entrenamiento al modelo. Usando el método escogido, XGBClassifier, se obtiene una precisión inicial de 60,95%. La meta es, variando distintos hiperparámetros del modelo, aumentar la precisión lo máximo posible. La métrica para evaluar el modelo es la función

‘mlogloss’. Esta función hace referencia a la perdida de **entropía cruzada múltiple**. “Esta es la función de pérdida utilizada en la regresión logística (multinomial) y sus extensiones, como las redes neuronales, definida como la verosimilitud logarítmica negativa de un modelo logístico que devuelve ‘y_pred’ probabilidades para sus datos de entrenamiento ‘y_true’” (Scikit-learn, s.f.).

```
In [7]: # fit model on training data
        model = XGBClassifier(random_state=42)
        eval_set = [(X_test, y_test)]
        modelo=model.fit(X_train, y_train, eval_metric=["mlogloss"], eval_set=eval_set, verbose=False)

        #verbose=True para que salgan todos los pasos de la cálculo de mlogloss para cada arbol
        print(modelo)
        # make predictions for test data
        predictions = modelo.predict(X_test)
        # evaluate predictions
        accuracy = accuracy_score(y_test, predictions)
        print("Accuracy: %.2f%%" % (accuracy * 100.0))

        XGBClassifier(objective='multi:softprob', random_state=42)
        Accuracy: 60.95%
```

Figura 12 Modelo inicial sin entrenamiento. Fuente: Elaboración propia

Las variables que se usarán posteriormente en el modelo son aquellas inicializadas por la función max: max_accuracy, max_learning_rate, max_colsample, max_subsample, max_max_depth y max_min_child. Estas variables se deben inicializar a 0, para posteriormente guardar los parámetros modificados y evitar posibles errores o confusiones. Como se indica en la Figura 13, a cada variable se le asignan distintos valores y el algoritmo ejecutará todas las combinaciones posibles y solo mostrará aquellas opciones que mejoren la precisión inicial.

```
import numpy
# grid search
max_accuracy=0
max_learning_rate=0
max_colsample=0
max_subsample=0
max_max_depth=0
max_min_child=0
learning_rate = [0.1, 0.2, 0.3, 0.4]
min_child_weight=[1,2]
max_depth = [3, 4,5]
subsample=[0.6,0.7,0.8]
colsample_bytree=[0.6,0.7,0.8]
```

Figura 13 Variables de entrenamiento. Fuente: Elaboración propia

A continuación (Figura 14), se muestra el algoritmo usado para mejorar el modelo inicial. El objetivo es 'multi:softprob', cuyo resultado contiene la matriz con las probabilidades de cada uno de los puntos de los datos de la muestra. Además, se observa cómo se evalúan las matrices 'X_test' y 'Y_test', y se entrena el modelo con 'X_train' y 'Y_train' mediante 'mlogloss' como se ha comentado anteriormente. Después de probar distintas combinaciones llega a la concusión que usando un 'learning_rate' de 0,30 aumenta la precisión ('max_accuracy') a 61,03%.

El modelo se hace mediante un **grid search** (búsqueda de cuadricula). Esta técnica consiste en ir probando distintos valores de cada parámetro y nos quedamos con el mejor valor antes de pasar al siguiente hiperparámetro. El modelo aplicado al trabajo se hace por pares, es decir de dos en dos, siendo así más eficiente y debido al clúster.

Llegados a este punto surge la necesidad de saber diferenciar que es un hiperparámetro y un parámetro. Pues un hiperparámetro es, según Rohan Joseph, "es una característica de un modelo que es externo al modelo y cuyo valor no se puede estimar a partir de los datos. El valor del hiperparámetro debe establecerse antes de que comience el proceso de aprendizaje." (Joseph, 2018) En cambio, un parámetro es una propiedad propia del modelo y sí que se puede estimar el valor que toma a partir de la muestra obtenida.

```
In [8]: import numpy
        # grid search
        max_accuracy=0
        max_learning_rate=0
        max_colsample=0
        max_subsample=0
        max_max_depth=0
        max_min_child=0
        learning_rate = [0.1, 0.2, 0.3, 0.4]
        min_child_weight=[1,2]
        max_depth = [3, 4,5]
        subsample=[0.6,0.7,0.8]
        colsample_bytree=[0.6,0.7,0.8]


        for i in learning_rate:
            model = XGBClassifier(objective='multi:softprob', learning_rate=i, random_state=42)
            #objective='multi:softprob'
            eval_set = [(X_test, y_test)]
            model.fit(X_train, y_train, early_stopping_rounds=10,  eval_metric=["mlogloss"],
            eval_set=eval_set, verbose=0)          #eval_metric=["logloss"]
            # make predictions for test data
            predictions = model.predict(X_test)
            # evaluate predictions
            accuracy = accuracy_score(y_test, predictions)
            if(accuracy>max_accuracy):
                max_accuracy=accuracy
                max_learning_rate=i
                print("learning_rate %.2f" % i)
                print("Max_Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
learning_rate 0.10
Max_Accuracy: 60.95%
learning_rate 0.30
Max_Accuracy: 61.03%
```

Figura 14 Algoritmo de mejora. Fuente: Elaboración propia

Una vez establecido el learning rate o tasa de aprendizaje, se busca optimizar los distintos parámetros previamente mencionados. Para ello, se realizan distintos bucles 'for', uno dentro de otro con la finalidad de que salga por pantalla la combinación de parámetros que maximiza la precisión del modelo. Por lo tanto, los valores que tomaran los parámetros principales son los siguientes:

- 'learning_rate': 0.30
- 'min_child_weight': 2.00
- 'max_depth': 3.00
- 'subsample': 0.70
- 'colsample_bytree': 0.70

Para una precisión final del modelo de 61,35% (Figura 15)

```
In [9]: for j in min_child_weight:
            for k in max_depth:
                for l in subsample:
                    for m in colsample_bytree:
                        model = XGBClassifier(objective='multi:softprob',random_state=42, learning_rate=max_learning_rate,
                                    min_child_weight=j,
                                     max_depth=k,
                                      subsample=l,
                                     colsample_bytree=m)
                        eval_set = [(X_test, y_test)]
                        model.fit(X_train, y_train,  early_stopping_rounds=10, eval_metric=["mlogloss"],
                          eval_set=eval_set, verbose=0)
                        # make predictions for test data
                        predictions = model.predict(X_test)
                        # evaluate predictions
                        accuracy = accuracy_score(y_test, predictions)
                        #print(j,",",k,",",l,"accuracy", accuracy)
                        if(accuracy>=max_accuracy):
                            max_accuracy=accuracy
                            max_subsample=l
                            max_colsample=m
                            max_min_child=j
                            max_max_depth=k
                            print("learning_rate %.2f" % max_learning_rate,"min_child_weight %.2f" % max_min_child, "max_depth
%.2f" % max_max_depth, "subsample %.2f" % max_subsample,"colsample_bytree %.2f" % max_colsample)
                            print("Max_Accuracy: %.2f%%" % (accuracy * 100.0))

learning_rate 0.30 min_child_weight 2.00 max_depth 3.00 subsample 0.70 colsample_bytree 0.70
Max_Accuracy: 61.35%
```

Figura 15 Modelo final. Elaboración propia

Para finalizar con la optimización de los parámetros queremos saber qué valor de 'reg_apha' y 'gamma' maximizan el modelo. El parámetro 'reg_alpha' (referencia a los pesos) cuanto mayor sea, más conservador será el modelo. Del mismo modo pasa con 'gamma' (referencia a los nodos de los árboles), cuando mayor sea su valor, más conservador será el modelo. Tras entrenar el modelo, ha resultado óptimo dejar ambas variables a 0, manteniendo la precisión en 61,35%.

En materia de calidad del modelo, según Hair et al. (Hair, 2014) un modelo de clasificación es suficientemente bueno si su precisión es superior al criterio de máxima probabilidad, esto es al valor que obtendríamos si siempre apostáramos por la clase/grupo que tiene una probabilidad más alta (en este caso el grupo 1, los que reciclan/reúsan pero no eco-innovan 5713/ total de la muestra final), y superior, en un 25% como mínimo al criterio de probabilidad proporcional (porcentaje grupo 1 al cuadrado por porcentaje del 2 al cuadrado por porcentaje del grupo 3 al cuadrado y luego multiplicado por 1,25). El mínimo

criterio de probabilidad proporcional resultante para el modelo es de 0,116%, resultando ser un valor muy pequeño. Por lo tanto, el modelo de clasificación es sobradamente bueno.

Tiene mucho sentido tener tanto 'reg_alpha' como 'gamma' a 0 ya que, como hemos comentado, cualquier valor superior indicaría que estamos ante un modelo más conservador y por lo tanto bajaría la precisión.

Una vez optimizado el modelo y obtenida la máxima precisión, vamos a observar la **matriz de confusión** (Figura 16). Esta matriz es muy común en el campo de la inteligencia artificial ya que permite observar el desempeño del algoritmo. Las columnas representan las predicciones mientras que las filas representan los casos reales. Los casos bien clasificados aparecen en la diagonal de la matriz, mientras los que no se han clasificado correctamente aparecen fuera de la diagonal.

En el caso del grupo 0 (empresas que no reciclan/reúsan ni innovan) el modelo predice 394 casos mientras que la realidad dice que hay 707. Para el caso de este conjunto de empresas tenemos una precisión de aproximadamente el 56%, el más bajo de los 3 grupos. Como se puede observar (Figura 17), el modelo predice mejor el grupo 2 (si reciclan/reúsan y si innovan), ya que en este grupo tiene una precisión del 69%. Finalmente, tanto el modelo real como la predicción coinciden que donde más casos hay es en el grupo 1 (si reciclan/reúsan, pero no innovan).

```
In [14]: # cargamos las librerias para poder sacar la confusion matrix
         from sklearn.metrics import confusion_matrix
         from sklearn.metrics import plot_confusion_matrix
         plot_confusion_matrix(model, X_test, y_test, values_format='d')
```

```
Out[14]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f77da320250>
```

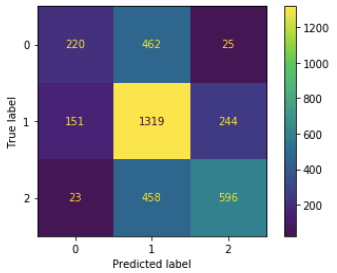

Figura 16 Matriz de confusión. Fuente: Elaboración propia

```
In [15]: from sklearn.metrics import classification_report, confusion_matrix
         print("Confusion Matrix:")
         print(confusion_matrix(y_test, predictions))

         print("Classification Report")
         print(classification_report(y_test, predictions))
```

```
Confusion Matrix:
[[ 220  462   25]
 [ 151 1319  244]
 [  23  458  596]]
Classification Report
              precision    recall  f1-score   support

           0       0.56      0.31      0.40       707
           1       0.59      0.77      0.67      1714
           2       0.69      0.55      0.61      1077

    accuracy                           0.61      3498
   macro avg       0.61      0.54      0.56      3498
weighted avg       0.61      0.61      0.60      3498
```

Figura 17 Informe clasificación. Fuente: Elaboración propia

Los valores finales de los hiperparámetos tras el "entrenamiento del modelo" son los siguientes (Figura 18):

```
In [14]: print('max_learning_rate=',max_learning_rate)
         print('max_min_child=',max_min_child)
         print('max_max_depth=',max_max_depth)
         print('max_subsample=',max_subsample)
         print('max_colsample=',max_colsample)
         print('max_reg_alpha=',max_reg_alpha)
         print('max_gamma=',max_gamma)

         max_learning_rate= 0.3
         max_min_child= 2
         max_max_depth= 3
         max_subsample= 0.7
         max_colsample= 0.7
         max_reg_alpha= 0
         max_gamma= 0
```

Figura 18 Valores finales hiperparámetros. Fuente: Elaboración propia

Visto que el modelo tiene es capaz de clasificar casos con una precisión suficientemente buena, un aspecto muy importante del estudio es evaluar el rol de las variables identificadas en el modelo. Esto es, conocer su importancia relativa y conocer en qué sentido influyen en que las empresas sean clasificadas en uno u otro grupo. Puesto que los modelos de machine learning son complejos, a efectos de interpretación son como una caja negra. Particularmente el modelo XGBoost generado tiene cientos de árboles de decisión que incluyen múltiples ramas de decisión con múltiples variables y límites de discriminación, por lo que la interpretación directa es inabordable por el ser humano. Es por ello que es necesario utilizar técnicas de interpretación. Existen numerosas técnicas de interpretación que se pueden aplicar a los modelos de machine learning (ej. LIME, SHAP,..). En este caso utilizaremos la técnica SHAP.

La **técnica SHAP** está totalmente relacionada con la clásica **teoría de juegos**. Mediante teoría de juegos cooperativos, las distintas variables se relacionan entre sí y finalmente se extrae una conclusión común. De esta conclusión surge el modelo de distribución empleado. La importancia que se da a cada variable en el modelo respecto al resto, se aplican los valores de Shapley.

A continuación, se muestra unas imágenes resumen que indican los valores medios de SHAP (Figura 19), lo que quiere decir, dar el valor medio que permita explicar la predicción para cualquier instancia como suma de todas sus aportaciones individuales. En la primera imagen podemos observar cómo afecta la media de cada valor shap según su variable y clase.

La variable que más impacta en la salida del modelo es la 'q24.7_Evaluating the impact of your enterprise on society'. Esta variable hace referencia a acción de las empresas de evaluar el impacto de su actividad en la sociedad en respuesta a la pregunta: En términos de sostenibilidad medioambiental y social, ¿cuáles de las siguientes acciones está llevando a cabo su empresa de forma activa? Tiene un reparto más o menos equitativo en el peso sobre las clases 0 y 2 pero muy poco en comparación a la clase 1.

De manera similar ocurre con las variables que más impacto tienen en el modelo: q24.6, q19.5, isocntry0 (países por debajo de la media europea en eco-innovación año 2020) y q24.5.

Este gráfico es muy intuitivo para modelos multiclase como el que tenemos entre manos ya que permite observar a simple vista y de manera muy intuitiva que variables influyen más al modelo en general además de la distribución en cada grupo. De esta manera podemos observar cómo en empresas fundadas antes del 2000 (‘q1_before2000’) apenas tiene impacto el valor medio del SHAP en la clase 2 (si reciclan/reúsan y si innovan) pero si lo tiene por igual en las clases 0 (ni reciclan/reúsan ni innovan) y 1 (si reciclan/reúsan, pero no innovan).

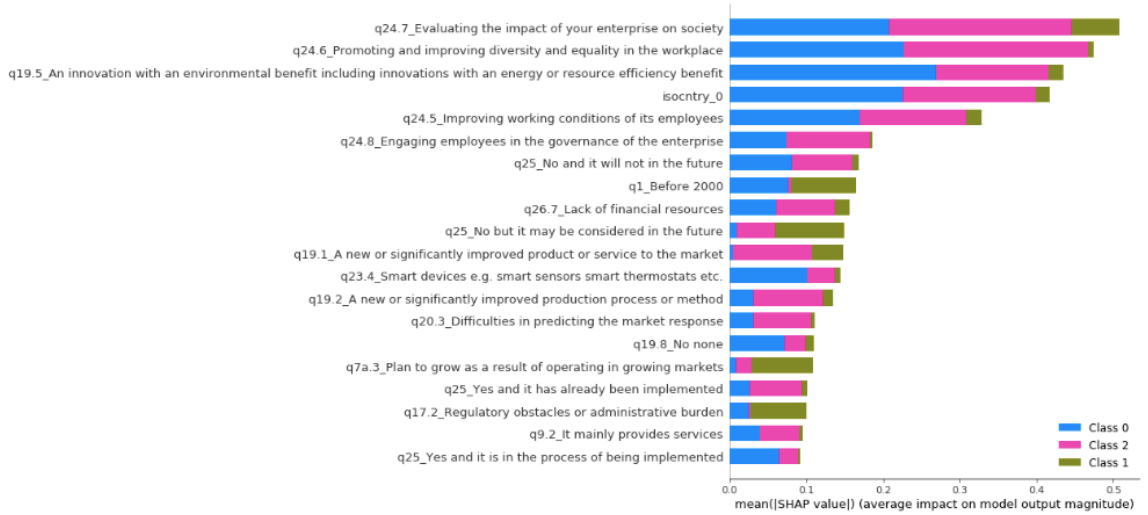

Figura 19 Feature importance SHAP, modelo variables reducido. Fuente: Elaboración propia

También se puede generar el mismo gráfico, pero excluyendo el resto de las clases y centrándose solo en aquella a analizar. Ese gráfico muestra la misma información que el anterior y es muy útil cuando quieres centrarse en una sola obviando el resto, pero este no es el objetivo final de la investigación. Gracias a esta nueva función podemos analizar el impacto de las variables solo para una clase, en este caso, la clase 0 sin verse el resto de las clases despejando así conclusiones que puedan llevar a un error.

En la clase 0 (Figura 20), las variables que más impactan en el modelo son: q19.5, isocntry_0, q24.6, q24.7 y q24.5. En cambio, las que menos influyen el modelo según el valor medio SHAP son: q19.1, 'q25_No but it may be considered in the future' y q7a3 entre otros.

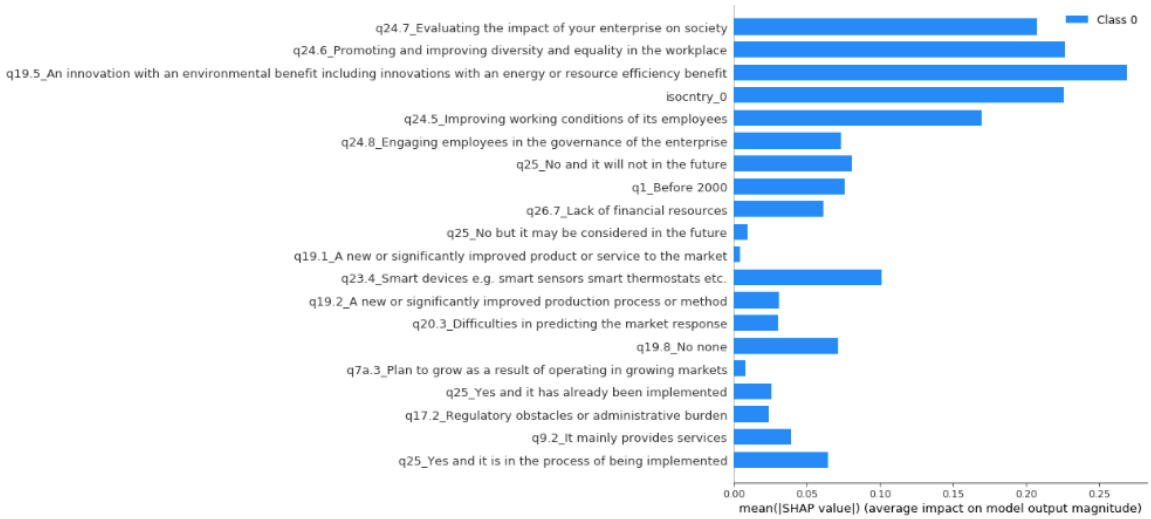

Figura 20 Feature importance SHAP, grupo 0. Fuente: Elaboración propia

En la clase 1 (Figura 21), las variables que más impactan en el modelo son: 'q25_No but it may be considered in the future', q1_before2000, q24.7, q7a3 y q17.2. En cambio, las que menos influyen el modelo según el valor medio SHAP son: 'q25_Yes and it has already been implemented', 'q25_Yes and it is in the process of being implemented' q24.8 y q20.3 y entre otros.

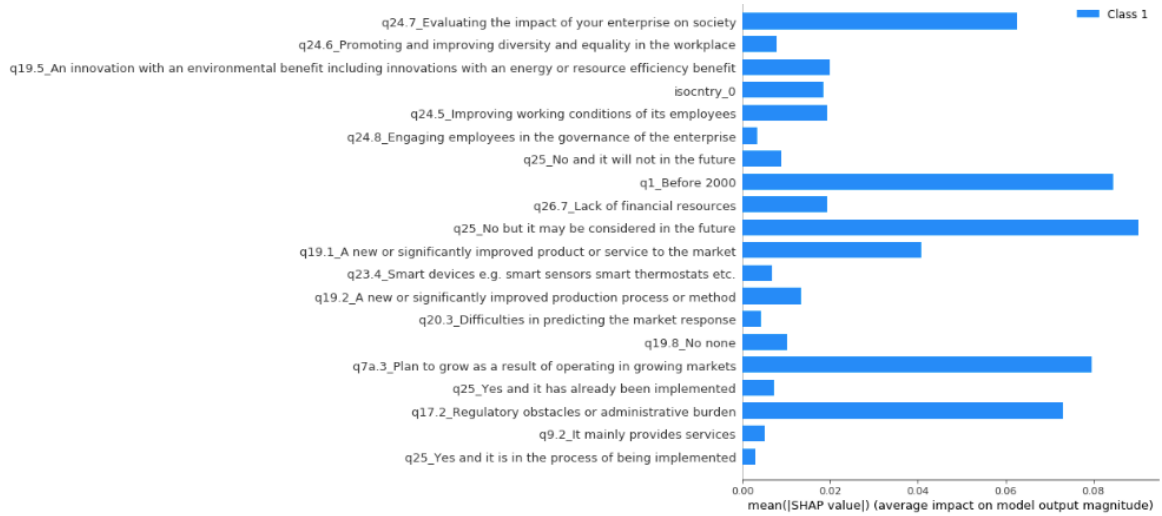

Figura 21 Feature importance SHAP, grupo 1. Fuente: Elaboración propia

En la clase 2 (Figura 22), las variables que más impactan en el modelo son: q19.5, isocntry_0, q24.6, q24.7 y q24.5. En cambio, las que menos influyen el modelo según el valor medio SHAP son: q17.2, q1_before 2000 y q7a3 entre otros.

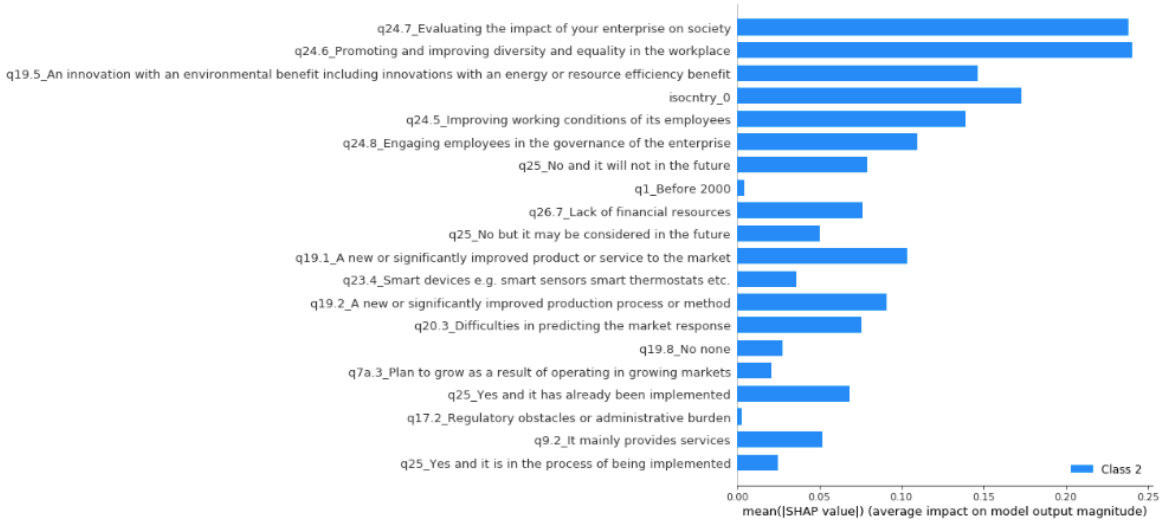

Figura 22 Feature importance SHAP, grupo 2. Fuente: Elaboración propia

Mediante los siguientes gráficos de dispersión de densidad de valores SHAP para cada variable conseguimos **identificar** tanto **el impacto** como la dirección de este que tiene cada variable o característica en la salida del modelo. "Las características se ordenan por la suma de las magnitudes del valor SHAP en todas las muestras" (Sukumar, 2020).

En el primer gráfico (Figura 23) tenemos como salida el grupo de empresas 0, es decir, aquellas que no realizan ni innovación y tampoco reciclan/reúsan. En este tipo de gráfico encontramos una barra lateral donde en los extremos encontramos los valores ''High" (alto) en color rojo y "Low" (bajo) en color azul. El color rojo representa valores altos de esa variable, es decir, si toma el valor 1. En cambio, el valor azul representa los bajos, en este caso si toman el valor 0. Luego si es un 1, el rojo aparece en la derecha de la variable respecto del eje (situado en el valor de SHAP 0.0 para la salida del modelo). Si el valor rojo está a la

derecha del eje quiere decir que incrementa la posibilidad de que sea clasificado, en este caso, como una empresa no eco-innovadora (ni recicla/reúsa ni innova). El ejemplo perfecto para este caso sería lo que ocurre con la variable: 'q25_No and it will not in the future'. En este caso, encontramos los valores rojos a la derecha del eje mientras que los azules se encuentran a la izquierda. La variable guarda las respuestas sobre las empresas que no realizan ninguna acción para convertirse en empresa sostenible (combinar éxito y rentabilidad a largo plazo con un impacto positivo en la sociedad y el medio ambiente) y tampoco tienen planteado hacerlo en el futuro. Tiene todo el sentido del mundo que los valores altos estén a la derecha ya que si están a la derecha aumente la posibilidad de que sea clasificado como una empresa de la clase 0, las no eco-innovadoras, y la variable representa a las empresas que no hacen ningún tipo de acción y tampoco lo plantean en su horizonte.

En cambio, en caso contrario tenemos numerosas variables en los cuales los valores altos se encuentran a la izquierda del eje. Al ser el caso contrario que el anterior, los valores rojos a la izquierda significan que queda reducida la posibilidad de ser clasificado en las empresas del grupo en cuestión, este caso las no eco-innovadoras. Es evidente este caso en las variables:

- 'q19.5_An innovation with an environmental benefit including innovations with an energy or resource efficiency benefit',
- 'q24.7_Evaluating the impact of your enterprise on society',
- 'q24.5_Improving working conditions of its employees',
- 'q23.4_Smart devices e.g. smart sensors smart thermostats etc.'
- 'q24.6_Promoting and improving diversity and equality in the workplace'
- 'q25_Yes and it is in the process of being implemented'

- 'q19.2_A new or significantly improved production process or method'

Las variables de la lista superior son solo algunas que nos permiten comprender perfectamente las características de este tipo de empresas.

Tiene todo el sentido que analizando el impacto que tiene las variables en la salida del modelo para las empresas no eco-innovadoras tengamos prácticas innovadoras o de desarrollo social con los valores altos a la izquierda del eje. Por lo tanto, es lógico pensar que empresas que evalúen su impacto en la sociedad, introduzcan/mejoren significativamente un método, usen dispositivos inteligentes, desarrollen innovaciones con un impacto medioambiental positivo o mejore las condiciones laborales de sus empleados se encuentren los unos en el lado negativo del eje. Ya que todas estas prácticas no son propias de empresas eco-innovadoras a priori y el estudio nos permite **confirmar nuestra hipótesis**.

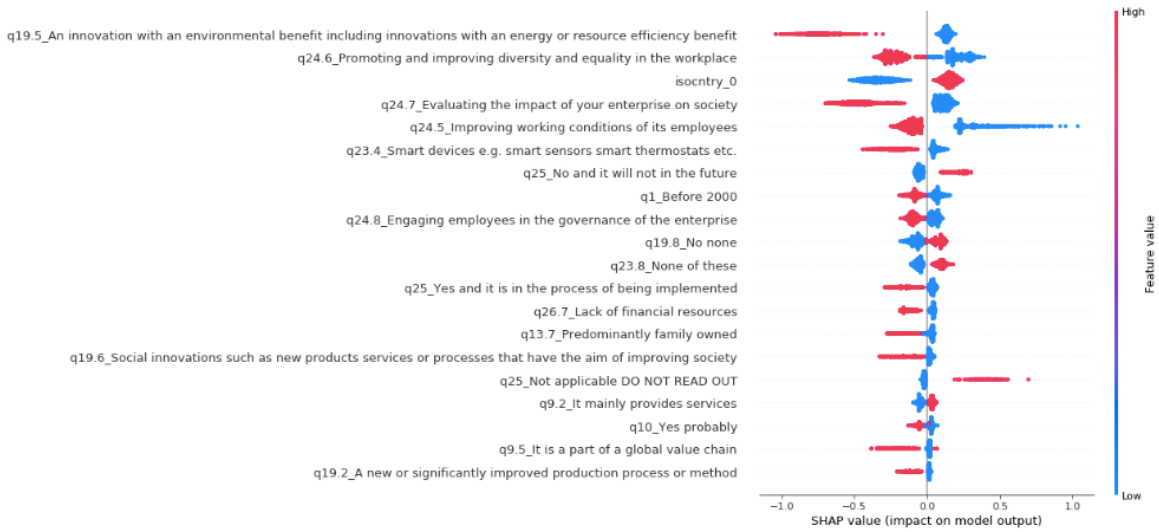

Figura 23 Valores SHAP, grupo 0. Fuente: Elaboración propia

A continuación, se presenta el mismo gráfico que el anterior, pero estamos estudiando el impacto del valor SHAP a la salida del modelo para las empresas que forman parte del grupo 1 (Figura 24). Se recuerda que las empresas que forman parte de la clase 1 son aquellas

que sí que reciclan/reúsan, pero no realizan innovación. Son empresas con tendencia eco-innovadora pero que aún no han decidido apostar al completo por su desarrollo sostenible interno por diversas razones.

Y esta afirmación anterior viene explicada por el modelo perfectamente, ya que hay tres variables que indican posibles razones por las cuales estas empresas no se deciden por innovar y formar parte de las empresas eco-innovadoras completas. Todas estas variables se encuentran con sus valores altos en la parte positiva del eje, por lo tanto, aumentan la posibilidad de que sea considerada una empresa con tendencia eco-innovadora. Las variables en cuestión son las siguientes:

- 'q25_No but it may be considered in the future'
- 'q26.4_It is not compatible with your current business model'
- 'q17.2_Regulatory obstacles or administrative burden'

Lo que se puede extraer de las tres variables superiores que a las empresas que no innovan, pero si reciclan tienen una tendencia eco-innovadora, pero por diversas razones se ven impedidos. Se llega a la conclusión que este tipo de empresas no realizan innovación porque no es compatible con su modelo actual de negocio o existen trabas burocráticas y barreras administrativas que lo impiden. **Pero la mayoría sí que considera una posibilidad innovar en el futuro**. Esto es clave, ya que podemos estar ante la respuesta a porque las empresas con un claro animo por el desarrollo sostenible no deciden innovar.

Destaca también que las empresas que fueron registradas en el registro mercantil de sus respectivos países antes del año 2000 y aquellas que forman parte de países que están por debajo de la media europea en el índice eco-innovador 2020 aumentan la probabilidad de pertenecer a este grupo de empresas. ¿Por qué tiene sentido esto? La gran mayoría de PYMEs

creadas antes del 2000 tienen un gran número de trabajadores de la vieja escuela y no están tan familiarizados con la innovación en términos ecológicos. Las pymes más jóvenes suelen ser ‘start-ups’ y es que “Un estudio realizado por Informa D&B (filial de Cesce) ha cifrado en 22.771 el número de 'start-ups' en España, lo que supone un 5% del total de empresas creadas entre 2015 y 2020 que continúan teniendo actividad. (Europapress, 2021)” Es normal que empresas con más de dos décadas de actividad a sus espaldas tenga menos tendencia innovadora que las ‘start-ups’, la cual la mayoría están pensadas para nacer ya de antemano con una mentalidad innovadora y sostenible. Es coherente que las empresas que formen parte de países como Malta o Hungría se encuentren en este grupo ya que estos países forman parte de los países europeos que menos innovan.

Por otro lado, encontramos aquellas características que reducen la posibilidad de que las empresas sean clasificadas como empresas poco eco-innovadoras. Las variables más destacadas son las siguientes:

- 'q4t_100000 euros or less'
- 'q26.7_Lack of financial resources'
- 'q26.8_None of the above'
- 'q12a_Less than 25'
- 'q7a.3_Plan to grow as a result of operating in growing markets'
- 'q24.7_Evaluating the impact of your enterprise on society'
- 'q9.6_It has a patent or patent application'
- 'q19.1_A new or significantly improved product or service to the market'
- 'q19.5_An innovation with an environmental benefit including innovations with an energy or resource efficiency benefit'

Las cuatro primeras hacen referencia reducen la posibilidad de que las empresas reciclen/reúsen, pero no innoven porque son propias de empresas no eco-innovadoras en ningún aspecto. Ya que empresas que facturen menos de 100.000€ al año o indiquen que tengan **problemas de acceso a recursos financieros suelen ser aspectos típicos de PYMEs de tamaño muy reducido** y tienen escasa capacidad ya no solo de innovar, si no de poder reciclar/reusar y causar un impacto significativo en la sociedad. La variable 'q26.8_None of the above' hace referencia a que obstáculos impiden a sus empresas ser sostenibles y contestan que ninguno de los mencionados, es decir, que hay una razón que les impide actuar en pro del medioambiente o la sociedad, pero no especifican la razón. La variable 'q12a_Less than 25' indica que empresas exportan menos de un 25% de sus servicios o bienes fuera de la UE. Las empresas con un carácter más interno (dentro de territorio UE) tienen menos tendencia a la innovación que aquellas más dadas al comercio exterior porque los procesos innovadores son demandados en todo el mundo. Además, como se ha comentado en el caso anterior, estas empresas no suelen realizar mejoras sociales para sus empleados.

El resto de las variables mencionadas anteriormente reducen la posibilidad de estar clasificadas como empresas de la clase 1, ya que son características más propias de empresas que se consideren eco-innovadoras. Las empresas que prevén crecer por trabajar con países emergentes, las que evalúan el impacto de su actividad en la sociedad, las que tienen una patente, aquellas que introducen un producto/servicio nuevo o innovador al mercado o una innovación con impacto ambiental positivo son características de empresas eco-innovadoras y de la clase 2 (empresas que reciclan/reúsan, pero también innovan).

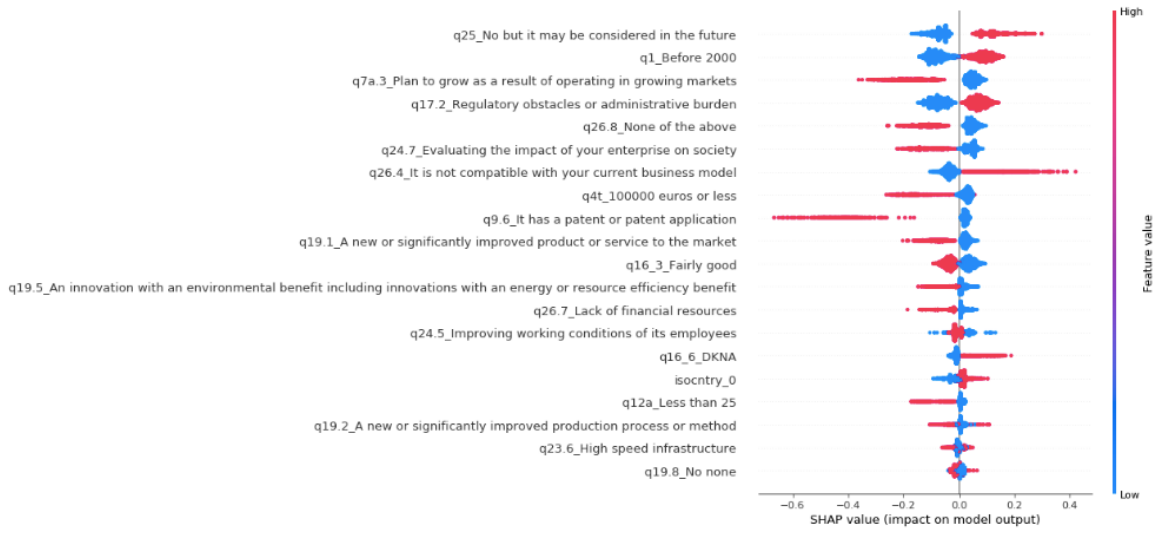

Figura 24 Valores SHAP, grupo 1. Fuente: Elaboración propia

Para el caso de las empresas que forman parte de la clase 2 (Figura 25), las totalmente eco-innovadoras se observan muchas características que evidencian un aumento de la posibilidad de ser clasificadas como empresas del grupo. Las principales variables que tienen un impacto positivo en la salida del modelo son:

- 'q24.6_Promoting and improving diversity and equality in the workplace'
- 'q24.7_Evaluating the impact of your enterprise on society'
- 'q24.5_Improving working conditions of its employees'
- 'q24.8_Engaging employees in the governance of the enterprise'
- 'q19.5_An innovation with an environmental benefit including innovations with an energy or resource efficiency benefit'
- 'q19.1_A new or significantly improved product or service to the market'
- 'q19.2_A new or significantly improved production process or method'
- 'q23.6_High speed infrastructure'
- 'q25_Yes and it has already been implemented'

- 'q9.8_It has a strategy or action plan to digitalise'
- 'q25_Yes and it is in the process of being implemented'

Las cuatro primeras variables identificadas todas tienen un carácter social claro. Ya sea promover la diversidad o la igualdad en el trabajo, mejorar sus condiciones, promocionar o implicar a los trabajadores… Sumado a la evaluación del impacto que puede tener la empresa en la sociedad en todos los ámbitos. Es deductivo pensar que las empresas que tengan una **vocación social clara** y promuevan **el bienestar de sus trabajadores**, posean asimismo una **clara intención de innovar en términos de desarrollo sostenible**. El estudio confirma esta hipótesis generalizada y no solo eso, si no que considera que es una de las características más notables de las PYMEs que presentan eco-novedades.

El resto de las variables que se encuentran en la lista superior hacen referencia explícita o implícita a la innovación. Destacan los planes de digitalización, ya haber realizado previamente innovaciones sostenibles, la mejora de productos o servicios, las innovaciones con un impacto positivo en el medioambiente y la implementación de una infraestructura de alta velocidad. Palmariamente estas variables representan a las empresas eco-innovadoras y aumentan la posibilidad de ser clasificadas en concordancia.

La investigación realizada y sus resultados plasman que la innovación y las acciones sociales tanto dentro como fuera de las empresas deben ir juntas para poder lograr el objetivo común marcado por las agendas internacionales: Economías prosperaras, pero comprometidas con el medio ambiente y la sociedad.

Existen dos variables que también representan unos a la derecha del eje y por tanto tienen un impacto positivo en la salida del modelo. Pero estas dos variables son especiales. Evidencian que las empresas eco-innovadoras también presentan problemas y no es todo

sencillo. Estas empresas encuentran como **principal barrera la falta de recursos financieros**. Y es que esta característica también está presente en los distintos grupos, pero de distinta forma. Aquellas empresas que no innovaban no destacaban que fuera por problemas financieros porque su tamaño es tan reducido que no tienen capacidad de innovación, por barreras más restrictivas como puede ser las administrativas o la imposibilidad de adaptar su modelo de negocio a lo que demanda la sociedad. En este caso, nos encontramos que la principal barrera que se encuentran las empresas que sí que innovan, es el acceso a los recursos financieros que les permita poder llevar a cabo con seguridad las inversiones necesarias para poder innovar con una perspectiva claramente sostenible. Asimismo, destacan también como posible barrera las dificultades de predecir la respuesta del mercado a las innovaciones. dos variables en cuestión son las siguientes:

- 'q26.7_Lack of financial resources'
- 'q20.3_Difficulties in predicting the market response'

Por el lado contrario, igualmente contamos con variables que nos indican que reducen la casualidad de ser una empresa comprometida con el medioambiente. Esas características son las siguientes:

- 'isocntry_0'
- 'q25_No and it will not in the future'
- 'q25_No but it may be considered in the future'
- 'q19.8_No none'
- 'q9.2_It mainly provides services'
- 'q16_6_DKNA'

Tal y como se ha comentado con anterioridad, la variable 'isocntry0' representa a los países cuyo índice eco-innovador para 2020 era inferior a la media europea. Por lo tanto, el estudio confirma que aquellas empresas que formen parte de los países que se incluye en dicho grupo, tengan menos facultades de ser catalogadas como eco-innovadoras. Esto no quiere decir que no existan empresas de países como Eslovenia que no sean eco-innovadoras, si no que tienen menos tendencia a serlo por diversos motivos los cuales no son objeto de estudio.

Destaca que tanto aquellas que contestan que en la actualidad no innovan y no plantean hacerlo en el futuro y las que sí que pueden considerarlo en el futuro expliquen lo mismo. Esto se debe porque en el presente no innovan, pero las causas que puedan hacer que innoven o no en el futuro será debido a la imposibilidad de hacer casar innovación medioambiental y rentabilidad de negocio o a barreras administrativas/financieras. Los valores representan el presente y permiten hacer lecturas de previsión el futuro, pero el modelo nos viene a decir que la posibilidad de poder innovar sosteniblemente en un futuro para las empresas que en la actualidad no lo hagan, son complicadas si no se hace nada desde los órganos competentes.

Las empresas cuyo principal negocio son los servicios es complicado que puedan realizar innovación ecológica debido a su naturaleza, ya que las principales innovaciones suelen ser de producto o en la cadena de producción. Es **normal las PYMEs del sector servicios se categoricen como empresas que no innoven**, pero no quiere decir que no reciclen/reúsen o realicen otras prácticas sostenibles como acciones sociales o medioambientales.

Para finalizar, la variable 'q16_6_DKNA' hace referencia a quienes que contestan no sabe/no contesta a como valorarían en su entorno empresarial la disponibilidad de apoyo (de cualquier tipo) a las empresas de su sector a ser más sostenibles. Esto hace ver la inseguridad, la desconfianza y el desconcierto que tienen las empresas respecto al apoyo que reciben. Aquellas que tienen dudas ya sea a nivel económico o administrativo, y no sabrían ni siquiera catalogar su entorno entre valores de muy bueno a muy malo, evidencia lo crítico del asunto. Es necesario cambiar el parecer de las empresas; darles facilidades, seguridades y un entorno próspero y cooperativo con el fin de desarrollar un modelo económico y sostenible en términos sociales y medioambientales.

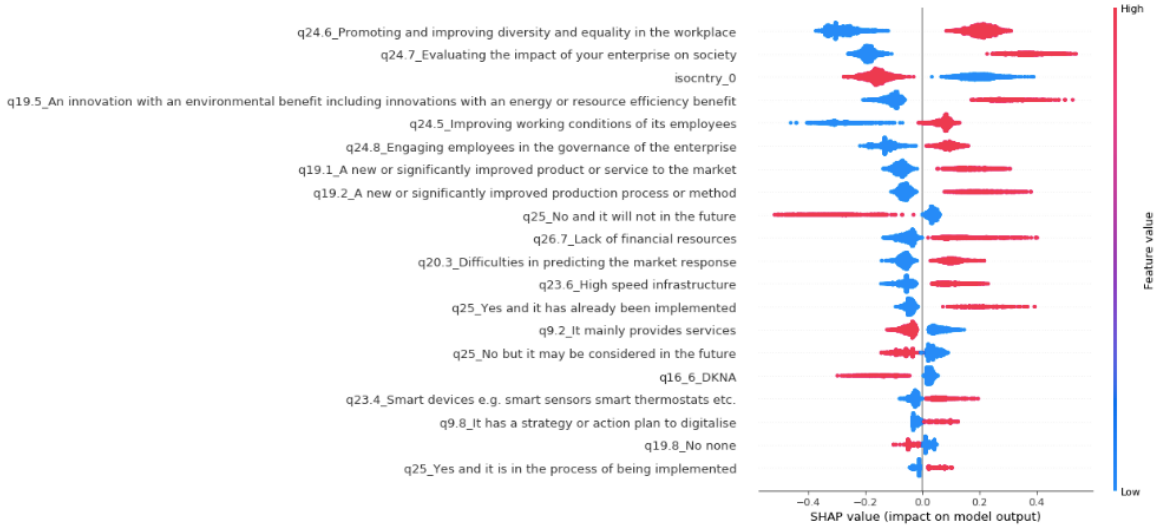

Figura 25 Valores SHAP, grupo 1. Fuente: Elaboración propia

Para finalizar con el apartado de resultamos contamos con este gráfico en cascada (Figura 26). “Los diagramas de cascada están diseñados para mostrar explicaciones de predicciones individuales, por lo que esperan una sola fila de un objeto explicación como entrada” (Lundberg, 2018). En el eje encontramos el valor esperado a la salida del modelo y luego el valor en que aparece en cada variable representa la contribución positiva al modelo

(rojo) y las negativas (azul). Este diagrama varía el valor de salida esperada del modelo completo sobre el conjunto total de datos a la salida del modelo para la predicción en concreto. Este gráfico se llama de interpretaciones locales. Esto es la interpretación para un caso particular (para una empresa concreta) cómo se han asignado los valores SHAP y como han afectado a su clasificación en un determinado grupo. En el caso de varios grupos se puede sacar un gráfico para cada grupo y por tanto nos daría el SHAP para ser clasificado en ese grupo concreto (el que más SHAP tiene sería el que nos marcaría en qué grupo se clasifica). Es decir que, si estuviéramos analizando una empresa del grupo 0, es decir, una empresa NO eco-innovadora se observarían valores SHAP altos para variables características de dicho grupo, y bajos (en azul) los que serían características del resto.

En el gráfico aparecen las 9 variables que más contribuyen a la explicación del modelo, y los 28 restantes que resultaron ser importantes según el algoritmo BorutaSHAP, pero menos contribuyen al modelo, han sido unidas en una sola variable al final del gráfico.

Como se puede observar, la variable con un impacto positivo más determinante ha sido la 'q24.5_Improving working conditions of its employees' y la segunda la variable 'q24.6_Promoting and improving diversity and equality in the workplace'. Ambas variables son de carácter social, lo que nos lleva a pensar que **las empresas más eco-innovadoras también son las más comprometidas con sus trabajadores y sus condiciones**. También destaca la variable que guarda las empresas que realizan innovaciones con un beneficio medioambiental. Por otro lado, el país del cual forman parte las empresas tiene su impacto positivo ya que ayuda a clasificar a priori a las empresas según su naturaleza eco-innovadora. Por lo general, hay un mayor número de variables con impacto positivo que negativo.

Asimismo, contamos con aquellas variables con un impacto negativo a la salida del modelo para todos los grupos. En este caso contamos con las variables: 'q25_Yes and it has already been implemented', 'q1_Before 2000' y 'q20.3_Difficulties in predicting the market response'. La desviación es mínima, pero todo parece indicar que la fecha de registro de las PYMEs o sus dificultades para predecir la respuesta del mercado no aportan evidencias claras para consideran a las empresas eco-innovadoras.

En el caso de la variable 'q25_Yes and it has already been implemented', es lógico que si una empresa ya tiene implementada innovaciones ecológicas en su empresa no expliquen el modelo correctamente de manera individual ya que también se tiene en cuenta en este gráfico el conjunto de empresas que no innovan y aparece un 0 en esta variable.

En este caso, estamos ante una empresa (la 48 en la lista) y queremos ver cómo influye el valor de SHAP para clasificar en la clase 0. La conclusión que sacamos, tal y como hemos comentado con anterioridad, esta empresa no corresponde con una del grupo 0. Las empresas de este grupo no tienen como características mejorar las condiciones de sus empleados ni innovaciones ecológicas.

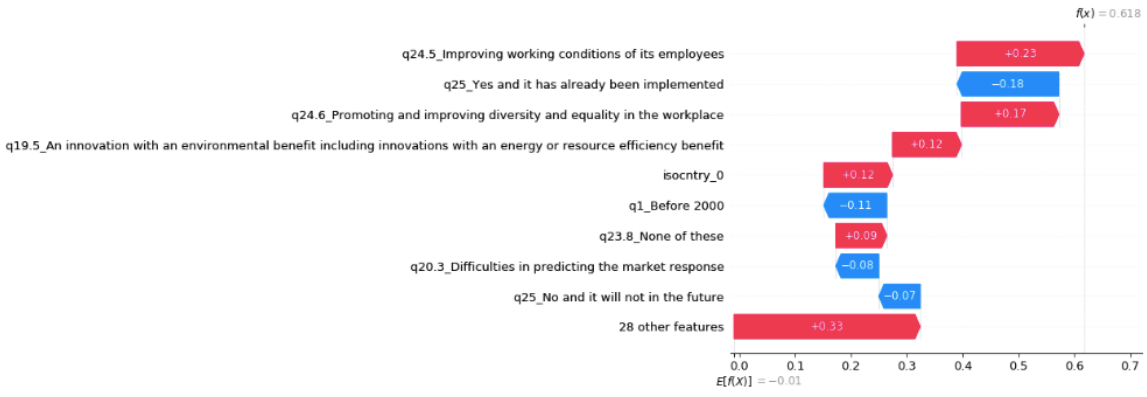

Figura 26 Interpretaciones locales valores de SHAP. Fuente: Elaboración propia

7. CONCLUSIÓN

El propósito principal del estudio era investigar y definir cuales son las características de las pequeñas y medianas empresas de la unión europea que realizan eco-innovación. De la mano de este propósito inicial está la intención de ser una investigación útil para las empresas y los organismos con el fin de conseguir una transición ecológica lo más rápida y eficiente posible. Resulta **esencial conocer el entorno y las características de las empresas** ante la innovación medioambiental y social para poder actuar posteriormente con precisión. De ahí la importancia de que los resultados extraídos y su análisis sean coherentes y clarificadores.

Basándonos en los resultados obtenidos a lo largo de la investigación, podemos afirmar que encontramos un variado abanico de características que definen con exactitud a las empresas eco-innovadoras. Como se ha podido comprobar, no solo a lo largo del estudio, si no la vida en general, las acciones positivas (ya sean internas o externas, sociales o no) que realizan las empresas tienen un impacto mayúsculo en la sociedad. Estas acciones no solo son buenas para el medio ambiente, si no que realzan la imagen de las pymes y les permite crecer. A continuación, se muestra una breve lista con las características principales que definen a las empresas eco-innovadoras:

- Empresas cuya sede social está en un país puntero en innovación ecológica
- Aquellas que mejoran o añaden nuevos productos o servicios al mercado
- Empresas que valoren los impactos sociales y ecológicos en la sociedad
- Que apliquen practicas sociales como: promover igualdad, salarios equitativos, condiciones óptimas para los trabajadores…

El estudio no solo se ha basado en conocer sus características, se ha ido más allá y también podemos extraer cuales son los problemas a los que se enfrentan estas sociedades de cara a poder innovar en términos ecológicos. Destacan barreras de entrada como podría ser la falta de financiación o problemas estratégicos como poder predecir la respuesta del mercado, ambos muy propios de las PYMEs. Sabiendo esto, se pueden tomar medidas para poder revertir la situación y poner al motor de Europa y sobre todo de España, en la **vanguardia mundial de la eco-innovación.**

Además de analizar las características de las PYMEs eco-innovadoras, asimismo se ha observado y sintetizado la situación de aquellas que no realizan en absoluto innovación ecológica y aquellas que se encuentran a mitad camino. La principal diferencia que encontramos es que aquellas que a fecha de estudio no realizan innovación, no consideran hacerlo en un futuro, pero las que si que realizan algo de eco-innovación tienen previsto lanzarse de lleno a ser empresas "ecofriendly". Los principales inconvenientes son las trabas burocráticas y su dificultad de hacer compatible la innovación con su modelo de negocio actual. Este último grupo de empresas debe ser objetivo para las autoridades competentes ya que, de la muestra empleada, aproximadamente el 50% de las empresas forman parte de este grupo. La muestra es representativa de las del junto de las PYMEs europeas, por lo tanto, debería ser objetivo para aquellas entidades competentes, inclinar a esta gran cantidad de sujetos hacia el lado puramente innovador de la balanza.

Se ha llegado a la conclusión de que las acciones sociales en relación a sus empleados y la sociedad están completamente relacionadas con la innovación ecológica. Queda demostrado que las empresas que invierten en sus empleados promueven la igualdad y se preocupa por su impacto en la sociedad son las que más invierten en el medioambiente. Es

lógico pensar que una empresa que no está comprometida con el planeta, pueda no estarlo con sus trabajadores o la sociedad. Se establece como objetivo buscar e incentivar empresas con prácticas sociales claras y que no innoven, ya ha quedado demostrado que estas empresas son proclives a ser empresas verdes de cara al futuro.

Cabe destacar la utilidad de la inteligencia artificial, concretamente del machine learning, para predecir con precisión el modelo planteado ante la gran cantidad de datos que se manejan en el estudio. La herramienta BorutaSHAP permite reducir la cantidad de variables sin dejar de definir correctamente el modelo, si no haciéndolo más concreto y exacto. XGBoost permite conocer la precisión del modelo y tras su entrenamiento, mejorar la precisión y nos extrae cuales son las características más relevantes a la salida del mismo. Sería inalcanzable para un único ser humano, incluso para un grupo de investigadores reducido poder abordar el problema sin usar herramientas de inteligencia artificial. Y es, por tanto, satisfactorio poder tener tales herramientas tan potentes, de manera totalmente gratuita y se invita, a todas aquellos investigadores, empresas o entidades que trabajen con una gran cantidad de datos a usar los métodos empleados en la investigación.

Para finalizar, la idea de este estudio es que sea dinámico y se actualice periódicamente. Ya que las empresas y sus entornos evolucionan. Si se decide emplear los resultados y los análisis de la investigación a largo plazo, serían válidos, pero sería conveniente tener en cuenta los tiempos y una actualización de los datos, aplicando los mismos métodos y procesos aplicados.

8. LIMITACIONES AL ESTUDIO

El presente estudio ha sido realizado con la última base de datos disponible a fecha de la realización del estudio, es decir, en una época concreta (es como una foto fija). El estudio es previo al COVID y a otros factores importantes a tener en cuenta como el conflicto ucraniano o los problemas energéticos. Todos estos sucesos están transformando el entorno al cual se enfrentan las empresas y las condiciones de hoy, puede diferenciar mucho a la de los datos del estudio. Sería interesante ver o evaluar cómo evolucionan las acciones eco-innovadoras a lo largo del tiempo y cómo evolucionan las características que definen ese comportamiento.

Pueden existir diferencias importantes en función del sector al que pertenezca la empresa, principalmente entre sectores de alta tecnología contra los que operan con baja tecnología o en función de si la empresa es de manufactura o de servicios (Peiro-Signes, sciencedirect, 2022). También puede haber diferencias destacables en función del nivel eco-innovador del país. Ya que como hemos demostrado, Isocountry es un elemento importante y diferencial. Podría ser que otras agrupaciones dentro de los países europeos revelen distintas conclusiones que podrían ser relevante. Como, por ejemplo: Países del sur de Europa vs Países del norte de Europa o en función del GDP per cápita.

Debido a que BorutaSHAP, XGBoost… son extensiones de Python al lanzador Anaconda y existen distintas versiones y modelos, se recomienda usar las mismas extensiones a todos aquellos miembros de un trabajo conjunto, ya que se podría llegar a conclusiones diferentes con los mismos datos.

9. REFERENCIAS

Anfevi. (s.f.). *anfevi*. Obtenido de Noticias: http://www.anfevi.com/news/la-tasa-de-reciclaje-alcanza-el-73-en-europa/#:~:text=Dinamarca%2C%20Suecia%2C%20B%C3%A9lgica%2C%20Luxemburgo,hasta%20un%2098%25%20en%20Dinamarca

Aznar, P. (21 de Julio de 2020). *Applesfera*. Obtenido de applesfera.com/apple-1/apple-anuncia-medidas-para-ser-compania-cero-impacto-medioambiente-2030#:~:text=Apple%20redujo%20su%20huella%20de,en%20un%2073%20por%20ciento

Cepal. (s.f.). *cepal*. Obtenido de https://www.cepal.org/es/temas/agenda-2030-desarrollo-sostenible/acerca-la-agenda-2030-desarrollo-sostenible

Comisión europea. (Septiembre de 2020). *europa.eu*. Obtenido de Eurobarometer: https://europa.eu/eurobarometer/surveys/detail/2244

CTMA consultores. (9 de Enero de 2020). *ctmaconsultores*. Obtenido de https://ctmaconsultores.com/la-empresa-y-el-medio-ambiente/

Euronews en español. (9 de Junio de 2022). *euronews*. Obtenido de Noticias de europa: https://es.euronews.com/my-europe/2022/06/09/prohibido-vender-coches-de-combustion-a-partir-de-2035-en-la-ue-un-objetivo-realmente-posi

Europapress. (16 de Junio de 2021). *europapress*. Obtenido de https://www.europapress.es/economia/noticia-numero-start-ups-espana-ascendio-23000-informa-db-20210616135407.html

Garrett, C. (29 de Junio de 2022). *climate.selectra*. Obtenido de https://climate.selectra.com/es/que-es/desarrollo-sostenible

Google. (s.f.). *sustainability.google*. Obtenido de https://sustainability.google/intl/es-419/progress/#

Hair, J. (Febrero de 2014). *research gate*. Obtenido de https://www.researchgate.net/publication/258046807_Partial_Least_Squares_Structural_Equation_Modeling_PLS-SEM_An_Emerging_Tool_for_Business_Research

itReseller Tech&Consulting. (29 de Julio de 2022). *itreseller*. Obtenido de https://www.itreseller.es/pyme/2022/07/8-de-cada-10-pymes-confian-en-la-digitalizacion-como-via-para-aumentar-sus-ingresos

Joseph, R. (30 de Diciembre de 2018). Obtenido de Towars Data Science: https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

Lundberg, S. (2018). *shap.readthedocs*. Obtenido de https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/waterfall.html

Masson-Delmotte. (2019). *ipcc.* Obtenido de https://www.ipcc.ch/site/assets/uploads/2019/11/SRCCL-Full-Report-Compiled-191128.pdf

Ministerio de trabajo y economia social. (27 de Mayo de 2022). *sepe*. Obtenido de https://www.sepe.es/HomeSepe/que-es-el-sepe/comunicacion-institucional/noticias/detalle-

noticia.html?folder=/2022/Mayo/&detail=Convocatoria_de_ayudas_destinadas_a_nuevos_proyectos_empresariales_de_empresas_innovadoras_Programa_NEOTEC

Ministerio para la transición ecologica, Gobierno de España. (2018). *miteco.* Obtenido de https://www.miteco.gob.es/es/ministerio/servicios/publicaciones/memoriaanualmiteco2018_tcm30-509805.pdf

Mohorte. (14 de Mayo de 2020). *magnet.xakata*. Obtenido de magnet: https://magnet.xataka.com/en-diez-minutos/100-personas-responsables-71-emisiones-contaminantes-mapa-1#:~:text=Lo%20ilustr%C3%B3%20hace%20dos%20a%C3%B1os,grandes%20de%20la%20econom%C3%ADa%20global

Nvidia. (s.f.). Obtenido de Nvidia: https://www.nvidia.com/en-us/glossary/data-science/xgboost/

Peiro-Signes, A. (Febrero de 2022). *sciencedirect*. Obtenido de https://www.sciencedirect.com/science/article/pii/S0038012121001373?via%3Dihub

Peiro-Signes, A., & Segarra-Oña, M. (2014, 2015). *ecommons*. Obtenido de https://ecommons.cornell.edu/bitstream/handle/1813/71896/Verma11_The_impact_of_environmental_certification.pdf?sequence=1

PWC (PriceWaterhouse&Coopers). (s.f.). *pwc*. Obtenido de https://www.pwc.es/es/fondos-europeos-next-generation/fondos-europeos-transicion-ecologica.html

Robledano, A. (23 de Septiembre de 2019). *Open Webinars*. Obtenido de Lenguajes de programación: https://openwebinars.net/blog/que-es-python/

rss (Responsabilidad social empresarial y sustenibilidad). (8 de Enero de 2022). *responsabilidadsocial*. Obtenido de https://responsabilidadsocial.net/3r-la-regla-de-las-tres-erres-reducir-reciclar-y-reutilizar/?amp

Scikit-learn. (s.f.). *scikit-learn*. Obtenido de http://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html

Significados. (s.f.). *significados*. Obtenido de https://www.significados.com/regla-de-las-3-r-reducir-reutilizar-reciclar/#:~:text=Qu%C3%A9%20es%20Regla%20de%20las,(Reducir%2C%20Reutilizar%2C%20Reciclar)%3A&text=Con%20esta%20serie%20de%20acciones,resultado%20de%20la%20acci%C3%B3n%20humana

Sukumar, R. (30 de Marzo de 2020). *medium*. Obtenido de https://medium.com/analytics-vidhya/shap-part-3-tree-shap-3af9bcd7cd9b

UN (Naciones Unidas). (s.f.). *un*. Obtenido de https://www.un.org/sustainabledevelopment/es/development-agenda/

Unión europea. (2021). Obtenido de europa.eu: https://ec.europa.eu/environment/ecoap/indicators/index_en

Yanes, J. (11 de Abril de 2022). *bbvaopenmind*. Obtenido de Tecnologia&Innovación: https://www.bbvaopenmind.com/tecnologia/innovacion/5-tecnologias-verdes-sostenibles/

10. ANEXOS

10.1 Anexo ODS

ANEXO

OBJETIVOS DE DESARROLLO SOSTENIBLE

Reflexión sobre la relación del TFG con los ODS en general y con el/los ODS más relacionados.

"Los **Objetivos de Desarrollo Sostenible** (ODS) constituyen un llamamiento universal a la acción para poner fin a la pobreza, proteger el planeta y mejorar las vidas y las perspectivas de las personas en todo el mundo". (UN (Naciones Unidas), s.f.) Pero para entender los ODS, no podemos dejar de lado a la agenda 2030 y es que "Los países miembros de la ONU acordaron 17 objetivos como parte de la agenda 2030. La agenda, aprobada en 2015 por la asamblea general de las Naciones Unidas, establece una visión transformadora hacia la sostenibilidad económica, social y ambiental además de ser la guía de referencia para la comunidad internacional hasta 2030." (Cepal, s.f.) La agenda pone la dignidad y la dignidad de todas las personas en el centro, junto al desarrollo sostenible necesario para progresar de la mano del planeta. De todos los objetivos planteados para cumplir con la agenda, los relacionados con el estudio son los siguientes:

- Objetivo 7 – Energía asequible y no contaminante
- Objetivo 8 – Trabajo decente y crecimiento económico
- Objetivo 9 – Industria, innovación e infraestructura

- Objetivo 11 – Ciudades y comunidades sostenibles
- Objetivo 12 – Producción y consumo responsables
- Objetivo 13 – Acción por el clima
- Objetivo 15 – Vida de ecosistemas terrestres
- Objetivo 17 – Alianzas para lograr los objetivos

Cualquier objetivo sostenible relacionado con el medioambiente, el bienestar del trabajador y la igualdad juega un papel importante a lo largo del trabajo. Haciendo hincapié en el objetivo 9 (Industria, innovación e infraestructura) y el objetivo 13 (Acción por el clima). Ya que el objetivo del estudio es dar solución a ambos, mediante lo que llamamos eco-innovación. No podemos dejar de lado el objetivo 8 (Trabajo decente y crecimiento económico), debido a que el estudio también se centra en las condiciones laborales, de la mano de un progreso económico.

Desde las instituciones públicas europeas y el gobierno español, se hace mucho eco sobre el pacto verde y la transición ecológica, pero ¿Qué es realmente la transición ecológica? "La transición ecológica se ocupa de la propuesta y ejecución de las políticas del gobierno en materia de energía y medioambiente para la transición a un modelo productivo y social más ecológico" (Ministerio para la transición ecologica, Gobierno de España, 2018). Tras la pandemia y los problemas geopolíticos y energéticos, se ha propulsado este proyecto con el fin de ser más sostenibles y menos dependientes de las energías no renovables.

En el capítulo 2 – "Descripción del contexto", concretamente en la sección 2.4 – "Desarrollo sostenible, ODS y agenda 2030" se comenta en profundidad los ODS y su relación con el trabajo.

10.2 Anexo códigos

STEP 1

PARTE 1. CARGAMOS LAS LIBRERÍAS

In [1]:

```
#https://amirali-n.github.io/BorutaFeatureSelectionWithShapAnalysis/

# First XGBoost model for Pitec dataset
import matplotlib.pyplot as plt
import xgboost as xgb
from numpy import loadtxt
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
# XGBoost kfold cross validation
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
# XGBoost stratified kfold cross validation
from sklearn.model_selection import StratifiedKFold
# one hot ecoding
from numpy import column_stack
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
# pandas
import pandas as pd
```

PARTE 2. CARGAMOS EL FICHERO DE DATOS Los ficheros tienen que estar en el mismo directorio En este caso tenemos un fichero para cada país. La mayoría de variables coinciden excepto algunas variables que habrá que tratar Por ejemplo la forma de agrupar los códigos nace: Algunos países utilizan nace_a y otros nace_b O la forma de ver el tamaño de la empresa size_1, sice_2, sice_3

In [2]:

```
# Read data from file 'filename.csv'
# (in the same directory that your python process is based)
# Control delimiters, rows, column names with read_csv (see later) sep=se
parador en el csv decimal=separador decimal de los números
df = pd.read_spss("ZA7637_v2-0-0.sav")
df.shape
```

Out[2]:

```
(16365, 385)
```

In [3]:

```
my_list = df.columns.values.tolist()
print(my_list)
```

```
['studyno', 'doi', 'version', 'edition', 'survey', 'caseid', 'uniqid',
'serialid', 'tnscntry', 'country', 'isocntry', 'size', 'nace_a', 'q1',
'vq1', 'q2a', 'vq2a', 'q2b', 'q2t', 'q3a', 'vq3a', 'q3b', 'q3t', 'q4a'
, 'vq4a', 'q4b', 'q4t', 'q5_1', 'q5_2', 'q6_1', 'q6_2', 'q7a.1', 'q7a.
2', 'q7a.3', 'q7a.4', 'q7a.5', 'q7a.6', 'q7a.7', 'q7a.8', 'q7a.9', 'q7
b.1', 'q7b.2', 'q7b.3', 'q7b.4', 'q7b.5', 'q7b.6', 'q7b.7', 'q7b.8', '
q7b.9', 'q7b.10', 'q8.1', 'q8.2', 'q8.3', 'q8.4', 'q8.5', 'q8.6', 'q8.
7', 'q9.1', 'q9.2', 'q9.3', 'q9.4', 'q9.5', 'q9.6', 'q9.7', 'q9.8', 'q
9.9', 'q9.10', 'q9.11', 'q10', 'q11.1', 'q11.2', 'q11.3', 'q11.4', 'q1
1.5', 'q11.6', 'q11.7', 'q11.8', 'q11.9', 'q12a', 'q12b', 'q13.1', 'q1
3.2', 'q13.3', 'q13.4', 'q13.5', 'q13.6', 'q13.7', 'q13.8', 'q13.9', '
q13.10', 'q14.1', 'q14.2', 'q14.3', 'q14.4', 'q14.5', 'q14.6', 'q15a.1
', 'q15a.2', 'q15a.3', 'q15a.4', 'q15a.5', 'q15a.6', 'q15a.7', 'q15a.8
', 'q15a.9', 'q15a.10', 'q15a.11', 'q15b.1', 'q15b.2', 'q15b.3', 'q15b
.4', 'q15b.5', 'q15b.6', 'q15b.7', 'q15b.8', 'q15b.9', 'q15b.10', 'q15
b.11', 'q16_1', 'q16_2', 'q16_3', 'q16_4', 'q16_5', 'q16_6', 'q16_7',
'q16_8', 'q17.1', 'q17.2', 'q17.3', 'q17.4', 'q17.5', 'q17.6', 'q17.7'
, 'q17.8', 'q17.9', 'q17.10', 'q18a', 'q18_open', 'q18.1', 'q18.2', 'q
18.3', 'q18.4', 'q18.5', 'q18.6', 'q18.7', 'q18.8', 'q18.9', 'q18.10',
'q18.11', 'q18.12', 'q18.13', 'q18.14', 'q18.15', 'q18.16', 'q18.17',
'q18.18', 'q18.19', 'q18.20', 'q18.21', 'q18.22', 'q18.23', 'q18.24',
'q18.25', 'q18.26', 'q18.27', 'q18.28', 'q18.29', 'q18.30', 'q18.31',
'q18.32', 'q18.33', 'q18.34', 'q18.35', 'q18.36', 'q18.37', 'q18.38',
'q18.39', 'q18.40', 'q18.41', 'q18.42', 'q18.43', 'q18.44', 'q18.45',
'q18.46', 'q18.47', 'q18.48', 'q18.49', 'q18.50', 'q18.51', 'q18.52',
'q18.53', 'q18.54', 'q18.55', 'q18.56', 'q18.57', 'q18.58', 'q18.59',
'q18.60', 'q18.61', 'q18.62', 'q18.63', 'q18.64', 'q18.65', 'q18.66',
'q18.67', 'q18.68', 'q18.69', 'q18.70', 'q18.71', 'q18.72', 'q18.73',
'q18.74', 'q18.75', 'q18.76', 'q18.77', 'q18.78', 'q18.79', 'q18.80',
'q18.81', 'q18.82', 'q18.83', 'q18.84', 'q18.85', 'q18.86', 'q18.87',
'q18.88', 'q18.89', 'q18.90', 'q18.91', 'q18.92', 'q18.93', 'q18.94',
'q18.95', 'q18.96', 'q18.97', 'q18.98', 'q18.99', 'q18.100', 'q18.101'
, 'q18.102', 'q18.103', 'q18.104', 'q18.105', 'q18.106', 'q18.107', 'q
18.108', 'q18.109', 'q18.110', 'q18.111', 'q18.112', 'q18.113', 'q18.1
14', 'q18.115', 'q18.116', 'q18.117', 'q18.118', 'q18.119', 'q18.120',
'q18.121', 'q18.122', 'q18.123', 'q18.124', 'q18.125', 'q18.126', 'q19
.1', 'q19.2', 'q19.3', 'q19.4', 'q19.5', 'q19.6', 'q19.7', 'q19.8', 'q
19.9', 'q20.1', 'q20.2', 'q20.3', 'q20.4', 'q20.5', 'q20.6', 'q20.7',
'q20.8', 'q20.9', 'q20.10', 'q20.11', 'q21.1', 'q21.2', 'q21.3', 'q21.
4', 'q21.5', 'q21.6', 'q21.7', 'q21.8', 'q21.9', 'q21.10', 'q21.11', '
q22', 'q23.1', 'q23.2', 'q23.3', 'q23.4', 'q23.5', 'q23.6', 'q23.7', '
q23.8', 'q23.9', 'q24.1', 'q24.2', 'q24.3', 'q24.4', 'q24.5', 'q24.6',
'q24.7', 'q24.8', 'q24.9', 'q24.10', 'q25', 'q26.1', 'q26.2', 'q26.3',
'q26.4', 'q26.5', 'q26.6', 'q26.7', 'q26.8', 'q26.9', 'q26.10', 'vp1m'
, 'vp1d', 'vp1', 'eu6', 'eu9', 'eu10', 'eu12', 'eu_nms3', 'eu15', 'eu_
nms10', 'eu25', 'eu_nms12', 'eu27', 'eu_nms13', 'eu28', 'eu27b', 'euro
z13', 'euronz13', 'euroz15', 'euronz15', 'euroz16', 'euronz16', 'euroz
17', 'euronz17', 'euroz18', 'euronz18', 'euroz19', 'euronzms', 'euronz
nm', 'euronz19', 'euroz', 'eu', 'w1', 'w5', 'w6', 'w7', 'w9', 'w10', '
w11', 'w13', 'w14', 'w24', 'w22', 'w94', 'w23', 'w29', 'w30', 'w81', '
```

```
w82', 'w87', 'w89', 'w90', 'w95', 'w96', 'w97', 'w83', 'w84', 'w98', '
wex']
```

```
df.head()
```

```
(…)
```

5 rows × 385 columns

```
#eliminamos las columnas que no nos interesan
df=df.drop(['studyno', 'doi', 'version', 'edition', 'survey', 'caseid', '
uniqid', 'serialid','tnscntry', 'country',
'q18a', 'q18_open','q18.1', 'q18.2', 'q18.3', 'q18.4', 'q18.5', 'q18.6',
'q18.7', 'q18.8', 'q18.9', 'q18.10', 'q18.11', 'q18.12', 'q18.13', 'q18.1
4', 'q18.15', 'q18.16', 'q18.17', 'q18.18', 'q18.19', 'q18.20', 'q18.21',
'q18.22', 'q18.23', 'q18.24', 'q18.25', 'q18.26', 'q18.27', 'q18.28', 'q1
8.29', 'q18.30', 'q18.31',
'q18.32', 'q18.33', 'q18.34', 'q18.35', 'q18.36', 'q18.37', 'q18.38', 'q1
8.39', 'q18.40', 'q18.41', 'q18.42', 'q18.43', 'q18.44', 'q18.45', 'q18.4
6', 'q18.47', 'q18.48', 'q18.49', 'q18.50', 'q18.51', 'q18.52', 'q18.53',
'q18.54', 'q18.55', 'q18.56', 'q18.57', 'q18.58', 'q18.59', 'q18.60', 'q1
8.61',
'q18.62', 'q18.63', 'q18.64', 'q18.65', 'q18.66', 'q18.67', 'q18.68', 'q1
8.69', 'q18.70', 'q18.71', 'q18.72', 'q18.73', 'q18.74', 'q18.75', 'q18.7
6', 'q18.77', 'q18.78', 'q18.79', 'q18.80', 'q18.81', 'q18.82', 'q18.83',
'q18.84', 'q18.85', 'q18.86', 'q18.87', 'q18.88', 'q18.89', 'q18.90',
'q18.91', 'q18.92', 'q18.93', 'q18.94', 'q18.95', 'q18.96', 'q18.97', 'q1
8.98', 'q18.99', 'q18.100', 'q18.101', 'q18.102', 'q18.103', 'q18.104', '
q18.105', 'q18.106', 'q18.107', 'q18.108', 'q18.109', 'q18.110', 'q18.111
', 'q18.112', 'q18.113', 'q18.114', 'q18.115', 'q18.116', 'q18.117', 'q18
.118',
'q18.119', 'q18.120', 'q18.121', 'q18.122', 'q18.123', 'q18.124', 'q18.12
5', 'q18.126',
 'eu6', 'eu9', 'eu10', 'eu12', 'eu_nms3', 'eu15', 'eu_nms10', 'eu25', 'eu
_nms12', 'eu27', 'eu_nms13', 'eu28', 'eu27b', 'euroz13', 'euronz13', 'eur
oz15', 'euronz15', 'euroz16', 'euronz16', 'euroz17', 'euronz17', 'euroz18
', 'euronz18', 'euroz19', 'euronzms', 'euronznm', 'euronz19', 'euroz', 'e
u',
          'w1', 'w5', 'w6', 'w7', 'w9', 'w10', 'w11', 'w13', 'w14', 'w24'
, 'w22', 'w94', 'w23', 'w29', 'w30', 'w81', 'w82', 'w87', 'w89', 'w90', '
w95', 'w96', 'w97', 'w83', 'w84', 'w98', 'wex',
'vq1', 'q2a', 'vq2a', 'q2b', 'q3a', 'vq3a', 'q3b','q4a', 'vq4a', 'q4b',
           'vp1m', 'vp1d', 'vp1', 'size'

          ],axis=1)
df.shape
```

Out[5]:

```
(16365, 177)
```

'vq1', 'q2a', 'vq2a', 'q2b', 'q3a', 'vq3a', 'q3b','q4a', 'vq4a', 'q4b' se elimina proque q2t recoge los valores de q2a y q2b, idem para el resto

In [6]:

```
df.head()
```

(...)

```
my_list = df.columns.values.tolist()
print(my_list)
```

```
['isocntry', 'nace_a', 'q1', 'q2t', 'q3t', 'q4t', 'q5_1', 'q5_2', 'q6_1',
'q6_2', 'q7a.1', 'q7a.2', 'q7a.3', 'q7a.4', 'q7a.5', 'q7a.6', 'q7a.7', 'q
7a.8', 'q7a.9', 'q7b.1', 'q7b.2', 'q7b.3', 'q7b.4', 'q7b.5', 'q7b.6', 'q7
b.7', 'q7b.8', 'q7b.9', 'q7b.10', 'q8.1', 'q8.2', 'q8.3', 'q8.4', 'q8.5',
'q8.6', 'q8.7', 'q9.1', 'q9.2', 'q9.3', 'q9.4', 'q9.5', 'q9.6', 'q9.7', '
q9.8', 'q9.9', 'q9.10', 'q9.11', 'q10', 'q11.1', 'q11.2', 'q11.3', 'q11.4
', 'q11.5', 'q11.6', 'q11.7', 'q11.8', 'q11.9', 'q12a', 'q12b', 'q13.1',
'q13.2', 'q13.3', 'q13.4', 'q13.5', 'q13.6', 'q13.7', 'q13.8', 'q13.9', '
q13.10', 'q14.1', 'q14.2', 'q14.3', 'q14.4', 'q14.5', 'q14.6', 'q15a.1',
'q15a.2', 'q15a.3', 'q15a.4', 'q15a.5', 'q15a.6', 'q15a.7', 'q15a.8', 'q1
5a.9', 'q15a.10', 'q15a.11', 'q15b.1', 'q15b.2', 'q15b.3', 'q15b.4', 'q15
b.5', 'q15b.6', 'q15b.7', 'q15b.8', 'q15b.9', 'q15b.10', 'q15b.11', 'q16_
1', 'q16_2', 'q16_3', 'q16_4', 'q16_5', 'q16_6', 'q16_7', 'q16_8', 'q17.1
', 'q17.2', 'q17.3', 'q17.4', 'q17.5', 'q17.6', 'q17.7', 'q17.8', 'q17.9'
, 'q17.10', 'q19.1', 'q19.2', 'q19.3', 'q19.4', 'q19.5', 'q19.6', 'q19.7'
, 'q19.8', 'q19.9', 'q20.1', 'q20.2', 'q20.3', 'q20.4', 'q20.5', 'q20.6',
'q20.7', 'q20.8', 'q20.9', 'q20.10', 'q20.11', 'q21.1', 'q21.2', 'q21.3',
'q21.4', 'q21.5', 'q21.6', 'q21.7', 'q21.8', 'q21.9', 'q21.10', 'q21.11',
'q22', 'q23.1', 'q23.2', 'q23.3', 'q23.4', 'q23.5', 'q23.6', 'q23.7', 'q2
3.8', 'q23.9', 'q24.1', 'q24.2', 'q24.3', 'q24.4', 'q24.5', 'q24.6', 'q24
.7', 'q24.8', 'q24.9', 'q24.10', 'q25', 'q26.1', 'q26.2', 'q26.3', 'q26.4
', 'q26.5', 'q26.6', 'q26.7', 'q26.8', 'q26.9', 'q26.10']
```

In [8]:

```
countries = ['AT', 'BE', 'BG', 'CY', 'CZ', 'DE', 'DK', 'EE', 'EL', 'ES',
'FI','FR', 'HR', 'HU', 'IE', 'IT', 'LT','LU', 'LV', 'MT','NL', 'PL', 'PT'
, 'RO', 'SE', 'SI', 'SK']
df=df[df.isocntry.isin(countries)]
df
```

(...)

12115 rows × 177 columns

In [9]:

```
# no hay países con alto nivel innovador en esta muestra
```

```
df['isocntry'] = df['isocntry'].replace(['AT','FI','DK','LU','SE','NL','DE','ES','FR','PT'],1)
df['isocntry'] = df['isocntry'].replace(['SK','CZ','BE','RO','PL','LT','LV','HU','EE','BG','IT', 'IE', 'HR','CY','MT','SI'],0)
```

In [10]:

```
for name in df.columns:
    print(name)
    print(df[name].value_counts())
    print("")
    print("")
```

```
isocntry
0    7409
1    4706
Name: isocntry, dtype: int64


nace_a
G - Wholesale and retail trade, repair of motor vehicles and    3257
C - Manufacturing                                               2332
F - Construction                                                1195
M - Professional, scientific and technical activities           1181
H - Transportation and storage                                   741
I - Accommodation and food service activities                    686
N - Administrative and support service activities                516
J - Information and communication                                463
Q - Human health and social work activities                      460
P - Education                                                    311
L - Real estate activities                                       294
K - Financial and insurance activities                           273
Arts, entertainment and recreation                               181
E - Water supply,sewerage,waste management/remediation activ     125
D - Electricity, gas, steam and air conditioningsupply            67
B - Mining and quarrying                                          33
Name: nace_a, dtype: int64


q1
Before 2000              5759
Between 2000 and 2014    4856
Between 2015 and 2018    1100
DK/NA                     279
2019 and after            121
Name: q1, dtype: int64
```

```
q2t
1 to 9 employees         6675
10 to 49 employees       3012
50 to 249 employees      1793
250 employees or more     635
Name: q2t, dtype: int64
```

```
q3t
1 to 9 employees                5937
Inap. (not 1 in q2a and q2b)    5440
0 employe                        345
10 to 49 employees               281
DK/NA                             89
250 employees or more             15
50 to 249 employees                8
Name: q3t, dtype: int64
```

```
q4t
More than 100,000 and up to 500,000 euros          3038
100,000 euros or less                              1913
More than 500,000 and up to 1 million euros        1530
More than 10 million and up to 50 million euros    1288
DK/NA                                              1177
More than 1 million and up to 2 million euros      1153
More than 2 million and up to 5 million euros      1038
More than 5 million and up to 10 million euros      693
More than 50 million euros                          285
Name: q4t, dtype: int64
```

```
q5_1
It has remained stable           1921
It has grown by less than 30%    1785
It has grown by at least 30%      951
It has decreased                  687
DK/NA                              96
Name: q5_1, dtype: int64
```

```
q5_2
```

```
It has grown by less than 30%    4037
It has remained stable           3377
It has grown by at least 30%     2387
It has decreased                 1774
DK/NA                             540
Name: q5_2, dtype: int64
```

```
q6_1
It does not plan to grow                5753
Grow by less than 10% per year          3182
Grow by between 10% and 20% per year    1867
Grow by more than 20% per year           764
DK/NA                                    549
Name: q6_1, dtype: int64
```

```
q6_2
Grow by less than 10% per year          3951
Grow by between 10% and 20% per year    3131
It does not plan to grow                3005
Grow by more than 20% per year          1252
DK/NA                                    776
Name: q6_2, dtype: int64
```

```
q7a.1
Not mentioned                   5408
Have a strategic growth plan    3276
Name: q7a.1, dtype: int64
```

```
q7a.2
Not mentioned                                                          5833
Plan to grow as a result of introducing some kind of innovation        2851
Name: q7a.2, dtype: int64
```

```
q7a.3
Not mentioned                                              5823
Plan to grow as a result of operating in growing markets   2861
Name: q7a.3, dtype: int64
```

```
q7a.4
Not mentioned                                          5875
Plan to grow as a result of entering new markets    2809
Name: q7a.4, dtype: int64


q7a.5
Not mentioned
6234
Plan to grow as a result of increased digitalisation in your enterprise
2450
Name: q7a.5, dtype: int64


q7a.6
Plan to grow in (OUR COUNTRY)    6468
Not mentioned                    2216
Name: q7a.6, dtype: int64


q7a.7
Not mentioned                                           6617
[EU] Plan to grow in other EU countries/ [Non-EU]    2067
Name: q7a.7, dtype: int64


q7a.8
Not mentioned                                 7467
Plan to grow in other non-EU countries    1217
Name: q7a.8, dtype: int64


q7a.9
Not mentioned    8167
DK/NA             517
Name: q7a.9, dtype: int64


q7b.1
Not mentioned
2077
There is no intention for your enterprise to grow beyond its current size
1354
Name: q7b.1, dtype: int64
```

```
q7b.2
Not mentioned
2941
Your enterprise does not have employees with the skills or expertise need
ed for it to grow     490
Name: q7b.2, dtype: int64


q7b.3
Not mentioned                                                        2744
Your enterprise does not have the financial resources to grow     687
Name: q7b.3, dtype: int64


q7b.4
Not mentioned
2224
There is decreasing demand for your enterprise's products or services or
the market is saturated    1207
Name: q7b.4, dtype: int64


q7b.5
Not mentioned
2362
Additional regulatory or administrative burdens and requirements would be
too high for your enterprise to grow    1069
Name: q7b.5, dtype: int64


q7b.6
Not mentioned
3132
Your enterprise does not want to grow because it would lose benefits link
ed to its SME status     299
Name: q7b.6, dtype: int64


q7b.7
Not mentioned
2699
```

```
The current location of your enterprise does not allow you to grow and yo
u do not wish to relocate elsewhere     732
Name: q7b.7, dtype: int64


q7b.8
Not mentioned
2425
Your enterprise relies on a few clients which are unlikely to increase th
eir demand    1006
Name: q7b.8, dtype: int64


q7b.9
Not mentioned               3174
Other (DO NOT READ OUT)      257
Name: q7b.9, dtype: int64


q7b.10
Not mentioned    3219
DK/NA             212
Name: q7b.10, dtype: int64


q8.1
Not mentioned               6421
In a large town or city    5694
Name: q8.1, dtype: int64


q8.2
Not mentioned                  7518
In a small town or village    4597
Name: q8.2, dtype: int64


q8.3
Not mentioned      10848
In a rural area     1267
Name: q8.3, dtype: int64


q8.4
```

```
Not mentioned            10574
In an industrial area     1541
Name: q8.4, dtype: int64


q8.5
Not mentioned                         10873
Near a border with an EU country      1242
Name: q8.5, dtype: int64


q8.6
Not mentioned                             11788
Near a border with a non-EU country         327
Name: q8.6, dtype: int64


q8.7
Not mentioned    12072
DK/NA               43
Name: q8.7, dtype: int64


q9.1
Not mentioned               7339
It mainly provides goods    4776
Name: q9.1, dtype: int64


q9.2
It mainly provides services    6887
Not mentioned                  5228
Name: q9.2, dtype: int64


q9.3
Not mentioned                                          11413
It sells goods online to buyers in EU countries          702
Name: q9.3, dtype: int64


q9.4
Not mentioned
10447
```

```
It is a member of an industry cluster or another SME business support org
anisation in the region     1668
Name: q9.4, dtype: int64


q9.5
Not mentioned                              10965
It is a part of a global value chain     1150
Name: q9.5, dtype: int64


q9.6
Not mentioned                               11375
It has a patent or patent application      740
Name: q9.6, dtype: int64


q9.7
Not mentioned                     11441
It is a non-profit enterprise      674
Name: q9.7, dtype: int64


q9.8
Not mentioned                                         9462
It has a strategy or action plan to digitalise    2653
Name: q9.8, dtype: int64


q9.9
Not mentioned              12015
Other (DO NOT READ OUT)      100
Name: q9.9, dtype: int64


q9.10
Not mentioned             11998
None (DO NOT READ OUT)      117
Name: q9.10, dtype: int64


q9.11
Not mentioned    12086
DK/NA               29
```

```
Name: q9.11, dtype: int64


q10
Yes, definitely                      4371
Yes, probably                        4130
No, probably not                     1624
No, definitely not                   1168
DK/NA                                 543
Not applicable (DO NOT READ OUT)      279
Name: q10, dtype: int64


q11.1
None, your enterprise only operates in (OUR COUNTRY)    7651
Not mentioned                                            4464
Name: q11.1, dtype: int64


q11.2
Not mentioned                                        8142
[EU] Other EU countries/ [Non-EU] EU countries       3973
Name: q11.2, dtype: int64


q11.3
Not mentioned                                                  10484
Other European countries outside of the EU (incl. Russia)       1631
Name: q11.3, dtype: int64


q11.4
Not mentioned    11370
North America      745
Name: q11.4, dtype: int64


q11.5
Not mentioned                       11627
Latin America and the Caribbean       488
Name: q11.5, dtype: int64


q11.6
```

```
Not mentioned    11576
China              539
Name: q11.6, dtype: int64


q11.7
Not mentioned                   11427
Rest of Asia and the Pacific      688
Name: q11.7, dtype: int64


q11.8
Not mentioned             11428
Middle East and Africa      687
Name: q11.8, dtype: int64


q11.9
Not mentioned    11907
DK/NA              208
Name: q11.9, dtype: int64


q12a
Less than 25%          1118
Between 25% and 50%     314
More than 50%           290
DK/NA                    89
Name: q12a, dtype: int64


q12b
Between 25% and 50%    0
DK/NA                  0
Less than 25%          0
More than 50%          0
Name: q12b, dtype: int64


q13.1
Not mentioned                 7537
Solely owned by one person    4578
Name: q13.1, dtype: int64
```

```
q13.2
Not mentioned                    6707
Owned by more than one person    5408
Name: q13.2, dtype: int64


q13.3
Not mentioned                                           10971
Part of a national or international enterprise group     1144
Name: q13.3, dtype: int64


q13.4
Not mentioned                   11698
Co-owned by a public entity       417
Name: q13.4, dtype: int64


q13.5
Not mentioned                        11983
Co-owned by venture capital firm       132
Name: q13.5, dtype: int64


q13.6
Not mentioned                 12018
Co-owned by business angel       97
Name: q13.6, dtype: int64


q13.7
Not mentioned                 9719
Predominantly family owned    2396
Name: q13.7, dtype: int64


q13.8
Not mentioned                                                          11588
Jointly owned by its members (e.g. cooperative, mutual society)          527
Name: q13.8, dtype: int64


q13.9
```

```
Not mentioned              11868
Other (DO NOT READ OUT)      247
Name: q13.9, dtype: int64


q13.10
Not mentioned    12054
DK/NA               61
Name: q13.10, dtype: int64


q14.1
Not mentioned                            9028
The sole founder of this enterprise    3087
Name: q14.1, dtype: int64


q14.2
Not mentioned                         9556
A co-founder of this enterprise    2559
Name: q14.2, dtype: int64


q14.3
Not mentioned                           9538
The sole owner of this enterprise    2577
Name: q14.3, dtype: int64


q14.4
Not mentioned                       8945
A co-owner of this enterprise    3170
Name: q14.4, dtype: int64


q14.5
Not mentioned        7747
None of the above    4368
Name: q14.5, dtype: int64


q14.6
Not mentioned    12069
DK/NA               46
```

```
Name: q14.6, dtype: int64


q15a.1
This is the first enterprise that you have ever established    3720
Not mentioned                                                  1926
Name: q15a.1, dtype: int64


q15a.2
Not mentioned                                                  3432
You have established or co-established other enterprise(s)    2214
Name: q15a.2, dtype: int64


q15a.3
Not mentioned
4749
You have closed - without bankruptcy - other enterprise(s) that you owned
or co-owned     897
Name: q15a.3, dtype: int64


q15a.4
Not mentioned
5444
You have closed - due to bankruptcy - other enterprise(s) that you owned
or co-owned     202
Name: q15a.4, dtype: int64


q15a.5
Not mentioned                                                  5016
You have sold other enterprise(s) that you owned or co-owned     630
Name: q15a.5, dtype: int64


q15a.6
Not mentioned
5525
You plan to relocate the headquarters of your enterprise to an EU country
in the future     121
Name: q15a.6, dtype: int64
```

```
q15a.7
Not mentioned
5627
You plan to relocate the headquarters of your enterprise to the USA in th
e future      19
Name: q15a.7, dtype: int64


q15a.8
Not mentioned
5587
You plan to relocate the headquarters of your enterprise to any other cou
ntry in the future      59
Name: q15a.8, dtype: int64


q15a.9
Not mentioned               5628
Other (DO NOT READ OUT)       18
Name: q15a.9, dtype: int64


q15a.10
Not mentioned             5578
None (DO NOT READ OUT)      68
Name: q15a.10, dtype: int64


q15a.11
Not mentioned    5632
DK/NA              14
Name: q15a.11, dtype: int64


q15b.1
Not mentioned                                              1269
You took this enterprise over from family member(s)     786
Name: q15b.1, dtype: int64


q15b.2
Not mentioned                                                 1425
You have established or co-established other enterprises     630
```

```
Name: q15b.2, dtype: int64


q15b.3
Not mentioned
1887
You have closed - without bankruptcy - other enterprise(s) that you owned
or co-owned     168
Name: q15b.3, dtype: int64


q15b.4
Not mentioned
2015
You have closed - due to bankruptcy - other enterprise(s) that you owned
or co-owned      40
Name: q15b.4, dtype: int64


q15b.5
Not mentioned                                                    1884
You have sold other enterprise(s) that you owned or co-owned     171
Name: q15b.5, dtype: int64


q15b.6
Not mentioned
2039
You plan to relocate the headquarters of your enterprise to an EU country
in the future      16
Name: q15b.6, dtype: int64


q15b.7
Not mentioned
2050
You plan to relocate the headquarters of your enterprise to the USA in th
e future       5
Name: q15b.7, dtype: int64


q15b.8
Not mentioned
2042
```

```
You plan to relocate the headquarters of your enterprise to any other cou
ntry in the future      13
Name: q15b.8, dtype: int64


q15b.9
Not mentioned              1955
Other (DO NOT READ OUT)     100
Name: q15b.9, dtype: int64


q15b.10
Not mentioned             1480
None (DO NOT READ OUT)     575
Name: q15b.10, dtype: int64


q15b.11
Not mentioned    2036
DK/NA              19
Name: q15b.11, dtype: int64


q16_1
Fairly good    7404
Very good      2165
Fairly poor    1561
DK/NA           622
Very poor       363
Name: q16_1, dtype: int64


q16_2
Fairly good    5389
Fairly poor    2310
DK/NA          2151
Very good      1421
Very poor       844
Name: q16_2, dtype: int64


q16_3
Fairly good    5997
Fairly poor    2336
```

```
DK/NA          1939
Very good      1060
Very poor       783
Name: q16_3, dtype: int64


q16_4
Fairly good    6923
Very good      2004
Fairly poor    1544
DK/NA          1257
Very poor       387
Name: q16_4, dtype: int64


q16_5
Fairly good    5230
Fairly poor    3618
Very good      1581
Very poor      1289
DK/NA           397
Name: q16_5, dtype: int64


q16_6
Fairly good    4725
Fairly poor    3171
DK/NA          2099
Very poor      1197
Very good       923
Name: q16_6, dtype: int64


q16_7
Fairly good    6708
Fairly poor    2223
Very good      1658
DK/NA           793
Very poor       733
Name: q16_7, dtype: int64


q16_8
Fairly good    6406
```

```
Very good      3924
Fairly poor    1061
DK/NA           451
Very poor       273
Name: q16_8, dtype: int64


q17.1
Not mentioned                   10982
Difficulties with innovation     1133
Name: q17.1, dtype: int64


q17.2
Regulatory obstacles or administrative burden    6294
Not mentioned                                    5821
Name: q17.2, dtype: int64


q17.3
Not mentioned     11270
Access to data      845
Name: q17.3, dtype: int64


q17.4
Not mentioned           11178
Internationalisation      937
Name: q17.4, dtype: int64


q17.5
Not mentioned        9887
Access to finance    2228
Name: q17.5, dtype: int64


q17.6
Not mentioned     8098
Payment delays    4017
Name: q17.6, dtype: int64


q17.7
```

```
Not mentioned                            9501
Skills, including managerial skills      2614
Name: q17.7, dtype: int64


q17.8
Not mentioned                           10781
Difficulties with digitalisation         1334
Name: q17.8, dtype: int64


q17.9
Not mentioned                 11423
Other (DO NOT READ OUT)         692
Name: q17.9, dtype: int64


q17.10
Not mentioned    10798
DK/NA             1317
Name: q17.10, dtype: int64


q19.1
Not mentioned                                                        8765
A new or significantly improved product or service to the market     3350
Name: q19.1, dtype: int64


q19.2
Not mentioned                                                       9729
A new or significantly improved production process or method        2386
Name: q19.2, dtype: int64


q19.3
Not mentioned                                                  10186
A new organisation of management or a new business model        1929
Name: q19.3, dtype: int64


q19.4
Not mentioned                                    9691
A new way of selling your goods or services      2424
```

```
Name: q19.4, dtype: int64


q19.5
Not mentioned
9334
An innovation with an environmental benefit, including innovations with a
n energy or resource efficiency benefit    2781
Name: q19.5, dtype: int64


q19.6
Not mentioned
9865
Social innovations, such as new products, services or processes that have
the aim of improving society    2250
Name: q19.6, dtype: int64


q19.7
Not mentioned                   11161
Any other type of innovation      954
Name: q19.7, dtype: int64


q19.8
Not mentioned    7527
No, none         4588
Name: q19.8, dtype: int64


q19.9
Not mentioned    12019
DK/NA               96
Name: q19.9, dtype: int64


q20.1
Not mentioned                        10541
Lack of technology infrastructure     1574
Name: q20.1, dtype: int64


q20.2
```

```
Not mentioned                                    9328
Lack of skills, including managerial skills    2787
Name: q20.2, dtype: int64


q20.3
Not mentioned                                      8124
Difficulties in predicting the market response    3991
Name: q20.3, dtype: int64


q20.4
Not mentioned
10486
Lack of collaboration partners, such as other enterprises, etc. for innov
ation projects     1629
Name: q20.4, dtype: int64


q20.5
Not mentioned                           9028
Legal or administrative environment    3087
Name: q20.5, dtype: int64


q20.6
Not mentioned
8667
Lack of financial resources, including from available support schemes
3448
Name: q20.6, dtype: int64


q20.7
Not mentioned                                          11118
Difficulties with protecting intellectual property      997
Name: q20.7, dtype: int64


q20.8
Not mentioned    9165
None of these    2950
Name: q20.8, dtype: int64
```

```
q20.9
Not mentioned                                                    11760
Your enterprise has no interest in innovating (DO NOT READ OUT)      355
Name: q20.9, dtype: int64


q20.10
Not mentioned              11935
Other (DO NOT READ OUT)      180
Name: q20.10, dtype: int64


q20.11
Not mentioned    11905
DK/NA              210
Name: q20.11, dtype: int64


q21.1
Not mentioned                  9434
Lack of financial resources    2681
Name: q21.1, dtype: int64


q21.2
Not mentioned                                     9674
Lack of skills, including managerial skills    2441
Name: q21.2, dtype: int64


q21.3
Not mentioned
10390
Lack of information technology infrastructure, such as high-speed interne
t connection     1725
Name: q21.3, dtype: int64


q21.4
Not mentioned           10016
Regulatory obstacles     2099
Name: q21.4, dtype: int64
```

```
q21.5
Not mentioned         10092
IT security issues     2023
Name: q21.5, dtype: int64


q21.6
Not mentioned                                     9756
Uncertainty about future digital standards    2359
Name: q21.6, dtype: int64


q21.7
Not mentioned                      9890
Internal resistance to change    2225
Name: q21.7, dtype: int64


q21.8
Not mentioned    7995
None of these    4120
Name: q21.8, dtype: int64


q21.9
Not mentioned                                                                  11
593
Your enterprise has no interest in digitalisation (DO NOT READ OUT)
522
Name: q21.9, dtype: int64


q21.10
Not mentioned              11947
Other (DO NOT READ OUT)      168
Name: q21.10, dtype: int64


q21.11
no  DK/NA    11839
DK/NA          276
Name: q21.11, dtype: int64
```

```
q22
Your enterprise has adopted/is planning to adopt basic digital technologi
es but not advanced digital technologies ...    3932
There is a need to introduce advanced digital technologies and your enter
prise has already started to adopt them         3272
Your enterprise does not need to adopt any digital technologies
1998
There is a need to introduce advanced digital technologies and your enter
prise is currently considering ...              1211
There is a need to introduce advanced digital technologies but your enter
prise does not have the knowledge ...            926
None (DO NOT READ OUT)
482
Other (DO NOT READ OUT)
149
DK/NA
145
Name: q22, dtype: int64


q23.1
Not mentioned
11247
Artificial intelligence, e.g. machine learning or technologies identifyin
g objects or persons, etc.      868
Name: q23.1, dtype: int64


q23.2
Not mentioned
6391
Cloud computing, i.e. storing and processing files or data on remote serv
ers hosted on the internet    5724
Name: q23.2, dtype: int64


q23.3
Not mentioned
11031
Robotics, i.e. robots used to automate processes for example in construct
ion or design, etc.     1084
Name: q23.3, dtype: int64
```

```
q23.4
Not mentioned                                                    8866
Smart devices, e.g. smart sensors, smart thermostats, etc.    3249
Name: q23.4, dtype: int64


q23.5
Not mentioned                                                     10489
Big data analytics, e.g. data mining and predictive analysis     1626
Name: q23.5, dtype: int64


q23.6
Not mentioned                8246
High speed infrastructure    3869
Name: q23.6, dtype: int64


q23.7
Not mentioned    11733
Blockchain         382
Name: q23.7, dtype: int64


q23.8
Not mentioned    8168
None of these    3947
Name: q23.8, dtype: int64


q23.9
Not mentioned    11936
DK/NA              179
Name: q23.9, dtype: int64


q24.1
Recycling or reusing materials    7182
Not mentioned                     4933
Name: q24.1, dtype: int64


q24.2
```

```
Reducing consumption of or impact on natural resources (e.g. saving water
or switching to sustainable resources)    6103
Not mentione
d
6012
Name: q24.2, dtype: int64


q24.3
Saving energy or switching to sustainable energy sources    6356
Not mentioned                                               5759
Name: q24.3, dtype: int64


q24.4
Not mentioned                                    8223
Developing sustainable products or services    3892
Name: q24.4, dtype: int64


q24.5
Improving working conditions of its employees    8432
Not mentioned                                     3683
Name: q24.5, dtype: int64


q24.6
Promoting and improving diversity and equality in the workplace    6445
Not mentioned                                                      5670
Name: q24.6, dtype: int64


q24.7
Not mentioned                                          8825
Evaluating the impact of your enterprise on society    3290
Name: q24.7, dtype: int64


q24.8
Not mentioned                                             6421
Engaging employees in the governance of the enterprise    5694
Name: q24.8, dtype: int64
```

```
q24.9
Not mentioned             11260
None (DO NOT READ OUT)      855
Name: q24.9, dtype: int64


q24.10
Not mentioned    11962
DK/NA              153
Name: q24.10, dtype: int64


q25
No, but it may be considered in the future            4519
Yes, and it is in the process of being implemented    2881
Yes, and it has already been implemented              1935
No, and it will not in the future                     1852
Not applicable (DO NOT READ OUT)                       542
DK/NA                                                  386
Name: q25, dtype: int64


q26.1
Not mentioned                                 11229
Lack of willingness among the management        886
Name: q26.1, dtype: int64


q26.2
Not mentioned                            8672
Lack of consumer or customer demand      3443
Name: q26.2, dtype: int64


q26.3
Not mentioned
9539
Lack of awareness about how to integrate sustainability into the enterpri
se's business model    2576
Name: q26.3, dtype: int64


q26.4
Not mentioned                                          9724
```

```
It is not compatible with your current business model    2391
Name: q26.4, dtype: int64


q26.5
Not mentioned                 10388
It would not be profitable     1727
Name: q26.5, dtype: int64


q26.6
Not mentioned                                     10204
Lack of skills, including managerial skills        1911
Name: q26.6, dtype: int64


q26.7
Not mentioned                  8991
Lack of financial resources    3124
Name: q26.7, dtype: int64


q26.8
Not mentioned        8669
None of the above    3446
Name: q26.8, dtype: int64


q26.9
Not mentioned              11794
Other (DO NOT READ OUT)      321
Name: q26.9, dtype: int64


q26.10
Not mentioned    11778
DK/NA              337
Name: q26.10, dtype: int64
```

In [ ]:

In [11]:

```
df_encoded= pd.get_dummies(df, columns=my_list)
df_encoded
```

(...)

12115 rows × 439 columns

In [12]:

```
list_columns= df_encoded.columns.tolist()
list_columns
```

Out[12]:

```
['isocntry_0',
 'isocntry_1',
 'nace_a_Arts, entertainment and recreation',
 'nace_a_B - Mining and quarrying',
 'nace_a_C - Manufacturing',
 'nace_a_D - Electricity, gas, steam and air conditioningsupply',
 'nace_a_E - Water supply,sewerage,waste management/remediation activ',
 'nace_a_F - Construction',
 'nace_a_G - Wholesale and retail trade, repair of motor vehicles and',
 'nace_a_H - Transportation and storage',
 'nace_a_I - Accommodation and food service activities',
 'nace_a_J - Information and communication',
 'nace_a_K - Financial and insurance activities',
 'nace_a_L - Real estate activities',
 'nace_a_M - Professional, scientific and technical activities',
 'nace_a_N - Administrative and support service activities',
 'nace_a_P - Education',
 'nace_a_Q - Human health and social work activities',
 'q1_2019 and after',
 'q1_Before 2000',
 'q1_Between 2000 and 2014',
 'q1_Between 2015 and 2018',
 'q1_DK/NA',
 'q2t_1 to 9 employees',
 'q2t_10 to 49 employees',
 'q2t_250 employees or more',
 'q2t_50 to 249 employees',
 'q3t_0 employe',
 'q3t_1 to 9 employees',
 'q3t_10 to 49 employees',
 'q3t_250 employees or more',
 'q3t_50 to 249 employees',
 'q3t_DK/NA',
 'q3t_Inap. (not 1 in q2a and q2b)',
```

```
 'q4t_100,000 euros or less',
 'q4t_DK/NA',
 'q4t_More than 1 million and up to 2 million euros',
 'q4t_More than 10 million and up to 50 million euros',
 'q4t_More than 100,000 and up to 500,000 euros',
 'q4t_More than 2 million and up to 5 million euros',
 'q4t_More than 5 million and up to 10 million euros',
 'q4t_More than 50 million euros',
 'q4t_More than 500,000 and up to 1 million euros',
 'q5_1_DK/NA',
 'q5_1_It has decreased',
 'q5_1_It has grown by at least 30%',
 'q5_1_It has grown by less than 30%',
 'q5_1_It has remained stable',
 'q5_2_DK/NA',
 'q5_2_It has decreased',
 'q5_2_It has grown by at least 30%',
 'q5_2_It has grown by less than 30%',
 'q5_2_It has remained stable',
 'q6_1_DK/NA',
 'q6_1_Grow by between 10% and 20% per year',
 'q6_1_Grow by less than 10% per year',
 'q6_1_Grow by more than 20% per year',
 'q6_1_It does not plan to grow',
 'q6_2_DK/NA',
 'q6_2_Grow by between 10% and 20% per year',
 'q6_2_Grow by less than 10% per year',
 'q6_2_Grow by more than 20% per year',
 'q6_2_It does not plan to grow',
 'q7a.1_Have a strategic growth plan',
 'q7a.1_Not mentioned',
 'q7a.2_Not mentioned',
 'q7a.2_Plan to grow as a result of introducing some kind of innovation',
 'q7a.3_Not mentioned',
 'q7a.3_Plan to grow as a result of operating in growing markets',
 'q7a.4_Not mentioned',
 'q7a.4_Plan to grow as a result of entering new markets',
 'q7a.5_Not mentioned',
 'q7a.5_Plan to grow as a result of increased digitalisation in your ente
rprise',
 'q7a.6_Not mentioned',
 'q7a.6_Plan to grow in (OUR COUNTRY)',
 'q7a.7_Not mentioned',
 'q7a.7_[EU] Plan to grow in other EU countries/ [Non-EU]',
```

```
 'q7a.8_Not mentioned',
 'q7a.8_Plan to grow in other non-EU countries',
 'q7a.9_DK/NA',
 'q7a.9_Not mentioned',
 'q7b.1_Not mentioned',
 'q7b.1_There is no intention for your enterprise to grow beyond its curr
ent size',
 'q7b.2_Not mentioned',
 'q7b.2_Your enterprise does not have employees with the skills or expert
ise needed for it to grow',
 'q7b.3_Not mentioned',
 'q7b.3_Your enterprise does not have the financial resources to grow',
 'q7b.4_Not mentioned',
 "q7b.4_There is decreasing demand for your enterprise's products or serv
ices or the market is saturated",
 'q7b.5_Additional regulatory or administrative burdens and requirements
would be too high for your enterprise to grow',
 'q7b.5_Not mentioned',
 'q7b.6_Not mentioned',
 'q7b.6_Your enterprise does not want to grow because it would lose benef
its linked to its SME status',
 'q7b.7_Not mentioned',
 'q7b.7_The current location of your enterprise does not allow you to gro
w and you do not wish to relocate elsewhere',
 'q7b.8_Not mentioned',
 'q7b.8_Your enterprise relies on a few clients which are unlikely to inc
rease their demand',
 'q7b.9_Not mentioned',
 'q7b.9_Other (DO NOT READ OUT)',
 'q7b.10_DK/NA',
 'q7b.10_Not mentioned',
 'q8.1_In a large town or city',
 'q8.1_Not mentioned',
 'q8.2_In a small town or village',
 'q8.2_Not mentioned',
 'q8.3_In a rural area',
 'q8.3_Not mentioned',
 'q8.4_In an industrial area',
 'q8.4_Not mentioned',
 'q8.5_Near a border with an EU country',
 'q8.5_Not mentioned',
 'q8.6_Near a border with a non-EU country',
 'q8.6_Not mentioned',
 'q8.7_DK/NA',
```

```
 'q8.7_Not mentioned',
 'q9.1_It mainly provides goods',
 'q9.1_Not mentioned',
 'q9.2_It mainly provides services',
 'q9.2_Not mentioned',
 'q9.3_It sells goods online to buyers in EU countries',
 'q9.3_Not mentioned',
 'q9.4_It is a member of an industry cluster or another SME business supp
ort organisation in the region',
 'q9.4_Not mentioned',
 'q9.5_It is a part of a global value chain',
 'q9.5_Not mentioned',
 'q9.6_It has a patent or patent application',
 'q9.6_Not mentioned',
 'q9.7_It is a non-profit enterprise',
 'q9.7_Not mentioned',
 'q9.8_It has a strategy or action plan to digitalise',
 'q9.8_Not mentioned',
 'q9.9_Not mentioned',
 'q9.9_Other (DO NOT READ OUT)',
 'q9.10_None (DO NOT READ OUT)',
 'q9.10_Not mentioned',
 'q9.11_DK/NA',
 'q9.11_Not mentioned',
 'q10_DK/NA',
 'q10_No, definitely not',
 'q10_No, probably not',
 'q10_Not applicable (DO NOT READ OUT)',
 'q10_Yes, definitely',
 'q10_Yes, probably',
 'q11.1_None, your enterprise only operates in (OUR COUNTRY)',
 'q11.1_Not mentioned',
 'q11.2_Not mentioned',
 'q11.2_[EU] Other EU countries/ [Non-EU] EU countries',
 'q11.3_Not mentioned',
 'q11.3_Other European countries outside of the EU (incl. Russia)',
 'q11.4_North America',
 'q11.4_Not mentioned',
 'q11.5_Latin America and the Caribbean',
 'q11.5_Not mentioned',
 'q11.6_China',
 'q11.6_Not mentioned',
 'q11.7_Not mentioned',
 'q11.7_Rest of Asia and the Pacific',
```

```
'q11.8_Middle East and Africa',
'q11.8_Not mentioned',
'q11.9_DK/NA',
'q11.9_Not mentioned',
'q12a_Between 25% and 50%',
'q12a_DK/NA',
'q12a_Less than 25%',
'q12a_More than 50%',
'q12b_Between 25% and 50%',
'q12b_DK/NA',
'q12b_Less than 25%',
'q12b_More than 50%',
'q13.1_Not mentioned',
'q13.1_Solely owned by one person',
'q13.2_Not mentioned',
'q13.2_Owned by more than one person',
'q13.3_Not mentioned',
'q13.3_Part of a national or international enterprise group',
'q13.4_Co-owned by a public entity',
'q13.4_Not mentioned',
'q13.5_Co-owned by venture capital firm',
'q13.5_Not mentioned',
'q13.6_Co-owned by business angel',
'q13.6_Not mentioned',
'q13.7_Not mentioned',
'q13.7_Predominantly family owned',
'q13.8_Jointly owned by its members (e.g. cooperative, mutual society)',
'q13.8_Not mentioned',
'q13.9_Not mentioned',
'q13.9_Other (DO NOT READ OUT)',
'q13.10_DK/NA',
'q13.10_Not mentioned',
'q14.1_Not mentioned',
'q14.1_The sole founder of this enterprise',
'q14.2_A co-founder of this enterprise',
'q14.2_Not mentioned',
'q14.3_Not mentioned',
'q14.3_The sole owner of this enterprise',
'q14.4_A co-owner of this enterprise',
'q14.4_Not mentioned',
'q14.5_None of the above',
'q14.5_Not mentioned',
'q14.6_DK/NA',
'q14.6_Not mentioned',
```

```
 'q15a.1_Not mentioned',
 'q15a.1_This is the first enterprise that you have ever established',
 'q15a.2_Not mentioned',
 'q15a.2_You have established or co-established other enterprise(s)',
 'q15a.3_Not mentioned',
 'q15a.3_You have closed - without bankruptcy - other enterprise(s) that
you owned or co-owned',
 'q15a.4_Not mentioned',
 'q15a.4_You have closed - due to bankruptcy - other enterprise(s) that y
ou owned or co-owned',
 'q15a.5_Not mentioned',
 'q15a.5_You have sold other enterprise(s) that you owned or co-owned',
 'q15a.6_Not mentioned',
 'q15a.6_You plan to relocate the headquarters of your enterprise to an E
U country in the future',
 'q15a.7_Not mentioned',
 'q15a.7_You plan to relocate the headquarters of your enterprise to the
USA in the future',
 'q15a.8_Not mentioned',
 'q15a.8_You plan to relocate the headquarters of your enterprise to any
other country in the future',
 'q15a.9_Not mentioned',
 'q15a.9_Other (DO NOT READ OUT)',
 'q15a.10_None (DO NOT READ OUT)',
 'q15a.10_Not mentioned',
 'q15a.11_DK/NA',
 'q15a.11_Not mentioned',
 'q15b.1_Not mentioned',
 'q15b.1_You took this enterprise over from family member(s)',
 'q15b.2_Not mentioned',
 'q15b.2_You have established or co-established other enterprises',
 'q15b.3_Not mentioned',
 'q15b.3_You have closed - without bankruptcy - other enterprise(s) that
you owned or co-owned',
 'q15b.4_Not mentioned',
 'q15b.4_You have closed - due to bankruptcy - other enterprise(s) that y
ou owned or co-owned',
 'q15b.5_Not mentioned',
 'q15b.5_You have sold other enterprise(s) that you owned or co-owned',
 'q15b.6_Not mentioned',
 'q15b.6_You plan to relocate the headquarters of your enterprise to an E
U country in the future',
 'q15b.7_Not mentioned',
```

```
 'q15b.7_You plan to relocate the headquarters of your enterprise to the
USA in the future',
 'q15b.8_Not mentioned',
 'q15b.8_You plan to relocate the headquarters of your enterprise to any
other country in the future',
 'q15b.9_Not mentioned',
 'q15b.9_Other (DO NOT READ OUT)',
 'q15b.10_None (DO NOT READ OUT)',
 'q15b.10_Not mentioned',
 'q15b.11_DK/NA',
 'q15b.11_Not mentioned',
 'q16_1_DK/NA',
 'q16_1_Fairly good',
 'q16_1_Fairly poor',
 'q16_1_Very good',
 'q16_1_Very poor',
 'q16_2_DK/NA',
 'q16_2_Fairly good',
 'q16_2_Fairly poor',
 'q16_2_Very good',
 'q16_2_Very poor',
 'q16_3_DK/NA',
 'q16_3_Fairly good',
 'q16_3_Fairly poor',
 'q16_3_Very good',
 'q16_3_Very poor',
 'q16_4_DK/NA',
 'q16_4_Fairly good',
 'q16_4_Fairly poor',
 'q16_4_Very good',
 'q16_4_Very poor',
 'q16_5_DK/NA',
 'q16_5_Fairly good',
 'q16_5_Fairly poor',
 'q16_5_Very good',
 'q16_5_Very poor',
 'q16_6_DK/NA',
 'q16_6_Fairly good',
 'q16_6_Fairly poor',
 'q16_6_Very good',
 'q16_6_Very poor',
 'q16_7_DK/NA',
 'q16_7_Fairly good',
 'q16_7_Fairly poor',
```

```
 'q16_7_Very good',
 'q16_7_Very poor',
 'q16_8_DK/NA',
 'q16_8_Fairly good',
 'q16_8_Fairly poor',
 'q16_8_Very good',
 'q16_8_Very poor',
 'q17.1_Difficulties with innovation',
 'q17.1_Not mentioned',
 'q17.2_Not mentioned',
 'q17.2_Regulatory obstacles or administrative burden',
 'q17.3_Access to data',
 'q17.3_Not mentioned',
 'q17.4_Internationalisation',
 'q17.4_Not mentioned',
 'q17.5_Access to finance',
 'q17.5_Not mentioned',
 'q17.6_Not mentioned',
 'q17.6_Payment delays',
 'q17.7_Not mentioned',
 'q17.7_Skills, including managerial skills',
 'q17.8_Difficulties with digitalisation',
 'q17.8_Not mentioned',
 'q17.9_Not mentioned',
 'q17.9_Other (DO NOT READ OUT)',
 'q17.10_DK/NA',
 'q17.10_Not mentioned',
 'q19.1_A new or significantly improved product or service to the market'
,
 'q19.1_Not mentioned',
 'q19.2_A new or significantly improved production process or method',
 'q19.2_Not mentioned',
 'q19.3_A new organisation of management or a new business model',
 'q19.3_Not mentioned',
 'q19.4_A new way of selling your goods or services',
 'q19.4_Not mentioned',
 'q19.5_An innovation with an environmental benefit, including innovation
s with an energy or resource efficiency benefit',
 'q19.5_Not mentioned',
 'q19.6_Not mentioned',
 'q19.6_Social innovations, such as new products, services or processes t
hat have the aim of improving society',
 'q19.7_Any other type of innovation',
 'q19.7_Not mentioned',
```

```
 'q19.8_No, none',
 'q19.8_Not mentioned',
 'q19.9_DK/NA',
 'q19.9_Not mentioned',
 'q20.1_Lack of technology infrastructure',
 'q20.1_Not mentioned',
 'q20.2_Lack of skills, including managerial skills',
 'q20.2_Not mentioned',
 'q20.3_Difficulties in predicting the market response',
 'q20.3_Not mentioned',
 'q20.4_Lack of collaboration partners, such as other enterprises, etc. f
or innovation projects',
 'q20.4_Not mentioned',
 'q20.5_Legal or administrative environment',
 'q20.5_Not mentioned',
 'q20.6_Lack of financial resources, including from available support sch
emes',
 'q20.6_Not mentioned',
 'q20.7_Difficulties with protecting intellectual property',
 'q20.7_Not mentioned',
 'q20.8_None of these',
 'q20.8_Not mentioned',
 'q20.9_Not mentioned',
 'q20.9_Your enterprise has no interest in innovating (DO NOT READ OUT)',
 'q20.10_Not mentioned',
 'q20.10_Other (DO NOT READ OUT)',
 'q20.11_DK/NA',
 'q20.11_Not mentioned',
 'q21.1_Lack of financial resources',
 'q21.1_Not mentioned',
 'q21.2_Lack of skills, including managerial skills',
 'q21.2_Not mentioned',
 'q21.3_Lack of information technology infrastructure, such as high-speed
internet connection',
 'q21.3_Not mentioned',
 'q21.4_Not mentioned',
 'q21.4_Regulatory obstacles',
 'q21.5_IT security issues',
 'q21.5_Not mentioned',
 'q21.6_Not mentioned',
 'q21.6_Uncertainty about future digital standards',
 'q21.7_Internal resistance to change',
 'q21.7_Not mentioned',
 'q21.8_None of these',
```

```
 'q21.8_Not mentioned',
 'q21.9_Not mentioned',
 'q21.9_Your enterprise has no interest in digitalisation (DO NOT READ OU
T)',
 'q21.10_Not mentioned',
 'q21.10_Other (DO NOT READ OUT)',
 'q21.11_DK/NA',
 'q21.11_no  DK/NA',
 'q22_DK/NA',
 'q22_None (DO NOT READ OUT)',
 'q22_Other (DO NOT READ OUT)',
 'q22_There is a need to introduce advanced digital technologies and your
enterprise has already started to adopt them',
 'q22_There is a need to introduce advanced digital technologies and your
enterprise is currently considering ...',
 'q22_There is a need to introduce advanced digital technologies but your
enterprise does not have the knowledge ...',
 'q22_Your enterprise does not need to adopt any digital technologies',
 'q22_Your enterprise has adopted/is planning to adopt basic digital tech
nologies but not advanced digital technologies ...',
 'q23.1_Artificial intelligence, e.g. machine learning or technologies id
entifying objects or persons, etc.',
 'q23.1_Not mentioned',
 'q23.2_Cloud computing, i.e. storing and processing files or data on rem
ote servers hosted on the internet',
 'q23.2_Not mentioned',
 'q23.3_Not mentioned',
 'q23.3_Robotics, i.e. robots used to automate processes for example in c
onstruction or design, etc.',
 'q23.4_Not mentioned',
 'q23.4_Smart devices, e.g. smart sensors, smart thermostats, etc.',
 'q23.5_Big data analytics, e.g. data mining and predictive analysis',
 'q23.5_Not mentioned',
 'q23.6_High speed infrastructure',
 'q23.6_Not mentioned',
 'q23.7_Blockchain',
 'q23.7_Not mentioned',
 'q23.8_None of these',
 'q23.8_Not mentioned',
 'q23.9_DK/NA',
 'q23.9_Not mentioned',
 'q24.1_Not mentioned',
 'q24.1_Recycling or reusing materials',
 'q24.2_Not mentioned',
```

```
 'q24.2_Reducing consumption of or impact on natural resources (e.g. savi
ng water or switching to sustainable resources)',
 'q24.3_Not mentioned',
 'q24.3_Saving energy or switching to sustainable energy sources',
 'q24.4_Developing sustainable products or services',
 'q24.4_Not mentioned',
 'q24.5_Improving working conditions of its employees',
 'q24.5_Not mentioned',
 'q24.6_Not mentioned',
 'q24.6_Promoting and improving diversity and equality in the workplace',
 'q24.7_Evaluating the impact of your enterprise on society',
 'q24.7_Not mentioned',
 'q24.8_Engaging employees in the governance of the enterprise',
 'q24.8_Not mentioned',
 'q24.9_None (DO NOT READ OUT)',
 'q24.9_Not mentioned',
 'q24.10_DK/NA',
 'q24.10_Not mentioned',
 'q25_DK/NA',
 'q25_No, and it will not in the future',
 'q25_No, but it may be considered in the future',
 'q25_Not applicable (DO NOT READ OUT)',
 'q25_Yes, and it has already been implemented',
 'q25_Yes, and it is in the process of being implemented',
 'q26.1_Lack of willingness among the management',
 'q26.1_Not mentioned',
 'q26.2_Lack of consumer or customer demand',
 'q26.2_Not mentioned',
 "q26.3_Lack of awareness about how to integrate sustainability into the
enterprise's business model",
 'q26.3_Not mentioned',
 'q26.4_It is not compatible with your current business model',
 'q26.4_Not mentioned',
 'q26.5_It would not be profitable',
 'q26.5_Not mentioned',
 'q26.6_Lack of skills, including managerial skills',
 'q26.6_Not mentioned',
 'q26.7_Lack of financial resources',
 'q26.7_Not mentioned',
 'q26.8_None of the above',
 'q26.8_Not mentioned',
 'q26.9_Not mentioned',
 'q26.9_Other (DO NOT READ OUT)',
 'q26.10_DK/NA',
```

```
 'q26.10_Not mentioned']
```

In [13]:

```
df_encoded=df_encoded.drop(['q7a.1_Not mentioned',
 'q7a.2_Not mentioned',
 'q7a.3_Not mentioned',
 'q7a.4_Not mentioned',
 'q7a.5_Not mentioned',
 'q7a.6_Not mentioned',
 'q7a.7_Not mentioned',
 'q7a.8_Not mentioned',
 'q7a.9_Not mentioned',
 'q7b.1_Not mentioned',
 'q7b.2_Not mentioned',
 'q7b.3_Not mentioned',
 'q7b.4_Not mentioned',
 'q7b.5_Not mentioned',
 'q7b.6_Not mentioned',
 'q7b.7_Not mentioned',
 'q7b.8_Not mentioned',
 'q7b.9_Not mentioned',
 'q7b.10_Not mentioned',
 'q8.1_Not mentioned',
 'q8.2_Not mentioned',
 'q8.3_Not mentioned',
 'q8.4_Not mentioned',
 'q8.5_Not mentioned',
 'q8.6_Not mentioned',
 'q8.7_Not mentioned',
 'q9.1_Not mentioned',
 'q9.2_Not mentioned',
 'q9.3_Not mentioned',
 'q9.4_Not mentioned',
 'q9.5_Not mentioned',
 'q9.6_Not mentioned',
 'q9.7_Not mentioned',
 'q9.8_Not mentioned',
 'q9.9_Not mentioned',
 'q9.10_Not mentioned',
 'q9.11_Not mentioned',
 'q11.1_Not mentioned',
 'q11.2_Not mentioned',
 'q11.3_Not mentioned',
 'q11.4_Not mentioned',
 'q11.5_Not mentioned',
```

```
'q11.6_Not mentioned',
'q11.7_Not mentioned',
'q11.8_Not mentioned',
'q11.9_Not mentioned',
'q13.1_Not mentioned',
'q13.2_Not mentioned',
'q13.3_Not mentioned',
'q13.4_Not mentioned',
'q13.5_Not mentioned',
'q13.6_Not mentioned',
'q13.7_Not mentioned',
'q13.8_Not mentioned',
'q13.9_Not mentioned',
'q13.10_Not mentioned',
'q14.1_Not mentioned',
'q14.2_Not mentioned',
'q14.3_Not mentioned',
'q14.4_Not mentioned',
'q14.5_Not mentioned',
'q14.6_Not mentioned',
'q15a.1_Not mentioned',
'q15a.2_Not mentioned',
'q15a.3_Not mentioned',
'q15a.4_Not mentioned',
'q15a.5_Not mentioned',
'q15a.6_Not mentioned',
'q15a.7_Not mentioned',
'q15a.8_Not mentioned',
'q15a.9_Not mentioned',
'q15a.10_Not mentioned',
'q15a.11_Not mentioned',
'q15b.1_Not mentioned',
'q15b.2_Not mentioned',
'q15b.3_Not mentioned',
'q15b.4_Not mentioned',
'q15b.5_Not mentioned',
'q15b.6_Not mentioned',
'q15b.7_Not mentioned',
'q15b.8_Not mentioned',
'q15b.9_Not mentioned',
'q15b.10_Not mentioned',
'q15b.11_Not mentioned',
'q17.1_Not mentioned',
'q17.2_Not mentioned',
```

```
'q17.3_Not mentioned',
'q17.4_Not mentioned',
'q17.5_Not mentioned',
'q17.6_Not mentioned',
'q17.7_Not mentioned',
'q17.8_Not mentioned',
'q17.9_Not mentioned',
'q17.10_Not mentioned',
'q19.1_Not mentioned',
'q19.2_Not mentioned',
'q19.3_Not mentioned',
'q19.4_Not mentioned',
'q19.5_Not mentioned',
'q19.6_Not mentioned',
'q19.7_Not mentioned',
'q19.8_Not mentioned',
'q19.9_Not mentioned',
'q20.1_Not mentioned',
'q20.2_Not mentioned',
'q20.3_Not mentioned',
'q20.4_Not mentioned',
'q20.5_Not mentioned',
'q20.6_Not mentioned',
'q20.7_Not mentioned',
'q20.8_Not mentioned',
'q20.9_Not mentioned',
'q20.10_Not mentioned',
'q20.11_Not mentioned',
'q21.1_Not mentioned',
'q21.2_Not mentioned',
'q21.3_Not mentioned',
'q21.4_Not mentioned',
'q21.5_Not mentioned',
'q21.6_Not mentioned',
'q21.7_Not mentioned',
'q21.8_Not mentioned',
'q21.9_Not mentioned',
'q21.10_Not mentioned',
'q23.1_Not mentioned',
'q23.2_Not mentioned',
'q23.3_Not mentioned',
'q23.4_Not mentioned',
'q23.5_Not mentioned',
'q23.6_Not mentioned',
```

```
 'q23.7_Not mentioned',
 'q23.8_Not mentioned',
 'q23.9_Not mentioned',
 'q24.1_Not mentioned',
 'q24.2_Not mentioned',
 'q24.3_Not mentioned',
 'q24.4_Not mentioned',
 'q24.5_Not mentioned',
 'q24.6_Not mentioned',
 'q24.7_Not mentioned',
 'q24.8_Not mentioned',
 'q24.9_None (DO NOT READ OUT)',
 'q24.9_Not mentioned',
 'q24.10_Not mentioned',
 'q26.1_Not mentioned',
 'q26.2_Not mentioned',
 'q26.3_Not mentioned',
 'q26.4_Not mentioned',
 'q26.5_Not mentioned',
 'q26.6_Not mentioned',
 'q26.7_Not mentioned',
 'q26.8_Not mentioned',
 'q26.9_Not mentioned',
 'q26.10_Not mentioned'
                 ],axis=1)
```

In [14]:

```
list_columns= df_encoded.columns.tolist()
list_columns
```

Out[14]:

```
['isocntry_0',
 'isocntry_1',
 'nace_a_Arts, entertainment and recreation',
 'nace_a_B - Mining and quarrying',
 'nace_a_C - Manufacturing',
 'nace_a_D - Electricity, gas, steam and air conditioningsupply',
 'nace_a_E - Water supply,sewerage,waste management/remediation activ',
 'nace_a_F - Construction',
 'nace_a_G - Wholesale and retail trade, repair of motor vehicles and',
 'nace_a_H - Transportation and storage',
 'nace_a_I - Accommodation and food service activities',
 'nace_a_J - Information and communication',
 'nace_a_K - Financial and insurance activities',
 'nace_a_L - Real estate activities',
```

```
'nace_a_M - Professional, scientific and technical activities',
'nace_a_N - Administrative and support service activities',
'nace_a_P - Education',
'nace_a_Q - Human health and social work activities',
'q1_2019 and after',
'q1_Before 2000',
'q1_Between 2000 and 2014',
'q1_Between 2015 and 2018',
'q1_DK/NA',
'q2t_1 to 9 employees',
'q2t_10 to 49 employees',
'q2t_250 employees or more',
'q2t_50 to 249 employees',
'q3t_0 employe',
'q3t_1 to 9 employees',
'q3t_10 to 49 employees',
'q3t_250 employees or more',
'q3t_50 to 249 employees',
'q3t_DK/NA',
'q3t_Inap. (not 1 in q2a and q2b)',
'q4t_100,000 euros or less',
'q4t_DK/NA',
'q4t_More than 1 million and up to 2 million euros',
'q4t_More than 10 million and up to 50 million euros',
'q4t_More than 100,000 and up to 500,000 euros',
'q4t_More than 2 million and up to 5 million euros',
'q4t_More than 5 million and up to 10 million euros',
'q4t_More than 50 million euros',
'q4t_More than 500,000 and up to 1 million euros',
'q5_1_DK/NA',
'q5_1_It has decreased',
'q5_1_It has grown by at least 30%',
'q5_1_It has grown by less than 30%',
'q5_1_It has remained stable',
'q5_2_DK/NA',
'q5_2_It has decreased',
'q5_2_It has grown by at least 30%',
'q5_2_It has grown by less than 30%',
'q5_2_It has remained stable',
'q6_1_DK/NA',
'q6_1_Grow by between 10% and 20% per year',
'q6_1_Grow by less than 10% per year',
'q6_1_Grow by more than 20% per year',
'q6_1_It does not plan to grow',
```

```
 'q6_2_DK/NA',
 'q6_2_Grow by between 10% and 20% per year',
 'q6_2_Grow by less than 10% per year',
 'q6_2_Grow by more than 20% per year',
 'q6_2_It does not plan to grow',
 'q7a.1_Have a strategic growth plan',
 'q7a.2_Plan to grow as a result of introducing some kind of innovation',
 'q7a.3_Plan to grow as a result of operating in growing markets',
 'q7a.4_Plan to grow as a result of entering new markets',
 'q7a.5_Plan to grow as a result of increased digitalisation in your ente
rprise',
 'q7a.6_Plan to grow in (OUR COUNTRY)',
 'q7a.7_[EU] Plan to grow in other EU countries/ [Non-EU]',
 'q7a.8_Plan to grow in other non-EU countries',
 'q7a.9_DK/NA',
 'q7b.1_There is no intention for your enterprise to grow beyond its curr
ent size',
 'q7b.2_Your enterprise does not have employees with the skills or expert
ise needed for it to grow',
 'q7b.3_Your enterprise does not have the financial resources to grow',
 "q7b.4_There is decreasing demand for your enterprise's products or serv
ices or the market is saturated",
 'q7b.5_Additional regulatory or administrative burdens and requirements
would be too high for your enterprise to grow',
 'q7b.6_Your enterprise does not want to grow because it would lose benef
its linked to its SME status',
 'q7b.7_The current location of your enterprise does not allow you to gro
w and you do not wish to relocate elsewhere',
 'q7b.8_Your enterprise relies on a few clients which are unlikely to inc
rease their demand',
 'q7b.9_Other (DO NOT READ OUT)',
 'q7b.10_DK/NA',
 'q8.1_In a large town or city',
 'q8.2_In a small town or village',
 'q8.3_In a rural area',
 'q8.4_In an industrial area',
 'q8.5_Near a border with an EU country',
 'q8.6_Near a border with a non-EU country',
 'q8.7_DK/NA',
 'q9.1_It mainly provides goods',
 'q9.2_It mainly provides services',
 'q9.3_It sells goods online to buyers in EU countries',
 'q9.4_It is a member of an industry cluster or another SME business supp
ort organisation in the region',
```

```
'q9.5_It is a part of a global value chain',
'q9.6_It has a patent or patent application',
'q9.7_It is a non-profit enterprise',
'q9.8_It has a strategy or action plan to digitalise',
'q9.9_Other (DO NOT READ OUT)',
'q9.10_None (DO NOT READ OUT)',
'q9.11_DK/NA',
'q10_DK/NA',
'q10_No, definitely not',
'q10_No, probably not',
'q10_Not applicable (DO NOT READ OUT)',
'q10_Yes, definitely',
'q10_Yes, probably',
'q11.1_None, your enterprise only operates in (OUR COUNTRY)',
'q11.2_[EU] Other EU countries/ [Non-EU] EU countries',
'q11.3_Other European countries outside of the EU (incl. Russia)',
'q11.4_North America',
'q11.5_Latin America and the Caribbean',
'q11.6_China',
'q11.7_Rest of Asia and the Pacific',
'q11.8_Middle East and Africa',
'q11.9_DK/NA',
'q12a_Between 25% and 50%',
'q12a_DK/NA',
'q12a_Less than 25%',
'q12a_More than 50%',
'q12b_Between 25% and 50%',
'q12b_DK/NA',
'q12b_Less than 25%',
'q12b_More than 50%',
'q13.1_Solely owned by one person',
'q13.2_Owned by more than one person',
'q13.3_Part of a national or international enterprise group',
'q13.4_Co-owned by a public entity',
'q13.5_Co-owned by venture capital firm',
'q13.6_Co-owned by business angel',
'q13.7_Predominantly family owned',
'q13.8_Jointly owned by its members (e.g. cooperative, mutual society)',
'q13.9_Other (DO NOT READ OUT)',
'q13.10_DK/NA',
'q14.1_The sole founder of this enterprise',
'q14.2_A co-founder of this enterprise',
'q14.3_The sole owner of this enterprise',
'q14.4_A co-owner of this enterprise',
```

```
 'q14.5_None of the above',
 'q14.6_DK/NA',
 'q15a.1_This is the first enterprise that you have ever established',
 'q15a.2_You have established or co-established other enterprise(s)',
 'q15a.3_You have closed - without bankruptcy - other enterprise(s) that
you owned or co-owned',
 'q15a.4_You have closed - due to bankruptcy - other enterprise(s) that y
ou owned or co-owned',
 'q15a.5_You have sold other enterprise(s) that you owned or co-owned',
 'q15a.6_You plan to relocate the headquarters of your enterprise to an E
U country in the future',
 'q15a.7_You plan to relocate the headquarters of your enterprise to the
USA in the future',
 'q15a.8_You plan to relocate the headquarters of your enterprise to any
other country in the future',
 'q15a.9_Other (DO NOT READ OUT)',
 'q15a.10_None (DO NOT READ OUT)',
 'q15a.11_DK/NA',
 'q15b.1_You took this enterprise over from family member(s)',
 'q15b.2_You have established or co-established other enterprises',
 'q15b.3_You have closed - without bankruptcy - other enterprise(s) that
you owned or co-owned',
 'q15b.4_You have closed - due to bankruptcy - other enterprise(s) that y
ou owned or co-owned',
 'q15b.5_You have sold other enterprise(s) that you owned or co-owned',
 'q15b.6_You plan to relocate the headquarters of your enterprise to an E
U country in the future',
 'q15b.7_You plan to relocate the headquarters of your enterprise to the
USA in the future',
 'q15b.8_You plan to relocate the headquarters of your enterprise to any
other country in the future',
 'q15b.9_Other (DO NOT READ OUT)',
 'q15b.10_None (DO NOT READ OUT)',
 'q15b.11_DK/NA',
 'q16_1_DK/NA',
 'q16_1_Fairly good',
 'q16_1_Fairly poor',
 'q16_1_Very good',
 'q16_1_Very poor',
 'q16_2_DK/NA',
 'q16_2_Fairly good',
 'q16_2_Fairly poor',
 'q16_2_Very good',
 'q16_2_Very poor',
```

```
 'q16_3_DK/NA',
 'q16_3_Fairly good',
 'q16_3_Fairly poor',
 'q16_3_Very good',
 'q16_3_Very poor',
 'q16_4_DK/NA',
 'q16_4_Fairly good',
 'q16_4_Fairly poor',
 'q16_4_Very good',
 'q16_4_Very poor',
 'q16_5_DK/NA',
 'q16_5_Fairly good',
 'q16_5_Fairly poor',
 'q16_5_Very good',
 'q16_5_Very poor',
 'q16_6_DK/NA',
 'q16_6_Fairly good',
 'q16_6_Fairly poor',
 'q16_6_Very good',
 'q16_6_Very poor',
 'q16_7_DK/NA',
 'q16_7_Fairly good',
 'q16_7_Fairly poor',
 'q16_7_Very good',
 'q16_7_Very poor',
 'q16_8_DK/NA',
 'q16_8_Fairly good',
 'q16_8_Fairly poor',
 'q16_8_Very good',
 'q16_8_Very poor',
 'q17.1_Difficulties with innovation',
 'q17.2_Regulatory obstacles or administrative burden',
 'q17.3_Access to data',
 'q17.4_Internationalisation',
 'q17.5_Access to finance',
 'q17.6_Payment delays',
 'q17.7_Skills, including managerial skills',
 'q17.8_Difficulties with digitalisation',
 'q17.9_Other (DO NOT READ OUT)',
 'q17.10_DK/NA',
 'q19.1_A new or significantly improved product or service to the market'
,
 'q19.2_A new or significantly improved production process or method',
 'q19.3_A new organisation of management or a new business model',
```

```
 'q19.4_A new way of selling your goods or services',
 'q19.5_An innovation with an environmental benefit, including innovation
s with an energy or resource efficiency benefit',
 'q19.6_Social innovations, such as new products, services or processes t
hat have the aim of improving society',
 'q19.7_Any other type of innovation',
 'q19.8_No, none',
 'q19.9_DK/NA',
 'q20.1_Lack of technology infrastructure',
 'q20.2_Lack of skills, including managerial skills',
 'q20.3_Difficulties in predicting the market response',
 'q20.4_Lack of collaboration partners, such as other enterprises, etc. f
or innovation projects',
 'q20.5_Legal or administrative environment',
 'q20.6_Lack of financial resources, including from available support sch
emes',
 'q20.7_Difficulties with protecting intellectual property',
 'q20.8_None of these',
 'q20.9_Your enterprise has no interest in innovating (DO NOT READ OUT)',
 'q20.10_Other (DO NOT READ OUT)',
 'q20.11_DK/NA',
 'q21.1_Lack of financial resources',
 'q21.2_Lack of skills, including managerial skills',
 'q21.3_Lack of information technology infrastructure, such as high-speed
internet connection',
 'q21.4_Regulatory obstacles',
 'q21.5_IT security issues',
 'q21.6_Uncertainty about future digital standards',
 'q21.7_Internal resistance to change',
 'q21.8_None of these',
 'q21.9_Your enterprise has no interest in digitalisation (DO NOT READ OU
T)',
 'q21.10_Other (DO NOT READ OUT)',
 'q21.11_DK/NA',
 'q21.11_no  DK/NA',
 'q22_DK/NA',
 'q22_None (DO NOT READ OUT)',
 'q22_Other (DO NOT READ OUT)',
 'q22_There is a need to introduce advanced digital technologies and your
enterprise has already started to adopt them',
 'q22_There is a need to introduce advanced digital technologies and your
enterprise is currently considering ...',
 'q22_There is a need to introduce advanced digital technologies but your
enterprise does not have the knowledge ...',
```

```
 'q22_Your enterprise does not need to adopt any digital technologies',
 'q22_Your enterprise has adopted/is planning to adopt basic digital tech
nologies but not advanced digital technologies ...',
 'q23.1_Artificial intelligence, e.g. machine learning or technologies id
entifying objects or persons, etc.',
 'q23.2_Cloud computing, i.e. storing and processing files or data on rem
ote servers hosted on the internet',
 'q23.3_Robotics, i.e. robots used to automate processes for example in c
onstruction or design, etc.',
 'q23.4_Smart devices, e.g. smart sensors, smart thermostats, etc.',
 'q23.5_Big data analytics, e.g. data mining and predictive analysis',
 'q23.6_High speed infrastructure',
 'q23.7_Blockchain',
 'q23.8_None of these',
 'q23.9_DK/NA',
 'q24.1_Recycling or reusing materials',
 'q24.2_Reducing consumption of or impact on natural resources (e.g. savi
ng water or switching to sustainable resources)',
 'q24.3_Saving energy or switching to sustainable energy sources',
 'q24.4_Developing sustainable products or services',
 'q24.5_Improving working conditions of its employees',
 'q24.6_Promoting and improving diversity and equality in the workplace',
 'q24.7_Evaluating the impact of your enterprise on society',
 'q24.8_Engaging employees in the governance of the enterprise',
 'q24.10_DK/NA',
 'q25_DK/NA',
 'q25_No, and it will not in the future',
 'q25_No, but it may be considered in the future',
 'q25_Not applicable (DO NOT READ OUT)',
 'q25_Yes, and it has already been implemented',
 'q25_Yes, and it is in the process of being implemented',
 'q26.1_Lack of willingness among the management',
 'q26.2_Lack of consumer or customer demand',
 "q26.3_Lack of awareness about how to integrate sustainability into the
enterprise's business model",
 'q26.4_It is not compatible with your current business model',
 'q26.5_It would not be profitable',
 'q26.6_Lack of skills, including managerial skills',
 'q26.7_Lack of financial resources',
 'q26.8_None of the above',
 'q26.9_Other (DO NOT READ OUT)',
 'q26.10_DK/NA']
```

In [15]:

```
variablenames = pd.DataFrame (list_columns, columns = ['variable_name'])
```

```
variablenames.to_csv ('variablenames.csv')
```

In [16]:

```
list_columns = [w.replace(',', '') for w in list_columns]
list_columns = [w.replace('/', '') for w in list_columns]
list_columns = [w.replace('%', '') for w in list_columns]
list_columns = [w.replace("(", '') for w in list_columns]
list_columns = [w.replace(")", '') for w in list_columns]
list_columns = [w.replace("[", '') for w in list_columns]
list_columns = [w.replace("]", '') for w in list_columns]

list_columns
```

Out[16]:

```
['isocntry_0',
 'isocntry_1',
 'nace_a_Arts entertainment and recreation',
 'nace_a_B - Mining and quarrying',
 'nace_a_C - Manufacturing',
 'nace_a_D - Electricity gas steam and air conditioningsupply',
 'nace_a_E - Water supplyseweragewaste managementremediation activ',
 'nace_a_F - Construction',
 'nace_a_G - Wholesale and retail trade repair of motor vehicles and',
 'nace_a_H - Transportation and storage',
 'nace_a_I - Accommodation and food service activities',
 'nace_a_J - Information and communication',
 'nace_a_K - Financial and insurance activities',
 'nace_a_L - Real estate activities',
 'nace_a_M - Professional scientific and technical activities',
 'nace_a_N - Administrative and support service activities',
 'nace_a_P - Education',
 'nace_a_Q - Human health and social work activities',
 'q1_2019 and after',
 'q1_Before 2000',
 'q1_Between 2000 and 2014',
 'q1_Between 2015 and 2018',
 'q1_DKNA',
 'q2t_1 to 9 employees',
 'q2t_10 to 49 employees',
 'q2t_250 employees or more',
 'q2t_50 to 249 employees',
 'q3t_0 employe',
 'q3t_1 to 9 employees',
 'q3t_10 to 49 employees',
 'q3t_250 employees or more',
```

```
 'q3t_50 to 249 employees',
 'q3t_DKNA',
 'q3t_Inap. not 1 in q2a and q2b',
 'q4t_100000 euros or less',
 'q4t_DKNA',
 'q4t_More than 1 million and up to 2 million euros',
 'q4t_More than 10 million and up to 50 million euros',
 'q4t_More than 100000 and up to 500000 euros',
 'q4t_More than 2 million and up to 5 million euros',
 'q4t_More than 5 million and up to 10 million euros',
 'q4t_More than 50 million euros',
 'q4t_More than 500000 and up to 1 million euros',
 'q5_1_DKNA',
 'q5_1_It has decreased',
 'q5_1_It has grown by at least 30',
 'q5_1_It has grown by less than 30',
 'q5_1_It has remained stable',
 'q5_2_DKNA',
 'q5_2_It has decreased',
 'q5_2_It has grown by at least 30',
 'q5_2_It has grown by less than 30',
 'q5_2_It has remained stable',
 'q6_1_DKNA',
 'q6_1_Grow by between 10 and 20 per year',
 'q6_1_Grow by less than 10 per year',
 'q6_1_Grow by more than 20 per year',
 'q6_1_It does not plan to grow',
 'q6_2_DKNA',
 'q6_2_Grow by between 10 and 20 per year',
 'q6_2_Grow by less than 10 per year',
 'q6_2_Grow by more than 20 per year',
 'q6_2_It does not plan to grow',
 'q7a.1_Have a strategic growth plan',
 'q7a.2_Plan to grow as a result of introducing some kind of innovation',
 'q7a.3_Plan to grow as a result of operating in growing markets',
 'q7a.4_Plan to grow as a result of entering new markets',
 'q7a.5_Plan to grow as a result of increased digitalisation in your ente
rprise',
 'q7a.6_Plan to grow in OUR COUNTRY',
 'q7a.7_EU Plan to grow in other EU countries Non-EU',
 'q7a.8_Plan to grow in other non-EU countries',
 'q7a.9_DKNA',
 'q7b.1_There is no intention for your enterprise to grow beyond its curr
ent size',
```

```
 'q7b.2_Your enterprise does not have employees with the skills or expert
ise needed for it to grow',
 'q7b.3_Your enterprise does not have the financial resources to grow',
 "q7b.4_There is decreasing demand for your enterprise's products or serv
ices or the market is saturated",
 'q7b.5_Additional regulatory or administrative burdens and requirements
would be too high for your enterprise to grow',
 'q7b.6_Your enterprise does not want to grow because it would lose benef
its linked to its SME status',
 'q7b.7_The current location of your enterprise does not allow you to gro
w and you do not wish to relocate elsewhere',
 'q7b.8_Your enterprise relies on a few clients which are unlikely to inc
rease their demand',
 'q7b.9_Other DO NOT READ OUT',
 'q7b.10_DKNA',
 'q8.1_In a large town or city',
 'q8.2_In a small town or village',
 'q8.3_In a rural area',
 'q8.4_In an industrial area',
 'q8.5_Near a border with an EU country',
 'q8.6_Near a border with a non-EU country',
 'q8.7_DKNA',
 'q9.1_It mainly provides goods',
 'q9.2_It mainly provides services',
 'q9.3_It sells goods online to buyers in EU countries',
 'q9.4_It is a member of an industry cluster or another SME business supp
ort organisation in the region',
 'q9.5_It is a part of a global value chain',
 'q9.6_It has a patent or patent application',
 'q9.7_It is a non-profit enterprise',
 'q9.8_It has a strategy or action plan to digitalise',
 'q9.9_Other DO NOT READ OUT',
 'q9.10_None DO NOT READ OUT',
 'q9.11_DKNA',
 'q10_DKNA',
 'q10_No definitely not',
 'q10_No probably not',
 'q10_Not applicable DO NOT READ OUT',
 'q10_Yes definitely',
 'q10_Yes probably',
 'q11.1_None your enterprise only operates in OUR COUNTRY',
 'q11.2_EU Other EU countries Non-EU EU countries',
 'q11.3_Other European countries outside of the EU incl. Russia',
 'q11.4_North America',
```

```
 'q11.5_Latin America and the Caribbean',
 'q11.6_China',
 'q11.7_Rest of Asia and the Pacific',
 'q11.8_Middle East and Africa',
 'q11.9_DKNA',
 'q12a_Between 25 and 50',
 'q12a_DKNA',
 'q12a_Less than 25',
 'q12a_More than 50',
 'q12b_Between 25 and 50',
 'q12b_DKNA',
 'q12b_Less than 25',
 'q12b_More than 50',
 'q13.1_Solely owned by one person',
 'q13.2_Owned by more than one person',
 'q13.3_Part of a national or international enterprise group',
 'q13.4_Co-owned by a public entity',
 'q13.5_Co-owned by venture capital firm',
 'q13.6_Co-owned by business angel',
 'q13.7_Predominantly family owned',
 'q13.8_Jointly owned by its members e.g. cooperative mutual society',
 'q13.9_Other DO NOT READ OUT',
 'q13.10_DKNA',
 'q14.1_The sole founder of this enterprise',
 'q14.2_A co-founder of this enterprise',
 'q14.3_The sole owner of this enterprise',
 'q14.4_A co-owner of this enterprise',
 'q14.5_None of the above',
 'q14.6_DKNA',
 'q15a.1_This is the first enterprise that you have ever established',
 'q15a.2_You have established or co-established other enterprises',
 'q15a.3_You have closed - without bankruptcy - other enterprises that yo
u owned or co-owned',
 'q15a.4_You have closed - due to bankruptcy - other enterprises that you
owned or co-owned',
 'q15a.5_You have sold other enterprises that you owned or co-owned',
 'q15a.6_You plan to relocate the headquarters of your enterprise to an E
U country in the future',
 'q15a.7_You plan to relocate the headquarters of your enterprise to the
USA in the future',
 'q15a.8_You plan to relocate the headquarters of your enterprise to any
other country in the future',
 'q15a.9_Other DO NOT READ OUT',
 'q15a.10_None DO NOT READ OUT',
```

```
 'q15a.11_DKNA',
 'q15b.1_You took this enterprise over from family members',
 'q15b.2_You have established or co-established other enterprises',
 'q15b.3_You have closed - without bankruptcy - other enterprises that yo
u owned or co-owned',
 'q15b.4_You have closed - due to bankruptcy - other enterprises that you
owned or co-owned',
 'q15b.5_You have sold other enterprises that you owned or co-owned',
 'q15b.6_You plan to relocate the headquarters of your enterprise to an E
U country in the future',
 'q15b.7_You plan to relocate the headquarters of your enterprise to the
USA in the future',
 'q15b.8_You plan to relocate the headquarters of your enterprise to any
other country in the future',
 'q15b.9_Other DO NOT READ OUT',
 'q15b.10_None DO NOT READ OUT',
 'q15b.11_DKNA',
 'q16_1_DKNA',
 'q16_1_Fairly good',
 'q16_1_Fairly poor',
 'q16_1_Very good',
 'q16_1_Very poor',
 'q16_2_DKNA',
 'q16_2_Fairly good',
 'q16_2_Fairly poor',
 'q16_2_Very good',
 'q16_2_Very poor',
 'q16_3_DKNA',
 'q16_3_Fairly good',
 'q16_3_Fairly poor',
 'q16_3_Very good',
 'q16_3_Very poor',
 'q16_4_DKNA',
 'q16_4_Fairly good',
 'q16_4_Fairly poor',
 'q16_4_Very good',
 'q16_4_Very poor',
 'q16_5_DKNA',
 'q16_5_Fairly good',
 'q16_5_Fairly poor',
 'q16_5_Very good',
 'q16_5_Very poor',
 'q16_6_DKNA',
 'q16_6_Fairly good',
```

```
 'q16_6_Fairly poor',
 'q16_6_Very good',
 'q16_6_Very poor',
 'q16_7_DKNA',
 'q16_7_Fairly good',
 'q16_7_Fairly poor',
 'q16_7_Very good',
 'q16_7_Very poor',
 'q16_8_DKNA',
 'q16_8_Fairly good',
 'q16_8_Fairly poor',
 'q16_8_Very good',
 'q16_8_Very poor',
 'q17.1_Difficulties with innovation',
 'q17.2_Regulatory obstacles or administrative burden',
 'q17.3_Access to data',
 'q17.4_Internationalisation',
 'q17.5_Access to finance',
 'q17.6_Payment delays',
 'q17.7_Skills including managerial skills',
 'q17.8_Difficulties with digitalisation',
 'q17.9_Other DO NOT READ OUT',
 'q17.10_DKNA',
 'q19.1_A new or significantly improved product or service to the market'
,
 'q19.2_A new or significantly improved production process or method',
 'q19.3_A new organisation of management or a new business model',
 'q19.4_A new way of selling your goods or services',
 'q19.5_An innovation with an environmental benefit including innovations
with an energy or resource efficiency benefit',
 'q19.6_Social innovations such as new products services or processes tha
t have the aim of improving society',
 'q19.7_Any other type of innovation',
 'q19.8_No none',
 'q19.9_DKNA',
 'q20.1_Lack of technology infrastructure',
 'q20.2_Lack of skills including managerial skills',
 'q20.3_Difficulties in predicting the market response',
 'q20.4_Lack of collaboration partners such as other enterprises etc. for
innovation projects',
 'q20.5_Legal or administrative environment',
 'q20.6_Lack of financial resources including from available support sche
mes',
 'q20.7_Difficulties with protecting intellectual property',
```

```
 'q20.8_None of these',
 'q20.9_Your enterprise has no interest in innovating DO NOT READ OUT',
 'q20.10_Other DO NOT READ OUT',
 'q20.11_DKNA',
 'q21.1_Lack of financial resources',
 'q21.2_Lack of skills including managerial skills',
 'q21.3_Lack of information technology infrastructure such as high-speed
internet connection',
 'q21.4_Regulatory obstacles',
 'q21.5_IT security issues',
 'q21.6_Uncertainty about future digital standards',
 'q21.7_Internal resistance to change',
 'q21.8_None of these',
 'q21.9_Your enterprise has no interest in digitalisation DO NOT READ OUT
',
 'q21.10_Other DO NOT READ OUT',
 'q21.11_DKNA',
 'q21.11_no  DKNA',
 'q22_DKNA',
 'q22_None DO NOT READ OUT',
 'q22_Other DO NOT READ OUT',
 'q22_There is a need to introduce advanced digital technologies and your
enterprise has already started to adopt them',
 'q22_There is a need to introduce advanced digital technologies and your
enterprise is currently considering ...',
 'q22_There is a need to introduce advanced digital technologies but your
enterprise does not have the knowledge ...',
 'q22_Your enterprise does not need to adopt any digital technologies',
 'q22_Your enterprise has adoptedis planning to adopt basic digital techn
ologies but not advanced digital technologies ...',
 'q23.1_Artificial intelligence e.g. machine learning or technologies ide
ntifying objects or persons etc.',
 'q23.2_Cloud computing i.e. storing and processing files or data on remo
te servers hosted on the internet',
 'q23.3_Robotics i.e. robots used to automate processes for example in co
nstruction or design etc.',
 'q23.4_Smart devices e.g. smart sensors smart thermostats etc.',
 'q23.5_Big data analytics e.g. data mining and predictive analysis',
 'q23.6_High speed infrastructure',
 'q23.7_Blockchain',
 'q23.8_None of these',
 'q23.9_DKNA',
 'q24.1_Recycling or reusing materials',
```

```
 'q24.2_Reducing consumption of or impact on natural resources e.g. savin
g water or switching to sustainable resources',
 'q24.3_Saving energy or switching to sustainable energy sources',
 'q24.4_Developing sustainable products or services',
 'q24.5_Improving working conditions of its employees',
 'q24.6_Promoting and improving diversity and equality in the workplace',
 'q24.7_Evaluating the impact of your enterprise on society',
 'q24.8_Engaging employees in the governance of the enterprise',
 'q24.10_DKNA',
 'q25_DKNA',
 'q25_No and it will not in the future',
 'q25_No but it may be considered in the future',
 'q25_Not applicable DO NOT READ OUT',
 'q25_Yes and it has already been implemented',
 'q25_Yes and it is in the process of being implemented',
 'q26.1_Lack of willingness among the management',
 'q26.2_Lack of consumer or customer demand',
 "q26.3_Lack of awareness about how to integrate sustainability into the
enterprise's business model",
 'q26.4_It is not compatible with your current business model',
 'q26.5_It would not be profitable',
 'q26.6_Lack of skills including managerial skills',
 'q26.7_Lack of financial resources',
 'q26.8_None of the above',
 'q26.9_Other DO NOT READ OUT',
 'q26.10_DKNA']
```

In [17]:

```
df_encoded.columns = list_columns
df_encoded = df_encoded.reset_index()
df_encoded
```

(...)

12115 rows × 286 columns

In [18]:

```
df_encoded['environmental_reduce_recible_reuse']=df_encoded['q24.2_Reduci
ng consumption of or impact on natural resources e.g. saving water or swi
tching to sustainable resources'] + df_encoded['q24.3_Saving energy or sw
itching to sustainable energy sources']+df_encoded['q24.1_Recycling or re
using materials']
df_encoded['environmental_reduce_recible_reuse'].value_counts()
```

Out[18]:

```
3    3824
0    2814
```

```
1    2785
2    2692
Name: environmental_reduce_recible_reuse, dtype: int64
```

In [19]:

```
df_encoded['environmental_reduce_recible_reuse']=df_encoded['environmenta
l_reduce_recible_reuse'].replace([1,2,3],1)
df_encoded['environmental_reduce_recible_reuse'].value_counts()
```

Out[19]:

```
1    9301
0    2814
Name: environmental_reduce_recible_reuse, dtype: int64
```

In [20]:

```
def frequency(ds, vars):
    if len(vars) > 1:
        c1 = ds[vars[0]]
        c2 = []
        for i in range(1,len(vars)):
            c2.append(ds[vars[i]])
        dfs = []
        dfs.append(pd.crosstab(c1,c2).unstack().reset_index().rename(colu
mns={0:'Count'}))
        dfs.append(pd.crosstab(c1,c2, normalize='all').unstack().reset_in
dex().rename(columns={0:'Percent'}))
        dfs.append(pd.crosstab(c1,c2, normalize='columns').unstack().rese
t_index().rename(columns={0:'Column Percent'}))
        dfs.append(pd.crosstab(c1,c2, normalize='index').unstack().reset_
index().rename(columns={0:'Row Percent'}))
        dfs = [df.set_index(vars) for df in dfs]
        df = dfs[0].join(dfs[1:]).reset_index()
        return df
```

In [21]:

```
data = df_encoded[['environmental_reduce_recible_reuse','q24.4_Developing
sustainable products or services']]
data
```

(...)

12115 rows × 2 columns

In [22]:

```
frequency(data,['environmental_reduce_recible_reuse','q24.4_Developing su
stainable products or services'])
```

(...)

```
df_encoded['environmental_reduce_recible_reuse'].dtypes
```

Out[23]:

```
dtype('int64')
```

In [ ]:

In [24]:

```
columna1=df_encoded.columns.get_loc('environmental_reduce_recible_reuse')
columna2=df_encoded.columns.get_loc('q24.4_Developing sustainable product
s or services')
print(columna1)
print(columna2)

df_encoded['cluster']=0
df_encoded['cluster'].dtypes
```

```
286
264
```

Out[24]:

```
dtype('int64')
```

In [25]:

```
for i in range(0,12115):
    if df_encoded.iloc[i, columna1]== 0 and df_encoded.iloc[i, columna2]=
= 0:
        df_encoded.loc[i, 'cluster']=0
    if df_encoded.iloc[i, columna1]==1 and df_encoded.iloc[i, columna2]==
0:
        df_encoded.loc[i, 'cluster']=1
    if df_encoded.iloc[i, columna1]==0 and df_encoded.iloc[i, columna2]==
1:
        df_encoded.loc[i, 'cluster']=3
    if df_encoded.iloc[i, columna1]==1 and df_encoded.iloc[i, columna2]==
1:
        df_encoded.loc[i, 'cluster']=2

df_encoded
```

(...)

12115 rows × 288 columns

In [26]:

```
cluster_selection=[0,1,2]
selection=[0]
```

In [27]:

```
df_encoded=df_encoded[df_encoded.cluster.isin(cluster_selection)]

df_encoded=df_encoded[df_encoded['q24.10_DKNA'].isin(selection)]
df_encoded
```

(...)

```
df_encoded['cluster']
```

Out[28]:

```
0         1
1         1
2         2
3         2
4         0
        ..
12110     1
12111     2
12112     1
12113     1
12114     1
Name: cluster, Length: 11658, dtype: int64
```

In [29]:

```
df_encoded['cluster'].value_counts()
```

Out[29]:

```
1    5713
2    3588
0    2357
Name: cluster, dtype: int64
```

In [30]:

```
df_encoded.to_csv ('Enterprises.csv')
```

.

STEP 2.1

PARTE 1. CARGAMOS LAS LIBRERÍAS

In [1]:

```
#https://amirali-n.github.io/BorutaFeatureSelectionWithShapAnalysis/

# First XGBoost model for Pitec dataset
import matplotlib.pyplot as plt
```

```
import xgboost as xgb
from numpy import loadtxt
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
# XGBoost kfold cross validation
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
# XGBoost stratified kfold cross validation
from sklearn.model_selection import StratifiedKFold
# one hot ecoding
from numpy import column_stack
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
# pandas
import pandas as pd
```

In [2]:

```
# Read data from file 'filename.csv'
# (in the same directory that your python process is based)
# Control delimiters, rows, column names with read_csv (see later) sep=se
parador en el csv decimal=separador decimal de los números. index_col=0 e
stablece la primera columna como índice
df_encoded = pd.read_csv("Enterprises.csv", sep=",", decimal=".", index_c
ol=0)
df_encoded.shape
```

Out[2]:

```
(11658, 288)
```

In [3]:

```
df_encoded.head()
```

(...)

5 rows × 288 columns

In [4]:

```
#list(df_encoded.columns)
```

In [5]:

```
df_encoded=df_encoded.drop(['environmental_reduce_recible_reuse',
 'q24.1_Recycling or reusing materials',
 'q24.2_Reducing consumption of or impact on natural resources e.g. savin
g water or switching to sustainable resources',
 'q24.3_Saving energy or switching to sustainable energy sources',
 'q24.4_Developing sustainable products or services',
],axis=1)
```

In [6]:

```
y=df_encoded['cluster']
```

In [7]:

```
#vemos la distribución de casos de y para ver si es imbalanced (grandes d
iferencias en el número de casos entre grupos
#o más o menos balanceado
from collections import Counter
# summarize the class distribution
counter = Counter(y)
for k,v in counter.items():
    per = v / len(y) * 100
    print('Class=%d, Count=%d, Percentage=%.3f%%' % (k, v, per))
```

```
Class=1, Count=5713, Percentage=49.005%
Class=2, Count=3588, Percentage=30.777%
Class=0, Count=2357, Percentage=20.218%
```

In [8]:

```
X=df_encoded.drop(['cluster', 'index'],axis=1)
X
```

Out[8]:

11658 rows × 281 columns

In [9]:

```
#list(X.columns)
```

In [10]:

```
my_list_encoded = df_encoded.columns.values.tolist()
#print(my_list_encoded)
```

for name in X.columns: print(name) print(X[name].value_counts()) print("") print("")

In [11]:

```
#El resultado parece bastante balanceado por lo que no es necesario tomar
medidas adicionales
# necesarias cuando es imbalance

#CARGAMOS LIBRERIAS para la siguiente fase
from numpy import mean
from numpy import std
from sklearn.metrics import brier_score_loss
from sklearn.metrics import make_scorer
from sklearn.dummy import DummyClassifier
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from matplotlib import pyplot
```

```
# evaluate a model
def evaluate_model(X, y, model):
  # define evaluation procedure
  cv = RepeatedStratifiedKFold(n_splits=5, n_repeats=3, random_state=1)
  # evaluate model
  scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs
=-1)
  return scores
```

In [12]:

```
#cargamos librerías de los distintos algoritmos
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import BaggingClassifier

# define models to test
def get_models():
    models, names = list(), list()
    # Dummy classifier
    models.append(DummyClassifier(strategy='most_frequent'))
    names.append('DC')
    #LR
    models.append(LogisticRegression(multi_class='multinomial', solver='l
bfgs'))
    names.append('LR')
     # LDA
    models.append(LinearDiscriminantAnalysis())
    names.append('LDA')
    # QDA
    models.append(QuadraticDiscriminantAnalysis())
    names.append('QDA')
    # GNB
    models.append(GaussianNB())
    names.append('GNB')
    # MNB
```

```
    models.append(MultinomialNB())
    names.append('MNB')
    # GPC
    #models.append(GaussianProcessClassifier())
    #names.append('GPC')
    # SVM
    models.append(SVC(gamma='auto'))
    names.append('SVM')
    # KNN
    models.append(KNeighborsClassifier())
    names.append('KNN')

    # Bagging
    models.append(BaggingClassifier(n_estimators=200))
    names.append('BAG')
    # RF
    models.append(RandomForestClassifier(n_estimators=200))
    names.append('RF')
    # ET
    models.append(ExtraTreesClassifier(n_estimators=200))
    names.append('ET')
    # XGBoost
    models.append(XGBClassifier())
    names.append('XGBC')
    return models, names
```

In [13]:

```
#vamos a calcular la precisión de los distintos métodos de machine learni
ng para ver cual es mas eficiente como clasificador
# define models
models, names = get_models()
results = list()
# evaluate each model
for i in range(len(models)):
 # evaluate the model and store results
 scores = evaluate_model(X, y, models[i])
 results.append(scores)
 # summarize performance
 print('>%s %.3f (%.3f)' % (names[i], mean(scores), std(scores)))
# plot the results
pyplot.boxplot(results, labels=names, showmeans=True)
pyplot.show()
```

```
>DC 0.490 (0.000)
>LR 0.592 (0.010)
```

```
>LDA 0.596 (0.010)
>QDA 0.375 (0.020)
>GNB 0.411 (0.011)
>MNB 0.514 (0.009)
/opt/anaconda3/lib/python3.7/site-packages/joblib/externals/loky/process_
executor.py:706: UserWarning: A worker stopped while some jobs were given
to the executor. This can be caused by a too short worker timeout or by a
memory leak.
  "timeout or by a memory leak.", UserWarning
>SVM 0.583 (0.007)
>KNN 0.503 (0.007)
>BAG 0.590 (0.008)
>RF 0.595 (0.010)
>ET 0.598 (0.009)
>XGBC 0.598 (0.010)
.
```

STEP 2.2

PARTE 1. CARGAMOS LAS LIBRERÍAS

In [1]:

```
#https://amirali-n.github.io/BorutaFeatureSelectionWithShapAnalysis/

# First XGBoost model for Pitec dataset
import matplotlib.pyplot as plt
import xgboost as xgb
from numpy import loadtxt
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
# XGBoost kfold cross validation
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
# XGBoost stratified kfold cross validation
from sklearn.model_selection import StratifiedKFold
# one hot ecoding
from numpy import column_stack
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
# pandas
import pandas as pd
```

In [2]:

```
# Read data from file 'filename.csv'
# (in the same directory that your python process is based)
# Control delimiters, rows, column names with read_csv (see later) sep=se
parador en el csv decimal=separador decimal de los números. index_col=0 e
stablece la primera columna como índice
df_encoded = pd.read_csv("Enterprises.csv", sep=",", decimal=".", index_c
ol=0)
df_encoded.shape
```

Out[2]:

```
(11658, 288)
```

In [3]:

```
df_encoded.head()
```

(…)

5 rows × 288 columns

In [4]:

```
#list(df_encoded.columns)
```

In [5]:

```
df_encoded=df_encoded.drop(['environmental_reduce_recible_reuse',
 'q24.1_Recycling or reusing materials',
 'q24.2_Reducing consumption of or impact on natural resources e.g. savin
g water or switching to sustainable resources',
 'q24.3_Saving energy or switching to sustainable energy sources',
 'q24.4_Developing sustainable products or services',
],axis=1)
```

In [6]:

```
y=df_encoded['cluster']
```

In [7]:

```
#vemos la distribución de casos de y para ver si es imbalanced (grandes d
iferencias en el número de casos entre grupos
#o más o menos balanceado
from collections import Counter
# summarize the class distribution
counter = Counter(y)
for k,v in counter.items():
    per = v / len(y) * 100
    print('Class=%d, Count=%d, Percentage=%.3f%%' % (k, v, per))
```

```
Class=1, Count=5713, Percentage=49.005%
Class=2, Count=3588, Percentage=30.777%
Class=0, Count=2357, Percentage=20.218%
```

In [8]:

```
X=df_encoded.drop(['cluster', 'index'],axis=1)
X
```

(...)

11658 rows × 281 columns

In [9]:

```
#list(X.columns)
```

In [10]:

```
my_list_encoded = df_encoded.columns.values.tolist()
#print(my_list_encoded)
```

for name in X.columns: print(name) print(X[name].value_counts()) print("") print("")

In [11]:

```
# split data into train and test sets. Primera aproximación al problema
#Separamos los datos en test y train (33% y 66% respectivamente)

test_size = 0.3
X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=test_size, random_state=0,stratify=y)


# fit model on training data mlogloss porque son varias categorías
model = XGBClassifier()
eval_set = [(X_test, y_test)]
modelo=model.fit(X_train, y_train, eval_metric=["mlogloss"], eval_set=eva
l_set, verbose=False)

#verbose=True para que salgan todos los pasos de la cálculo de mlogloss p
ara cada arbol
print(model)
# make predictions for test data
predictions = model.predict(X_test)
# evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
XGBClassifier(objective='multi:softprob')
Accuracy: 61.58%
```

In [12]:

```
#Vamos ahora con el algoritmo definitivo para elegir las variables releva
nt
from BorutaShap import BorutaShap
```

```
# load X and y
# NOTE BorutaPy accepts numpy arrays only, hence the .values attribute y
= np.array(y_train)
#X = X_train
#y = y_train

# no model selected default is Random Forest, if classification is True i
t is a Classification problem
model = XGBClassifier()

# if classification is False it is a Regression problem
Feature_Selector = BorutaShap(model=model,
                              importance_measure='shap',
                              classification=True)

Feature_Selector.fit(X=X_train, y=y_train, n_trials=100, sample=False,
train_or_test = 'test', normalize=True, verbose=True)


'''
Sample: Boolean
        if true then a rowise sample of the data will be used to calculate
the feature importance values

sample_fraction: float
        The sample fraction of the original data used in calculating the f
eature importance values only
        used if Sample==True.

train_or_test: string
        Decides whether the feature improtance should be calculated on out
of sample data see the dicussion here.
        https://compstat-lmu.github.io/iml_methods_limitations/pfi-data.h
tml#introduction-to-test-vs.training-data

normalize: boolean
            if true the importance values will be normalized using the z-
score formula

verbose: Boolean
        a flag indicator to print out all the rejected or accepted feature
s.
'''
```

```
#the decision which kind of data you want to use depends on the question
you are interested in:
#How much does the model rely on the respective variable to make predicti
ons? This question leads to a calculation based on the training data.
#The second possible question is as follows: How much does the feature co
ntribute to model performance on unknown data? In this case, the test dat
a would be used.
```

```
100%|██████████| 100/100 [17:33<00:00, 10.53s/it]
37 attributes confirmed important: ['q23.6_High speed infrastructure', 'q
19.8_No none', 'q23.8_None of these', 'q19.6_Social innovations such as n
ew products services or processes that have the aim of improving society'
, 'q26.7_Lack of financial resources', 'q12a_Less than 25', 'q23.5_Big da
ta analytics e.g. data mining and predictive analysis', 'q17.2_Regulatory
obstacles or administrative burden', 'q24.5_Improving working conditions
of its employees', 'q9.6_It has a patent or patent application', 'q10_Yes
probably', 'q9.2_It mainly provides services', 'q16_3_Fairly good', 'q23.
4_Smart devices e.g. smart sensors smart thermostats etc.', 'q24.8_Engagi
ng employees in the governance of the enterprise', 'q24.6_Promoting and i
mproving diversity and equality in the workplace', 'q20.3_Difficulties in
predicting the market response', 'q26.4_It is not compatible with your cu
rrent business model', 'q9.8_It has a strategy or action plan to digitali
se', 'q26.8_None of the above', 'q25_Yes and it has already been implemen
ted', 'q19.1_A new or significantly improved product or service to the ma
rket', 'q19.5_An innovation with an environmental benefit including innov
ations with an energy or resource efficiency benefit', 'q25_Yes and it is
in the process of being implemented', 'q16_6_DKNA', 'q25_No but it may be
considered in the future', 'q25_Not applicable DO NOT READ OUT', 'q25_No
and it will not in the future', 'isocntry_0', 'q1_Before 2000', 'q24.7_Ev
aluating the impact of your enterprise on society', 'q7a.3_Plan to grow a
s a result of operating in growing markets', 'q19.2_A new or significantl
y improved production process or method', 'q4t_100000 euros or less', 'q1
3.7_Predominantly family owned', 'q11.3_Other European countries outside
of the EU incl. Russia', 'q9.5_It is a part of a global value chain']
237 attributes confirmed unimportant: ['q16_6_Fairly poor', 'q16_8_Fairly
good', 'q5_1_It has remained stable', 'q14.4_A co-owner of this enterpris
e', 'q4t_More than 2 million and up to 5 million euros', 'q16_5_DKNA', 'q
17.3_Access to data', 'q6_2_Grow by between 10 and 20 per year', 'q13.4_C
o-owned by a public entity', 'q14.3_The sole owner of this enterprise', '
q16_3_Fairly poor', 'q16_7_Very good', 'q22_There is a need to introduce
advanced digital technologies but your enterprise does not have the knowl
edge ...', 'q15b.4_You have closed - due to bankruptcy - other enterprise
s that you owned or co-owned', 'q15b.11_DKNA', 'q26.6_Lack of skills incl
uding managerial skills', 'q7a.9_DKNA', 'q7b.6_Your enterprise does not w
ant to grow because it would lose benefits linked to its SME status', 'q1
```

5b.10_None DO NOT READ OUT', 'q4t_DKNA', 'q14.6_DKNA', 'q15a.3_You have c
losed - without bankruptcy - other enterprises that you owned or co-owned
', 'q15a.5_You have sold other enterprises that you owned or co-owned', '
q16_5_Very poor', 'q16_7_Fairly good', 'q20.2_Lack of skills including ma
nagerial skills', 'q6_1_It does not plan to grow', 'q4t_More than 50 mill
ion euros', 'q16_7_Very poor', 'q3t_50 to 249 employees', 'q12b_Less than
25', 'q20.9_Your enterprise has no interest in innovating DO NOT READ OUT
', 'q13.6_Co-owned by business angel', 'q2t_1 to 9 employees', 'q15a.7_Yo
u plan to relocate the headquarters of your enterprise to the USA in the
future', 'q10_No probably not', 'q7b.5_Additional regulatory or administr
ative burdens and requirements would be too high for your enterprise to g
row', 'q14.2_A co-founder of this enterprise', 'q15a.8_You plan to reloca
te the headquarters of your enterprise to any other country in the future
', 'q17.10_DKNA', 'q4t_More than 5 million and up to 10 million euros', '
q17.1_Difficulties with innovation', 'q15a.4_You have closed - due to ban
kruptcy - other enterprises that you owned or co-owned', 'q21.6_Uncertain
ty about future digital standards', 'q3t_1 to 9 employees', 'q8.1_In a la
rge town or city', 'q8.2_In a small town or village', 'q19.7_Any other ty
pe of innovation', 'q8.7_DKNA', 'q20.8_None of these', 'q21.11_no DKNA',
'q6_1_DKNA', 'q16_6_Very good', 'q23.7_Blockchain', 'q9.10_None DO NOT RE
AD OUT', 'q4t_More than 1 million and up to 2 million euros', 'q6_2_It do
es not plan to grow', 'q16_1_Very poor', 'q20.7_Difficulties with protect
ing intellectual property', 'q10_Not applicable DO NOT READ OUT', 'q21.5_
IT security issues', 'q22_DKNA', 'isocntry_1', "q7b.4_There is decreasing
demand for your enterprise's products or services or the market is satura
ted", 'q16_8_Fairly poor', 'q26.9_Other DO NOT READ OUT', 'q16_7_DKNA', '
q8.6_Near a border with a non-EU country', 'q2t_50 to 249 employees', 'na
ce_a_G - Wholesale and retail trade repair of motor vehicles and', 'q16_4
_Very good', 'nace_a_H - Transportation and storage', 'nace_a_N - Adminis
trative and support service activities', 'q16_4_Fairly good', 'q7b.2_Your
enterprise does not have employees with the skills or expertise needed fo
r it to grow', 'q12a_Between 25 and 50', 'q11.6_China', 'q12a_DKNA', 'q17
.9_Other DO NOT READ OUT', 'q11.4_North America', 'q4t_More than 500000 a
nd up to 1 million euros', 'q12b_DKNA', 'q16_8_DKNA', 'q5_2_It has remain
ed stable', 'q15a.6_You plan to relocate the headquarters of your enterpr
ise to an EU country in the future', 'q16_2_Fairly good', 'q15b.7_You pla
n to relocate the headquarters of your enterprise to the USA in the futur
e', 'nace_a_K - Financial and insurance activities', 'nace_a_B - Mining a
nd quarrying', 'q19.3_A new organisation of management or a new business
model', 'q3t_Inap. not 1 in q2a and q2b', 'q7b.10_DKNA', 'q16_3_Very good
', 'q17.6_Payment delays', 'q13.2_Owned by more than one person', 'q16_8_
Very poor', 'q2t_10 to 49 employees', 'q11.9_DKNA', 'q21.2_Lack of skills
including managerial skills', 'q11.1_None your enterprise only operates i
n OUR COUNTRY', 'q23.2_Cloud computing i.e. storing and processing files

or data on remote servers hosted on the internet', 'q25_DKNA', 'q5_2_It h
as grown by less than 30', 'q21.8_None of these', 'q23.9_DKNA', 'q20.1_La
ck of technology infrastructure', 'q5_1_It has decreased', 'q16_2_Very po
or', 'q4t_More than 10 million and up to 50 million euros', 'q12b_Between
25 and 50', 'q16_1_Fairly good', 'q22_None DO NOT READ OUT', 'q17.7_Skill
s including managerial skills', 'q3t_0 employe', 'q5_2_It has grown by at
least 30', 'q1_DKNA', 'q15b.6_You plan to relocate the headquarters of yo
ur enterprise to an EU country in the future', 'q13.5_Co-owned by venture
capital firm', 'q15b.2_You have established or co-established other enter
prises', 'q21.9_Your enterprise has no interest in digitalisation DO NOT
READ OUT', 'q15a.11_DKNA', 'q10_No definitely not', 'q5_1_It has grown by
at least 30', 'q16_6_Very poor', 'q7b.1_There is no intention for your en
terprise to grow beyond its current size', 'q22_There is a need to introd
uce advanced digital technologies and your enterprise is currently consid
ering ...', 'q9.11_DKNA', 'q15a.2_You have established or co-established
other enterprises', 'nace_a_J - Information and communication', 'nace_a_D
- Electricity gas steam and air conditioningsupply', 'q7a.6_Plan to grow
in OUR COUNTRY', 'q20.10_Other DO NOT READ OUT', 'q16_8_Very good', 'q1_B
etween 2015 and 2018', 'q20.4_Lack of collaboration partners such as othe
r enterprises etc. for innovation projects', 'nace_a_Arts entertainment a
nd recreation', 'q21.10_Other DO NOT READ OUT', 'q8.4_In an industrial ar
ea', 'q21.1_Lack of financial resources', 'q9.9_Other DO NOT READ OUT', '
q11.5_Latin America and the Caribbean', 'q16_4_Fairly poor', 'q20.5_Legal
or administrative environment', 'q6_1_Grow by more than 20 per year', 'q1
3.10_DKNA', 'q5_1_It has grown by less than 30', 'nace_a_P - Education',
'nace_a_M - Professional scientific and technical activities', 'q5_2_DKNA
', 'q7a.7_EU Plan to grow in other EU countries Non-EU', 'q17.4_Internati
onalisation', 'q15b.9_Other DO NOT READ OUT', 'q15b.1_You took this enter
prise over from family members', 'q13.8_Jointly owned by its members e.g.
cooperative mutual society', 'q22_Your enterprise does not need to adopt
any digital technologies', 'q11.7_Rest of Asia and the Pacific', 'q16_3_D
KNA', 'q6_1_Grow by less than 10 per year', 'q12a_More than 50', 'q6_2_DK
NA', 'q16_1_Very good', 'q4t_More than 100000 and up to 500000 euros', 'n
ace_a_E - Water supplyseweragewaste managementremediation activ', 'q15b.3
_You have closed - without bankruptcy - other enterprises that you owned
or co-owned', 'nace_a_Q - Human health and social work activities', 'q14.
5_None of the above', 'q15b.8_You plan to relocate the headquarters of yo
ur enterprise to any other country in the future', 'q21.11_DKNA', 'q16_3_
Very poor', 'q16_5_Very good', 'q9.7_It is a non-profit enterprise', 'q26
.10_DKNA', 'q15a.1_This is the first enterprise that you have ever establ
ished', 'q13.9_Other DO NOT READ OUT', 'q13.3_Part of a national or inter
national enterprise group', 'q23.3_Robotics i.e. robots used to automate
processes for example in construction or design etc.', 'q7a.1_Have a stra
tegic growth plan', 'q26.2_Lack of consumer or customer demand', 'q3t_250

```
employees or more', 'nace_a_I - Accommodation and food service activities
', 'q7a.5_Plan to grow as a result of increased digitalisation in your en
terprise', 'q5_1_DKNA', 'q16_4_DKNA', 'q1_2019 and after', 'q10_Yes defin
itely', 'q7b.7_The current location of your enterprise does not allow you
to grow and you do not wish to relocate elsewhere', 'q26.5_It would not b
e profitable', 'q16_5_Fairly poor', 'q1_Between 2000 and 2014', 'q9.4_It
is a member of an industry cluster or another SME business support organi
sation in the region', 'q16_2_Very good', 'nace_a_L - Real estate activit
ies', 'q16_1_DKNA', 'q21.4_Regulatory obstacles', 'q24.10_DKNA', 'q17.5_A
ccess to finance', 'q16_6_Fairly good', 'q11.2_EU Other EU countries Non-
EU EU countries', 'q19.9_DKNA', 'q15a.10_None DO NOT READ OUT', 'q8.5_Nea
r a border with an EU country', 'q9.3_It sells goods online to buyers in
EU countries', 'q12b_More than 50', 'q22_Other DO NOT READ OUT', 'q15b.5_
You have sold other enterprises that you owned or co-owned', 'q7a.8_Plan
to grow in other non-EU countries', 'q6_2_Grow by less than 10 per year',
'q16_4_Very poor', 'q11.8_Middle East and Africa', 'q16_1_Fairly poor', '
q16_2_Fairly poor', 'q21.3_Lack of information technology infrastructure
such as high-speed internet connection', 'q7a.4_Plan to grow as a result
of entering new markets', 'q26.1_Lack of willingness among the management
', 'q3t_DKNA', 'q8.3_In a rural area', 'q13.1_Solely owned by one person'
, 'q15a.9_Other DO NOT READ OUT', 'q16_5_Fairly good', 'q22_There is a ne
ed to introduce advanced digital technologies and your enterprise has alr
eady started to adopt them', 'q10_DKNA', 'nace_a_C - Manufacturing', 'q6_
2_Grow by more than 20 per year', 'q2t_250 employees or more', 'nace_a_F
- Construction', 'q9.1_It mainly provides goods', 'q16_2_DKNA', 'q20.6_La
ck of financial resources including from available support schemes', 'q20
.11_DKNA', 'q16_7_Fairly poor', 'q6_1_Grow by between 10 and 20 per year'
, 'q7b.8_Your enterprise relies on a few clients which are unlikely to in
crease their demand', 'q3t_10 to 49 employees', 'q5_2_It has decreased',
'q23.1_Artificial intelligence e.g. machine learning or technologies iden
tifying objects or persons etc.', 'q7b.3_Your enterprise does not have th
e financial resources to grow', 'q7b.9_Other DO NOT READ OUT']
7 tentative attributes remains: ['q21.7_Internal resistance to change', '
q14.1_The sole founder of this enterprise', 'q7a.2_Plan to grow as a resu
lt of introducing some kind of innovation', 'q22_Your enterprise has adop
tedis planning to adopt basic digital technologies but not advanced digit
al technologies ...', "q26.3_Lack of awareness about how to integrate sus
tainability into the enterprise's business model", 'q17.8_Difficulties wi
th digitalisation', 'q19.4_A new way of selling your goods or services']
```

Out[12]:

```
'\nSample: Boolean\n\tif true then a rowise sample of the data will be us
ed to calculate the feature importance values\n\nsample_fraction: float\n
\tThe sample fraction of the original data used in calculating the featur
e importance values only\n        used if Sample==True.\n\ntrain_or_test:
```

```
string\n\tDecides whether the feature improtance should be calculated on
out of sample data see the dicussion here.\n        https://compstat-lmu.
github.io/iml_methods_limitations/pfi-data.html#introduction-to-test-vs.t
raining-data\n\nnormalize: boolean\n            if true the importance va
lues will be normalized using the z-score formula\n\nverbose: Boolean\n\t
a flag indicator to print out all the rejected or accepted features.\n'
```

In [13]:

```
 # Returns a subset of the original data with the selected features
subset = Feature_Selector.Subset()
train_BorutaShap=pd.concat([subset, y_train], axis=1)
# guardamos el fichero con los datos subset. Estos son con las variables
finales de X_train c
train_BorutaShap.to_csv ('train_BorutaShap_enterprises.csv')

#filtramos estas variables también en la matriz X_test
# filtramos solo las columnas que se encuentran en la lista de variables
seleccionadas
select_X_test=X_test.filter(list(subset.columns))
test_BorutaShap=pd.concat([select_X_test, y_test], axis=1)
test_BorutaShap.to_csv ('test_BorutaShap_enterprises.csv')
```

.

STEP 3

PARTE 1. CARGAMOS LAS LIBRERÍAS

In [1]:

```
#https://amirali-n.github.io/BorutaFeatureSelectionWithShapAnalysis/

# First XGBoost model for Pitec dataset
import matplotlib.pyplot as plt
import xgboost as xgb
from numpy import loadtxt
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
# XGBoost kfold cross validation
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score


# XGBoost stratified kfold cross validation
```

```
from sklearn.model_selection import StratifiedKFold

# one hot ecoding
from numpy import column_stack
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder

# pandas
import pandas as pd
```

In [2]:

```
# Read data from file 'filename.csv'
# (in the same directory that your python process is based)
# Control delimiters, rows, column names with read_csv (see later) sep=se
parador en el csv decimal=separador decimal de los números
train = pd.read_csv("train_BorutaShap_enterprises.csv", sep=",", decimal=
".",index_col=0)

test = pd.read_csv("test_BorutaShap_enterprises.csv", sep=",", decimal=".
",index_col=0)


train.shape
```

Out[2]:

```
(8160, 38)
```

In [3]:

```
train.head()
```

(…)

5 rows × 38 columns

In [4]:

```
#cogemos la última columna como variable objetivo uno Objet11 como categó
rica
#Vemos el número de columnas de la matriz
numcolumns=len(train.columns)
numcolumns

y_train = train.iloc[0:,numcolumns-1] #Asigna las última columna a la mat
riz Y
# seleccionamos el resto como matriz de variables para la predicción y a
esa matriz le quitaremos aquellas que no queramos
X_train = train.iloc[0:,0:numcolumns-1]  #Asigna las primeras  columnas a
la matriz X
```

```
y_test = test.iloc[0:,numcolumns-1] #Asigna las última columna a la matri
z Y
# seleccionamos el resto como matriz de variables para la predicción y a
esa matriz le quitaremos aquellas que no queramos
X_test = test.iloc[0:,0:numcolumns-1]  #Asigna las primeras  columnas a l
a matriz X

X_train.head()
```

(…)

5 rows × 37 columns

In [5]:

```
y_train
```

Out[5]:

```
72      1
663     0
1893    2
8248    0
6120    1
       ..
3195    1
9740    0
9155    0
9534    1
8776    2
Name: cluster, Length: 8160, dtype: int64
```

In [6]:

```
X_test
```

(…)

3498 rows × 37 columns

In [7]:

```
# fit model on training data
model = XGBClassifier(random_state=42)
eval_set = [(X_test, y_test)]
modelo=model.fit(X_train, y_train, eval_metric=["mlogloss"], eval_set=eva
l_set, verbose=False)

#verbose=True para que salgan todos los pasos de la cálculo de mlogloss p
ara cada arbol
print(modelo)
```

```
# make predictions for test data
predictions = modelo.predict(X_test)
# evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
XGBClassifier(objective='multi:softprob', random_state=42)
Accuracy: 60.95%
```

VAMOS A AJUSTAR LOS HIPERPARÁMETROS DEL MODELO

In [8]:

```
import numpy
# grid search
max_accuracy=0
max_learning_rate=0
max_colsample=0
max_subsample=0
max_max_depth=0
max_min_child=0
learning_rate = [0.1, 0.2, 0.3, 0.4]
min_child_weight=[1,2]
max_depth = [3, 4,5]
subsample=[0.6,0.7,0.8]
colsample_bytree=[0.6,0.7,0.8]


for i in learning_rate:
    model = XGBClassifier(objective='multi:softprob', learning_rate=i, ra
ndom_state=42)
    #objective='multi:softprob'
    eval_set = [(X_test, y_test)]
    model.fit(X_train, y_train, early_stopping_rounds=10,  eval_metric=["
mlogloss"],
    eval_set=eval_set, verbose=0)          #eval_metric=["logloss"]
    # make predictions for test data
    predictions = model.predict(X_test)
    # evaluate predictions
    accuracy = accuracy_score(y_test, predictions)
    if(accuracy>max_accuracy):
        max_accuracy=accuracy
        max_learning_rate=i
        print("learning_rate %.2f" % i)
        print("Max_Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
learning_rate 0.10
Max_Accuracy: 60.95%
```

```
learning_rate 0.30
Max_Accuracy: 61.03%
```

In [9]:

```
for j in min_child_weight:
    for k in max_depth:
        for l in subsample:
            for m in colsample_bytree:
                model = XGBClassifier(objective='multi:softprob',random_s
tate=42, learning_rate=max_learning_rate,
                              min_child_weight=j,
                               max_depth=k,
                                subsample=l,
                               colsample_bytree=m)
                eval_set = [(X_test, y_test)]
                model.fit(X_train, y_train,  early_stopping_rounds=10, ev
al_metric=["mlogloss"],
                  eval_set=eval_set, verbose=0)
                # make predictions for test data
                predictions = model.predict(X_test)
                # evaluate predictions
                accuracy = accuracy_score(y_test, predictions)
                #print(j,",",k,",",l,"accuracy", accuracy)
                if(accuracy>=max_accuracy):
                    max_accuracy=accuracy
                    max_subsample=l
                    max_colsample=m
                    max_min_child=j
                    max_max_depth=k
                    print("learning_rate %.2f" % max_learning_rate,"min_c
hild_weight %.2f" % max_min_child, "max_depth %.2f" % max_max_depth, "sub
sample %.2f" % max_subsample,"colsample_bytree %.2f" % max_colsample)
                    print("Max_Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
learning_rate 0.30 min_child_weight 2.00 max_depth 3.00 subsample 0.70 co
lsample_bytree 0.70
Max_Accuracy: 61.35%
```

In [10]:

```
max_accuracy=0
reg_alpha=[0, 0.01, 0.02, 0.03]
gamma=[0, 0.1,0.2,0.3,0.4,0.5]
#for i, value in enumerate(learning_rate):

for n in reg_alpha:
    for p in gamma:
```

```
        model = XGBClassifier(objective='multi:softprob', random_state=42
,
                              learning_rate=max_learning_rate, reg_alpha=n,
                              gamma=p,
                              min_child_weight=max_min_child,
                               max_depth=max_max_depth,
                                subsample=max_subsample,
                               colsample_bytree=max_colsample)

        eval_set = [(X_test, y_test)]
        model.fit(X_train, y_train,  early_stopping_rounds=10, eval_metri
c=["mlogloss"],
        eval_set=eval_set, verbose=0)
        # make predictions for test data
        predictions = model.predict(X_test)
        # evaluate predictions
        accuracy = accuracy_score(y_test, predictions)
        if(accuracy>max_accuracy):
            max_accuracy=accuracy
            max_reg_alpha=n
            max_gamma=p
            print("reg_alpha %.2f" % n,"gamma %.2f" % p)
            print("Max_Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
reg_alpha 0.00 gamma 0.00
Max_Accuracy: 61.35%
```

max_accuracy=0 scale_pos_weight=[1,2,3,4,5, 6,10, 20] for o in scale_pos_weight: model = XGBClassifier(objectiveo='multi:softprob', random_state=42, learning_rate=max_learning_rate, min_child_weight=max_min_child, max_depth=max_max_depth, subsample=max_subsample, colsample_bytree=max_colsample, reg_alpha=max_reg_alpha, gamma=max_gamma) eval_set = [(X_test, y_test)] model.fit(X_train, y_train, early_stopping_rounds=10, eval_metric=["mlogloss"], eval_set=eval_set, verbose=0) # make predictions for test data predictions = model.predict(X_test) # evaluate predictions accuracy = accuracy_score(y_test, predictions) if(accuracy>max_accuracy): max_accuracy=accuracy max_scale=o print("scale_pos_weight %.2f" % o) print("Max_Accuracy: %.2f%%" % (accuracy * 100.0))

In [11]:

```
#modelo definitivo
model = XGBClassifier(objective='multi:softprob', random_state=42,
                              learning_rate=max_learning_rate)
#min_child_weight=max_min_child,
                               #max_depth=max_max_depth,
                                #subsample=max_subsample,
                               #colsample_bytree=max_colsample,
                               #reg_alpha=max_reg_alpha,
                              #gamma=max_gamma)
```

```
eval_set = [(X_test, y_test)]
model.fit(X_train, y_train, early_stopping_rounds=10,  eval_metric=["mlog
loss"], eval_set=eval_set, verbose=0)
# make predictions for test data
predictions = model.predict(X_test)
# evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print(model)
print("MAccuracy: %.2f%%" % (accuracy * 100.0))
```

```
XGBClassifier(learning_rate=0.3, objective='multi:softprob', random_state
=42)
MAccuracy: 61.03%
```

In [12]:

```
# cargamos las librerias para poder sacar la confusion matrix
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix

plot_confusion_matrix(modelo, X_test, y_test, values_format='d')
```

Out[12]:

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f966
ee27410>
```

(...)

```
from sklearn.metrics import classification_report, confusion_matrix
#imprimimos la confusion matrix y el informe de clasificación.
print("Confusion Matrix:")
print(confusion_matrix(y_test, predictions))

print("Classification Report")
print(classification_report(y_test, predictions))
```

```
Confusion Matrix:
[[ 220  462   25]
 [ 151 1319  244]
 [  23  458  596]]
Classification Report
              precision    recall  f1-score   support

           0       0.56      0.31      0.40       707
           1       0.59      0.77      0.67      1714
           2       0.69      0.55      0.61      1077

    accuracy                           0.61      3498
   macro avg       0.61      0.54      0.56      3498
```

```
weighted avg       0.61      0.61      0.60      3498
```

In [14]:

```
print('max_learning_rate=',max_learning_rate)
print('max_min_child=',max_min_child)
print('max_max_depth=',max_max_depth)
print('max_subsample=',max_subsample)
print('max_colsample=',max_colsample)
print('max_reg_alpha=',max_reg_alpha)
print('max_gamma=',max_gamma)
```

```
max_learning_rate= 0.3
max_min_child= 2
max_max_depth= 3
max_subsample= 0.7
max_colsample= 0.7
max_reg_alpha= 0
max_gamma= 0
```

.

STEP 4

PARTE 1. CARGAMOS LAS LIBRERÍAS

In [1]:

```
#https://amirali-n.github.io/BorutaFeatureSelectionWithShapAnalysis/

# First XGBoost model for Pitec dataset
import matplotlib.pyplot as plt
import xgboost as xgb
from numpy import loadtxt
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
# XGBoost kfold cross validation
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score


# XGBoost stratified kfold cross validation
from sklearn.model_selection import StratifiedKFold
```

```
# one hot ecoding
from numpy import column_stack
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder

# pandas
import pandas as pd
```

In [2]:

```
# Read data from file 'filename.csv'
# (in the same directory that your python process is based)
# Control delimiters, rows, column names with read_csv (see later) sep=se
parador en el csv decimal=separador decimal de los números
train = pd.read_csv("train_BorutaShap_enterprises.csv", sep=",", decimal=
".",index_col=0)

test = pd.read_csv("test_BorutaShap_enterprises.csv", sep=",", decimal=".
",index_col=0)
```

```
train.shape
```

Out[2]:

```
(8160, 38)
```

In [3]:

```
train.head()
```

(…)

5 rows × 38 columns

In [4]:

```
train.columns
```

Out[4]:

```
Index(['q23.6_High speed infrastructure', 'q19.8_No none',
       'q23.8_None of these',
       'q19.6_Social innovations such as new products services or process
es that have the aim of improving society',
       'q26.7_Lack of financial resources', 'q12a_Less than 25',
       'q23.5_Big data analytics e.g. data mining and predictive analysis
',
       'q17.2_Regulatory obstacles or administrative burden',
       'q24.5_Improving working conditions of its employees',
       'q9.6_It has a patent or patent application', 'q10_Yes probably',
       'q9.2_It mainly provides services', 'q16_3_Fairly good',
```

```
       'q23.4_Smart devices e.g. smart sensors smart thermostats etc.',
       'q24.8_Engaging employees in the governance of the enterprise',
       'q24.6_Promoting and improving diversity and equality in the workp
lace',
       'q20.3_Difficulties in predicting the market response',
       'q26.4_It is not compatible with your current business model',
       'q9.8_It has a strategy or action plan to digitalise',
       'q26.8_None of the above',
       'q25_Yes and it has already been implemented',
       'q19.1_A new or significantly improved product or service to the m
arket',
       'q19.5_An innovation with an environmental benefit including innov
ations with an energy or resource efficiency benefit',
       'q25_Yes and it is in the process of being implemented', 'q16_6_DK
NA',
       'q25_No but it may be considered in the future',
       'q25_Not applicable DO NOT READ OUT',
       'q25_No and it will not in the future', 'isocntry_0', 'q1_Before 2
000',
       'q24.7_Evaluating the impact of your enterprise on society',
       'q7a.3_Plan to grow as a result of operating in growing markets',
       'q19.2_A new or significantly improved production process or metho
d',
       'q4t_100000 euros or less', 'q13.7_Predominantly family owned',
       'q11.3_Other European countries outside of the EU incl. Russia',
       'q9.5_It is a part of a global value chain', 'cluster'],
      dtype='object')
```

In [5]:

```
test.columns
```

Out[5]:

```
Index(['q23.6_High speed infrastructure', 'q19.8_No none',
       'q23.8_None of these',
       'q19.6_Social innovations such as new products services or process
es that have the aim of improving society',
       'q26.7_Lack of financial resources', 'q12a_Less than 25',
       'q23.5_Big data analytics e.g. data mining and predictive analysis
',
       'q17.2_Regulatory obstacles or administrative burden',
       'q24.5_Improving working conditions of its employees',
       'q9.6_It has a patent or patent application', 'q10_Yes probably',
       'q9.2_It mainly provides services', 'q16_3_Fairly good',
       'q23.4_Smart devices e.g. smart sensors smart thermostats etc.',
       'q24.8_Engaging employees in the governance of the enterprise',
```

```
       'q24.6_Promoting and improving diversity and equality in the workp
lace',
       'q20.3_Difficulties in predicting the market response',
       'q26.4_It is not compatible with your current business model',
       'q9.8_It has a strategy or action plan to digitalise',
       'q26.8_None of the above',
       'q25_Yes and it has already been implemented',
       'q19.1_A new or significantly improved product or service to the m
arket',
       'q19.5_An innovation with an environmental benefit including innov
ations with an energy or resource efficiency benefit',
       'q25_Yes and it is in the process of being implemented', 'q16_6_DK
NA',
       'q25_No but it may be considered in the future',
       'q25_Not applicable DO NOT READ OUT',
       'q25_No and it will not in the future', 'isocntry_0', 'q1_Before 2
000',
       'q24.7_Evaluating the impact of your enterprise on society',
       'q7a.3_Plan to grow as a result of operating in growing markets',
       'q19.2_A new or significantly improved production process or metho
d',
       'q4t_100000 euros or less', 'q13.7_Predominantly family owned',
       'q11.3_Other European countries outside of the EU incl. Russia',
       'q9.5_It is a part of a global value chain', 'cluster'],
      dtype='object')
```

In [6]:

```
#cogemos la última columna como variable objetivo uno Objet11 como categó
rica
#Vemos el número de columnas de la matriz
numcolumns=len(train.columns)
numcolumns

y_train = train.iloc[0:,numcolumns-1] #Asigna las última columna a la mat
riz Y
# seleccionamos el resto como matriz de variables para la predicción y a
esa matriz le quitaremos aquellas que no queramos
X_train = train.iloc[0:,0:numcolumns-1]  #Asigna las primeras  columnas a
la matriz X

y_test = test.iloc[0:,numcolumns-1] #Asigna las última columna a la matri
z Y
# seleccionamos el resto como matriz de variables para la predicción y a
esa matriz le quitaremos aquellas que no queramos
```

```
X_test = test.iloc[0:,0:numcolumns-1]  #Asigna las primeras  columnas a l
a matriz X
```

```
X_train.head()
```

(…)

5 rows × 37 columns

In [7]:

```
y_train
```

Out[7]:

```
72      1
663     0
1893    2
8248    0
6120    1
       ..
3195    1
9740    0
9155    0
9534    1
8776    2
Name: cluster, Length: 8160, dtype: int64
```

In [8]:

```
X_test
```

(…)

3498 rows × 37 columns

In [9]:

```
max_learning_rate= 0.3
max_min_child= 2
max_max_depth= 3
max_subsample= 0.7
max_colsample= 0.7
max_reg_alpha= 0
max_gamma= 0


model = XGBClassifier(objective='multi:softprob', random_state=42,
                              learning_rate=max_learning_rate)
#                              min_child_weight=max_min_child,
 #                               max_depth=max_max_depth,
  #                                subsample=max_subsample,
```

```
   #                           colsample_bytree=max_colsample,
    #                          reg_alpha=max_reg_alpha,
     #                        gamma=max_gamma)
eval_set = [(X_test, y_test)]
model.fit(X_train, y_train, early_stopping_rounds=10,  eval_metric=["mlog
loss"], eval_set=eval_set, verbose=0)
# make predictions for test data
predictions = model.predict(X_test)
# evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print(model)
print("MAccuracy: %.2f%%" % (accuracy * 100.0))
```

```
XGBClassifier(learning_rate=0.3, objective='multi:softprob', random_state
=42)
MAccuracy: 61.03%
```

In [10]:

```
#FEATURE IMPORTANCE SHAP del modelo de variables reducido
import shap
import matplotlib.pyplot as plt
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values, X_test, plot_type="bar",show=False)
plt.savefig('figure1a.png',format = "png",dpi = 300,bbox_inches = 'tight'
)
#para cada categoría
```

(…)

```
#https://medium.com/analytics-vidhya/shap-part-3-tree-shap-3af9bcd7cd9
b
shap.summary_plot(shap_values, X_test, class_inds=[0],plot_type="bar")

(…)
        shap.summary_plot(shap_values, X_test, class_inds=[1],plot_typ
e="bar")
(…)
shap.summary_plot(shap_values, X_test, class_inds=[2],plot_type="bar")
(…)

# cargamos las librerias para poder sacar la confusion matrix
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
plot_confusion_matrix(model, X_test, y_test, values_format='d')
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f
77da320250>
(…)
```

```
from sklearn.metrics import classification_report, confusion_matrix
print("Confusion Matrix:")
print(confusion_matrix(y_test, predictions))

print("Classification Report")
print(classification_report(y_test, predictions))
```

```
Confusion Matrix:
[[ 220  462   25]
 [ 151 1319  244]
 [  23  458  596]]
Classification Report
              precision    recall  f1-score   support

           0       0.56      0.31      0.40       707
           1       0.59      0.77      0.67      1714
           2       0.69      0.55      0.61      1077

    accuracy                           0.61      3498
   macro avg       0.61      0.54      0.56      3498
weighted avg       0.61      0.61      0.60      3498
```

In [16]:

```
# para las distintas clases
shap.summary_plot(shap_values[0], X_test)
shap.summary_plot(shap_values[1], X_test)
shap.summary_plot(shap_values[2], X_test)
```

(…)

```
# Ahora podemos coger el caso 48 y ver cómo influye el shap value para cl
asificar en la clase 0
shap.force_plot(explainer.expected_value[0], shap_values[0][48,:], X_test
.iloc[48,:])
```

Out[17]:

Visualization omitted, Javascript library not loaded!

Have you run `initjs()` in this notebook? If this notebook was from another user you must also trust this notebook (File -> Trust notebook). If you are viewing this notebook on github the Javascript has been stripped for security. If you are using JupyterLab this error is because a JupyterLab extension has not yet been written.

In [18]:

```
from shap import Explanation
shap.waterfall_plot(Explanation(shap_values[0][48],explainer.expected_val
ue[0], feature_names=X_test.columns.tolist())
```

```
import numpy as np
```

```
vals= np.abs(shap_values).mean(0)#which amounts to compute the average of
the absolute value of the shap values).
feature_names = X_test.columns.tolist()
```

In [20]:

```
vals
```

Out[20]:

```
array([[0.03061534, 0.02755125, 0.02298734, ..., 0.02896348, 0.06741755,
        0.00312483],
       [0.02135167, 0.05620652, 0.02587927, ..., 0.01527776, 0.00909621,
        0.00985598],
       [0.04697992, 0.05943299, 0.02526912, ..., 0.01265181, 0.00660696,
        0.00803756],
       ...,
       [0.02534986, 0.01892113, 0.01783303, ..., 0.06345036, 0.05463015,
        0.00504887],
       [0.0244181 , 0.05700393, 0.03704061, ..., 0.01217837, 0.0063838 ,
        0.08715863],
       [0.01784497, 0.02772326, 0.02328998, ..., 0.01723056, 0.00937423,
        0.00814073]], dtype=float32)
```

In [21]:

```
feature_names
```

Out[21]:

```
['q23.6_High speed infrastructure',
 'q19.8_No none',
 'q23.8_None of these',
 'q19.6_Social innovations such as new products services or processes tha
t have the aim of improving society',
 'q26.7_Lack of financial resources',
 'q12a_Less than 25',
 'q23.5_Big data analytics e.g. data mining and predictive analysis',
 'q17.2_Regulatory obstacles or administrative burden',
 'q24.5_Improving working conditions of its employees',
 'q9.6_It has a patent or patent application',
 'q10_Yes probably',
 'q9.2_It mainly provides services',
 'q16_3_Fairly good',
 'q23.4_Smart devices e.g. smart sensors smart thermostats etc.',
 'q24.8_Engaging employees in the governance of the enterprise',
 'q24.6_Promoting and improving diversity and equality in the workplace',
 'q20.3_Difficulties in predicting the market response',
 'q26.4_It is not compatible with your current business model',
 'q9.8_It has a strategy or action plan to digitalise',
 'q26.8_None of the above',
```

```
 'q25_Yes and it has already been implemented',
 'q19.1_A new or significantly improved product or service to the market'
,
 'q19.5_An innovation with an environmental benefit including innovations
with an energy or resource efficiency benefit',
 'q25_Yes and it is in the process of being implemented',
 'q16_6_DKNA',
 'q25_No but it may be considered in the future',
 'q25_Not applicable DO NOT READ OUT',
 'q25_No and it will not in the future',
 'isocntry_0',
 'q1_Before 2000',
 'q24.7_Evaluating the impact of your enterprise on society',
 'q7a.3_Plan to grow as a result of operating in growing markets',
 'q19.2_A new or significantly improved production process or method',
 'q4t_100000 euros or less',
 'q13.7_Predominantly family owned',
 'q11.3_Other European countries outside of the EU incl. Russia',
 'q9.5_It is a part of a global value chain']
```

In [22]:

```
#podemos poner los valores de importancia de cada clase en una matriz. (n
o tener en cuenta aquí la leyenda de las columnas)
#creamos una matriz de con los valores shap medios de cada uno de las obj
etivos.
feature_importance = pd.DataFrame(list(zip(feature_names, vals[0],vals[1]
, vals[2])))
```

In [23]:

```
# valores shap numéticos
feature_importance
```

Out[23]:

	0	1	2	3
0	q23.6_High speed infrastructure	0.030615	0.021352	0.046980
1	q19.8_No none	0.027551	0.056207	0.059433
2	q23.8_None of these	0.022987	0.025879	0.025269
3	q19.6_Social innovations such as new products ...	0.017170	0.010163	0.010706

	0	1	2	3
4	q26.7_Lack of financial resources	0.018106	0.029658	0.058548
5	q12a_Less than 25	0.082631	0.012429	0.006379
6	q23.5_Big data analytics e.g. data mining and ...	0.006873	0.006742	0.010741
7	q17.2_Regulatory obstacles or administrative b...	0.028922	0.044449	0.024735
8	q24.5_Improving working conditions of its empl...	0.067626	0.052189	0.043164
9	q9.6_It has a patent or patent application	0.020790	0.009079	0.011063
10	q10_Yes probably	0.012182	0.014724	0.015720
11	q9.2_It mainly provides services	0.042688	0.022996	0.028753
12	q16_3_Fairly good	0.018412	0.013492	0.009427
13	q23.4_Smart devices e.g. smart sensors smart t...	0.034185	0.028350	0.024125
14	q24.8_Engaging employees in the governance of ...	0.054297	0.066373	0.054107
15	q24.6_Promoting and improving diversity and eq...	0.130416	0.205381	0.171778
16	q20.3_Difficulties in predicting the market re...	0.040062	0.051075	0.025546
17	q26.4_It is not compatible with your current b...	0.019394	0.012059	0.061539
18	q9.8_It has a strategy or action plan to digit...	0.058616	0.012318	0.007667
19	q26.8_None of the above	0.025639	0.010994	0.012746
20	q25_Yes and it has already been implemented	0.016809	0.020040	0.013879

		0	1	2	3
21	q19.1_A new or significantly improved product ...	0.151372	0.031712	0.089161
22	q19.5_An innovation with an environmental bene...	0.353495	0.073148	0.097776
23	q25_Yes and it is in the process of being impl...	0.028253	0.013803	0.015218
24	q16_6_DKNA	0.008960	0.014461	0.014897
25	q25_No but it may be considered in the future	0.068402	0.031813	0.033581
26	q25_Not applicable DO NOT READ OUT	0.014059	0.012023	0.007552
27	q25_No and it will not in the future	0.038858	0.038525	0.187485
28	isocntry_0	0.169807	0.128328	0.123786
29	q1_Before 2000	0.028151	0.047555	0.056143
30	q24.7_Evaluating the impact of your enterprise...	0.111152	0.118691	0.113687
31	q7a.3_Plan to grow as a result of operating in...	0.021687	0.024154	0.022662
32	q19.2_A new or significantly improved producti...	0.072058	0.033677	0.028654
33	q4t_100000 euros or less	0.013366	0.019705	0.020190
34	q13.7_Predominantly family owned	0.028963	0.015278	0.012652
35	q11.3_Other European countries outside of the ...	0.067418	0.009096	0.006607
36	q9.5_It is a part of a global value chain	0.003125	0.009856	0.008038