UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Combinación de modelos heterogéneos para la identificación de variantes genéticas implicadas en la diabetes mellitus tipo 2 (DT2)

Trabajo Fin de Grado

Grado en Ingeniería Informática

AUTOR/A: Molto Molto, Jorge Ramon

Tutor/a: Navarro Cerdán, José Ramón

Cotutor/a externo: ARNAL BENEDICTO, LAURA

CURSO ACADÉMICO: 2022/2023

A mis tutores e ITI por brindarme la oportunidad,
recursos, orientación y conocimiento.

A mi familia, mis amigos y mi pareja,
su inquebrantable apoyo y comprensión durante este proceso.
Su aliento y amor fueron mi mayor fortaleza.

Resum

La genòmica és la disciplina que estudia el genoma humà. Un dels principals reptes és la detecció de variables genètiques rellevants que determinin un genotip o una malaltia. En aquest àmbit d'estudi ens trobem amb una peculiaritat en les dades: la quantitat de mostres és limitada a causa dels costos i la variabilitat de cada observació és molt elevada a causa de la naturalesa del genoma humà.

Per abordar aquesta problemàtica i trobar les variables genòmiques rellevants per a una malaltia, utilitzarem un algorisme ensemble. Aquí, la selecció de variables es regeix pel consens de múltiples execucions de models d'aprenentatge aplicats a diverses particions de les dades. Sobre cadascuna de les particions, un inductor realitzarà una selecció de característiques basant-se en una variable objectiu, que en el cas que ens ocupa seria detectar si es tracta d'un cas o un control. Cadascuna d'aquestes seleccions es consensua mitjançant un esquema de vots, on s'espera que la rellevància d'una variable es vegi reflectida en el nombre de vots obtinguts mitjançant la suma de les vegades que una variable ha estat seleccionada.

Aquest algorisme està basat en l'ús de models heterogenis, per tant, el rendiment dependrà de la capacitat de l'algorisme d'aprenentatge automàtic per adaptar-se a la distribució condicional real dels gens implicats respecte a la presència o absència de la malaltia estudiada. Com que la distribució és a priori desconeguda en aquest tipus de problemes i amb l'objectiu que l'aproximació proposada sigui agnòstica a la distribució, es proposa la implementació de diferents combinacions de models en l'execució de l'ensemble. El seu potencial radica en què les votacions dependran de diferents models amb les seves respectives maneres d'interpretar l'espai. S'espera que els inductors basats en la combinació de models ofereixin millors seleccions de variables rellevants per a la malaltia, independentment de quina sigui la distribució condicional dels gens rellevants respecte a la presència o absència de la malaltia.

Paraules clau: Ciéncia de dades, variables genètiques, ensemble, selecció de característiques, dades de microarrays, espais d'alta dimensionalitat.

Resumen

La genómica es la disciplina que estudia el genoma humano. Uno de los principales desafíos es: la detección de variables genéticas relevantes que determinen un genotipo o enfermedad. En este ámbito de estudio nos encontramos con una peculiaridad en los datos: la cantidad de muestras es limitada debido a los costos y la variabilidad de cada observación es muy elevada debido a la naturaleza del genoma humano.

Para abordar esta problemática, encontrar las variables genómicas relevantes para una enfermedad, utilizaremos un algoritmo ensemble. Aquí, la selección de variables se rige por el consenso de múltiples ejecuciones de modelos de aprendizaje aplicados a diversas particiones de los datos. Sobre cada una de las particiones un inductor realizará una selección de características basándose en una variable objetivo, que en el caso que nos ocupa sería detectar si se trata de un caso o un control. Cada una de estas selecciones se consensúa en base a un esquema

de votos donde se espera que la relevancia de una variable se vea reflejada en el número de votos obtenidos a través de la suma de las veces que una variable ha sido seleccionada.

Este algoritmo esta basado en el uso de modelos heterogéneos, por tanto el rendimiento dependerá de la capacidad del algoritmo de machine learning para adaptarse a la distribución condicional real de los genes implicados respecto a la presencia o no de la enfermedad estudiada. Puesto que la distribución es a priori desconocido en este tipo de problemas y con el objetivo de que la aproximación propuesta sea agnóstica a la distribución, se propone la implementación de distintas combinaciones de modelos en la ejecución del ensemble. Su potencial radica en que la votaciones dependerán de distintos modelos con sus respectivas maneras de interpretar el espacio. Cabe esperar que los inductores basados en combinación de modelos ofrezcan mejores selecciones de variables relevantes para la enfermedad con independencia de cual sea la distribución condicional de los genes relevantes respecto a la presencia o ausencia de la enfermedad.

Palabras clave: Salud, ciencia de datos, variables genéticas, ensemble, selección de características, datos de microarrays, espacios de alta dimensionalidad.

Abstract

Genomics is the discipline that studies the human genome. One of the main challenges is the detection of relevant genetic variables that determine a genotype or disease. In this field of study, we encounter a peculiarity in the data: the quantity of samples is limited due to costs, and the variability of each observation is very high due to the nature of the human genome.

To address this issue and find the relevant genomic variables for a disease, we will use an ensemble algorithm. Here, the selection of variables is guided by the consensus of multiple runs of learning models applied to various data partitions. On each of these partitions, an inducer will perform a feature selection based on a target variable, which in the case at hand would be to detect whether it is a case or a control. Each of these selections is consensually decided through a voting scheme, where it is expected that the relevance of a variable is reflected in the number of votes obtained by summing the times a variable has been selected.

This algorithm is based on the use of heterogeneous models; therefore, performance will depend on the machine learning algorithm's ability to adapt to the actual conditional distribution of the genes involved regarding the presence or absence of the studied disease. Since the distribution is a priori unknown in this type of problem, and with the aim of making the proposed approach distribution-agnostic, we suggest implementing different combinations of models in the ensemble execution. Its potential lies in the fact that the votes will depend on different models with their respective ways of interpreting the space. It is expected that inductors based on model combinations will offer better selections of relevant variables for the disease, regardless of the conditional distribution of the relevant genes concerning the presence or absence of the disease.

Key words: Health, data science, genetic variables, ensemble, Feature selection, Microarray data, High dimensionality spaces

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

A nivel mundial, se dedican grandes esfuerzos en el tratamiento de los datos para poder mejorarlos y extraer conclusiones valiosas de ellos. Entre algunas disciplinas desarrolladas con ese propósito, podemos encontrar el Machine Learning, la Inteligencia Artificial y la Ciencia de Datos, para lograr este propósito.

Estas herramientas se utilizan para mejorar la eficiencia de múltiples ámbitos, algunos de los más significativos serían la biología y la medicina. El enfoque de este trabajo se centrará en la genómica, que según el National Human Genome Research Institute: "La genómica es un campo de la biología que se centra en el estudio de todo el ADN de un organismo, es decir, su genoma. Esa tarea incluye identificar y caracterizar todos los genes y elementos funcionales del genoma de un organismo, así como la forma en que interactúan. "[3]

La identificación de genes en genómica consiste en detectar que variables genéticas son las que determinan o propician ciertas enfermedades. En Data Mining se clasificaría como un problema de selección de características.

El objetivo primordial del presente trabajo es la identificación de variables genéticas que desempeñan un papel en el desarrollo de enfermedades, concretamente de la diabetes tipo 2.

El genoma presenta un desafío debido a su alta dimensionalidad, tienen una gran número de variables que se pueden medir como expresiones genéticas, mutaciones o cadenas de proteínas. El desafío aumenta con el hecho de que en biología es complicado obtener grandes conjuntos de muestras. Por tanto, nos enfrentamos a condiciones muy particulares: alta dimensionalidad con un número limitado de muestras. En estas circunstancias, donde el número de variables es muy superior al de muestras es común que los modelos de aprendizaje encuentren relaciones falsas entre variables [2]. Para abordar estas problemáticas, se proponen enfoques de computación basada en combinatoria y computación paralela.

El ITI (Instituto Tecnológico de Informática) desarrolló en 2022 un algoritmo Ensemble llamado SEQENS [1] para poder paliar los problemas de alta dimensionalidad y se utilizará como base para este trabajo.

1.1 Motivación

Hoy en día, nos enfrentamos a la realidad de que la obtención de muestras es un proceso difícil y costoso, normalmente los dataset con los que se trabajan son reducidos. Sin embargo, esta situación podría cambiar con el transcursos de los años.

Existen varios factores que podrían facilitar la obtención de muestras de calidad, como la disminución de los costes asociados o el crecimiento en la confianza y la consolidación de técnicas de procesamiento de datos. Sin embargo, hay dos circunstancias que no cambiarán: en primer lugar, la naturaleza del genoma humano conlleva una cantidad significativamente alta de variables, que puede oscilar entre miles o decenas de miles; en segundo lugar, en el caso de enfermedades raras, la disponibilidad de muestras siempre será limitada

En definitiva, aunque con el transcurso de los años pueda haber un aumento en la cantidad de muestras de calidad disponibles, la variabilidad inherente a estas muestras seguirá siendo significativamente mayor que la cantidad de muestras disponibles.

La motivación principal de este trabajo es mejorar la selección de características cuando la muestras presenta una dimensionalidad muy por encima de la cantidad de observaciones. En el caso de este trabajo, mejorar la selección de variables genéticas que estén asociadas a una enfermedad, concretamente trabajando sobre la diabetes mellitus tipo 2.

"La diabetes es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre (o azúcar en sangre), que con el tiempo conduce a daños graves en el corazón, los vasos sanguíneos, los ojos, los riñones y los nervios"[4]. En 2022 se estimaba que la diabetes afectaba al 9,3 % de la población mundial y según los registros históricos, este porcentaje aumenta anualmente [5]. Según NIDDK, los factores de riesgo para la diabetes tipo 2 son tanto ambientas, genéticos y conductuales [6]. Entre ellos encontramos: sobrepeso, ser mayor de 35 años, tener antecedentes familiares, no mantenerse activo físicamente, tener prediabetes, pertenecer a ciertas razas y tener antecedentes de diabetes gestacional.

La mayoría de avances a nivel de computación y mejora de los datos suelen suponer una mejora de la sociedad y un aumento de la eficiencia en los ámbitos de aplicación. Con este trabajo se pretende aportar un grano de arena al progreso de la Ciencia de Datos en la biología.

Mejorar la detección de variables genéticas que influyen en enfermedades tiene un gran potencial. Estos avances pueden tener repercusiones en áreas como la prevención de enfermedades, un tratamiento preventivo más efectivo y la agilización de procesos al poder determinar con mayor precisión qué partes del genoma de un individuo son relevantes para diagnosticar una enfermedad.

El algoritmo en el que se basa este trabajo, SEQENS [1], es sensible a la distribución de los datos y a la correlación entre variables y casos. Además dicha distribución es siempre desconocida, por lo tanto, es crucial investigar cómo lograr mayor estabilidad independientemente de forma agnóstica a la distribución de los datos.

Aunque esta problemática inicialmente se aborda en el contexto del estudio del genoma, hay múltiples disciplinas que también se enfrentas a *la maldición de la dimensionalidad*. Estas disciplinas, como el procesamiento de imágenes, el análisis de texto, la química, a nivel industrial, la astronomía y otros problemas de biología, podrían beneficiarse de los enfoques y avances desarrollados en este trabajo.

1.2 Objetivos

El propósito central de este trabajo es mejorar la selección de características, en situaciones de alta dimensionalidad, maximizando su número y posicionándolas según su relevancia, para ello se va a hacer uso del algoritmo SEQENS [1]. La imparcialidad de este algoritmo se ve afectada por la distribución seguida en el espacio de características en relación a la variable dependiente. En esencia, el trabajo propuesto aquí busca reducir la incertidumbre que surge debido a la distribución real de las enfermedades en el espacio.

El funcionamiento de este algoritmo se basa en las votaciones de múltiples ejecuciones en particiones diferentes de los datos en estudio. El descubrimiento de las variables relevantes se basa en el uso de diferentes modelos inductores que relacionan la variable objetivo con las variables explicativas en un esquema forward-backward. Estos modelos inductores aplicados muestran distintos rendimientos dependiendo de la distribución de los datos en el problema. En general, la distribución real de los daatos es desconocida, lo que impide seleccionar de antemano el modelo que logrará los resultados más óptimos.

Es por ello que la elección del inductor aplicado influirá significativamente sobre la calidad de la lista final de variables influyentes. Es este contexto, la calidad implica identificar las variables influyentes y su posición correcta en cuanto a su relevancia a la hora de explicar una determinada enfermedad.

Con el objetivo de reducir la dependencia del rendimiento de la distribución de los datos, se plantea la idea de crear inductores que combinen múltiples modelos en lugar de un único modelo, intentando así crear un algoritmo de selección de característicos cuyo rendimiento sea lo más agnóstico posible a la distribución rea.

Se espera que al considerar diferentes enfoques de interpretación del espacio en las votaciones de cada subconjunto de datos, a través de diferentes inductores que ven el espacio de forma diferente, los resultados sean más robustos y coherentes, y que la distribución condicional de los datos respecto a la enfermedad estudiada impacte menos sobre el rendimiento máximo posible para el problema estudiado en cada caso.

Buscando lograr el objetivo final, se plantean varios objetivos secundarios a abordar:

- Creación de observaciones.
- Creación de enfermedades sintéticas.
- Creación de inductores basados en múltiples modelos.

- Evaluación de distintas ejecuciones de SEQENS utilizando como inductores los modelos individuales y los creados con múltiples modelos como propuestas para mejora del rendimiento. Sobre conjuntos de datos ficticios.

- Experimentación en conjunto de datos reales.

1.3 Estructura memoria

En esta sección, se proporciona una descripción detallada de cómo está estructurada la memoria, con el propósito de facilitar la comprensión y el seguimiento por parte del lector.

Capítulo 2: Estado del Arte

Se presentan los avances más relevantes que hay en el área de estudio de la detección de variables. Se abordan los trabajos previos que han influido en el campo, así como las técnicas ya existentes para la selección de características en situaciones de alta dimensionalidad

Capítulo 3: Metodología

Este capítulo detalla que métodos y enfoques que se han empleado para llevar a cabo la investigación. Se proporciona una explicación detallada sobre el funcionamiento del algoritmo de selección de características SEQENS, que constituye una parte fundamental de este trabajo. Además, se introduce la métrica utilizada para evaluar la calidad de las listas de variables seleccionadas, Cliff. Se detallan los modelos utilizados para llevar a cabo la investigación, se explican los procesos de generación de datos sintéticos y se describe cómo se combinan diferentes modelos para formar las propuestas de solución.

Capítulo 4: Resultados

En el cuarto capítulo, se presentan y analizan los resultados obtenidos por los diferentes experimentos realizados. Se abordan dos experimentos fundamentales. El primero, evalúa del rendimiento de los modelos propuestos en comparación con los modelos básicos, utilizando un conjunto de datos sintéticos con una única distribución. El segundo, sigue una estructura similar, pero se ejecuta sobre un conjunto de datos reales y distintas distribuciones del objetivo respecto de las características estudiadas.

Capítulo 5: Discusión

Se dedica a un análisis profundo de los resultados presentados en el capítulo anterior. Los hallazgos obtenidos en los experimentos se discuten en relación con el estado del arte y las expectativas planteadas al inicio del estudio.

Capítulo 6 y 7: Conclusiones y Trabajos Futuros

Los últimos dos capítulos, “Conclusiones ”y “Trabajos Futuros ”, ofrecen cierres y proyecciones respectivamente. En “Conclusiones ”, se presentan las conclusiones finales del estudio, resumiendo los logros alcanzados, los desafíos enfrentados y las contribuciones realizadas. Por otro lado, “Trabajos Futuros ”explora las posibles extensiones y continuaciones de la investigación, señalando áreas de

interés para futuros estudios y como se podría ampliar esta misma investigación para mejorar la propuesta.

Esta estructura pretende ayudar al lector a adentrarse desde la comprensión del problema hasta la solución propuesta y los resultados obtenidos con ella.

CAPÍTULO 2
Estado del arte

En el contexto de análisis de datos de alta dimensionalidad, como los microarrays, se han desarrollado una seria de métodos y enfoques para abordar la selección de características relevantes. En esta sección se revisa y sintetiza información de diversas fuentes para establecer el estado del arte en este campo, con un especial enfoque en el algoritmo SEQENS y su optimización.

Un microarray es una tecnología que permite analizar la expresión génica en células. En su núcleo, el ADN contiene genes que codifican proteínas, y estas proteínas realizan funciones esenciales en el organismo. Los mircroarrays permiten medir muchos genes al mismo tiempo. Se utiliza para obtener información valiosa sobre expresiones genéticas y desarrollo de enfermedades. En la Fig.2.1 se puedo observar el proceso de obtención de un DNA microarray Dataset.[7]

Introducción al problema

Los microarrays presentan una gran complejidad debido a estar compuestos por datos con muchas variables (genes), pero un número limitado de muestras. La *maldición de la dimensionalidad* [2] se convierte en un desafío crítico al analizar estos datos, ya que aumenta la probabilidad de sobreajuste y dificulta la significación estadística.

Enfoques y soluciones

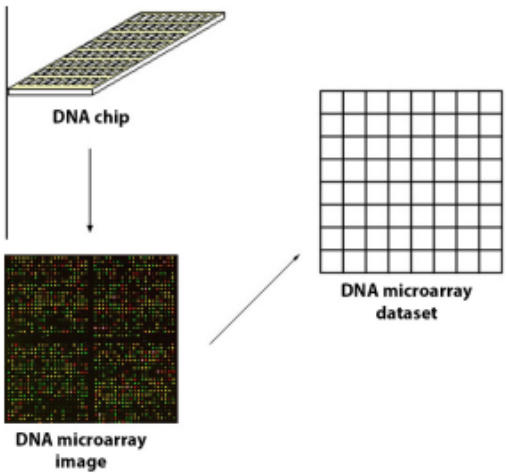

Figura 2.1: Proceso de obtención de un set de datos de microarrays. (Fuente:[7])

Este problema se enfoca de dos maneras: con métodos de selección de característica y con métodos de extracción.

Los métodos de selección de un subconjunto de características eliminan las características irrelevantes o redundantes. El objetivo principal es identificar y conservar un subconjunto de características del conjunto de datos original. Estos métodos de selección se dividen en filtros, wrappers y técnicas integradas. Por otro lado, los métodos de extracción de características buscan transformar las características originales en un nuevo conjunto de características. Estas transformaciones pueden ser lineales o no lineales y se aplican con el propósito de crear nuevas representaciones de los datos que conserven la información más relevante.

Según [8], tras analizar las principales implementación de estos enfoques obtiene las tabla 2.1 donde expresa las ventajas y desventajas de cada enfoque.

En este apartado, nos centraremos en los métodos de selección de características, puesto que SEQENS pertenece a este enfoque, siendo el algoritmo que se propone mejorar.

Tabla 2.1: Ventajas y Desventajas de la Selección y Extracción de Características en el Análisis de Datos de Microarrays de Cáncer. (Fuente: [8])

Método	**Ventajas**	**Desventajas**
Selección	Preservación de las características de los datos para interpretación. Poder discriminativo. Tiempos de entrenamiento más cortos. Reducción del sobreajuste	Limitado por las características existentes. Puede omitir información relevante
Extracción	Mayor poder discriminatorio. Control del sobreajuste cuando es no supervisado	Pérdida de la interpretabilidad de los datos. La transformación puede ser computacionalmente costosa

Los primeros enfoques de selección de características se basaban en métodos univariados como t-tests, ANOVA y pruebas estadísticas similares. Estos métodos son eficientes pero no consideran las interacciones entre características. A medida que los datos aumentaron la complejidad, surgieron problemas con la dependencia entre características y la necesidad de abordar datos de múltiples clases. Para ello, se exploraron técnicas de Análisis de Varianza (ANOVA) y métodos más complejos ([9], [7]).

En el contexto de los microarrays, la selección de características se ha dividido en enfoques de filtrado y envoltura. Los enfoques de filtrado, como la prueba t y ANOVA, son rápidos pero no consideran el clasificador. Por otro lado, los enfoques de envoltura, como el Sequential Forward Selection (SFS) y el método SEQENS, consideran el clasificador pero son computacionalmente costosos ([8], [7], [1]).

El método SEQENS es un algoritmo de ensemble que utiliza múltiples instancias de SFS para mejorar la selección de características. Al combinar los resultados de diferentes instancias de SFS, SEQENS busca mejorar la estabilidad y el rendi-

miento general de la selección de características. Al emplear múltiples instancias de SFS generadas a partir de diferentes divisiones de datos o inductores, SEQENS introduce diversidad en las selecciones de características, mejorando así la robustez del método [1].

Comparación de Métodos de Selección de Características

A medida que aumenta el número de métodos de selección de características de calidad, la tarea de seleccionar el método más adecuado se vuelve aún más compleja. Las únicas guías disponibles son las experiencias anteriores y los estudios comparativos de la literatura [10]. Se puede observar que los resultados obtenidos en varios estudios como [7] son altamente dependientes del clasificador, el método de selección de características y, en particular, del conjunto de datos. Los autores recomiendan el estudio cuidadoso de las particularidades de cada problema.

En el estudio [10] se comparan diez algoritmos de selección de características populares, incluyendo enfoques univariados (*t-test*, distancia Bhattacharyya, ANOVA y *entropy*), *tree ensembles least absolute shrinkage and selection operator* (LASSO), *maximum relevance and minimum redundancy* (MID y MIQ), *iterative Relief algorithm* (iRelief) y *linear support vector machine* (SVM). La conclusión es que las técnicas basadas en entropía lograron los valores de estabilidad más altos para la mayoría de las bases de datos investigadas, seguidas de otras técnicas univariadas (ANOVA FS, *Bhattacharyya distance-based feature selection* y *t-test* FS).

El estudio [7] evaluó los métodos de selección de características CFS, FCBF, INTERACT, *Information Gain*, ReliefF, mRMR y SVM-RFE. Aunque no se analizaron en detalle los resultados, parece que los mejores resultados (en general) se obtuvieron con el clasificador SVM, combinado con un filtro de subconjunto que busca identificar un subconjunto óptimo de características relevantes del conjunto completo, y el método de validación DOB-SCV.

En relación al algoritmo que trata este trabajo, el propio artículo en el que se presenta [1] compara su rendimiento con otros nueve métodos de selección de características: ANOVA, Pearson, Welch, mRMR, MultiSURF, Lasso, SVM-RFE, *Random Forest* y *Gradient Boost*.

Los algoritmos se evalúan en tres escenarios distintos que representan enfermedades distribuidas. En la Figura 2.2, se puede observar la comparación entre SEQENS y los otros métodos de selección de características.

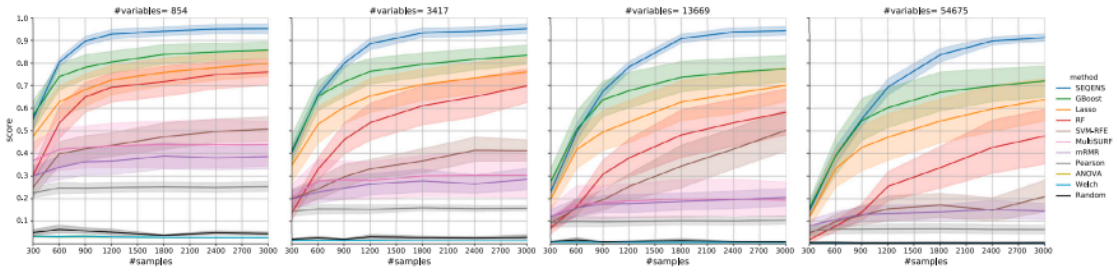

Figura 2.2: SEQENS vs métodos de selección de características. (Fuente: [1])

Ninguno de los métodos presentados supera a los demás en todos los escenarios. Sin embargo, en promedio, SEQENS ha obtenido el mejor rendimiento al

identificar variables relevantes en las tres enfermedades ficticias propuestas, especialmente cuando el número de muestras es bajo. En otras palabras, SEQENS generaliza mejor en diferentes escenarios que los otros métodos con los que se compara. También muestra una mayor estabilidad. Dado que en la realidad es difícil conocer el tipo de distribución de las variables relevantes, este hecho proporciona un sólido argumento a favor del uso de SEQENS en la identificación de genes.

Optimización del Algoritmo SEQENS

Aunque en promedio, a lo largo de tres enfermedades, SEQENS logra los mejores resultados [1], se observa que en diversos escenarios y configuraciones, Lasso, GBoost y Random Forest están mejor posicionados. El mismo documento sugiere pautas para mejorar el rendimiento de SEQENS, entre las cuales se encuentra la posibilidad de combinar diferentes inductores. El algoritmo SEQENS utilizado en el artículo se basa en un único inductor, k-nearest neighbours. Es esta línea de investigación la que este trabajo propone explorar y ampliar.

Conclusión

La selección de características en datos de alta dimensionalidad, como los microarrays, sigue siendo un desafío en la investigación biomédica y otros campos. La optimización de algoritmos como SEQENS es una vía de investigación prometedora para mejorar la estabilidad, el rendimiento y la confiabilidad en la identificación de características. El uso de enfoques de ensemble, como el implementado en SEQENS, muestra una mayor robustez en comparación con métodos individuales y puede ser útil para abordar la *maldición de la dimensionalidad* en el análisis de datos complejos.

CAPÍTULO 3
Metodología

En este capítulo, se describe la metodología utilizada en el presente estudio para llevar a cabo la investigación y obtener resultados significativos. Se detallan las diferentes técnicas y herramientas empleadas para cumplir con los objetivos planteados.

3.1 Modelos inductores utilizados

Para el proceso de machine learning se han elegido 3 modelos. De entre todas las posibilidad, han sido seleccionado por eficiencia y tiempo de cómputo. A continuación se perfilan los modelos seleccionados.

K-Neighbors Regressor: Esta clasificado como un modelo de aprendizaje supervisado no paramétrico. El algoritmo de este modelo evalúa cada nueva observación según las votaciones de un determinado número de observaciones que se encuentras más próximas a ellas. El número de observaciones que influirán en la decisión será el valor K [11] . Para los experimentos realizados, se ha utilizado el modelo con los valores por defectos definidos por scikit-learn.[12]

Es la figura 3.1 se puede observar el funcionamiento de KNN en un escenario de tres clases y cinco vecinos. Cuando una nueva observación es analizada, se consulta la etiqueta de los cinco vecinos más cercanos. En el caso de un problema de regresión, se asigna un valor continuo determinado por las clases cercanas.

Support Vector Regression: El modelo Support Vector Regression (SVR) es un tipo de algoritmo supervisado que busca optimizar un hiperplano con un margen de error (épsilon) que delimite la distribución de una clase, de manera que

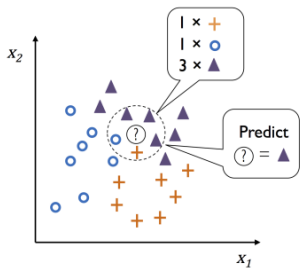

Figura 3.1: Ejemplo visual de KNN. Problema de 3 clases y K = 5. (Fuente:[11])

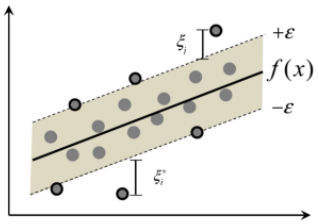

Figura 3.2: Ejemplo visual SVR (Fuente: [14])

se aproxime a la relación entre las variables de entrada y una variable objetivo continua. Al mismo tiempo, se minimiza el error de predicción [13].

Como se observa en la figura 3.2, SVR tratar de optimizar la función f(x) y sus márgenes de error para ajustarse a la distribución de las observaciones pertenecientes a una clase.

Los kernels son funciones matemáticas que determinan cómo se transforman los datos en el espacio de características, y la elección de un kernel puede influir en cómo se adapta el hiperplano a los datos. [15] Para la ejecución de los experimentos se ha decido utilizar los instancias de SVR con dos modelos distintos: Polinomial y Radial Basis Function (Función de Base Radial). Al medir el rendimiento de dos kernels distintos, estamos replicando la incertidumbre de los problemas reales, donde no conocemos la distribución exacta de los datos. Esta estrategia nos permite evaluar cómo se comporta el modelo SVR en diferentes escenarios y cómo afecta la elección del kernel a la precisión y fiabilidad de las predicciones.

Decision Tree Regressor: Para explicar este modelo, utilizaremos el ejemplo de la figura 3.3. Este algoritmo crea un árbol donde, en función del valor de una variable, se dirige el flujo hacia niveles inferiores o hacia hojas. Cuando llega a una hoja, se le asigna una clase. En la imagen, se puede observar cómo, dependiendo del valor de variables como la edad o el número de imágenes, se dirige el flujo hasta llegar a hojas que clasifican en diferentes categorías como “unknown ”, “tech ”, o “bus ”.

Para la ejecución de estos modelos se ha seleccionado el criterio de error *friedman_mse* y una profundidad máxima de tres.

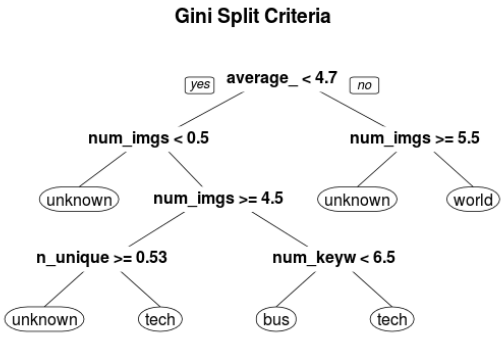

Figura 3.3: Ejemplo visual DT (Fuente: [16])

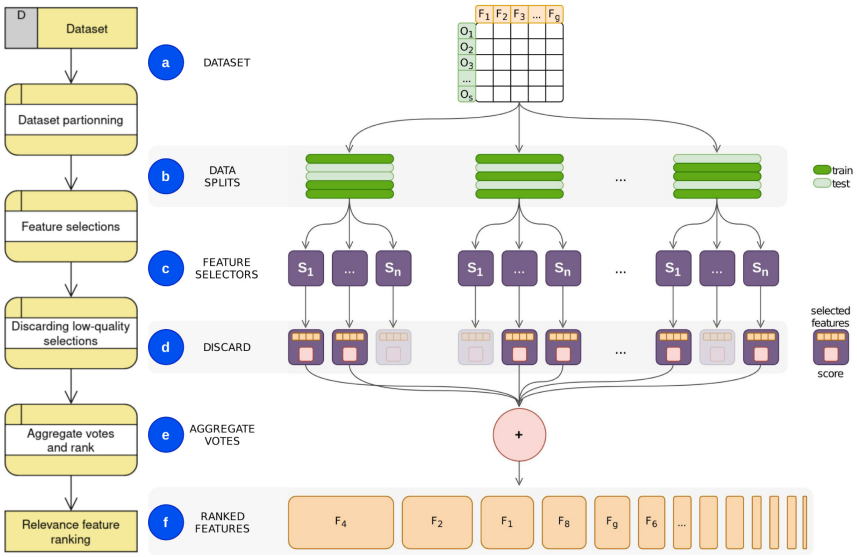

Figura 3.4: Esquema de los componentes de SEQENS (Fuente: [1])

3.2 SEQENS

SEQENS es un algoritmo de ensemble diseñado por el ITI (Instituto Tecnológico de Informática) [1] de Valencia que pretende paliar el problema inherente a los dataset con una alta dimensionalidad conocido como *curse of dimensionality* [2]. En estas circunstancias, es común que debido al alto número de variables, se encuentren relaciones falsas entre ellas debido a problemas como la cuasicolinealidad o la colinearidad entre las variables.

El funcionamiento de SEQENS se basa en una técnica llamada "Sequential Feature Search "(Búsqueda Secuencial de Características). Esta técnica técnica busca seleccionar las variables más relevantes para predecir un objetivo en un modelo predictivo. En lugar de comprobar todas las combinaciones, SEQENS tiene un enfoque distinto. Para explicar su funcionamiento, se ilustrará el proceso utilizando la Figura 3.4.

El proceso comienza con el dataset completo a analizar (A). SEQENS divide el conjunto de datos en múltiples particiones (B),cada una con su propio subconjunto de datos de entrenamiento y prueba. En cada partición, se aplica la técnica de Búsqueda Secuencial de Características (C), que añaden características de manera secuencial al subconjunto de características seleccionadas. En cada paso, se evalúa el rendimiento del modelo con el subconjunto actual de características y se selecciona la característica que más mejora dicho rendimiento.

Este proceso continúa iterativamente hasta que cumple con un criterio de detención: un número predefinido de características o cuando agregar más características no mejora el rendimiento del modelo. Una vez finalizado el proceso, se descartan las selecciones de variables que tengan baja calidad (D). Luego, se agregan las veces que cada característica aparece en los mejores subconjuntos (E) y se crea un ranking basado en la frecuencia de aparición (F). Esta última parte del proceso se puede interpretar como una votación, en la que cada subpartición de los datos vota por las variables que considera más relevantes.

La Búsqueda Secuencial de Características (Sequential Feature Search o SFS) es el núcleo del algoritmo SEQENS, y su funcionamiento se basa en dos pasos principales: un paso hacia adelante (Forward) y un paso hacia atrás (Backward). A continuación, se describen estos dos pasos en detalle: (este proceso pertenece a paso C del esquema 3.4.)

1. Paso hacia adelante (Fordward): Este paso comienza con un conjunto de variables vacío. En cada iteración, SEQENS evalúa todas las variables que aún no están en el conjunto actual y selecciona la variable que proporciona la mejora más significativa en la métrica de evaluación utilizada, R^2 en nuestro caso. Esta variable seleccionada se agrega al conjunto actual de variables. Llegados a este punto se pasa a un segundo nivel en el que se evaluarán todos los modelos con dos variables que incluyen a la primera ganadora en la primera etapa. Este proceso se repite iterativamente hasta que se alcance un cierto número de pasos (en la configuración predeterminada de SEQENS, este número es 10). Una vez llegado a ese nivel se evalúan todos los modelos entrenados en el recorrido del camino elegido y se elige aquella opción con mejor score obteniéndose así las mejores características que se incluyeron en ese momento.

2. Paso hacia atrás (Backward): n cada iteración de este paso hacia atrás, SEQENS evalúa todas las variables en el conjunto actual y elimina la variable que tiene el menor impacto en la métrica de evaluación utilizada. Este proceso también se realiza iterativamente, y el número de pasos hacia atrás es igual al número de pasos hacia adelante que se realizaron.

El resultado final de este proceso es un conjunto de variables que se consideran las más predictivas según la métrica de evaluación utilizada. Es importante destacar que el SFS se aplica de manera independiente a cada división del conjunto de datos (denominada “secuencial ”en SEQENS). Durante cada ejecución del algoritmo, se utilizan los datos de entrenamiento para ajustar el modelo inductor y los datos de prueba para evaluar su capacidad predictiva.

El papel del inductor es fundamental, ya que se ejecuta en cada nodo para cada combinación guiada de un número diferente de variables sobre cada uno de los secuenciales. La función objetivo se utiliza para guiar al algoritmo hacia la selección del mejor subconjunto de variables predictivas. En última instancia, el conjunto final de variables seleccionadas dependerá de cuántas variables se agregaron durante los pasos hacia adelante y cuántas de ellas se eliminaron durante los pasos hacia atrás. El objetivo principal es encontrar un subconjunto óptimo de variables que sea altamente predictivo, todo ello según la métrica de evaluación utilizada en función de los datos y el problema en cuestión.

La Fig.3.5 ejemplifica el funcionamiento de un secuencial. Como se puede observar, en la primera capa se evalúan todas las variables y se elige la que tiene mayor poder explicativo. Posteriormente, en capas inferiores se arrastra la variable de cada capa que ha formado un conjunto con mejor puntaje. Al final, el secuencial selecciona como lista final aquella que ha presentado el mayor valor. Sobre cada nodo del secuencial se aplica el inductor seleccionado.

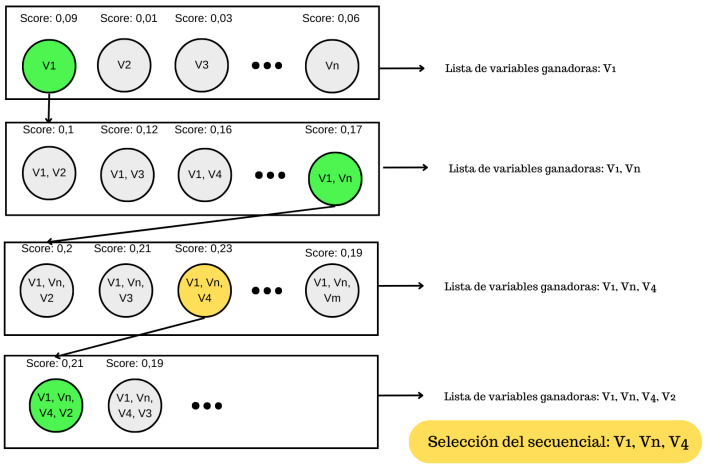

Figura 3.5: Funcionamiento de un secuencial. Elaboración propia

El algoritmo SEQENS destaca por su capacidad para abordar la falta de estabilidad en las selecciones de características en entornos de alta dimensionalidad, como es el análisis genómico, donde el número de variables es significativamente mayor que el número de observaciones. Su enfoque se basa en dividir los datos en múltiples secuenciales y aplicar la técnica de selección de características de manera independiente en cada una de estas divisiones. Esto permite obtener selecciones de características más sólidas y coherentes, afrontando así los desafíos asociados con la *maldición de la dimensionalidad*.

3.3 Cliff

La métrica Cliff, presentada en [1], se utiliza para evaluar la calidad de una lista de características en una clasificación, especialmente cuando se conoce un conjunto de variables relevantes de antemano. Esta métrica proporciona una puntuación que oscila entre 0 y 1. Un valor de 1 se alcanza cuando todas las variables relevantes están en las primeras posiciones de la lista, y disminuye a medida que estas variables relevantes se encuentran más abajo en la clasificación.

La fórmula para calcular la puntuación de Cliff se basa en las posiciones de las variables relevantes en la lista de características. En otras palabras, no solo tiene en cuenta si las características relevantes están presentes en la lista, sino también dónde están posicionadas dentro de la lista. La puntuación de Cliff se calcula utilizando la siguiente fórmula:

$$s(R,p)) = \begin{cases} \frac{1}{R} - \frac{\alpha}{2}(p-R-1) & \text{si } p < R \\ s(R,R)(p-R+1)^{-\beta} & \text{si } p \geq R \end{cases} \tag{3.1}$$

Las variables en la ecuación son fundamentales para comprender cómo afectan al comportamiento de la puntuación, como se ilustra en la Figura 3.6:

- R: En la ecuación, R es un valor fijo que representa el número de variables relevantes para una enfermedad específica. En la ejecución mostrada en la

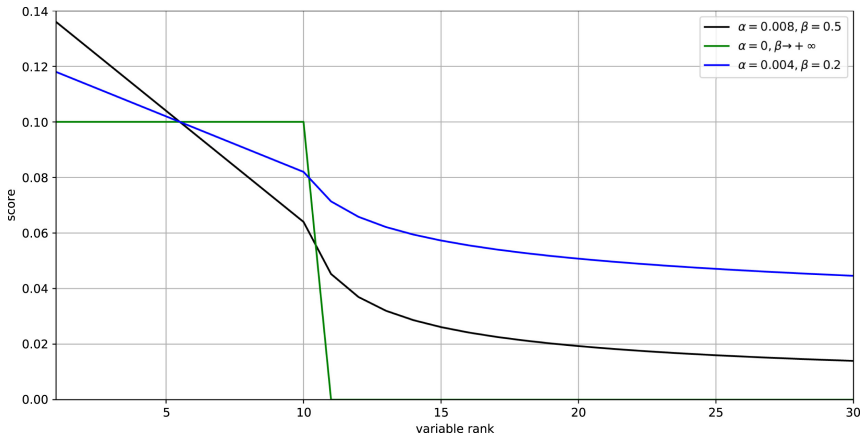

Figura 3.6: Comportamiento de Cliff. (Fuente: [1])

Figura 3.6, se ha fijado en 10. Se puede observar que a partir de la posición 10 en el eje X, la forma de la función cambia significativamente.

- *p*: En la ecuación, p es una variable que representa la posición de una característica específica en el ranking. Dependiendo de si p es menor que R o mayor o igual a R, se aplicará una de las dos fórmulas para calcular s(R,p). Se evaluará de manera diferente si la variable se encuentra entre las posiciones R (es decir, si puede ser una variable relevante) o no.

- α y β: Estos son parámetros ajustables en la ecuación que afectan cómo se calcula s(R,p). α influye en la puntuación cuando p es menor que R, mientras que β afecta la puntuación cuando p es mayor o igual a R. Dependiendo de los valores de α y β, verás diferentes patrones en cómo cambian las puntuaciones a medida que te mueves a lo largo del ranking. Se puede observar en la Fig.3.6 como influye a la puntuación distintos valores de estos parámetros.

3.4 Simulación datos

La generación de datos sintéticos desempeña un papel fundamental en la realización de nuestros experimentos. Con ello, se aborda la dificultad de obtener conjuntos de datos completos y tener un control preciso sobre qué variables se consideran relevantes y cuáles no. Estos datos sintéticos representan una *ground-truth* conocida, lo que nos facilita llevar a cabo evaluaciones precisas. Además, esta capacidad de generar datos nos permite proponer distintas distribuciones, lo que, nos brinda la flexibilidad necesaria para evaluar el algoritmo en diversos escenarios.

Puesto que vamos a trabajar en un espacio multidimensional de características formadas por SNPs, es crucial entenderlos [17]. Los polimorfismos de nucleótido único son variaciones genéticas debidas a diferencias en un solo nucleótido en la secuencia de SNP. Cada uno representa un cambio específico en una ubicación particular del ADN. En la práctica, se generarán arrays para representar numéricamente la información genética. Cada posición en el array correspondería a un

SNP específico, y el valor en esa posición indicaría la variante genética para ese SNP en ese individuo. Los posibles valores son 0, 1 y 2:

- 0: No se observan mutaciones en esta posición.
- 1: En esta posición hay una solo mutación presente en una de los alelos.
- 2: En esta posición se observan mutaciones en ambos alelos.

Los alelos son las dos copias de un gen que se heredan de los padres. En el caso de los SNPs, los dos alelos pueden ser iguales o diferentes. Si los alelos son iguales, el SNP se denomina homocigoto. Si los alelos son diferentes, el SNP se denomina heterocigoto.

Una vez generadas un determinado número de observaciones SNP con un determinado número de variables, se seleccionarán las variables relevantes para generar una enfermedad sintética con una determinada distribución. De esta manera, se conoce el *groundtruth*, es decir, qué variables son relevantes para la enfermedad sintética. De esta manera, podremos medir el rendimiento del algoritmo utilizado para seleccionar las variables relevantes. Las variables elegidas serán consideradas informativas, mientras que el resto se considerará ruido.

Tenemos la capacidad de generar enfermedades sintéticas con tres tipos de distribuciones en el espacio: hiperplano, hiperesfera y k-clusters.

- Escenario hiperplano: Los casos y los controles se separan en el espacio con un hiperplano en el subespacio definido por las variables elegidas como relevantes. Se observa en la Fig.3.7(Izq.) como se distribuye en el espacio los casos y los controles.
- Escenario de la hiperesfera: En este contexto, los controles se concentran alrededor de un punto central de origen, mientras que los casos se sitúan más allá de un radio. Como resultado, la frontera que separa los casos de los controles toma la forma de una hiperesfera. Puedes observar la distribución de esta frontera en el espacio en la figura 3.7 (Cen.).
- Escenario k-clusters: En el escenario de k-clusters, se analiza la distribución de los datos en un espacio multidimensional. En este proceso, las observaciones se organizan en grupos diferentes según las similitudes entre las variables. La figura 3.7 (Der.) ilustra la disposición de estos grupos en el espacio.

3.5 Inductores basados en la combinación de modelos

Se propone la combinación de modelos como solución para mejorar el rendimiento de SEQENS, con el objetivo de obtener puntuaciones medias superiores y aumentar la estabilidad del algoritmo en diversos escenarios.

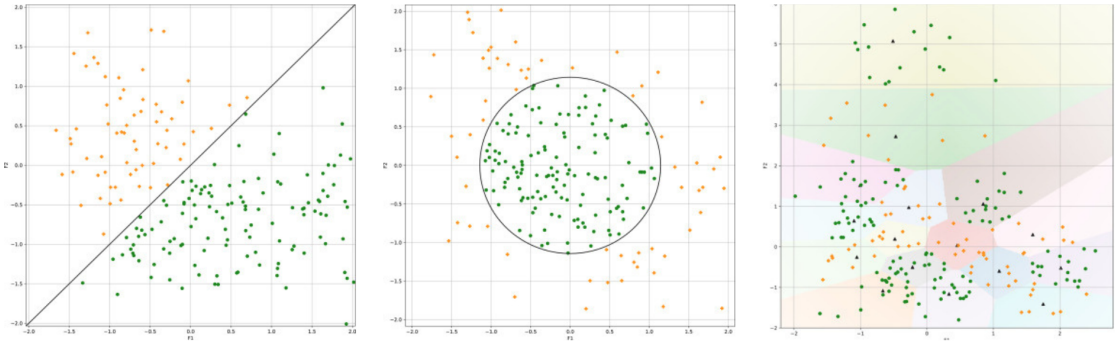

Figura 3.7: Escenario hiperplano (Izq.), escenario hiperesfera (Cen.) y escenario k-clusters (Der.). (Fuente:[1])

En un trabajo reciente de realizado por ITI [1], se comparó SEQENS con otros métodos de selección de características. Los resultados demostraron que la posición de SEQENS en el ranking de rendimiento variaba según la distribución específica de la enfermedad en los datos analizados. La capacidad de SEQENS para independizar su rendimiento de la distribución real y desconocida de los datos podría constituir una ventaja significativa, lo que mejoraría su rendimiento global.

Para lograr este objetivo, se propone hacer uso de una combinación de modelos, creando un modelo inductor único que tenga en cuenta el rendimiento de varios modelos. Este inductor considerará diversas perspectivas sobre la realidad y las posibles distribuciones de las enfermedades.

Con este fin, se han diseñado modelos para evaluar diferentes modos de combinación. La combinación de modelos se centra en la forma en que se manejan los puntajes obtenidos por cada uno de los modelos para calcular un puntaje global. Este puntaje global será fundamental para la selección de características en cada partición de datos (secuencial). Cada secuencial, a través de una serie de pasos Backward-Fordward, buscará optimizar este puntaje global.

Este enfoque busca que el algoritmo se adapte de manera óptima a la distribución real de los datos. Al tratar de optimizar un puntaje que tiene en cuenta el rendimiento de varios modelos, se espera que el algoritmo ajuste su comportamiento en función de los resultados que cada uno de estos modelos presenta.

Cada uno de los modelos propuestos se origina a partir de una interpretación y combinación única de los puntajes generados por los modelos que lo conforman. La eficiencia de cada modelo se refleja en el valor de R^2 que produce, y este valor varía según cómo se haya gestionado el R^2 de cada modelo para obtener el puntaje del inductor. Se han desarrollado e implementado tres inductores distintos en función de diferentes combinaciones:

- Inductor Máximo: En este enfoque, se calcula la puntuación final seleccionando la puntuación más alta entre las métricas de cada modelo. Esto resalta la métrica en la que un modelo se destaca más y la utiliza como medida unificada de rendimiento. (Rn), representa a los R^2 obtenido por cada modelos que compone el inductor. Este modelo se denomina MIX-MAX.

$$R_{\text{Final}} = \text{máx}(R_1, R_2, \ldots, R_n)$$

- Inductor Media Aritmética Ponderada: Aquí, las predicciones (Yp_i) ponderadas por su rendimiento se suman y luego se dividen por la suma total de ponderaciones. Esto permite tener en cuenta la contribución relativa de cada métrica en el resultado final. (Ri), representa a los R^2 obtenido por cada modelos que compone el inductor. Este modelo se denomina MIX-MAP.

$$y_{\text{res}} = \frac{\sum_{i=1}^{n} y_{p_i} \cdot r_i}{\sum_{i=1}^{n} r_i} \tag{3.2}$$

  El valor R^2 final se obtiene comparando $Yres$) con los datos reales utilizando la función r2_score. Este valor representa la calidad de ajuste del conjunto de modelos combinados a los datos observados.

- Media aritmética geométrica: Las puntuaciones ponderadas se combinan utilizando la media geométrica ponderada. Esta fórmula permite dar más peso a las métricas que son más importantes, al tiempo que considera la contribución relativa de cada métrica en el resultado final. (Ri), representa a los R^2 obtenido por cada modelos que compone el inductor. Este modelo se denomina MIX-MGP.

$$y_{\text{res}} = \exp\left(\frac{\sum_{i=1}^{n} r_i \cdot \log(y_{p_i})}{\sum_{i=1}^{n} r_i}\right) \tag{3.3}$$

  El valor R^2 final se obtiene comparando y_res con los datos reales utilizando la función r2_score. Este valor representa la calidad de ajuste del conjunto de modelos combinados a los datos observados.

CAPÍTULO 4
Resultados

4.1 Experimento uno: Selección de características basado en datos sintéticos

El propósito de este experimento es analizar el rendimiento de SEQENS al implementar distintos inductores para la selección de características. Se pretende comparar cómo afecta al rendimiento el uso de inductores *simples* frente al uso de los inductores propuestos, que están formados por la combinación de distintos modelos inductores. El experimento se centra en la evaluación comparativa del rendimiento, medido a través del indicador Cliff (Sección 3), del algoritmo SEQENS implementado con siete inductores distintos. Tres de los inductores serán los propuestos basados en la combinación de modelos, y los otros cuatro serán modelos inductores simples.

Los diferentes inductores serán evaluados en términos de su capacidad para identificar las variables genéticas influyentes en una enfermedad sintética. Esta enfermedad sintética se genera para controlar las variables que seleccionamos como significativas, así como la distribución de la enfermedad con respecto a las variables elegidas. Esto permitirá evaluar los resultados del algoritmo en diferentes situaciones, comparando diferentes modelos inductores con diferentes técnicas de combinación.

Diseño del experimento

El diseño del experimento implica la utilización de tres enfoques combinatorios de modelos. Cada uno de estos enfoques se basa en la combinación de cuatro modelos: Decision Tree Regressor, kNN y Support Vector Regressor con Kernel polinomial y con kernel rbf. En cada uno de los tres enfoques, la determinación del score final del inductor se realiza mediante la agregación de los puntajes obtenidos de los modelos que los conforman (Sección 3). Esta determinación se ha realizado de tres formas diferentes: mediante la selección del valor máximo, la media aritmética ponderada o la aplicación de la media geométrica ponderada.

Los modelos inductores que serán objeto de comparación son los siguientes:

1. Modelos basado en combinatoria de modelos mediante la media ponderada de sus scores. (MIX-MAP)
2. Modelos basado en combinatoria de modelos mediante la media geométrica de sus scores. (MIX-MGP)
3. Modelos basado en combinatoria de modelos mediante el score máximo. (MIX-MAX)
4. kNN con un valor de K igual a 5. (KNN)
5. SVR con Kernel Polinomial de Grado 2. (SVR-POLY)
6. SVR con Kernel RBF. (SVR-RBF)
7. Decision Tree con una profundidad máxima de 3 y el criterio friedman_mse. (DT)

El funcionamiento de los modelos *MIX* está explicado en la sección 3. Los modelos 1, 2 y 3 están fundamentados en los modelos 4, 5, 6 y 7, utilizando los mismos parámetros y configuraciones, para posibilitar la comparación de su rendimiento. SEQENS [1], se basa en un esquema de votos que se realiza sobre la selección de características mediante un proceso SBFS (Sequential Backward Floating Selection) en un conjunto de particiones de datos. Estas particiones se denominan "secuenciales ", y para este experimento en particular, se han utilizado un total de 100 secuenciales.

Datos

Los datos empleados en este estudio han sido generados de forma sintética (ver Capítulo 3). Se ha procedido a generar un total de mil doscientas observaciones SNP (polimorfismos de nucleótido único), cada una con 854 variables correspondientes. A partir de esta información, se ha modelado de manera sintética una enfermedad que presenta una distribución en forma de hiperesfera.

Los intervalos de tamaños de muestra seleccionados para evaluar el rendimiento son: 300, 600, 900 y 1200.

Con el propósito de simular enfermedades, se ha definido previamente un conjunto de 20 variables de significativas. El conocimiento del *groundtruth* permite calcular el KPI Cliff y evaluar la performance de la aproximación estudiada lo que nos va a facilitar comparaciones de rendimiento entre aproximaciones.

Resultados y análisis

A continuación, se presentan y analizan los resultados del experimento uno, que se ha realizado para evaluar el rendimiento de diversas implementaciones de SEQENS, específicamente en función del modelo utilizado como inductor

En la figura 4.1 se puede observar el rendimiento de distintas implementaciones de SEQENS según el modelo que se ha utilizado como inductor.

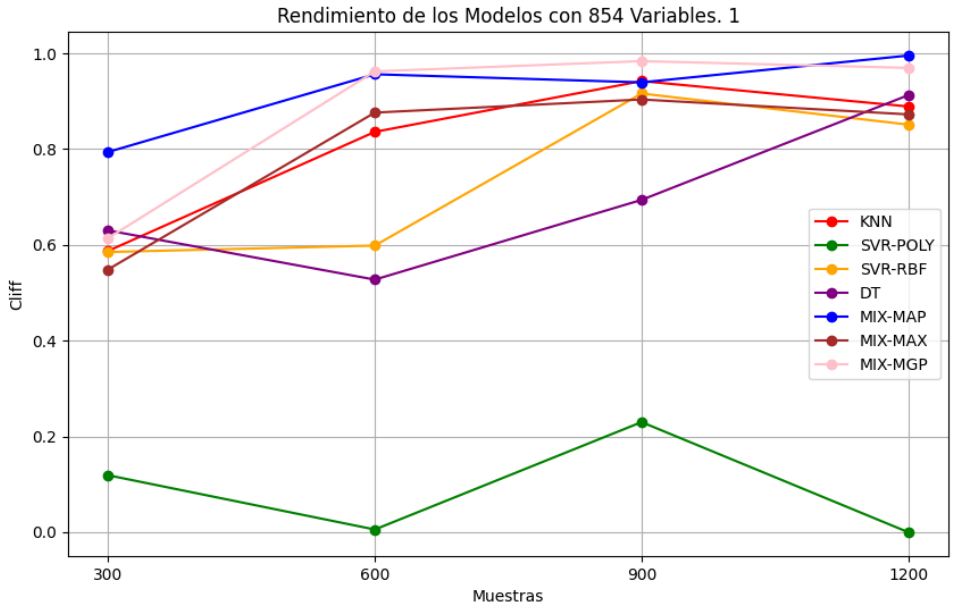

Figura 4.1: Resultado experimento uno. Elaboración propia

Podemos destacar el desempeño que muestran las ejecuciones de SEQENS con los inductores *mix* propuestos para mejorar el rendimiento. Estos destacan sobre el uso de inductores simples.

En concreto, el modelo MIX-MAX exhibe la mejor puntuación de Cliff con 300 muestras. Para 600 muestras, dos modelos liderar las puntuaciones, MIX-MAX y MIX-MGP. Para 1200 muestras, MIX-MAX y MIX-MGP destacan por encima de los demás inductores. Por último, cabe destacar que el inductor MIX-MGP alcanza un valor de Cliff superior a 0.95 en todos los rangos de muestras por encima de 600.

Como conclusión, los modelos inductores basados en la combinación de modelos han obtenido mayoritariamente mejores resultado los modelos individualmente. Específicamente, MIX-MAP se distingue por obtener el mejor rendimiento tanto en la situación de máximo como de mínimo número de muestras.

En la figura 4.2 detalla los tiempos de ejecución de cada modelo en distintos números de muestras. Cabe mencionar que las ejecuciones han sido en paralelo, considerando la alta carga computacional. No obstante, los tiempos se presentan como si se hubiesen ejecutado de forma secuencial.

La Figura 4.2 muestra claramente que el avance en rendimiento va acompañado de un incremento en el consumo de recursos. Dado que los modelos mixtos son una combinación de modelos individuales, el tiempo de ejecución de los mixes se extiende en proporción al tiempo de cada modelo individual.

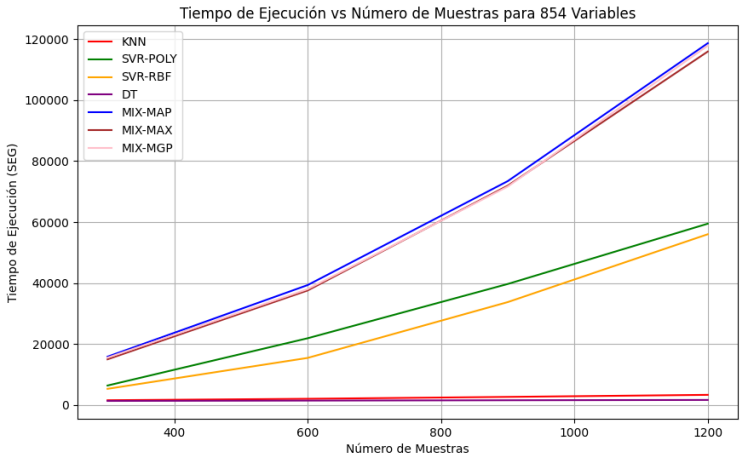

Figura 4.2: Tiempos de ejecución según la cantidad de muestras. Elaboración propia.

Discusión

Los resultados obtenidos hasta el momento parecen alentadores. Sin embargo, es crucial tener en cuenta el contexto y la metodología empleada, esto resalta algunas circunstancias que requieren ser consideradas.

El experimento se ha realizado con datos sintéticos, la utilización de los datos reales podría alterar el rendimiento observada. Además, es importante destacar que se ha utilizado únicamente el escenario de la hiperesfera. El objetivo principal es que los modelos propuestos tengan un impacto positivo en la estabilidad general, independientemente del contexto específico. A pesar de haber observado un mejor rendimiento en este escenario particular, que fue elegido de manera arbitraria, es fundamental replicar estos experimentos en diferentes escenarios con distribuciones de enfermedades diferentes. Esto permitirá evaluar el rendimiento de la aproximación de inductores combinados en comparación con la de un inductor único de manera más generalizada.

Además, conviene recordar que los resultados mostrados provienen de una sola ejecución. Aunque se observe que generalmente los modelos propuestos mejoran el rendimiento, se requieren múltiples experimentos para asegurar la robustez de estos resultados. Por ejemplo, el excelente resultado del MIX-MAX con pocas muestras podría ser el resultado de un caso excepcional en el que tanto los modelos como las muestras tuvieron un desempeño excepcional.

En el próximo experimento, se buscará abordar los desafíos anteriormente mencionados.

Conclusión

Al comparar el rendimiento del algoritmo SEQENS con distintos modelos como inductores observamos que los modelos combinados propuestos para mejorar el rendimiento han mostrado que de forma general mejoran el rendimiento o igualan al mejor de los modelos (Fig. 4.1). Las mejoras en términos del indicador

Cliff son evidentes, si bien es importante señalar que estas mejoras se obtienen a costa de un gasto adicional de recursos. (Fig. 4.2).

Además, cabe destacar que los modelos *mix* han considerado las predicciones del modelo SVR con Kernel polinomial, el cual tenía las predicciones de menor calidad. Sin embargo, dichos modelos asignan mayor importancia a las predicciones que han demostrado un mejor desempeño y un mal desempeño no disminuye notablemente su rendimiento. Este hallazgo nos sugiere que los modelos *mix* podrían ser capaces de lograr resultados consistentes y favorables, independientemente de la distribución real de los datos en el espacio multivariante. Esto implica un grado de agnosticismo con respecto al rendimiento final en relación con la distribución real de la enfermedad. Esto se debe a que estos modelos se beneficiarán de las predicciones del modelo que interprete de manera más efectiva la relación entre las variables y los casos.

4.2 Experimento dos: Selección de características basado en datos reales

Este experimento pretende ampliar la óptica y estresar los resultados del primero. Para ello, se utilizarán datos reales y se realizarán varias ejecuciones del mismo experimento para comprobar la estabilidad de los resultados. Posteriormente, se añadirán experimentos complementarios con el fin de explorar la solidez de los mismo.

Diseño del experimento

Para este experimento se han utilizado los mismos modelos que se han presentado en el experimento comentado en 4.1. Por tanto, serán siete los inductores puestos a prueba. Cuatro serán modelos *simples* y tres serán modelos creados a base de la combinación de los cuatro modelos simples.

La idea del experimento es observar como se comporta la combinación de modelo en distintos escenarios y cual sería el rendimiento medio de los tres escenarios propuestos: Hiperesfera, Hiperplano y k-cluster (Sección 3).

Para este experimento se han realizado cinco ejecuciones SEQENS con cada uno de los siete modelos inductores sobre cada uno de los tres escenario para obtener la media y sus respectivas desviaciones.

Datos

Para la realización de este experimento se han contado con datos reales proporcionados por la empresa ITI. Los datos de los que se dispone y que se han hecho uso son los siguientes:

- Un conjunto de datos sensibles de 488 muestras del genoma de 488 individuos con 15581 variables. Para cumplir con la RGPD, el acceso ha sido restringido y anonimizando los datos a nivel de individuo y variable.
- Lista de variables relevantes para el desarrollo de Diabetes mellitus tipo 2.
- Tres tipos de targets: k-cluster, hiperplano, hiperesfera.

Cabe destacar que, a pesar de que los datos son reales, los targets se han generado sintéticamente para disponer de un *groundtruth* con el que evaluar el rendimiento de las distintas ejecuciones.

En este experimento, respecto al anterior, tenemos una limitación en el número de muestras puesto que son 488. De manera sintética creamos hasta 1200. Esto, confirma la problemática a resolver, en este caso real, vemos que la cantidad de variables de las que se dispone es 31,92 veces mayor que el número de muestras y en casos reales este ratio puede llegar a ser mucho más grande.

Resultados y análisis

Se han ejecutado las variantes de SEQENS sobre tres tipos de escenario: Hiperesfera, Hiperplano y k-cluster (ver Capítulo 3). Durante los gráficos se encuentran dos tipos de bandas. En los gráficos que muestran todos los modelos, la banda representada se ajusta a $1{,}96 \times (\sigma/\sqrt{n})$, siendo n el número de muestras obtenidas. Esta banda muestra el comportamiento de la media con un nivel de confianza del 95 %. Por otro lado, las bandas que se presentan en los gráficos que muestran un único modelo se ajustan a dos veces la desviación típica en ambos sentidos. Esta banda representa la dispersión de los puntajes obtenidos.

Escenario Hiperesfera

En la Fig. 4.3, se proporcionan los resultados obtenidos con las bandas que muestran el comportamiento de la media con un nivel de confianza del 95 %. Por otra parte, se presentan los resultados de cada modelo de manera individual, con las bandas que muestran la variabilidad de los puntajes en las Figuras 4.4 y 4.5. Estas figuras muestran las ejecuciones de SEQENS con los modelos *simples* y los modelos *mix* respectivamente. Además, la figura 4.6 muestra los boxplots para los diferentes inductores con diferentes número de muestras en el escenario hiperesfera.

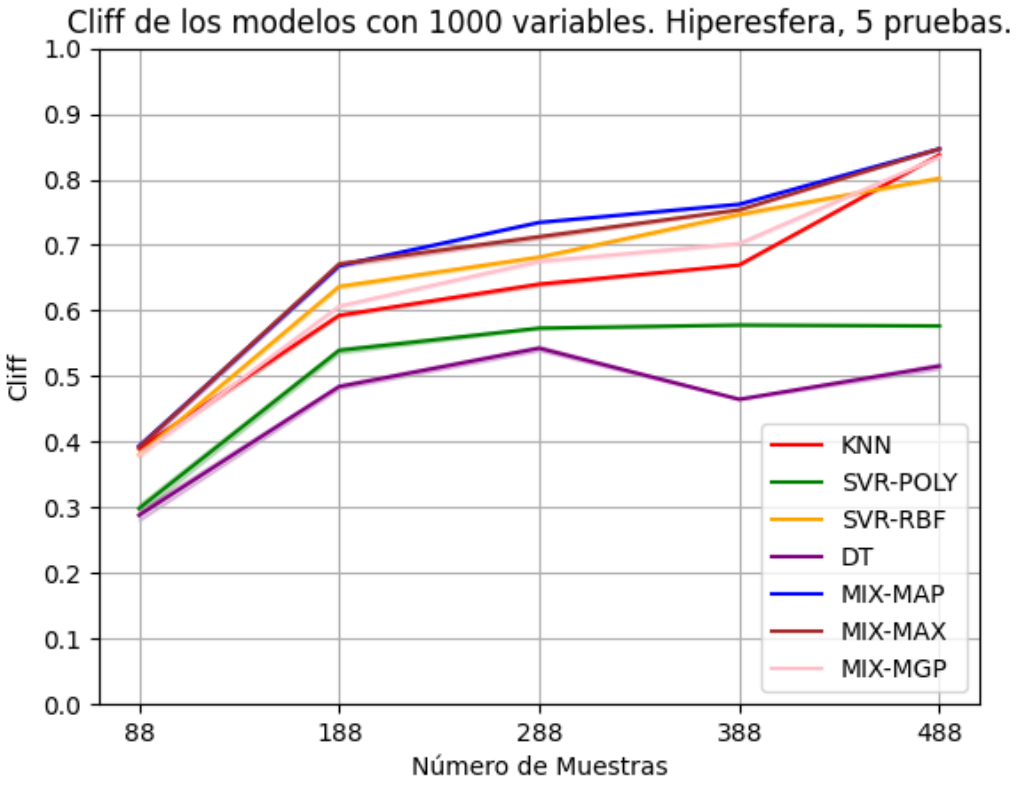

Figura 4.3: Resultado en el escenario Hiperesfera. Elaboración propia.

Se puede confirmar que en la mayoría de circunstancias, los modelos propuestos, han obtenido mejores y resultados más estables que la ejecución con modelos

simples (Fig. 4.3). En particular, los tres enfoques propuestos superan a las alternativas cuando se trata de conjuntos pequeños de muestras, y esta tendencia se mantiene al aumentar el número de muestras. Si observamos la Figura 4.6, se hace evidente que, para cada grupo de diagramas de caja asociados a diferentes números de muestras, los resultados de los modelos *mix* se sitúan consistentemente en la parte superior del grupo, y la variabilidad generalmente parece ser menor que la media del grupo.

Un aspecto destacado en la Figura 4.5 es la notable estabilidad que el modelo MIX-MAP mantiene a lo largo de todas sus iteraciones. Esta coherencia a lo largo del tiempo es un indicativo positivo de la robustez del modelo.

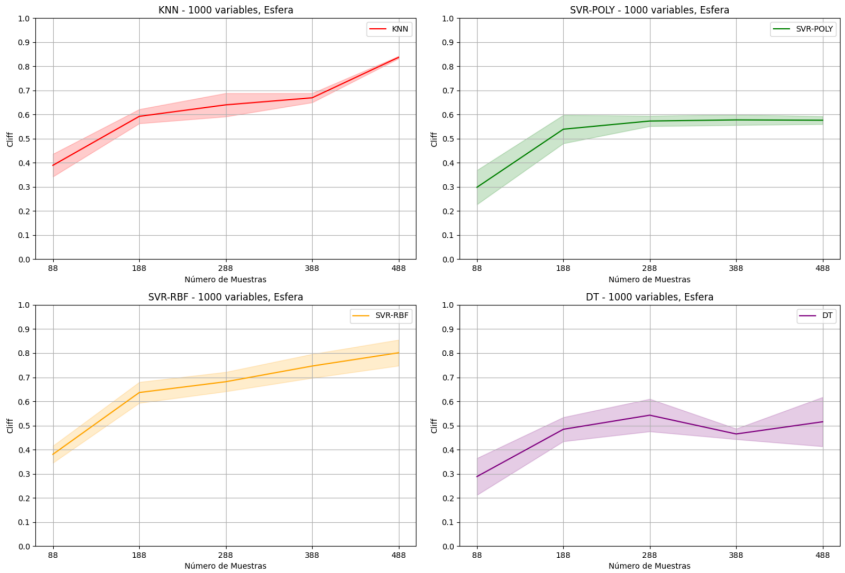

Figura 4.4: Rendimiento modelos simples en el escenario Hiperesfera. Elaboración propia.

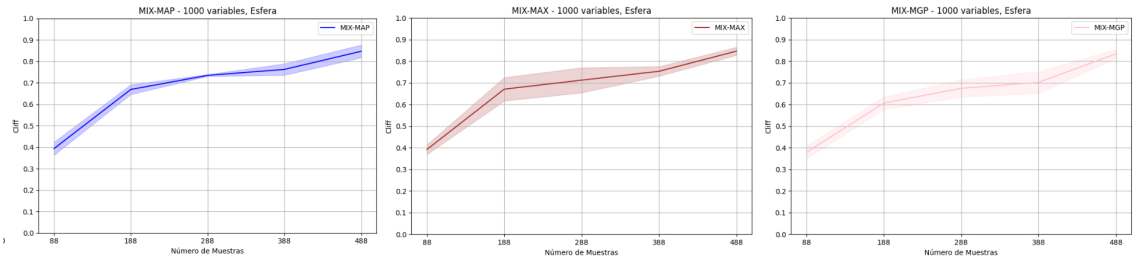

Figura 4.5: Rendimiento modelos mix en el escenario Hiperesfera. Elaboración propia.

Se puede concluir que en este escenario, los modelos *mix* propuestos han demostrado una relación positiva entre el número de muestras y el rendimiento, así como una notable estabilidad. Son hallazgos positivos que sugieren que los modelos propuestos podrían ser una solución óptima para abordar el problema planteado. Es importante resaltar que los hallazgos de este experimento refuerzan y amplían las conclusiones obtenidas en el experimento anterior, lo que aporta robustez a los resultados.

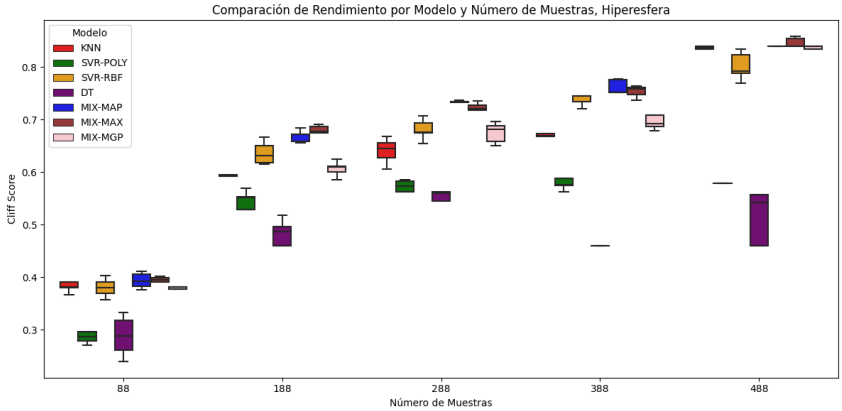

Figura 4.6: Diagrama de cajas, Hiperesfera. Elaboración propia.

Escenario Hiperplano

En la Figura 4.7 se presenta un resumen de los resultados obtenidos en este escenario. Se muestra el rendimiento y la variabilidad de cada modelo de manera individual en las Figuras 4.8 y 4.9. Además, se incluye un diagrama de barras en la Figura 4.10 que podría ayudar a la comparación integral de los resultados.

Al analizar la Fig.4.7 vemos que las diferencias entre los modelos serán menos marcadas. Podemos observar en la gráfica que los modelos que tienen un rendimiento inferior son DT, SVR y SVR POLY. Por otro lado, los demás modelos siguen una tendencia similar. Es notable el destacado desempeño de MIXMAP, que logra el mejor resultado con 88 muestras.

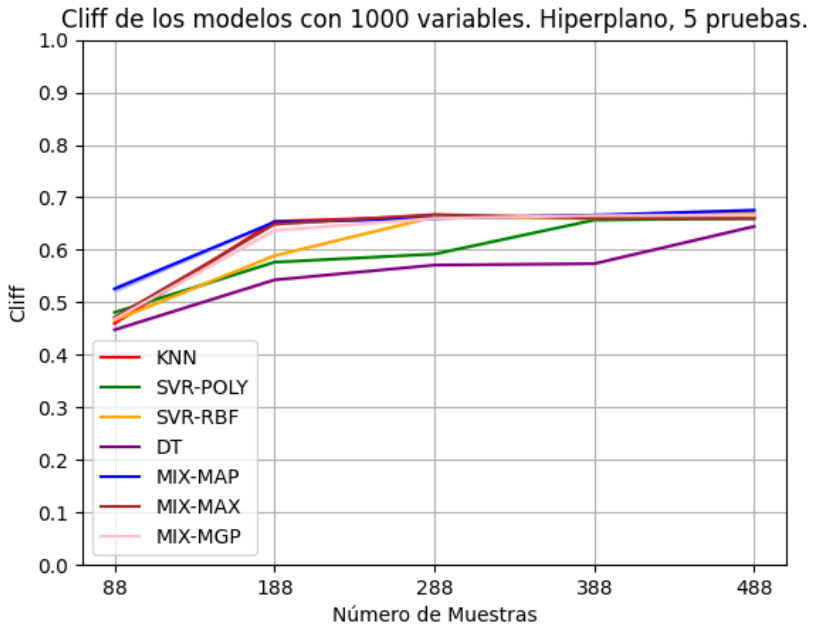

Figura 4.7: Resultado en el escenario Hiperplano. Elaboración propia.

En la Fig.4.8 y 4.9 vemos que la variabilidad de los inductores es bastante similar entre todos los modelos.

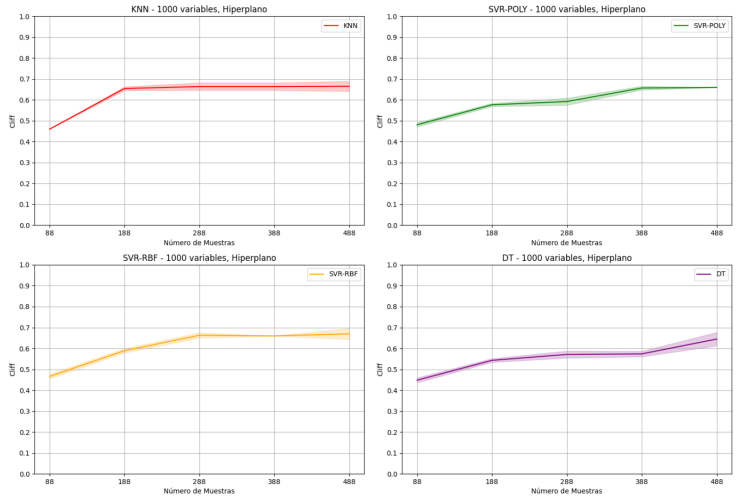

Figura 4.8: Rendimiento modelos simples en el escenario Hiperplano. Elaboración propia.

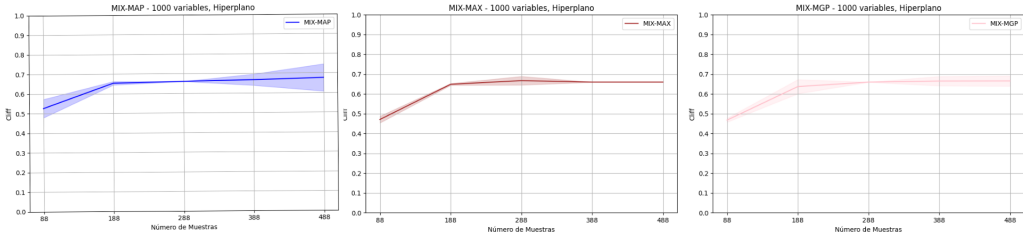

Figura 4.9: Rendimiento modelos mix en el escenario Hiperplano. Elaboración propia.

Al profundizar en el análisis de la Figura 4.10, se puede apreciar que, en general, los modelos *mix* están ubicados en la parte superior del grupo en casi todas las situaciones, y presentan variabilidades más restringidas.

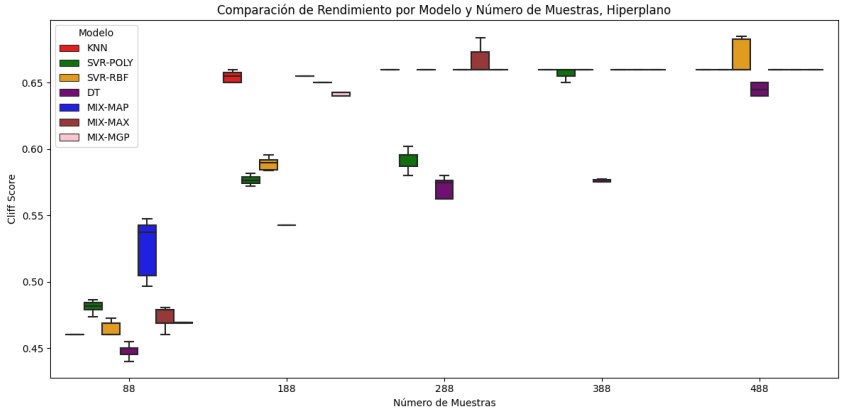

Figura 4.10: Diagrama de cajas, Hiperplano. Elaboración propia

Aunque las disparidades con respecto a los demás modelos no sean tan marcadas como en el escenario de la esfera, podemos afirmar que los modelos *mix* muestran una tendencia a ser más estables y a presentar mejores resultados en este escenario.

Escenario K-Cluster

Los resultados se presentan de manera similar a los escenarios anteriores. En la Figura 4.11, se muestran los resultados completos, mientras que para facilitar la visualización de los modelos *simples*, se presenta la Figura 4.12, y para los modelos *mix* se muestra la Figura 4.13. Finalmente, en la Figura 4.14, se presentan los diagramas de caja del experimento.

En la Figura 4.11, se observa un comportamiento relativamente similar entre los modelos, con algunas excepciones. SVR-POLY es el inductor con peor rendimiento, seguido de SVR-RBF y DT. Estos dos inductores muestran un rendimiento inferior al grupo entre las 88 y 288 muestras. A excepción de SVR-POLY, el resto de inductores presenta un comportamiento similar a partir de 288 muestras.

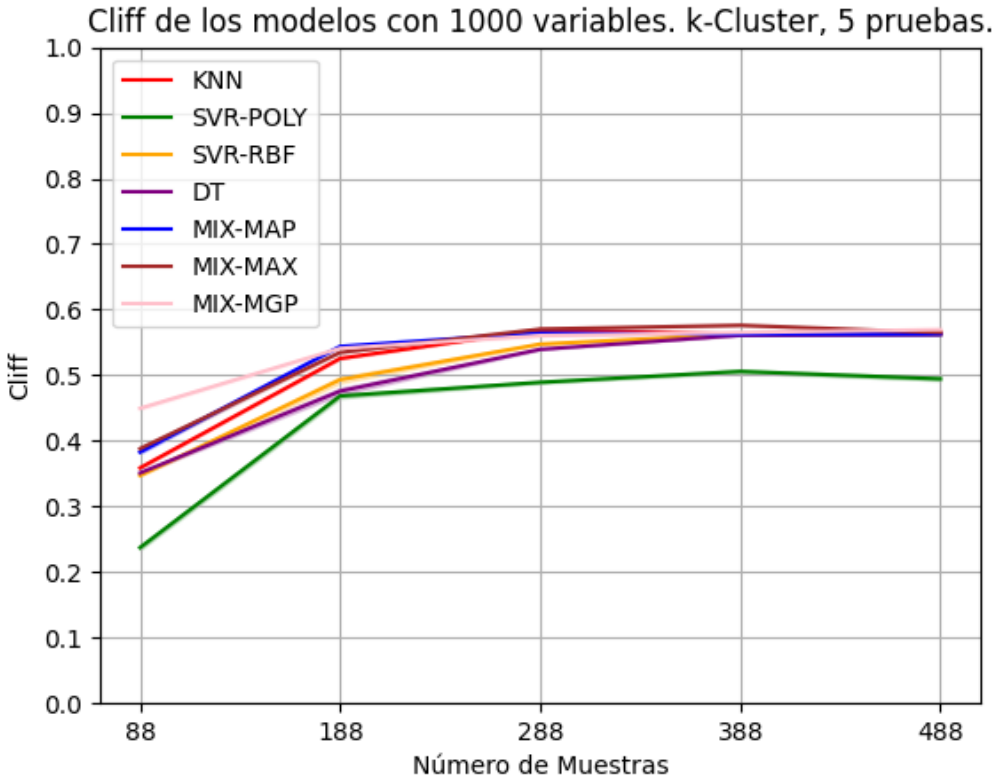

Figura 4.11: Resultado en el escenario K-Cluster. Elaboración propia.

Destacamos que para niveles de muestras bajos, los mejores modelos son los *mix*, siendo el mejor MIX-MGP para 88 muestras. Para 188 muestras, los tres *mix* y KNN muestran un rendimiento muy similar. En 288 muestras, los tres *mix* y KNN muestran un rendimiento equivalente.

Al comparar la estabilidad de los modelos *simples* (Figura 4.12) con los modelos *mix* (Figura 4.13), se observa que los modelos *mix* presentan una estabilidad a lo largo de todo el rango, algo que los modelos *simples* no logran, a excepción de KNN que muestra la mejor estabilidad.

Para concluir, en los diagramas de caja (Figura 4.14), los modelos *mix* vuelven a ubicarse en la esquina superior derecha de cada grupo de cajas. La diferencia es menos notable a medida que aumenta el número de muestras.

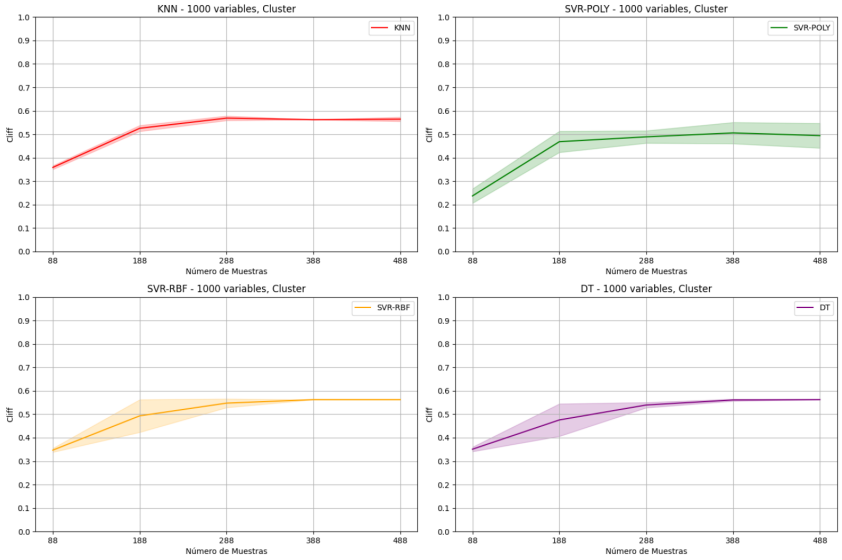

Figura 4.12: Rendimiento modelos simples en el escenario K-Cluster. Elaboración propia.

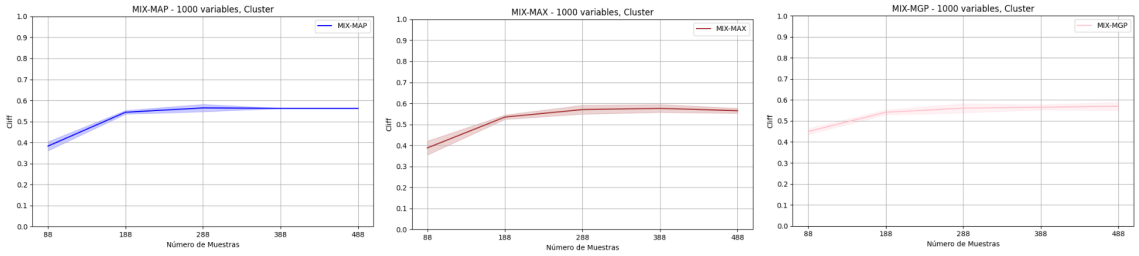

Figura 4.13: Rendimiento modelos mix en el escenario K-Cluster. Elaboración propia.

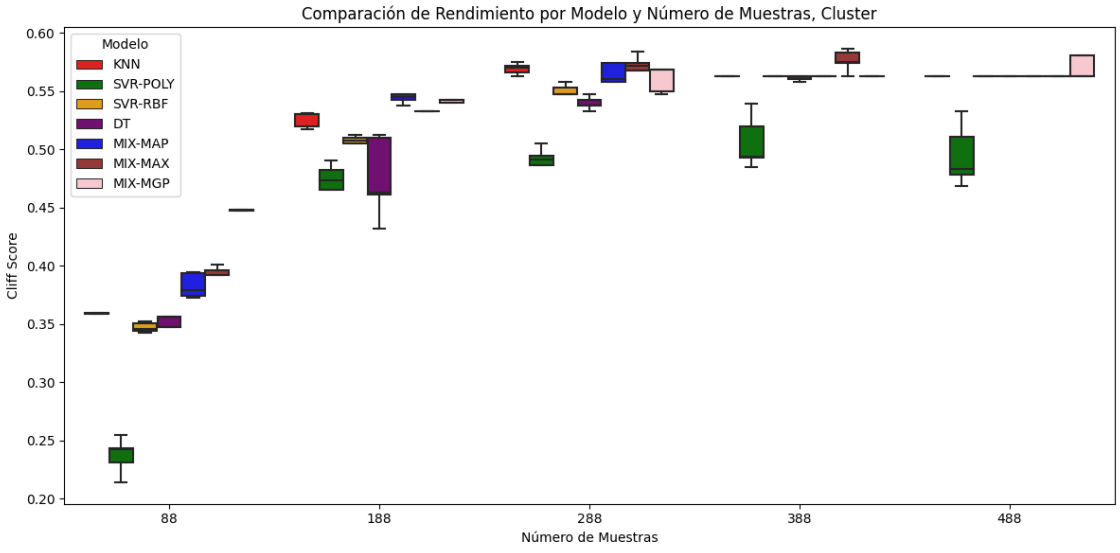

Figura 4.14: Diagrama de cajas, K-Cluster. Elaboración propia.

Resultados combinados

En esta sección se presentan los resultados de los tres escenarios juntos. Estos son los mismos resultados que se han presentado anteriormente, pero al agregarlos por media, podemos obtener una visión integral del funcionamiento de los modelos en situaciones reales. Como se mencionó previamente, la distribución real de los datos es desconocida, por lo que la combinación de resultados puede ayudar a obtener una idea de su comportamiento en situaciones del mundo real y eliminar el impacto de la distribución respecto del rendimiento final del algoritmo al realizar la identificación de variables.

En la Fig.4.15, podemos observar que los modelos forman dos grupos por similitud de rendimiento. Los que han obtenido generalmente un rendimiento más bajo son DT y SVR-POLY. Dentro del grupo que ha obtenido un mejor rendimiento, se destacan ligeramente las líneas marrón (MIX-MAX) y azul (MIX-MAP) cuando la cantidad de muestras está entre 88 y 288.

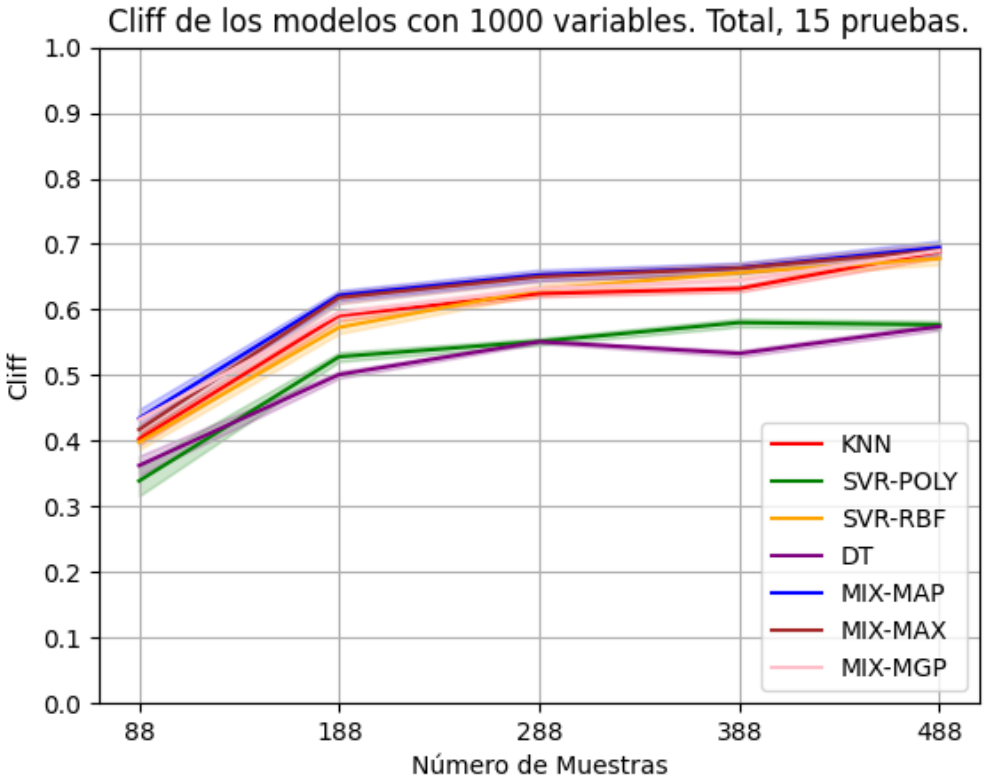

Figura 4.15: Resultado en el total de escenarios. Elaboración propia.

Observando la Figura 4.15 y 4.16, podemos obtener conclusiones más sólidas respecto a los grupos de muestras:

- Con 88 muestras, en general, los modelos *mix* han obtenido los mejores resultados y presentan la menor variabilidad. Es importante destacar que MIX-MAP alcanza los mejores resultados. El modelo que más se asemeja es KNN, aunque no logra puntuaciones tan elevadas.

- Con 188 muestras, seguimos observando el mismo patrón que con 88 muestras. Los modelos *mix* se ubican en la parte superior del grupo, siendo MIX-MAX y MIX-MAP los que obtienen los resultados más altos. El rendimiento de los modelos *mix* no solapa con KNN que es el modelo *simple* con mejor rendimiento.
- Con 288 muestras, sigue existiendo una ligera superioridad de los modelos mix. Sus rendimiento no solapan con ningún otro modelo.
- Con 488 muestras, las diferencias se acortan y aunque los modelos *mix* continúan obteniendo los mejores resultados, KNN muestra un rendimiento muy similar.

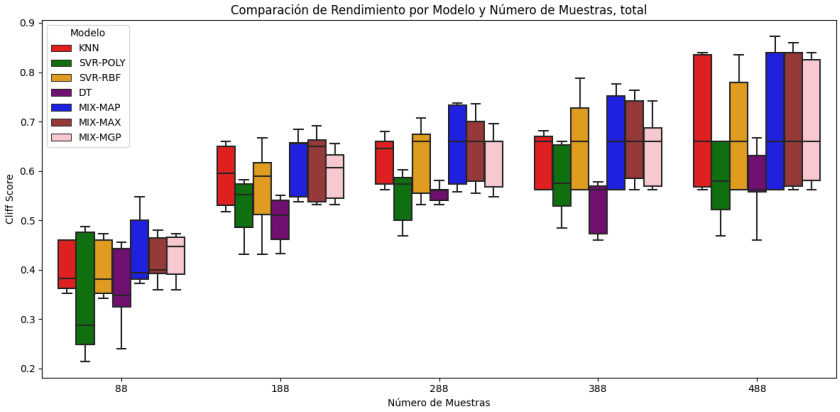

Figura 4.16: Diagrama de cajas, Total. Elaboración propia.

En lo que respecta a la estabilidad de los modelos, que se puede comparar en las Figuras 4.17 y 4.18, no podemos observar ningún modelo que destaque especialmente.

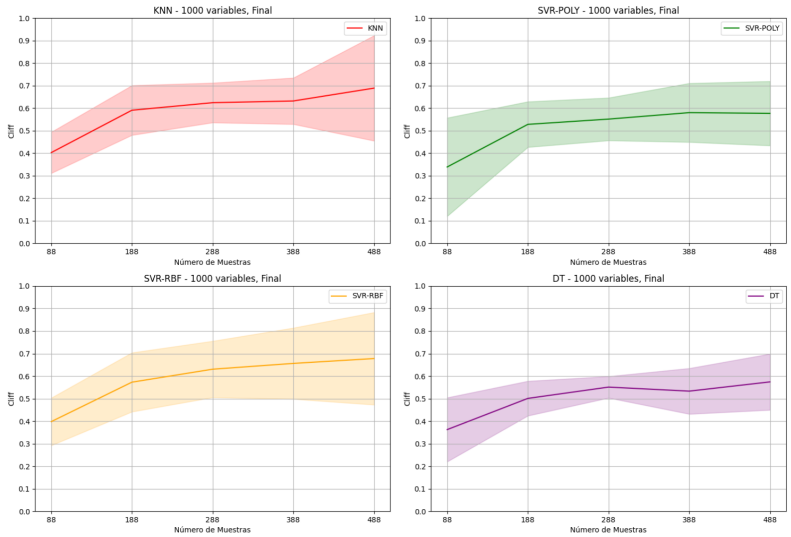

Figura 4.17: Rendimiento modelos simples en el total de escenarios. Elaboración propia.

En conclusión, podemos afirmar que los modelos *mix* obtienen en general resultados superiores a los modelos simples. No obstante, a medida que aumenta el número de muestras, las diferencias se reducen.

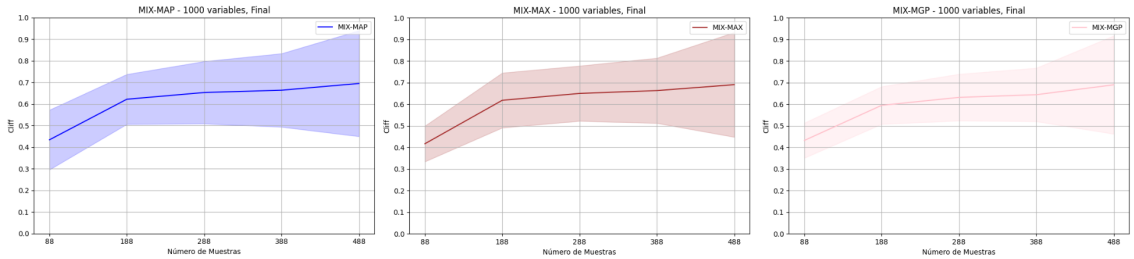

Figura 4.18: Rendimiento modelos mix en el total de escenarios. Elaboración propia

Discusión y conclusión

Este experimento ha ampliado el alcance del primero (4.1), las diferencias son: se han aumentado el número de experimentos para poder obtener un rango de rendimiento teniendo en cuenta distintas ejecuciones, se han utilizado ratos reales (488 muestra de genomas la respectiva separación entre casos y control) , se han testeado tres escenarios distintos y se han agrupado los datos.

A través de la comparación de los modelos *simples* y *mix* en escenarios de Hiperesfera, Hiperplano y K-Cluster, se pueden extraer conclusiones significativas que arrojan luz sobre la efectividad y la robustez de los enfoques propuestos.

En el escenario de Hiperesfera, es donde hemos encontrado las mayores diferencias entre los modelos *simples* y los *mix* a favor de los *mix*. Es especialmente evidente en los conjuntos de muestras más pequeños. Además, el modelo MIX-MAP, destaca por su estabilidad constante en todas las iteraciones. Estos hallazgos respaldan la hipótesis de que la combinación de modelos puede resultar en un mejor rendimiento y confirma las limitadas conclusiones que se obtuvieron en el experimento 1.

En el escenario de Hiperplano, los modelos *mix* nuevamente se posicionan como los más efectivos. Las diferencias entre los modelos no son tan pronunciadas como en el escenario de la Hiperesfera; sin embargo, se mantiene la tendencia de los modelos *mix* en obtener mejores resultados y una mayor estabilidad.

El el escenario de K-Cluster, se reafirma la superioridad de los modelos *mix* en términos de rendimiento y estabilidad. A pesar de obtener un rendimiento medio no tan distinto entre modelos, los *mix* presentan una variabilidad inferior a lo largo del rango de muestras.

En el análisis conjunto de los tres escenarios y teniendo en cuenta las conclusiones obtenidos por los resultados individuales, podemos concluir que los modelos *mix* superan a los modelos *simples* en términos de rendimiento y estabilidad en distintos escenarios. A medida que el número de muestras aumenta, tiende a reducirse las diferencias entre modelos *mix* y simples. Los resultados sugieren que la combinación de los modelos resulta ser siempre beneficiosa especialmente en conjuntos de datos más pequeños.

Entre los modelos *mix*, se aprecia un comportamiento similar. No obstante, MIX-MAP parece tener una estabilidad mejor y destacar en algunas ocasiones, especialmente en rangos de muestras bajos.

4.3 Experimento complementario 1: Distribución del ruido e información.

Cuando un secuencial realiza una selección de características en SEQENS (ver Sección 3), devuelve un puntaje asociado a dicha lista. El punjate devuelto puede servir para establecer dos distribuciones una que modele la distribución de la información donde las etiquetas son las originales y otras que modele la distribución del ruido habiéndose barajado previamente las etiquetas. Este experimento permite observar cómo se distribuye el ruido y la información dependiendo del inductor.

Una manera de mejorar SEQENS sería encontrar un umbral más certero que elimine los secuenciales de baja calidad. Esto permitiría incrementar la calidad del voto, ya que descartaría aquellos secuenciales cuya opinión bien podría proceder del ruido, pues no alcanzan un umbral mínimo de calidad.

Además, al descartar secuenciales de baja calidad, la calidad de los votos crecerá pues tan sólo contaran en la decisión aquellos secuenciales que han producido un valor de score mínimo que está dentro de la calidad mínima estipulada. Para ello se observarán la distribución de los scores obtenidos suponiendo que el target es el original (información) y la misma distribución cuando se ha producido un barajado previo de la etiqueta (ruido). La distribución del ruido permitirá establecer un valor mínimo de calidad a exigir a la selección con el target original basado en un percentil alto del ruido donde se observe poco o nulo solapamientro entre ambas distribuciones.

Diseño del experimento y datos

Para este experimento se han utilizado los modelos expuestos en el Experimento 1 y 2. Cuatro modelos *simples* y tres modelos *mix*. Por tanto, serán siete los inductores puestos a prueba para la obtención de ruido e información

Para poder obtener la distribución de los puntajes de la información y el ruido se han realizado dos ejecuciones de cada inductor por número de muestra. Una ejecución tiene el target establecido de manera correcta y la otra tiene el target barajado. De esta manera, se obtienen claramente los puntajes de los secuenciales que corresponden a la información y al ruido respectivamente.

Los datos que se han utilizado para el experimento son los mismos que en el Experimento 2 en el escenario de k-cluster. Un conjunto de datos reales de 488 individuos anonimizados y un target generado a partir de ellos con distribución de k-cluster (ver Sección 3).

El experimento, al ser complementario, se ha realizado en un solo escenario, con una sola iteración y con un número concreto de muestras. Estas variables se han elegido arbitrariamente.

Resultados y análisis

Tras la ejecución del experimento se obtienen los resultados mostrados en la Figura 4.19. Se puede observar en ella la distribución de los puntajes de los secuenciales de la ejecución con el target correcto (información) y con el target barajado (ruido).

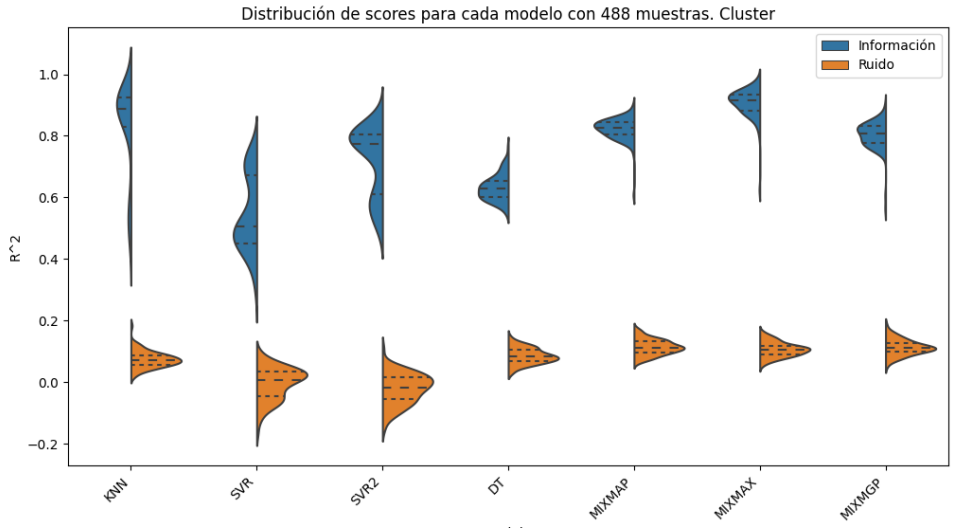

Figura 4.19: Distribución de ruido e información según modelos. Elaboración propia.

Se observar que los modelos *mix* muestran unas distribuciones de ruido e información más distanciadas y por lo tanto con un menor grado de solapamiento.. Además se puede observar que la distribución de la información es más compacta en los modelos *mix* que en el resto de modelos.

Discusión y conclusión

El experimento muestra limitaciones claras, solo se ha ejecutado una vez, en un escenario concreto y con un número de muestras determinados. Aunque las conclusiones no puedan ser rotundas, se pueden extraer algunas ideas para realizar trabajos futuros.

Es este experimento concreto se podría proponer un umbral de calidad de 0.7. Esto, en los modelos *mix* descartaría todos los secuenciales que pertenecen al ruido y aprovecharía la mayoría de secuenciales de información. Sin embargo, en otros modelos, aunque eliminaría todos los secuenciales de ruido, también eliminaría una gran parte de los que pertenecen a la información. En estos casos, el cálculo del puntaje se habría realizado en vano, ya que no alcanzaría el umbral de calidad mínimo.

También podemos extraer otra conclusión valiosa, los *mixs*, al presentar un una distribución de los puntajes más alta y compacta podemos asumir que la predicción será generalmente mejor y más fiable, puesto que la capacidad que tiene cada secuencial de realizar una selección de variables adecuada es en media superior que la mayor parte de los inductores.

En conclusión, se observa que la separación entre ruido e información en los modelos propuestos **mix**, es significativamente superior que en el resto de modelos. Esto sugiere que la calidad de la predicción será más confiable y que mediante la optimización del umbral de calidad de los secuenciales se podría obtener buenos resultados y optimizar el cálculo, ya que se descartarían los secuenciales que presentan ruido. A pesar de ser un experimento con limitaciones, abre las puertas a nuevas investigaciones futuras y supone otro argumento a favor del uso de los inductores combinados propuestos para mejorar el rendimiento.

4.4 Experimento complementario 2: Influencia del número de secuenciales.

SEQENS es un algoritmo de selección de características que se basa en la idea de realizar múltiples particiones del conjunto de datos original. Esto permite entrenar modelos de distintas maneras, obteniendo distintas opiniones y votos. De esta forma, el algoritmo pretende paliar el efecto de la *maldición de la dimensionalidad*, que dificulta la selección de características más relevantes [2]. Esto se consigue obteniendo la opinión de múltiples inductores, cada uno entrenado en una partición diferente del conjunto de datos.

En teoría, podría pensarse que un mayor número de particiones llevaría a un mejor rendimiento, ya que se contarían con más inductores que darían su opinión sobre las variables relevantes. En este experimento se pretende poner a prueba cómo afecta el número de secuenciales al rendimiento del algoritmo, implementado con distintos modelos y para diferentes tamaños de muestras.

Diseño del experimento y datos

Para llevar a cabo este experimento, se realizaron múltiples ejecuciones del algoritmo SEQENS. Se utilizaron dos tamaños de muestra: 188 y 488.

Se eligieron dos inductores para este experimento: el *mix* MIXMAP, representante de los modelos mix, y el inductor *simple* KNN. Estos modelos fueron seleccionados debido a que, según los resultados de los Experimentos 1 y 2, MIXMAP parecía ser el más estable de los *mix*, y KNN presentaba un rendimiento más cercano al de los modelos *mix* en comparación con otros modelos *simples*.

Los datos utilizados para los experimentos son los mismos que se emplearon en el Experimento 2, en el escenario de k-cluster. Un conjunto de datos de 488 individuos anonimizados y un target generado a partir de ellos con distribución de hiperesfera (ver Sección 3). Se han seleccionado dos escenarios de tamaño de muestra: 188 y 488, con el fin de obtener conclusiones en dos rangos de muestras.

Se ha explorado un rango de secuenciales desde 100 hasta 600, con incrementos de 50 secuenciales.

Resultados y análisis

Tras la ejecución del experimento, se obtienen los resultados mostrados en la Figura 4.20. En el eje X se representa el número de secuenciales, mientras que en el eje Y se presenta el valor de Cliff (ver Sección 3) obtenido en cada ejecución de SEQENS.

Tras el análisis de los resultados, se obtuvieron resultados interesantes. El rendimiento de MIXMAP es generalmente superior al de KNN en cualquier circunstancia del experimento.

Se observó una relación directamente proporcional entre el número de secuenciales y el rendimiento, con la excepción de KNN con 188 muestras.Este resultado

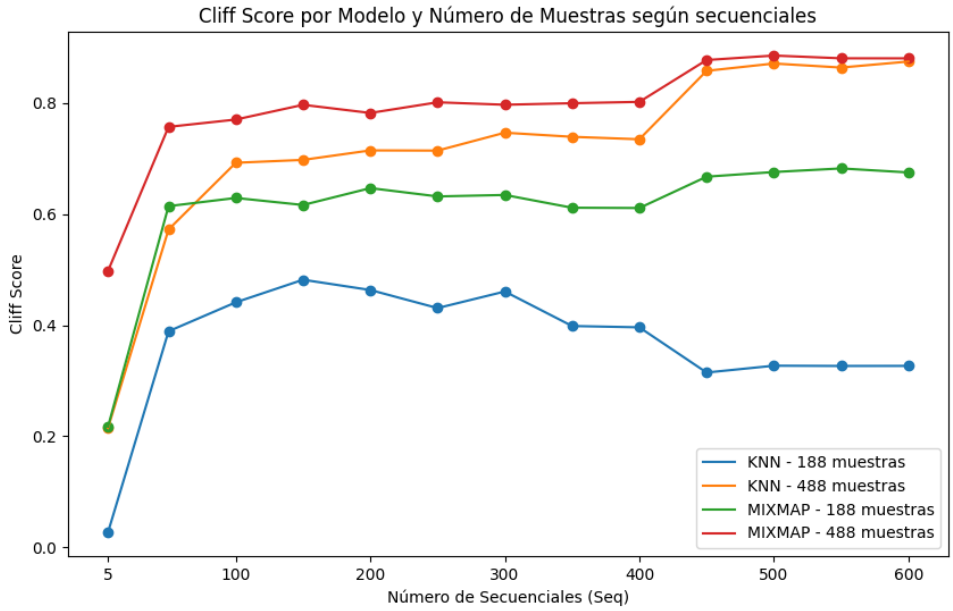

Figura 4.20: Variación de rendimiento según número de secuenciales. Elaboración propia.

sugiere que KNN puede no ser tan robusto como MIXMAP ante la escasez de datos.

Se identificaron dos niveles de rendimiento: uno entre 50 secuenciales y 400 secuenciales, y otro a partir de 450 secuenciales, mostrándose un cambio significativo en el rendimiento con 450 secuenciales en todos los escenarios.

Discusión y conclusión

A pesar de las limitaciones del experimento, que se ejecutó solo una vez en un escenario específico, se pueden extraer conclusiones valiosas.

El inductor *mix* propuesto, MIXMAP, mostró consistentemente un rendimiento superior al modelo KNN, lo cual respalda el uso de los modelos *mix* propuestos. En los experimentos anteriores, solo se habían ejecutado con 100 secuenciales.

La relación entre el número de secuenciales y la mejora del rendimiento es constante en los modelos *mix*. Se puede observar que en KNN el rendimiento es errático con 188 muestras; a partir de 150 secuenciales, parece que el rendimiento se vuelve inversamente proporcional al número de secuenciales. Esto proporciona cierta seguridad en el hecho de que con los modelos *mix* podría mantenerse siempre esta relación positiva entre más recursos y mejor rendimiento ante cualquier número de muestras cicunstancia que parece no observarse en el caso de bajo número de muestras con el modelo único KNN probado en este experimento.

En el momento que KNN se aproxima más al inductor *mix* es a partir de 450 secuenciales y 488 muestras, en ese punto el salto que rendimiento es mayor en KNN. Esto sugiere que para que KNN muestre un rendimiento similar necesita del uso de más recursos. Además este comportamiento sólo se observa en el caso

de la muestra grande (488) pues en el caso de muestra pequeña (188) KNN no sigue este patrón, mientras que *mix* si que lo hace en todo momento.

En conclusión, este experimento es un argumento sólido para el uso de los inductores *mixs* y abre nuevas oportunidades de investigación para mejorar el rendimiento. El modelo MIXMAP presenta apróximadamente un comportamiento monótono creciente del rendimiento ante el incremento del número de secuenciales mientras que KNN no. Se debe estudiar cuál puede ser un número óptimo de secuenciales en cada circunstancia, pero en el caso que nos ocupa el experimento indica un rango de indiferencia entre los 100 y 400 secuenciales y un salto significativo con 450 secuenciales.

CAPÍTULO 5
Discusión

Este trabajo se ha planteado con la intención de abordar el desafío que representa la selección de características en entornos de alta dimensionalidad, enfocándose de manera específica en la identificación de variables genéticas relevantes para enfermedades. Para afrontar la problemática inherente a la alta dimensionalidad, se ha empleado el algoritmo SEQENS. Su optimización ha sido propuesta como objetivo central de este estudio.

Al analizar los resultados de los experimentos 1 y 2, se puede concluir que los resultados obtenidos son alentadores. En ambos casos, se evidencia que el rendimiento de los inductores basados en la combinación de modelos supera a la utilización de modelos individuales.

Es relevante considerar diversos aspectos. Tanto para la selección de la combinación como para la comparación de los modelos, la elección ha sido realizada de forma arbitraria, tomando en cuenta aspectos como el tiempo de ejecución debido a las limitaciones temporales y al gran volumen de cálculos requeridos. Se espera que, con una selección más exhaustiva de los modelos, el rendimiento pueda ser mejorado aún más. Los resultados de estos experimentos indican que la combinación de modelos supera la efectividad de los modelos individuales.

No se ha realizado una optimización de los hiperparámetros de los modelos; se han empleado configuraciones preestablecidas y, en algunos casos, se han reducido los parámetros en el ámbito de los árboles de decisión para reducir el tiempo de ejecución. Una optimización de dichos hiperparámetros podría conducir a resultados más satisfactorios.

Si bien los resultados han sido prometedores, no se puede pasar por alto el aumento considerable en la carga computacional. SEQENS ya implicaba una carga sustancial, y al ejecutar los modelos *mix*, esta carga se incrementa aún más. Sería necesario estudiar si este incremento en la carga computacional está justificado por la mejora en los resultados obtenidos. Aunque cabe decir que el aumento de tiempo de ejecución no debería ser una limitación, pues la tarea de identificación de genes debe realizarse solo una vez y es de gran utilidad para la investigación médica. En este trabajo, se han realizado experimentos con datos reales, a pesar del alto costo computacional, demostrando que es posible llevar a cabo incluso con las restricciones temporales propias de un Trabajo de Fin de Grado y gracias también a que dicho algoritmo de selección de características es altamente paralelizable.

Al comparar los resultados con el artículo que presenta el algoritmo SEQENS, llegamos a una conclusión prometedora. En dicho artículo, SEQENS se compara con otros nueve algoritmos, y demuestra ser el más estable entre todos ellos. El propio artículo reconoce que no se han explorado las versiones más óptimas de los algoritmos comparativos, y que podrían existir implementaciones que ofrezcan un rendimiento aún mejor. Sin embargo, es importante señalar que SEQENS tampoco ha sido optimizado al máximo y su investigación es incipiente, lo que sugiere que aún existen diversas áreas de mejora por explorar.

El artículo presenta SEQENS con la implementación de un solo inductor, el KNN. En este trabajo, se ha seguido esta configuración y se ha superado el rendimiento de SEQENS con KNN al emplear modelos mixtos, particularmente cuando se trata de conjuntos de datos de muestras inferiores. Si SEQENS con KNN ya había demostrado su eficacia en ese contexto, se puede inferir que los modelos mixtos propuestos han logrado mejorar su rendimiento aún más. En consecuencia, se podría argumentar que si se repitiera la misma comparación que en el artículo, la diferencia en rendimiento sería más pronunciada y favorecería mucho más a SEQENS.

En resumen, a través de la experimentación realizada en este trabajo, se ha logrado demostrar la viabilidad de la idea de combinar modelos con el fin de optimizar el rendimiento de SEQENS y mejorar su adaptabilidad ante diversas distribuciones de datos. Este enfoque representa una contribución significativa en el estado del arte. SEQENS había demostrado ser una opción válida para la selección de características en entornos de alta dimensionalidad, y los descubrimientos presentados aquí presentan la posibilidad de potenciar su rendimiento y plantean nuevos caminos para futuras investigaciones en esta área.

CAPÍTULO 6
Trabajos futuros

A través de la experimentación, se ha logrado demostrar la validez de la estrategia de combinar modelos para potenciar el rendimiento de SEQENS, obteniendo resultados consistentes en muestras reales. A lo largo de este trabajo, han surgido dudas y oportunidades para continuar avanzando tanto en la mejora de la implementación de la combinación de modelos como en la optimización del propio algoritmo. Estas ideas abren caminos para futuras investigaciones en esta área, con el potencial de contribuir al avance de la selección de características en entornos de alta dimensionalidad y su aplicación en campos como la investigación biomédica y otros como química o industria. A continuación, se presentan algunas líneas de investigación que podrían explorarse:

- **Optimización de parámetros de SEQENS**: Los parámetros como training_size, max_interactions y n_sequential en SEQENS ofrecen oportunidades de optimización. Sería valioso determinar el punto óptimo de secuencias que encuentre un punto óptimo entre rendimiento y tiempo de ejecución. El experimento complementario 2 muestra resultados de cómo las secuenciales tienen influencia directa en el rendimiento del algoritmo.
- **Selección de modelos para la combinación**: En los experimentos presentados, la elección de modelos para la combinación ha sido en gran parte arbitraria. Una selección más fundamentada podría llevarnos a mejorar aún más los resultados. Por ejemplo, se podría investigar qué modelos son adecuados para diferentes distribuciones de datos y utilizarlos para la combinación.
- **Optimización de modelos individuales**: Existe margen en la mejora de la optimización de los modelos seleccionados para la combinatoria. Diferentes configuraciones de sus hiperparámetros internos podrían llevar a mejoras en el rendimiento global.
- **Investigar otras opciones de combinación**: Además de las estrategias de combinación interna utilizadas, como el máximo, la media ponderada y la media geométrica, sería interesante investigar otras formas de combinar los resultados de los modelos. La exploración de enfoques alternativos podría revelar combinaciones más efectivas.
- **Combinación externa de resultados**: Si bien la combinación de modelos se ha abordado internamente en este trabajo, es decir, cada inductor es un

modelo compuesto por una combinación de modelos individuales, otra posibilidad sería realizar ejecuciones de los modelos individuales y analizar diferentes formas de combinar los resultados obtenidos por separado.

- **Introducción de umbrales como filtro de calidad de los secuenciales**: El experimento complementario 1 muestra cómo se distribuyen los puntajes de los secuenciales en términos de proporcionar ruido o información. La introducción de umbrales de calidad podría ayudar a eliminar los secuenciales que no son relevantes, lo que podría mejorar la precisión de las predicciones, aumentar la fiabilidad y mejorar la eficiencia de cálculo.

Estas áreas de investigación podrían mejorar el rendimiento de SEQENS y ampliar el conocimiento para la mejora de la selección de características.

CAPÍTULO 7
Conclusiones

Este trabajo se ha dedicado a abordar una problemática de gran relevancia: la selección de características en entornos de alta dimensionalidad, como los que se encuentran en el campo de la genómica. Los microarrays, son estructuras de datos utilizadas para analizar la genómica del ADN, presentan un gran desafío debido a la gran cantidad de variables y la limitada posibilidad de obtener muestras en contextos médicos. Esta disparidad entre variables y muestras da lugar al problema conocido como *la maldición de la dimensionalidad*, donde el elevado número de variables aumenta la complejidad de identificar las más relevantes, el riesgo de sobreajuste y la detección de falsas relaciones.

Para abordar esta problemática, se ha utilizado SEQENS, un algoritmo que ha demostrado tener un papel relevante en la selección de características relevantes en contextos de alta dimensionalidad, en comparación con otras opciones en el estado del arte. SEQENS aplica Sequential Feature Selection utilizando un modelo de machine learning como inductor, operando en diversas particiones de los datos y combinando los resultados mediante un proceso de votación. El objetivo es que las variables relevantes sean seleccionadas con mayor frecuencia en las diferentes ejecuciones de SFS, buscando mejorar la estabilidad y lidiar con la complejidad de la dimensionalidad.

A pesar de los resultados previos favorables obtenidos con SEQENS, se ha observado que el algoritmo, al basarse en un único modelo como inductor, se ve influido en gran medida por la distribución de los casos y la capacidad interpretativa de dicho modelo. Dado que en situaciones reales no se dispone del conocimiento de la distribución real de las enfermedades, surge la posibilidad de mejorar el rendimiento de SEQENS a través de la combinación de modelos.

Así pues con el fin de mejorar el rendimiento del algoritmo, se ha propuesto la implementación de un inductor para la selección de características compuesto por una combinación de distintos modelos. Esta estrategia busca hacer a SEQENS más agnóstico ante la incertidumbre de la distribución de los datos, evitando depender de una única interpretación del espacio al usar un único modelo y considerando la visión de modelos con sus respectivas formas de interpretar el espacio de características. La expectativa es que los inductores combinados se adapte mejor a la distribución de los datos.

La implementación práctica involucró la creación de tres modelos mix, cada uno formado por cuatro modelos: KNN, SVR con kernel POLY, SVR con kernel

KFB y Decision Tree. Cada uno de estos modelos *mix* combina los resultados de formas distintas, asignando puntuaciones a las predicciones de cada secuencial mediante diversas combinaciones: el puntaje máximo, la media geométrica ponderada y la media artimética ponderada.

En la fase experimental (Exp. 1 y Exp. 2), se ha comparado SEQENS implementado con cada uno de los modelos mix, en contraposición a SEQENS implementado con cada uno de los modelos individuales que conforman los modelos mix. Se han realizado pruebas utilizando datos sintéticos, datos reales y distintas distribuciones de los datos.

Los resultados obtenidos han sido evaluados sobe tres escenarios distintos (hiperesfera, hiperplano y k-clusters) y en general sugieren que los modelos *mix*, en general, alcanzan mejores resultados y mayor estabilidad que las alternativas de modelo único. Asimismo, al considerar estos tres escenarios conjuntamente, los modelos *mix* también muestran un mejor desempeño y mayor estabilidad.

A pesar de los logros obtenidos, es importante señalar que esta mejora en el rendimiento ha conllevado un aumento sustancial en los recursos computacionales necesarios para la ejecución. Sin embargo, se argumenta que esta carga adicional podría estar justificada por la mejora de las prestaciones y la elevada capacidad de paralelización del algoritmo. La tarea de detección de variables genéticas relevantes solo se realiza una vez y es gran utilidad para la investigación biológica y clínica.

Además, esta investigación ofrece posibles direcciones para futuros trabajos con el fin de mejorar tanto el algoritmo SEQENS como la estrategia de combinación de modelos. Entre las oportunidades a explorar se encuentra la optimización de la selección de modelos y sus respectivos parámetros, así como la investigación sobre la cantidad óptima de secuenciales para lograr un equilibrio entre tiempo de computación y rendimiento.

En resumen, este trabajo ha demostrado que la estrategia de combinar modelos para construir un inductor en la implementación de SEQENS es una táctica válida y efectiva para mejorar el rendimiento del algoritmo. Los resultados obtenidos han sido consistentes en muestras reales y han revelado posibles nuevos caminos de mejora del rendimiento.

Bibliografía

[1] François Signol, Laura Arnal, J. Ramón Navarro-Cerdán, Rafael Llobet, Joaquim Arlandis, Juan-Carlos Perez-Cortes. SEQENS: An ensemble method for relevant gene identification in microarray data. *Computers in Biology and Medicine*, 152:106413, 2023.

[2] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, Amparo Alonso-Betanzos. *Feature selection for high-dimensional data*. Springer, 2015.

[3] National Human Genome Research Institute, Genómica Consultado el 08/2023 en https://www.genome.gov/es/genetics-glossary/Genomica

[4] Organización Panamericana de la Salud, Diabetes Consultado el 08/2023 en https://www.paho.org/es/temas/diabetes

[5] statista, Ranking de los países con mayor número de enfermos de diabetes en 2021 Consultado el 08/2023 en https://es.statista.com/estadisticas/612458/paises-con-mayor-numero-de-personas-con-diabetes/

[6] NIH, Factores de riesgo para la diabetes tipo 2 Consultado el 08/2023 en https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/factores-riesgo-tipo-2

[7] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111-135, 2014.

[8] Z.M. Hira, D.F. Gillies. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, 2015:198363, 2015. doi: 10.1155/2015/198363.

[9] S. Selvaraj, J. Natarajan. Microarray data analysis and mining tools. *Bioinformation*, 6(3):95-9, 2011. doi: 10.6026/97320630006095.

[10] P. Drotár, J. Gazda, Z. Smékal. An experimental comparison of feature selection methods on two-class biomedical datasets. *Computers in Biology and Medicine*, 66:1-10, 2015. doi: 10.1016/j.compbiomed.2015.08.010.

[11] Machine Learning Lecture Notes, por Sebastian Raschka de University of Wisconsin–Madison. Publicado en 2008 Consultado en https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/02_knn_notes.pdf.

[12] Documentación de KNeighborsRegressor por scikit learn Consultado en https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html.

[13] Support Vector Regression Tutorial for Machine Learning, publicado en Analytics Vidhya Consultado en https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/#h-introduction-to-support-vector-regression-sv.

[14] SUPPORT VECTOR REGRESSION:PROPIEDADES Y APLICACIONES, trabajo de fin de grado de Juan José Martín Guareño Consultado en https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/#h-introduction-to-support-vector-regression-sv.

[15] Support Vector Machine - Regression (SVR), Notas de Ph.D Saed Sayad Consultado en http://www.saedsayad.com/support_vector_machine_reg.htm.

[16] Classification and regression decision trees explained Publicado en Learn by Marketing, Consultado en https://www.learnbymarketing.com/methods/classification-and-regression-decision-trees-explained/.

[17] ¿Qué son los polimorfismos de nucleótido único (SNP)? Publicado en MedliinePlus, Consultado en https://medlineplus.gov/spanish/genetica/entender/investigaciongenomica/snp/.

APÉNDICE A
Entorno de trabajo

El entorno de trabajo es una parte fundamental de la estructura de proyecto. Organizado de manera efectiva permite trabajar adecuada y organizadamente. Hay varias condiciones específicas del trabajo que dan sentido a la estructura que se presenta en esta sección.

- Colaboración con ITI: Al trabajar conjuntamente con ITI requiere una comunicación constante a nivel personal y de trabajo. Es necesario compartir constantemente archivos y documentos. La estructura del entorno permite hacerlo de manera bidireccional.
- Realización del computo remotamente: ITI pone a mi disposición capacidad de computo. El trabajo requiere de realizar cálculos con datos de gran volumen.
- TFG de desarrollo: la propia naturaleza de un trabajo de desarrollo necesita una estructura donde programar, organizar archivos y tener respaldos de seguridad.

Veamos como esta organizado el entorno de trabajo. En la figura 1. podemos ver una organización general del entorno de trabajo.

Como se ha aclarado, hay 2 usuarios principales que han hecho posible la realización del trabajo. Se comentará que relación tienen con cada elemento.

A.1 Máquina local

Es el corazón del entorno de trabajo, desde el cual se acceden a las diferentes herramientas.

Al iniciar prácticamente cualquier sesión de trabajo se realiza una conexión SSH a la máquina remota proporcionada por ITI.

A.2 Máquina remota

ITI pone a mi disposición una máquina para trabajar remotamente y ejecutar el código de los experimentos.

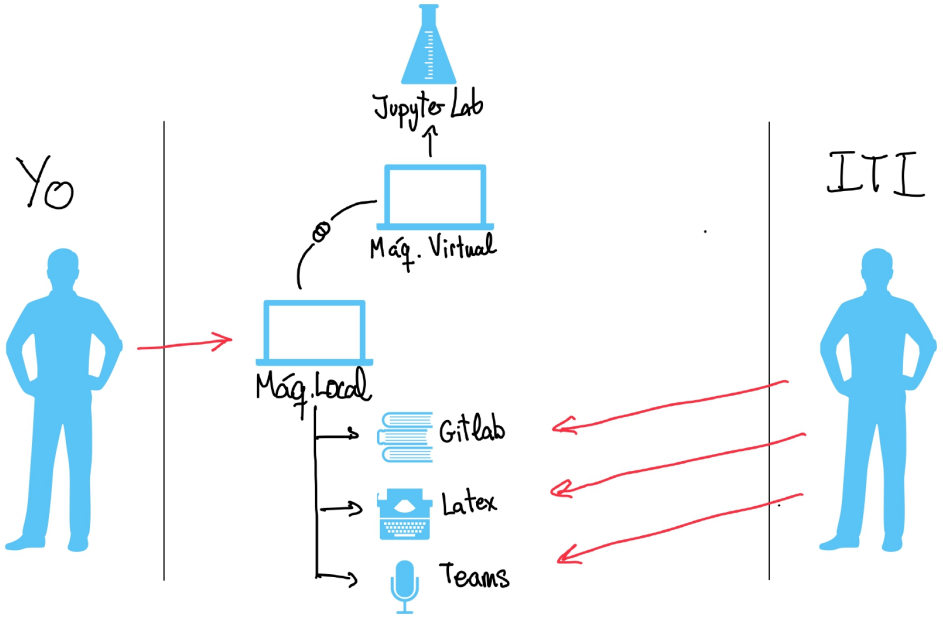

Figura A.1: Representación visual del entorno de trabajo

Se accede a la máquina virtual mediante una conexión SSH.

SSH es una tipo de conexión donde el cliente establece una sesión remota en un servidor a través de la red. Además, es un protocolo de red seguro q que proporciona un mecanismo de autenticación y encriptación.

Una vez conectado a la máquina se ejecuta el entorno de desarrollo Jupyter Lab y es asignado a un puerto para poder acceder desde el navegador local.

A.3 Navegador local

El Navegador permite acceder a distintas herramientas: Jypyter Lab, Gitlab, Latex y Teams.

A.4 Jupyter Lab

Jupyter Lab es un entorno interactivo para desarrollar y ejecutar código en diferentes lenguajes de programación, como Python.

Esta herramienta es excelente para trabajar de manera organizada, ya que permite la ejecución de Jupyter notebooks. Estos notebooks funcionan mediante celdas, que se pueden crear con formato Markdown para explicar lo que se está realizando, así como celdas de código para ejecutarlo.

Una vez que Jupyter Lab se esta ejecutando en la máquina y se le ha asignado un puerto, podemos acceder a él desde nuestro navegador local con una URL con la siguiente estructura: “http://localhost:puerto/?token ”

A.5 GitLab

GitLab una plataforma web basada en Git que proporciona herramientas para gestionar proyectos y llevar un control de versiones.

A esta herramienta tenemos acceso las 2 entidades del proyecto. GitLab trabaja por repositorios, por tanto, hay alguno donde ambos tenemos acceso y otros que son privados. Estas son las utilidades que nos proporciona:

- Control de versiones y back up: con forme se avanza en el desarrollo se realiza un control de versiones y se suben a modo de copia de seguridad
- Comunicación y coordinación: Aquí se encuentra nuestro centro de operaciones, donde llevamos un seguimiento de las reuniones, tareas por realizar, ideas para las próximas sesiones...

  También esta el acceso a Overleaf.
- Compartir documentos: para poder realizar este trabajo ha sido necesaria la transmisión de conocimientos por parte de ITI. Para ello hemos utilizado esta herramienta donde se ha compartido documentación de las librerías a utilizar, tutoriales, instrucciones para configurar el entorno de trabajo, archivos con datos...

A.6 Latex

Latex es un sistema que permite construir documentos de calidad centrándose principalmente en la estructura y el formato. Proporciona herramientas como: la gestión de referencias y bibliografías, portabilidad, personalización, alta calidad tipográfica y soporte para fórmulas matemáticas. Estas características hacen que sea una herramienta por encima de la competencia a la hora de desarrollar documentos relevantes o de aspecto académico.

Nosotros usamos Overleaf, una herramienta web implementa Latex y permite el trabajo en equipo. Tanto yo como ITI tiene acceso.

A.7 Teams

Teams en una herramienta de Microsoft diseñada para realizar y organizar reuniones. Mediante ella realizamos sesiones de seguimiento cada cierto tiempo.

APÉNDICE B
Implementación de los inductores mix

En esta sección se presenta la implementación de los inductores basados en la combinación de modelos propuestos como mejora para SEQENS. El concepto implementado se explica en la sección 3. Se han propuesto tres modelos basados en diferentes combinatorias: valor máximo, media aritmética ponderada y media geométrica ponderada. Estos modelos se han denominado MIX-MAX, MIX-MAP y MIX-MGP, respectivamente.

La implementación del regresor es común a los tres modelos. El código de la implementación es el siguiente:

```
class CustomRegressor(BaseEstimator, RegressorMixin):
def __init__(self):

    # iniciacion Modelos de regresion
    self.knn_regressor = KNeighborsRegressor(n_neighbors=5)
    self.svr_regressor = SVR(C=1.0, kernel='poly', gamma='scale',
        degree=2)
    self.dt_regressor = DecisionTreeRegressor(criterion='
        friedman_mse', max_depth=3, random_state=42)
    self.svr2_regressor = SVR(C=1.0, kernel='rbf', gamma='scale')

def fit(self, X, y):
    # Entrenamiento de los modelos
    self.knn_regressor.fit(X, y)
    self.svr_regressor.fit(X, y)
    self.dt_regressor.fit(X, y)
    self.svr2_regressor.fit(X, y)
    return self

def predict(self, X):
    # Predicciones de los modelos
    knn_pred = self.knn_regressor.predict(X)
    svr_pred = self.svr_regressor.predict(X)
    dt_pred = self.dt_regressor.predict(X)
    srv2_pred = self.svr2_regressor.predict(X)
    yr = np.column_stack((knn_pred, svr_pred, dt_pred, srv2_pred))

    return yr
```

Listing B.1: Implementación de regresor común a los inductores combinados.

La función de score es diferente para cada modelo, a continuación se presentan:

```
1
2  def myScoreMAX(y_true, y_pred):
3      y_p1 = y_pred[:,0]
4      y_p2 = y_pred[:,1]
5      y_p3 = y_pred[:,2]
6      y_p4 = y_pred[:,3]
7
8
9      r21 = r2_score(y_true, y_p1)
10     r22 = r2_score(y_true, y_p2)
11     r23 = r2_score(y_true, y_p3)
12     r24 = r2_score(y_true, y_p4)
13
14
15     return max(r21,r22,r23,r24)
```

Listing B.2: Implementación función score del inductor MIX-MAX.

```
1
2  def myScoreMAP(y_true, y_pred): #media aritmetica ponderada.
3
4      y_p1 = y_pred[:,0]
5      y_p2 = y_pred[:,1]
6      y_p3 = y_pred[:,2]
7      y_p4 = y_pred[:,3]
8
9      r21 = r2_score(y_true, y_p1)
10     r22 = r2_score(y_true, y_p2)
11     r23 = r2_score(y_true, y_p3)
12     r24 = r2_score(y_true, y_p4)
13
14     if(r21 < 0): r21 = 0
15     if(r22 < 0): r22 = 0
16     if(r23 < 0): r23 = 0
17     if(r24 < 0): r24 = 0
18
19     r2t = r21 + r22 + r23 + r24
20
21     if r2t == 0:
22         r2t = 0.0001
23
24     y_res = (y_p1 * r21 / r2t) + (y_p2 * r22 / r2t) + (y_p3 * r23 / r2t
           ) + (y_p4 * r24 / r2t)
25
26     score_result = r2_score(y_true, y_res)
27
28     return score_result
```

Listing B.3: Implementación función score del inductor MIX-MAP.

```
1
2  def myScoreMGP(y_true, y_pred): #media geometrica ponderada.
3      y_p1 = y_pred[:,0]
4      y_p2 = y_pred[:,1]
5      y_p3 = y_pred[:,2]
6      y_p4 = y_pred[:,3]
7      y_p5 = y_pred[:,4]
8
9
10     r21 = r2_score(y_true, y_p1)
11     r22 = r2_score(y_true, y_p2)
12     r23 = r2_score(y_true, y_p3)
13     r24 = r2_score(y_true, y_p4)
14
15     r21 = max(r21, 0.0001)
16     r22 = max(r22, 0.0001)
17     r23 = max(r23, 0.0001)
18     r24 = max(r24, 0.0001)
19
20     y_p1 = np.maximum(y_p1, 0.0001)
21     y_p2 = np.maximum(y_p2, 0.0001)
22     y_p3 = np.maximum(y_p3, 0.0001)
23     y_p4 = np.maximum(y_p4, 0.0001)
24
25     r2t = r21 + r22 + r23 + r24
26
27     if r2t == 0:
28         r2t = 0.0001
29
30     y_res = np.exp((r21 * np.log(y_p1 + 0.0001) + r22 * np.log(y_p2 +
           0.0001) + r23 * np.log(y_p3 + 0.0001) + r24 * np.log(y_p4 +
           0.0001) / r2t)
31
32     score_result = r2_score(y_true, y_res)
33
34     return score_result
```

Listing B.4: Implementación función score del inductor MIX-MGP.

ANEXO

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. **Fin de la pobreza.**				x
ODS 2. **Hambre cero.**				x
ODS 3. **Salud y bienestar.**	x			
ODS 4. **Educación de calidad.**				x
ODS 5. **Igualdad de género.**				x
ODS 6. **Agua limpia y saneamiento.**				x
ODS 7. **Energía asequible y no contaminante.**				x
ODS 8. **Trabajo decente y crecimiento económico.**				x
ODS 9. **Industria, innovación e infraestructuras.**		x		
ODS 10. **Reducción de las desigualdades**.				
ODS 11. **Ciudades y comunidades sostenibles**.				x
ODS 12. **Producción y consumo responsables.**				x
ODS 13. **Acción por el clima.**				x
ODS 14. **Vida submarina.**				x
ODS 15. **Vida de ecosistemas terrestres.**				x
ODS 16. **Paz, justicia e instituciones sólidas.**				x
ODS 17. **Alianzas para lograr objetivos.**				x

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

Mi Trabajo de Fin de Grado (TFG) está íntimamente relacionado con dos Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas: el ODS 3, que se enfoca en Salud y Bienestar, y el ODS 9, que se centra en Industria, Innovación e Infraestructuras. Mi investigación se adentra en el emocionante campo de la genómica para mejorar significativamente la forma en que manejamos y analizamos los datos genómicos humanos, con el objetivo de identificar las variables responsables de enfermedades. Esta investigación tiene un potencial impacto transformador en el bienestar y la salud de la sociedad en su conjunto.

En el mundo actual, la detección y comprensión de las variables que desencadenan enfermedades son fundamentales para el sistema de atención médica. Al mejorar nuestra capacidad para identificar estas variables, podemos fortalecer la prevención y el tratamiento de una amplia gama de enfermedades. Este enfoque es especialmente relevante para enfermedades raras y poblaciones con un número limitado de pacientes, donde la atención y el tratamiento personalizado son críticos.

Para lograr estos objetivos, he trabajado en la optimización de un algoritmo de vanguardia diseñado para la selección de características en conjuntos de datos genómicos. Este enfoque se alinea directamente con el concepto de innovación, ya que estamos utilizando el aprendizaje automático (machine learning) y la bioinformática para desarrollar soluciones de vanguardia en el campo de la salud.

A pesar de que la implementación de los ODS puede parecer compleja y abstracta en ocasiones, es fundamental comprender que estos objetivos encierran valores esenciales que pueden guiar nuestras acciones en la sociedad. Para los estudiantes universitarios, el conocimiento y la adopción de los ODS pueden servir como brújula moral a medida que avanzan hacia sus carreras profesionales.

Seguir los ODS no solo es una práctica ética, sino que también abre un mundo de oportunidades de emprendimiento. Los desafíos que abordan los ODS ofrecen un terreno fértil para la innovación y el desarrollo de soluciones sostenibles. Esto presenta oportunidades para emprendedores y profesionales comprometidos con el bienestar de la sociedad y el planeta.

Además de contribuir a los ODS 3 y 9, mi trabajo también aborda indirectamente otros Objetivos de Desarrollo Sostenible, como el ODS 4 (Educación de Calidad) y el ODS 17 (Alianzas para Lograr los Objetivos). A medida que avanzamos hacia un futuro impulsado por la ciencia y la tecnología, la educación desempeña un papel crucial. Mi investigación no solo avanza en la comprensión de la genómica, sino que también contribuye al desarrollo de habilidades técnicas y científicas necesarias para abordar los desafíos

globales. Además, colaborar con otros investigadores y profesionales en este campo demuestra la importancia de las alianzas interdisciplinarias para lograr avances significativos en la medicina y la salud.

En resumen, mi TFG representa un esfuerzo apasionado por contribuir al ODS 3 y al ODS 9 al mejorar la capacidad de identificar las variables responsables de enfermedades a través de avanzadas técnicas de análisis de datos genómicos. Este enfoque no solo es un acto de innovación científica, sino también un compromiso con la salud y el bienestar de las personas, así como con el desarrollo sostenible.

ETS Enginyeria Informàtica
Camí de Vera, s/n. 46022. València
T +34 963 877 210
F +34 963 877 219
etsinf@upvnet.upv.es - www.inf.upv.es