UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Departamento de Sistemas Informáticos y Computación

Universitat Politècnica de València

# APLICACIÓN DE MODELOS DE CLASIFICACIÓN A DATOS INCIERTOS

Trabajo Fin de Máster

**Máster Universitario en Ingeniería y Tecnología de Sistemas Software**

**Autor**: Yandry Paúl Quiroz Marcillo

**Tutores**: Mª Jose Ramírez Quintana
Cèsar Ferri Ramírez

2016 - 2017

# Agradecimientos

A mi familia por su apoyo incondicional en mis decisiones, A Silvia por comprenderme y aceptar este reto juntos, y a tres personas que contribuyeron a que esto fuera posible LLovany, Tía Sandra y Willian.

A mis tutores por el seguimiento y la dedicación brindada para la culminación de este trabajo.

A la empresa Marbelize S.A, donde laboro por la oportunidad y confianza brindada.

A mis compañeros de clases con los cuales formamos un buen equipo de trabajo y por las vivencias compartidas.

# Resumen

En el ámbito del aprendizaje de modelos a partir de datos, la calidad de los modelos depende en gran medida de la calidad de los datos usados en su entrenamiento. Es por ello, que de las etapas de que consta cualquier proceso de extracción de conocimiento, la de preparación y limpieza de los datos es la que ocupa un tiempo mayor. De hecho, es habitual que los modelos se generen con datos "limpios" bajo condiciones casi "perfectas". Sin embargo, la situación suele ser bastante diferente durante la etapa de aplicación del modelo, ya que los datos reales suelen contener ruido, puede haber valores faltantes, erróneos, o incluso ser inciertos en el sentido de que no conocemos su valor exacto, sino que solo tenemos un conocimiento aproximado de su valor.

En este TFM vamos a estudiar este último caso, cómo aplicar modelos a datos inciertos. Vamos a centrarnos en problemas de clasificación en los que la incertidumbre solo está presente en atributos numéricos.

**Palabras clave:** Incertidumbre en los datos, aplicación de modelos, aprendizaje automático, clasificación.

# Abstract

In the field of learning models from data, the quality of models depends to a large extent on the quality of the training data. That is the reason why the preparation and cleaning of data is one of the stages of the knowledge extraction process in which more time is investing. In fact, the most common scenario in the model training phase is that models be created under almost "perfect" conditions (and using "perfect" training data). However, the situation is often quite different during the model deployment phase, since, in the "real world", data usually contain noise, there may are missing or incorrect values, or even be uncertain, in the sense that we do not know their exact value, we only have an approximate knowledge of its value.

In this Master Thesis, we will study how to apply the learning models to uncertain data. We will focus on classification problems in which uncertainty is only present in numerical attributes.

**Keywords:** Uncertain data, model application, machine learning, classification

# Tabla de contenidos

# Índice

# Índice de figuras

# Índice de tablas

# 1. Introducción.

En la actualidad los avances tecnológicos permiten generar y almacenar grandes volúmenes de datos haciendo un uso generalizado de herramientas informáticas para la extracción adecuada del conocimiento que encierra la información. Este hecho ha transformado el análisis de datos, orientándolos hacia determinadas técnicas especializadas, las cuales se encuentran englobadas bajo el nombre de minería de datos.

La minería de datos hace uso del aprendizaje automático (del inglés, "Machine Learning") y tiene como objetivo utilizar datos y experiencias pasadas para resolver un problema que se plantee en la actualidad. Para ello se lleva a cabo un proceso de aprendizaje sobre un conjunto de datos, cuya clase ya se conoce (conjunto de entrenamiento), permitiendo así generar un modelo en base de relaciones, patrones o reglas, para poder clasificar nuevos elementos.

La calidad de los datos es vital, puesto que de ellos depende el correcto análisis para obtener así una información útil. Es ahí cuando la depuración y limpieza de los datos cobra un aspecto fundamental en el análisis, siendo una de las tareas que mayor esfuerzo requiere en la generación de modelos de minería de datos y es necesario adoptar estrategias productivas para tomar decisiones correctas [1].

La principal problemática que encontramos en el análisis, es que los datos no están en el formato y la forma adecuada puesto que existen grandes volúmenes de datos que presentan valores anómalos que se pueden haber generado por sensores que no han funcionado correctamente, información contradictoria porque los datos provienen de distintas fuentes de información, datos que contengan ruido o presentar valores faltantes que se han omitido por diversos factores, errores técnicos o humanos, averías en dispositivos [2], entre otros.

En este trabajo vamos a centrarnos en el tema de los valores faltantes, se debe tratar de su detección y su tratamiento previo a la fase de entrenamiento de los modelos, y tomar una decisión acerca de cómo serán resueltos, ya que muchas técnicas de aprendizaje son sensibles a la ausencia de valores.

Existen diferentes formas de tratar el problema de los valores faltantes [3] [4] [5], el objetivo de estas técnicas es que los datos estén completos en condiciones casi perfectas para obtener un conocimiento de calidad. Una de las formas más sencillas de resolver el tema de los valores faltantes es la eliminación de estas instancias, el inconveniente de

esta técnica es que si existen muchas instancias que presentan valores faltantes se puede perder mucha información degradándose el conocimiento esperado de los datos.

Otra forma alternativa para el tratamiento es rellenar los valores faltantes con algún valor, esto se puede realizar por medio de la técnica de imputación que a menudo implica procesos y métodos, es importante elegir un método de imputación acorde a las características del conjunto de datos para minimizar el ruido y el sesgo en ellos [6].

Una de las técnicas de la imputación es la media, es una de las formas más sencillas de estimar valores faltantes para obtener una muestra completa, consiste en reemplazar cada uno de los valores faltantes con la media de los datos observados para esa variable, tiene como desventaja que modifica la distribución de la variable reduciendo artificialmente su varianza [7].

Una manera mas sofisticada de estimar un valor es predecirlo a partir de otros ejemplos, utilizando cualquier técnica predictiva del aprendizaje automático (clasificación o regresión) [8], donde se estima la variable como un objetivo y con las variables restantes, se los emplea como predictores de la técnica empleada; o se pueden emplear técnicas más específicas, como por ejemplo determinar el sexo a partir del nombre.

También se pueden segmentar las tuplas por los valores que están disponibles, se obtienen modelos diferentes para cada segmento y luego se combinan. O bien modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

Los valores faltantes es un ejemplo de incertidumbre máxima donde no se conoce su valor, pero muchas veces se tiene un poco más de conocimientos sobre ellos, donde conocemos un valor aproximado teniendo un grado de incertidumbre presente en los datos, estos se pueden generar en la recolección de datos por diferentes factores, en el campo de la medicina se puede determinar la edad de un paciente cuando no se conoce y es factible estimar que la misma oscila entre 20 y 30 años, porque el paciente es joven, así mismo puede darse el caso de salarios de empleados expresados en intervalos, datos tomados con sensores los cuales tienen un grado de error que vienen determinados por las características.

Como se ve hay varios trabajos desarrollados para realizar la limpieza de los datos en la fase del aprendizaje para generar modelos óptimos, pero qué pasa si hemos aprendido el modelo con datos limpios y procesados; y, en la aplicación del modelo, los datos presentan incertidumbre. De acuerdo con la revisión de la literatura no se presenta un método donde los modelos se aprenden con datos limpios, en condiciones casi perfectas, pero que al momento de realizar su aplicación los datos no se encuentran bajo

las mismas condiciones presentando incertidumbre en sus datos. Este trabajo se centra en cómo aplicar técnicas de minería de datos a datos que contienen incertidumbre, solo nos vamos a centrar en atributos numéricos donde la incertidumbre es fácilmente representable en intervalos.

Por lo tanto, el objetivo de esta investigación es definir una nueva aproximación para aplicar modelos de clasificación aprendidos con datos "limpios y sin incertidumbre", a datos inciertos durante la etapa de test. Hemos desarrollado dos métodos diferentes: uno que hace uso de las etiquetas de clase predichas por el modelo (es decir, usando clasificadores de los llamados "crisp") y la segunda usando las probabilidades de pertenencia a cada clase predichas por un clasificador "soft". Evaluamos estos dos métodos experimentalmente usando diversas técnicas de generación de los clasificadores y varios conjuntos de datos.

El resto del documento está organizado de la siguiente manera. En el segundo capítulo se revisan los conceptos básicos sobre el proceso de extracción del conocimiento (del inglés, “Knowledge Discovery in Databases” (KDD)), sus diferentes fases, así como las tareas y técnicas aplicadas en la etapa de minería de datos. En el tercer capítulo, presentamos los dos métodos que proponemos para aplicar modelos de clasificación a datos que contienen incertidumbre. En el cuarto capítulo se detalla la evaluación experimental de los métodos propuestos en el capítulo anterior, incluyendo la metodología empleada y los diferentes escenarios de experimentación realizados, y se comentan los resultados obtenidos. En el quinto capítulo se presentan las conclusiones y se proponen trabajos futuros. Finalmente, en los anexos se incluyen los resultados experimentales desglosados por cada dataset.

# 2. Preliminares.

**Resumen**: En este capítulo se introducirán definiciones fundamentales referentes al proceso KDD, describiendo cada una de sus etapas. Se realizará una revisión detallada a cerca de la etapa de minería de datos, abordando tópicos complementarios a cerca de sus tareas y técnicas las cuales son fundamentales para el para el desarrollo del presente trabajo.

## 2.1. Proceso de Extracción del conocimiento (KDD)

El proceso KDD consiste en un conjunto de técnicas y tecnologías, diseñado para resolver problemas relacionados con la cantidad masiva de datos en la tarea de análisis de información de forma digital. Este proceso tiene como objetivo la identificación de patrones, que son capaces de explicar un problema y auxiliar en su solución [9]. Ha sido de gran aporte para los sistemas de información ya que los datos que almacenan son utilizados como materia prima, y al hacer uso de estos mediante modelos que ayudan a la interpretación y el análisis generan un valor agregado obteniendo conocimiento necesario para la toma de decisiones.

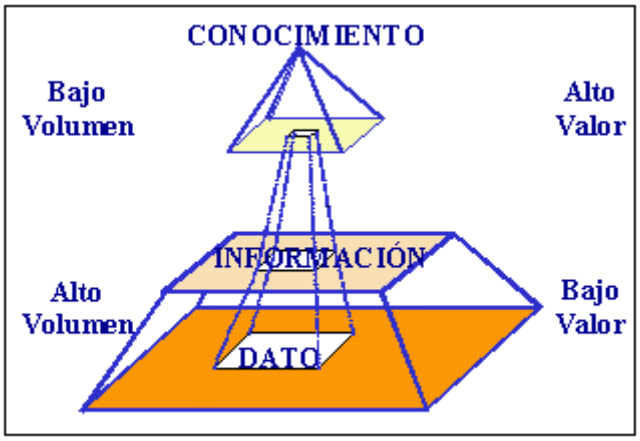



*Figura 1 Relación entre dato, información y conocimiento [10].*

Surge para ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

Existe una metodología llamada CRISP-DM (Cross Industry Standard Process for Data Mining) que se puso en marcha con un proyecto de la Unión Europea. El proyecto fue dirigido por cinco empresas: PSS, Teradata, Daimler AG, NCR Corporation y Ohra. Es

un proceso estándar que han adoptado las industrias para la minería de datos [11]. Consta de 6 fases definidas para su implementación:

- Comprensión del negocio
- Comprensión de Datos
- Preparación de datos
- Modelado
- Evaluación
- Despliegue

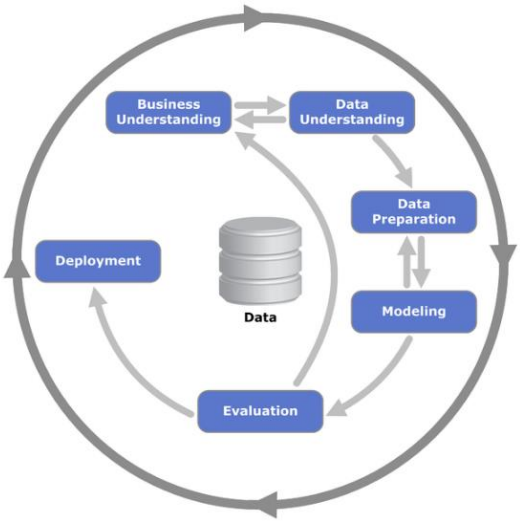



*Figura 2 Metodología CRISP-MD [7]*

Las flechas en el diagrama indican las dependencias más importantes y frecuentes entre fases. El círculo exterior en el diagrama simboliza la naturaleza cíclica de la minería de datos en sí. Un proceso de minería de datos continua después del despliegue de una solución. Las lecciones aprendidas durante el proceso pueden provocar nuevas preguntas de negocio, a menudo más centradas y posteriores procesos de minería de datos se beneficiarán de la experiencia de los anteriores.

### 2.1.1. Fases del proceso de extracción del conocimiento.

KDD es un proceso iterativo donde las salidas de algunas de sus fases pueden hacer volver a pasos anteriores donde se requieren de varias iteraciones para obtener conocimiento de calidad. También se considera un proceso interactivo donde un experto en el dominio del tema debe prestar ayuda para llevar a cabo procesos en algunas de las fases. A continuación, se describen las fases del proceso KDD [8].

### 2.1.1.1. Integración y recopilación de datos.

Las bases de datos y las aplicaciones basadas para el proceso tradicional en línea son suficientes para cubrir necesidades diarias de una organización como, por ejemplo: control de inventarios, facturación, etc. Pero resultan complejas para determinar funciones como el análisis, la planificación y la predicción, que nos sean útil para la toma de decisiones.

Para poder realizar un proceso KDD lo más normal es que los datos se encuentren distribuidos en diferentes fuentes de datos, lo cual es un reto muy grande ya que pueden utilizar diferentes claves primarias, formatos de registros, entre otros motivos. Lo primero es lograr la integración de las múltiples bases de datos, dando lugar a la tecnología de almacenes de datos (data warehousing).

En la Figura 3 se muestra un esquema sobre el proceso de integración de un almacén de datos a partir de distintas fuentes.

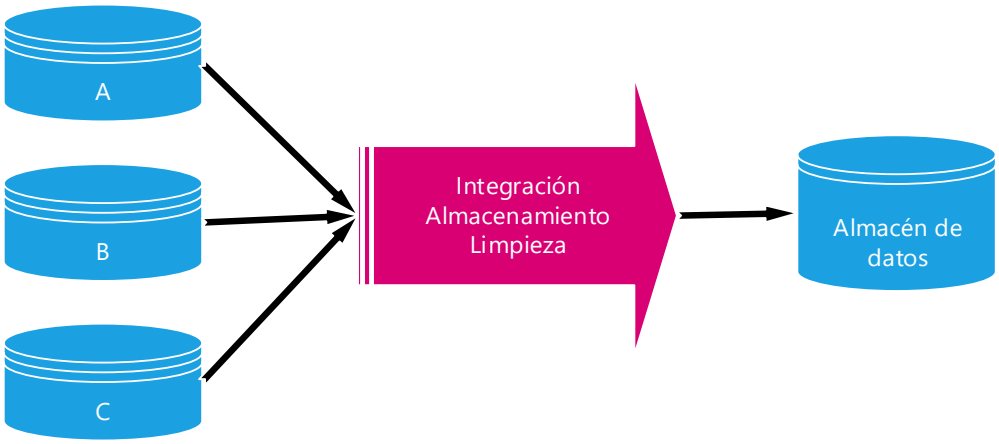



*Figura 3 Integración de en un almacén de datos [8].*

### 2.1.1.2. Selección, limpieza y transformación.

La calidad del conocimiento no solo depende del algoritmo de minería utilizado, también es importante la calidad de los datos, para ello después de integrar las diferentes fuentes se debe seleccionar, limpiar y trasformar para generar el conjunto o vista minable con el fin de detectar datos coleccionados en la etapa anterior que sean irrelevantes para la tarea de minería de datos que se desea realizar.

Una vez que se identifican los datos a utilizar, se debe entender el significado de los atributos, pudiendo detectar posibles errores en la integración, como datos repetidos o datos expresados en diferentes formatos los cuales pueden surgir al recolectarse de datos de fuentes distintas. Se debe tener en cuenta que puede haber datos anómalos, los cuales pueden representar errores en los datos, o bien, pueden ser valores correctos que son

simplemente diferentes a los demás. Algunos algoritmos de minería de datos ignoran estos datos, pero en ocasiones no es muy recomendable eliminarlos ya que dependiendo el contexto puede que estos sean relevantes para el análisis.

Ahora bien, analizado el tipo de problema y los datos disponibles se selecciona el algoritmo o los algoritmos que se van a emplear. Los datos deben ser ajustados al formato de entrada que requiere cada algoritmo. Con esta fase se consigue tener un conjunto de datos adecuados para la aplicación de las siguientes fases.

#### 2.1.1.3. Minería de datos.

El objetivo de esta fase es producir nuevo conocimiento que pueda utilizar el usuario. Se realiza construyendo un modelo sobre los datos preparados y transformados en las fases anteriores. El modelo es una descripción de patrones y relaciones entre los datos que se usan para hacer predicciones. En esta fase de deben tener algunos puntos claro sobre el problema a abordar, enfocándolo correctamente sobre las tareas de la minería de datos como, por ejemplo, si es un problema de clasificación o regresión y así poder determinar la técnica a utilizar con respecto al tipo de tarea, por citar un ejemplo dentro de las tareas de clasificación existen la técnica de árboles de decisión.

#### 2.1.1.4. Evaluación e interpretación.

Medir la calidad de los patrones descubiertos por los algoritmos de minería de datos no es un problema trivial, se interpretan y se evalúan los resultados obtenidos en la etapa anterior, utilizando técnicas que permitan un mejor análisis. Los patrones descubiertos deben tener cualidades: ser precisos, comprensibles e interesantes, si no cumplen con las expectativas se debe aplicar nuevamente los algoritmos con cambios de parámetros o ejecutar nuevos algoritmos a fin de obtener resultados más deseables. Todo esto hace que el proceso de minería de datos sea un proceso interactivo.

#### 2.1.1.5. Difusión y uso.

Construido y validado el modelo puede usarse para que un analista recomiende acciones basándose en los resultados generados por el modelo, o proceder a integrarse con otras aplicaciones que ayuden en la toma de decisiones, como en el caso del correo electrónico determinar si un correo es spam, o asistir en las predicciones del clima, la aplicabilidad del KDD es muy amplia y se puede enfocar hacia diversos contextos.

### 2.2. Tareas de la minería de datos.

Se distinguen varios tipos de tareas cada una de las cuales se puede considerar como un problema a ser resuelto por un algoritmo de la minería de datos. Cada tarea va a tener su

propio enfoque y requisitos que necesitan cada tarea. En este apartado vamos a diferencias las tareas de la minería de datos en predictivas y descriptivas y explicaremos brevemente algunas de ellas.

### 2.2.1. Predictivas

También conocidos como aprendizaje supervisado se basan en entrenar a un modelo o método por medio de diferentes datos para poder predecir una variable partiendo de estos mismos datos.

#### 2.2.1.1. Clasificación

Es quizás la más utilizada de las tareas [8] tiene por objetivos asignar objetos a uno o varios grupos definidos denominados clase de la instancia. [13] El proceso general para generar un modelo de clasificación se resume en la Figura 4.

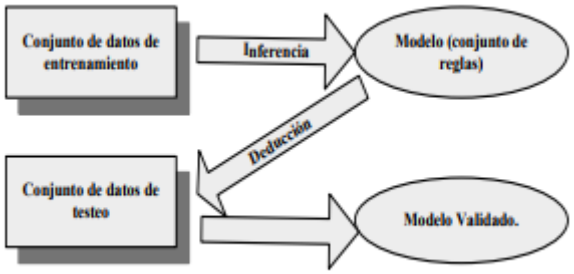



*Figura 4 Etapas para la generación de un modelo de clasificación [13]*

Los problemas de clasificación parten de un conjunto de datos la cual tiene un conjunto de características y conocemos la clase a la cual pertenece llamándose a este conjunto de entrenamiento o aprendizaje, creando un conjunto de reglas el cual nos permiten validar con un conjunto de datos diferente, permitiendo estimar la precisión del modelo de clasificación [14].

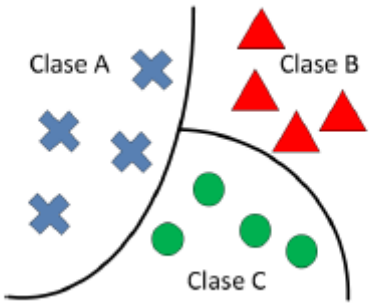



*Figura 5 Ejemplo de Clasificación.*

Hay varias técnicas que se pueden utilizar para la clasificación, a continuación, mencionamos algunas de ellas:

- Árboles de decisión
- Regresión Logística
- Máquinas de Soporte Vectorial (SVM)
- Clasificadores Bayesianos

#### 2.2.1.2. Regresión

El objetivo de las tareas de regresión es la predicción de un número, construyendo un modelo basado en uno o más predictores (variables numéricas y categóricas), que comienza con un conjunto de datos en el que se conocen los valores objetivos. En el proceso de construcción de modelos, un algoritmo de regresión estima el valor del objetivo en función de los predictores para cada caso de los datos de compilación. Estas relaciones entre predictores y objetivo se resumen en un modelo, que puede aplicarse a un conjunto de datos diferente en el que los valores objetivo son desconocidos [15].

El modelado de regresión tiene muchas aplicaciones en análisis de tendencias, planificación de negocios, marketing, pronóstico financiero, predicción de series de tiempo, modelado de respuesta biomédica y de drogas y modelado ambiental.

El análisis de regresión busca determinar los valores de parámetros para una función que hacen que la función se ajuste mejor a un conjunto de observaciones de datos. La siguiente ecuación expresa estas relaciones en símbolos. Muestra que la regresión es el proceso de estimar el valor de un objetivo continuo ( y ) como una función ( F ) de uno o más predictores ( x 1 , x 2 , ..., x n ), un conjunto de parámetros ($\theta$ 1 , $\theta$ 2 , ..., $\theta$ n ), y una medida del error ( e ).

$$y = F\,(x\,, \theta) + e$$

### 2.2.1. Descriptivas.

También conocidos como aprendizaje no supervisado permiten formar grupos de datos rápidamente, las observaciones son clasificadas en grupos que no son conocidas con anterioridad. Los elementos de las variables pueden estar conectados entre sí, de acuerdo con vínculos desconocidos de antemano.

#### 2.2.1.1. Agrupamiento (Clustering)

Se caracteriza por descubrir grupos y la identificación de distribuciones similares. es el proceso de dividir un conjunto de datos en grupos de tal manera que los miembros de cada grupo sean lo más similares posible entre sí y los diferentes grupos sean tan

diferentes como sea posible. El objetivo final es fragmentar el conjunto de elementos dados en regiones homogéneas. El agrupamiento puede descubrir relaciones previamente no detectadas en un conjunto de datos siendo aplicable a varios contextos en los negocios, deportes, literatura, etc. [16].

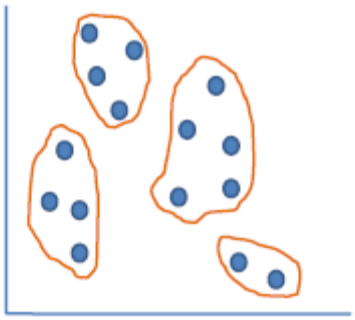


*Figura 6 Ejemplo de Agrupamiento*

Una cuestión importante en el agrupamiento es cómo determinar la similitud entre dos objetos, de modo que los conglomerados se pueden formar a partir de objetos con alta similitud dentro de los clústeres y baja similitud entre los clústeres. Comúnmente, para medir similitud o disimilitud entre objetos, se utiliza una medida de distancia como Euclidean, Manhattan y Minkowski. Una función de distancia devuelve un valor inferior para pares de objetos que son más similares entre sí.

Clustering también puede usarse para detección de anomalías. Una vez que los datos se han segmentado en grupos, es posible que algunos casos no encajan bien en los clústeres. Estos casos son anomalías o valores atípicos.

#### 2.2.1.2. Reglas de asociación

Las reglas de asociación encuentran todos los conjuntos de elementos que tienen soporte mayor que el soporte mínimo y luego utilizan los conjuntos de elementos grandes para generar las reglas deseadas que tienen confianza mayor que la confianza mínima. La elevación de una regla es la relación entre el soporte observado y el esperado si X e Y fueran independientes. Un ejemplo típico y ampliamente utilizado de aplicación de reglas de asociación es el análisis de la cesta de mercado.


| Rule | Support | Confidence | Lift |
|---|---|---|---|
| $A \Rightarrow D$ | 2/5 | 2/3 | 10/9 |
| $C \Rightarrow A$ | 2/5 | 2/4 | 5/6 |
| $A \Rightarrow C$ | 2/5 | 2/3 | 5/6 |
| $B \& C \Rightarrow D$ | 1/5 | 1/3 | 5/9 |

*Figura 7 Ejemplo de regla de asociación*

## 2.3. Técnicas de la minería de datos.

En la sección 2.2 hemos realizado una breve descripción sobre las tareas de la minería de datos, en el presente trabajo abordamos temas de clasificación que pertenecen a las tareas predictivas, para los cuales existen varias técnicas que permiten resolver problemas de este tipo, para la ejecución del experimento a realizar utilizaremos las técnicas de Árboles de Decisión, Naive Bayes y Random Forest.

Este último es un algoritmo que pertenecen a los llamados combinadores de clasificadores o ensambles [17], se desarrollan bajo el principio de generar más de una modelo, variando el conjunto de datos o la técnica empleada para entrenar cada uno de los clasificadores, posteriormente proceden a combinar sus predicciones lo que permite que pueda dar mejores resultados que cada uno de los modelos utilizados individualmente. A continuación, se realiza una breve descripción de cada uno de ellos.

### 2.3.1. Árboles de Decisión

Los árboles de decisión son una serie de decisiones o condiciones organizadas de forma jerárquica, a modo de árbol, en donde a los nodos terminal se les llaman hojas y a cada nodo no terminal del árbol se asocia un atributo y este a su vez a una condición, que determina cuáles datos de la muestra entran en esa rama [18], de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta algunas de sus hojas permitiendo una fácil interpretación [9].

Los árboles de decisión que se usan para predecir variables categóricas se llaman árboles de clasificación, mientras que los árboles de decisión que se utilizan para predecir variables continuas se llaman árboles de regresión [19].


*Figura 8 Representación de un árbol de decisión.*

El algoritmo básico para la construcción de árboles de decisión llamado ID3 por JR Quinlan que emplea una búsqueda de arriba hacia abajo, codiciosa a través del espacio de posibles ramas sin retroceso [16]. ID3 utiliza Entropy y Gain de información para construir un árbol de decisión.

**Entropía.**

Un árbol de decisión se construye de arriba hacia abajo desde un nodo raíz e implica particionar los datos en subconjuntos que contienen instancias con valores similares (homogéneos). Permite calcular la homogeneidad de una muestra. Si la muestra es completamente homogénea la entropía es cero y si la muestra es una división igual tiene entropía de uno.

$$\text{Entropy (D)} = \Sigma i \sum i = 1..c \ (-p_i \,.\, log_2 \, p_i)$$

$$p_i = \frac{|D_i|}{|D|}$$

**Ganancia de información**

La ganancia de información se basa en la disminución de entropía después de dividir un conjunto de datos en un atributo. La construcción de un árbol de decisión tiene que ver con encontrar un atributo que devuelva la ganancia de información más alta (es decir, las ramas más homogéneas).

$$\text{Gain (D, A)} = \text{Entropy (D)} - \sum a \in \text{V(A)} p_a \,.\, \text{Entropy } (D_a)$$

$$p_a = \frac{|D_a|}{|D|}$$

$$D_a = \{d \in \mathrm{D} \mid A(d) = a$$

### 2.3.2. Random Forest.

Random forest es una técnica que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución [20].

La fase de aprendizaje consiste en crear muchos árboles de decisión independientes, construyéndolos a partir de datos de entrada ligeramente distintos. Se altera, por tanto, el conjunto inicial de partida, haciendo lo siguiente:

- Se selecciona aleatoriamente con reemplazamiento un porcentaje de datos de la muestra total.

Es habitual incluir un segundo nivel aleatoriedad, esta vez afectando los atributos:

- En cada nodo, al seleccionar la partición óptima, se tiene en cuenta sólo una porción de los atributos, elegidos al azar en cada ocasión.

Una vez que se generan muchos árboles, la fase de clasificación se lleva a cabo de la siguiente forma:

- Cada árbol se evalúa de forma independiente y la predicción del bosque será la media de todos sus árboles en caso de que sea un problema de regresión, cuando se trate de un problema de clasificación realizara un voto mayoritario sobre todos los arboles del bosque es decir la clase con mayor voto [21]. En la Figura 9 podremos ver de manera conceptual la descripción del modelo.

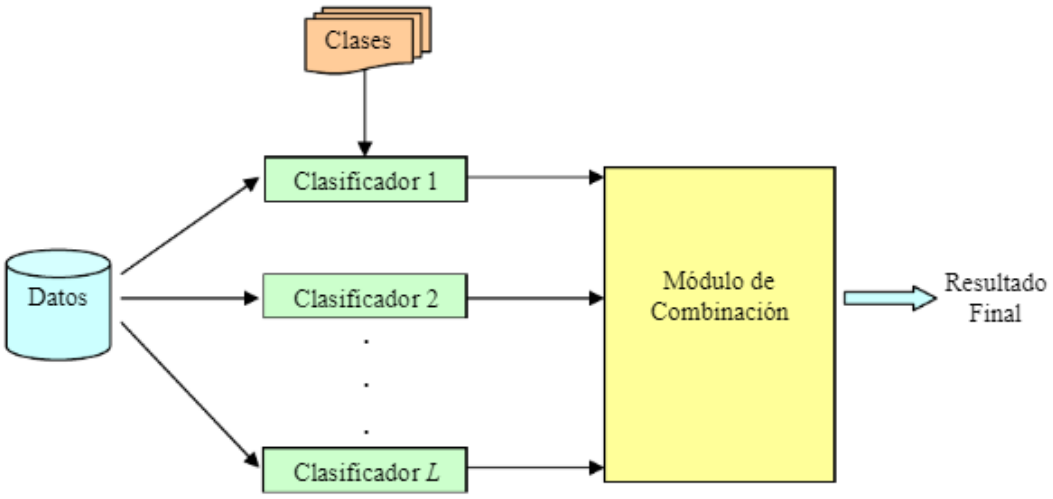



*Figura 9 Representación conceptual del modelo Random Forest [17].*

### 2.3.3. Naive Bayes.

Los clasificadores bayesianos son clasificadores estadísticos que se basa en el teorema de Bayes. Suponen que el efecto de un valor de atributo en una clase dada es independiente de los valores de los otros atributos [22]. Esta suposición se denomina independencia condicional de clase. Está hecho para simplificar la computación involucrada y, en este sentido, se considera "ingenuo" [23].

El teorema de Bayes muestra la relación entre una probabilidad condicional y su inversa. Proporciona un método para calcular P (Y | X) usando P (X), P (Y) y P (X | Y), que se define [24]

$$P(c \mid x) = \frac{P(x \mid c)\, P(c)}{P(x)}$$

El teorema de Bayes proporciona una forma de calcular la probabilidad posterior, $P(c|x)$ , de $P(c)$ , $P(x)$ y $P(x \mid c)$ . El clasificador asume que el efecto del valor de un predictor ( $x$ ) en una clase dada ( $c$ ) es independiente de los valores de otros predictores. Esta suposición se llama independencia condicional de clase.

- P (c | x) es la probabilidad posterior de la clase (objetivo) dado predictor (atributo).
- P (c) es la probabilidad previa de clase o probabilidad marginal, es "anterior" en el sentido de que no tiene en cuenta ninguna información sobre x.
- P (x | c) es la probabilidad condicional de x dado c.
- P (x) es la probabilidad previa de predictor.

La probabilidad posterior se puede calcular primero, construyendo una tabla de frecuencia para cada atributo contra el objetivo. Luego, transformar las tablas de frecuencia a tablas de verosimilitud y finalmente utilizar la ecuación Bayesiana Naive para calcular la probabilidad posterior para cada clase. La clase con mayor probabilidad posterior es el resultado de la predicción [16].

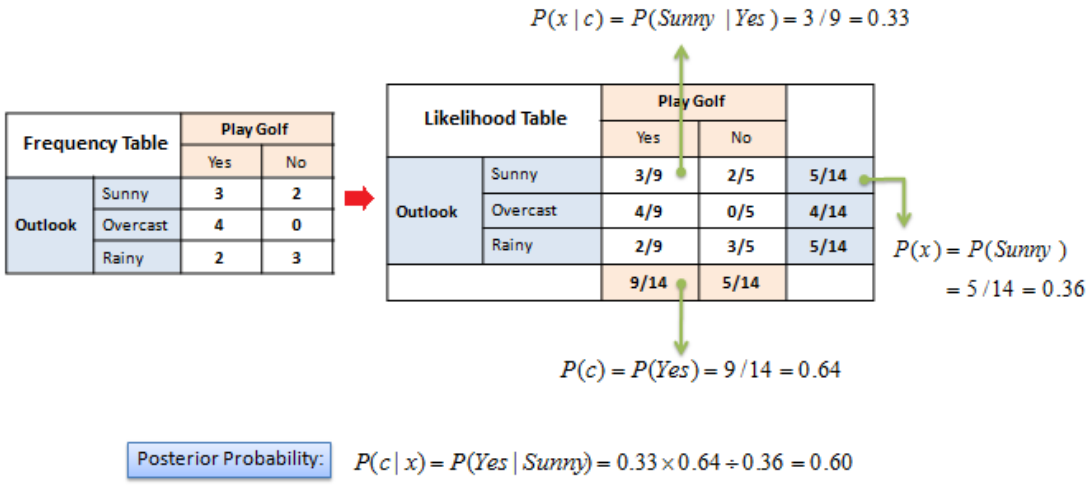



*Figura 10 Aplicación del teorema de Bayes [16].*

# 3. Aproximación

**Resumen**: En este capítulo se describe el proceso propuesto que se va a realizar para la aplicación de modelos de clasificación a datos inciertos. Para lo cual se han definidos tres fases con la aplicación de dos aproximaciones. Además, se presenta un procedimiento general que permita generar incertidumbre en atributos numéricos, que nos servirá para evaluar experimentalmente las aproximaciones presentadas.

Una vez que se ha revisado los conceptos necesarios para enmarcar el presente trabajo, el cual consiste en abordar problemas de clasificación, donde los modelo han sido aprendidos con datos perfectos, pero que al momento de realizar su aplicación los datos no se encuentran bajo las mismas condiciones, presentando incertidumbre en datos numéricos, representada a través de intervalos, por lo que no son capaces de poder ejecutarse debido a que esta incertidumbre no fue parte de su aprendizaje. Se han definido tres fases para la ejecución del presente problema y se definen en la Figura 11.


*Figura 11 Fases del proceso de experimentación*

Haciendo un resumen de cada una de las etapas, consisten en generar N instancias a partir de la instancia con incertidumbre y se obtendrán a base de cubrir valores en el intervalo. Luego se aplicará a esas N instancias el clasificador, y a continuación se combinarán las predicciones para poder dar una única predicción a la instancia.

La Figura 12 hace referencia al problema descrito donde se puede observar que el atributo “Monto crédito”, presenta incertidumbre expresado a través de intervalos, con este ejemplo se va a realizar de manera detallada la explicación de cada una de las etapas.

| Peso | Sexo | Edad | Monto Crédito |
|---|---|---|---|
| 185 | M | 25 | 1500 – 2000 |
| 150 | F | 23 | 500 – 1000 |
| 168 | M | 30 | 1000 – 1500 |

| CREDITO (CLASE) |
|---|
| 1 |
| 0 |

*Figura 12 Ejemplo del problema de clasificación a abordar.*

## 3.1. Generación de instancias.

El objetivo de esta fase consiste en generar N valores de manera uniforme en un atributo numérico expresado en intervalos, se deben tomar en cuenta los valores extremos de cada intervalo, ya que son los datos que se tiene certeza que se conocen y deben formar parte ellos. N estará representado para la aplicación del experimento por 5,7,9,11,13 y 15 para la realización de la explicación se toma el valor de N=5.

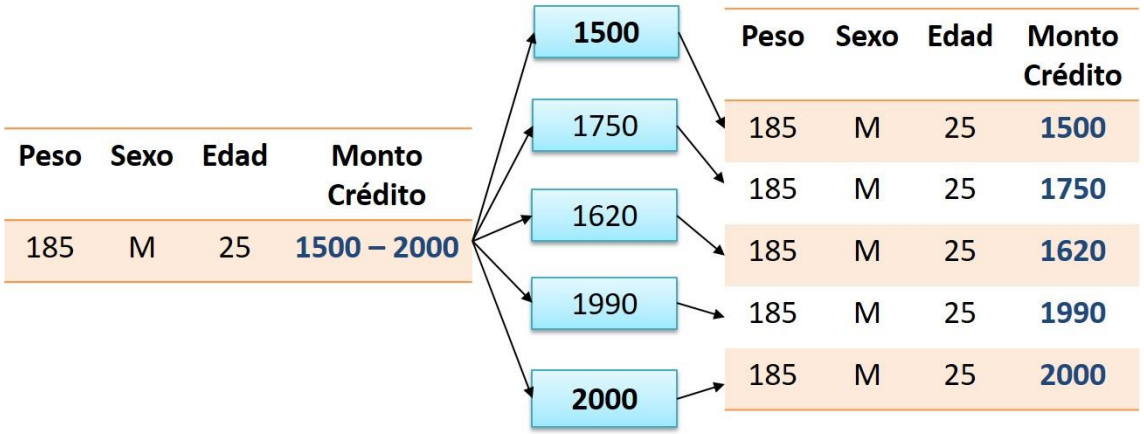



*Figura 13 Generación de N instancias.*

Como se puede observar en la Figura 13 para la instancia que presenta incertidumbre en el atributo "Monto Crédito" se han generado 5 instancias en base a los valores generados dentro del intervalo, ahora que el atributo numérico se encuentra en condiciones que el modelo sea capaz de poder ejecutarse, a continuación, se detalla la aplicación de los clasificadores.

## 3.2. Aplicación del Clasificador.

Para realizar la aplicación de los clasificadores una vez que se han generado las N instancias se debe tener en cuenta que el valor que tomará el atributo del "Monto crédito", será cada uno de los valores generados dentro del intervalo. Se han aplicado dos tipos de clasificadores para la predicción de las instancias.

### 3.2.1. Aplicación de clasificadores “Crisp”

Este tipo de clasificadores al ejecutarse da como resultado una etiqueta de clase [25]. La figura 14 muestra que el clasificador ha sido ejecutado para cada instancia generada en la fase 1.

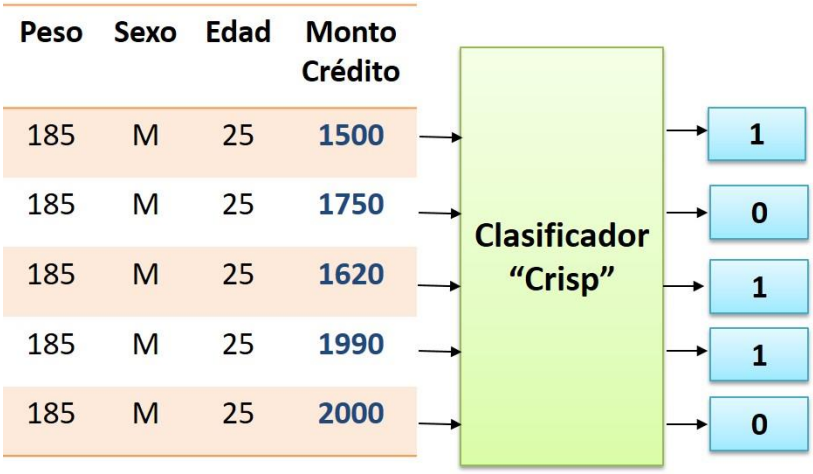



*Figura 14 Aplicación clasificadores "Crisp"*

### 3.2.1. Aplicación de clasificadores “Soft”

Los clasificadores “Soft” al ejecutarse se obtiene como resultado la probabilidad de pertenencia de cada clase [25]. La figura 15 muestra un ejemplo de los resultados que se obtienen al aplicar este tipo de clasificadores.

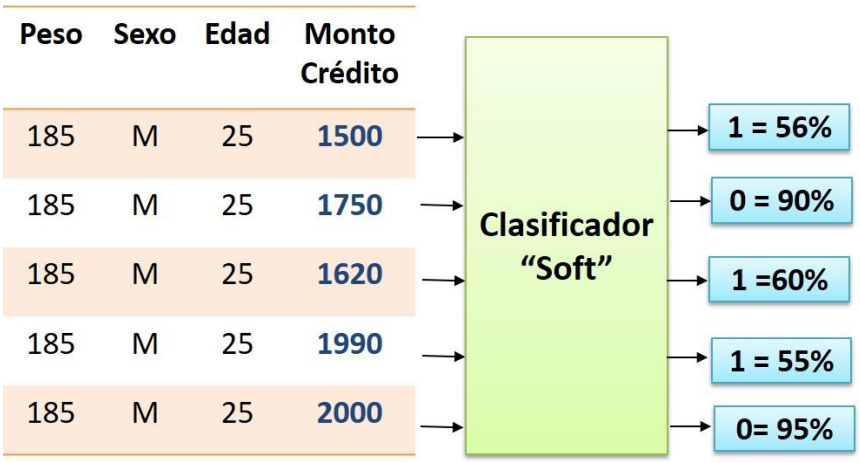



*Figura 15 Aplicación clasificadores "Soft"*

## 3.3. Combinación de predicciones.

Para realizar las combinaciones de las predicciones obtenidas por los clasificadores "Crips" y "Soft" se define un método para cada una de ellas y se muestra a continuación su manera de proceder.

### 3.3.1. Aproximación de voto mayoritario.

Para realizar la combinación de los resultados obtenidos por los clasificadores "Crisp" usaremos la aproximación de voto mayoritario, la cual consiste en tomar como predicción de la instancia la que mayor vez se repita para cada clase, la Figura 16 muestra cómo se realiza la predicción en base a los resultados obtenidos. Se puede ver que la clase "1" tiene 3 predicciones y la clase "0" tiene 2 predicciones, tomando como predicción final la clase 1.

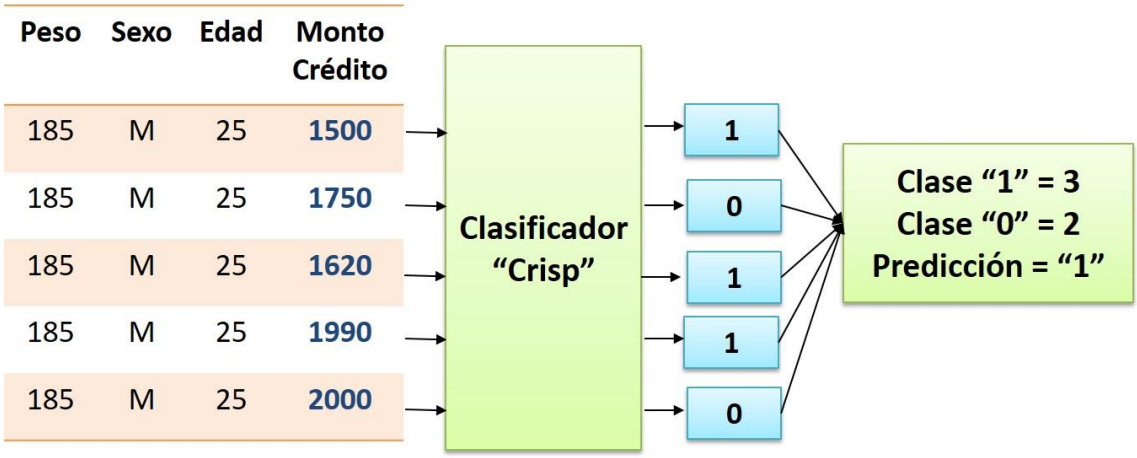



*Figura 16 Predicción por voto mayoritario*

### 3.3.2. Aproximación de probabilidad de pertenencia.

La siguiente aproximación que permite obtener una predicción, es la probabilidad de pertenencia de la clase, que procede a combinar los resultados obtenidos por los clasificadores "Soft". En la cual se realiza la suma de probabilidad de cada clase y se obtiene como predicción la que mayor sea. En la Figura 17 se puede ver de que manera se realiza su predicción, la suma de sus probabilidades de la clase "1" es 175 mientras que la suma de la clase "0" es 185 con esto la predicción de la instancia será la clase "0".

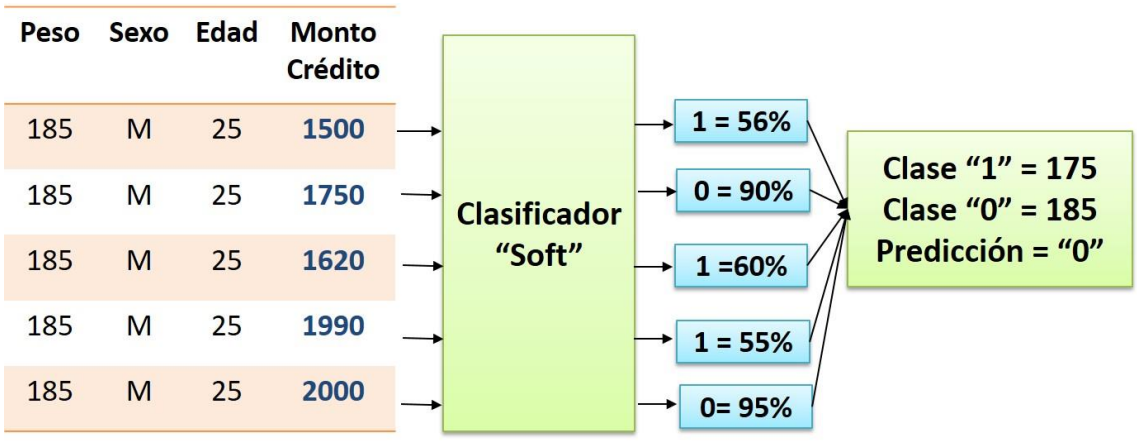



*Figura 17 Predicción por probabilidad de pertenencia*

La diferencia en los resultados entre las aproximaciones puede generarse por motivo de que, si el clasificador predice 3 instancias de la clase "1" y 2 instancias de la clase "0", por voto mayoritario se definirá la clase "1", pero puede darse el caso que en la probabilidad de pertenencia la suma de probabilidades de las 2 instancias sea mayor que la suma de probabilidades de las 3 instancias clasificadas tomando como predicción la clase "0", como se puede observar en los ejemplos ilustrados en las Figuras 16 y 17, a pesar de tener el mismo número clases predichas obtienen como predicción final diferentes clases.

## 3.4. Generación de Incertidumbre

Para poder evaluar experimentalmente las aproximaciones propuestas se necesita disponer de datos de test que presenten incertidumbre en atributos numéricos expresados en intervalos, se procederá a generar esta incertidumbre de manera artificial. Para efectos de explicación el atributo a considerar para generar incertidumbre es "Monto del crédito". En las Figura 18 se puede ver la descripción de los conjuntos de datos que servirán de base para la aplicación, es importante comentar que se hicieron varias revisiones para generar conjunto de datos con incertidumbre. Las cuales se describen a continuación.

| Peso | Sexo | Edad | Monto del crédito | Credibilidad (Clase) |
|---|---|---|---|---|
| **185** | M | 45 | 1049 | 1 |
| **150** | M | 50 | 500 | 1 |
| **168** | F | 33 | 841 | 0 |
| **80** | M | 25 | 2122 | 0 |
| **95** | F | 18 | 2171 | 1 |
| **60** | F | 38 | 8534 | 1 |
| **68** | F | 25 | 3398 | 0 |
| **78** | M | 31 | 1361 | 1 |
| **102** | M | 28 | 1098 | 0 |
| 77 | M | 29 | 10010 | 1 |

Conjunto de Entrenamiento

Conjunto de Validación

| Peso | Sexo | Edad | Monto del crédito |
|---|---|---|---|
| **1** | 18 | 55 | 3040 |
| **2** | 6 | 35 | 700 |
| **4** | 24 | 18 | 6780 |
| **1** | 3 | 25 | 300 |

*Figura 18 Conjunto de datos de entrenamiento y validación.*

### 3.4.1. Generación de incertidumbre con K-means

K-means es un algoritmo que permite agrupar las instancia de un conjunto datos en K grupos de acuerdo con la relación entre ellos [26]. Para la ejecución del modelo no se han tenido en cuenta la clase y el atributo numérico donde se generará incertidumbre, donde el número de grupos K es determinado por el usuario en este caso se realiza el reemplazo de K = 10 y K = 20.

| Peso | Sexo | Edad |
|---|---|---|
| **185** | M | 25 |
| **150** | F | 23 |
| **168** | M | 30 |
| **Peso** | Sexo | Edad |

Datos para generar K-means

| Peso | Sexo | Edad | Grupo K-means | Monto del crédito | Intervalo Generado |
|---|---|---|---|---|---|
| **185** | M | 25 | 1 | 3040 | 300-3040 |
| **150** | F | 23 | 2 | 700 | 700 - 6780 |
| **168** | M | 30 | 2 | 6780 | 700 - 6780 |
| **Peso** | Sexo | Edad | 1 | 300 | 300-3040 |

Matriz con grupos generados

*Figura 19 Proceso K-means*

Con este proceso cada instancia del conjunto de validación pertenece a uno de los grupos generados (Grupo K-means), ahora se debe tomar el valor mínimo y el valor máximo del atributo numérico que no se utilizó para realizar la agrupación (Monto del crédito) de cada grupo, como se puede ver en la Figura 19, el grupo 1 tiene 2 instancias que pertenecen a él, tomando su valor máximo y valor mínimo el intervalo quedaría definido por (300 - 3040), el cual se debe asignar a todas las instancias al mismo grupo, y realizar este proceso para los demás grupos que se generen.

Al realizar este proceso se presentaron algunos inconvenientes, los grupos generados tenían números variados de instancias pertenecientes a ellos. Como se puede observar en las Figuras 20 y 21, la cual representan la relación entre el valor máximo y valor mínimo de cada grupo aplicado para K = 10 y K = 20 respectivamente, no muestran valores uniformes en la amplitud del intervalo. Para la aplicación de la aproximación presentada, la realización de este proceso no será de gran ayuda porque al tener distintos rangos en la amplitud del intervalo no se podría controlar la incertidumbre que estos presentan, por lo que no se hará uso de esta técnica.

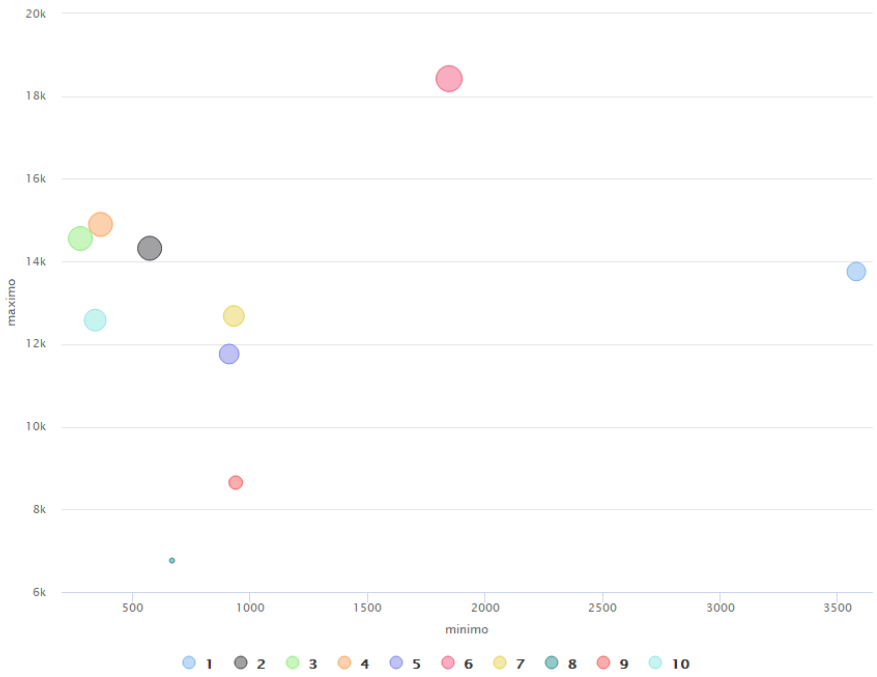



*Figura 20 Agrupación de instancias K = 10*

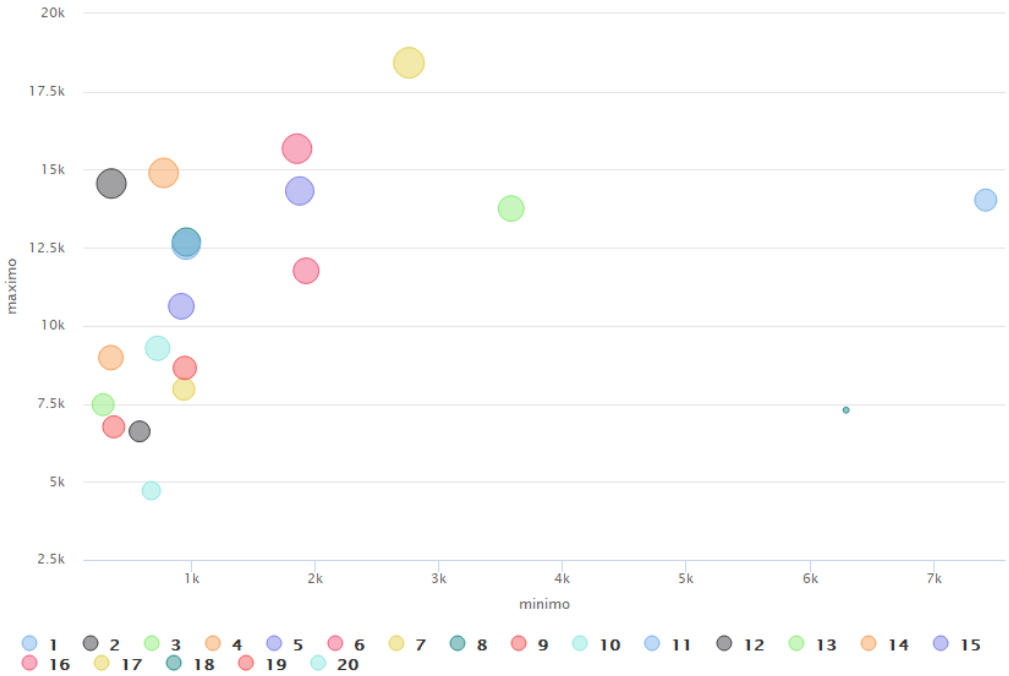



*Figura 21 Agrupación de instancias K = 20*

La Figuras 20, 21 fueron generadas con el conjunto de datos (German Credit) las cuales, muestran la información que se generó en el momento de la validación del proceso.

### 3.4.1. Generación de incertidumbre por cálculo de la media

Descartado el proceso de K-means para su aplicación, se realiza el siguiente proceso, el cual consiste en tomar el atributo numérico del conjunto de datos de entrenamiento en que se va a generar incertidumbre, y calcular la media de sus valores. Se emplean porcentajes I de la media calculada y dicho valor se sumará y restará al valor conocido que tenemos en el conjunto de validación, asegurándonos que no se generen valores negativos, con esto conseguimos representar la incertidumbre a base de intervalos, donde I se debe ir incrementando para ir teniendo un mayor grado de incertidumbre presente. Para efectos del experimento I varia en un 10%, 20%, 30%. A continuación, se muestra una descripción grafica del proceso realizado.

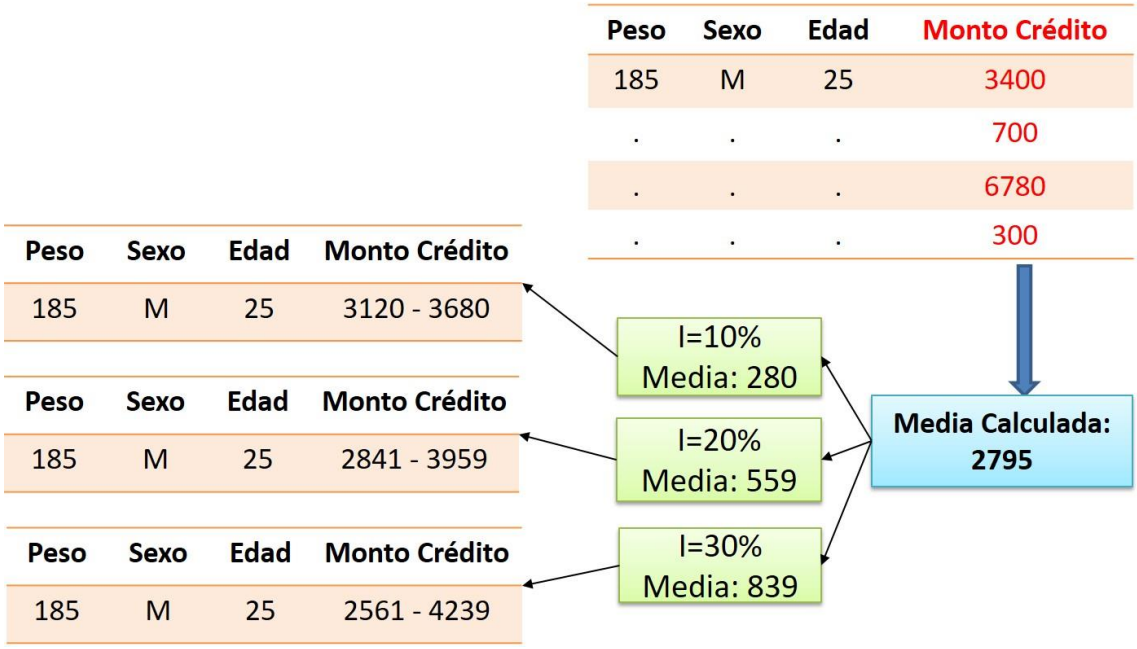



*Figura 22 Generación de incertidumbre con cálculo de la media*

Como se puede observar en la Figura 22 por cada instancia del conjunto de validación se han creado niveles de incertidumbre en el atributo numérico "Monto del crédito", siguiendo con el ejemplo se tiene en la primera instancia, donde el valor del atributo es 3400 se procede a generar los intervalos en base al porcentaje I de la media, cuando I = 10% el intervalo queda definido por (3120 - 3680 ), I = 20% (2841 – 3959) e I = 30% (2561 - 4239) repitiendo este proceso para todas las instancias del conjunto de validación. Con este proceso el atributo "Monto del crédito" presenta intervalos con rangos de amplitud uniformes, y será la manera de proceder para generar incertidumbre en los diferentes conjuntos de datos que se apliquen para el experimento.

Ahora bien, una vez que se ha definido la manera de generar incertidumbre en el atributo numérico, y como se va a tratar en base a las fases descritas, en el siguiente capítulo se explica la manera de realizar la experimentación de los diferentes escenarios con respecto a las variaciones de I y de N.

# 4. Experimentos

**Resumen.** Este capítulo tiene por objetivo llevar a cabo diferentes pruebas mediante el análisis experimental del proceso planteado en el capítulo 3. Se definen las técnicas de minería de datos a utilizar para la experimentación junto con los dataset y los recursos computaciones usados. Finalmente se muestran y analizan los resultados obtenidos.

## 4.1. Metodología.

En el capítulo 3 se describe la manera de proceder para la aplicación de los modelos. La Figura 23 muestra un esquema referente al proceso realizado.

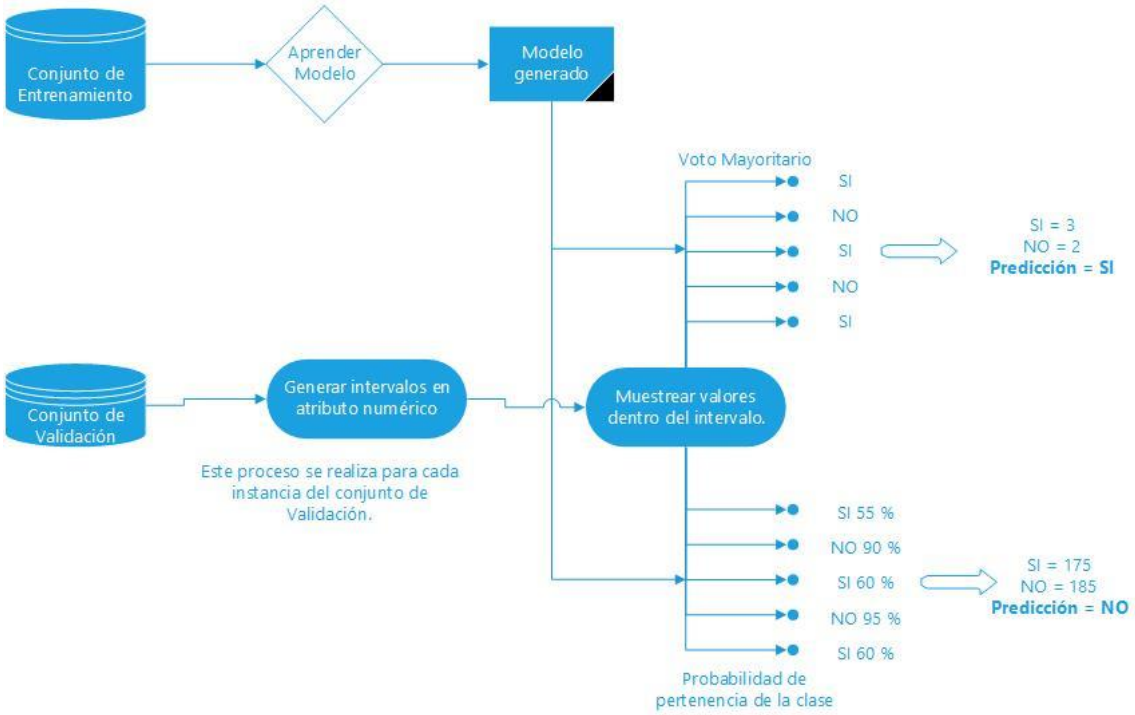



*Figura 23 Descripción de la metodología a seguir para aplicar los modelos de clasificación*

### 4.1.1. Modelos de clasificación

Para la ejecución de la metodología se aplican las siguientes técnicas de clasificación: Arboles de decisión (AD), Naive Bayes (NB) y Random Forest (RF) de las cuales ya se hizo una breve descripción en la sección 2.3.

### 4.1.2. Medida de evaluación.

Para determinar la precisión de los clasificadores se ha utilizado la métrica de Accurary (exactitud), la cual hace uso de la matriz de confusión que es una tabla de orden 2 donde se muestra el número de predicciones correctas e incorrectas hechas por el modelo en comparación con las clasificaciones reales en los datos de prueba, donde, las filas representan las clases actuales de una instancia y las columnas las clases predichas o estimadas.

| | | Predicción | |
|---|---|---|---|
| | | $C_P$ | $C_N$ |
| Clase real | $C_P$ | TP: True positive | FN: False negative |
| | $C_N$ | FP: False positive | TN: True negative |

*Figura 24 Matriz de Confusión [13]*

$$accuracy = (TP + TN)/(TP + TN + FP + FN)$$

### 4.1.3. Escenarios de experimentación

Se procede a explicar la manera de realizar la experimentación en base a la variación de los parámetros I y de N, donde I se define como el porcentaje que se aplica a la media calculada del conjunto de entrenamiento para generar incertidumbre en el conjunto de validación, y N son los valores que se deben generar en los intervalos.

Definido I = (10%, 20% y 30%) y N = (5,7,9,11,13 y 15), para cada valor de I se generan todos los valores de N, cada combinación se aplica a todo el conjunto de validación, obteniendo un accuracy por cada uno de ella y para cada aproximación. El modelo también se ejecutará para cuando se conoce su valor y cuando se tiene incertidumbre máxima, a continuación, se presentada cada escenario de manera visual.

**Valores conocidos (Incertidumbre cero)**

Este proceso sería un proceso normal de ejecución donde sus datos de entrenamiento y de test se encuentran bajo las mismas condiciones, es decir que el atributo numérico no presenta incertidumbre y se conoce su valor. Siendo este el primer resultado de accuracy en el experimento.

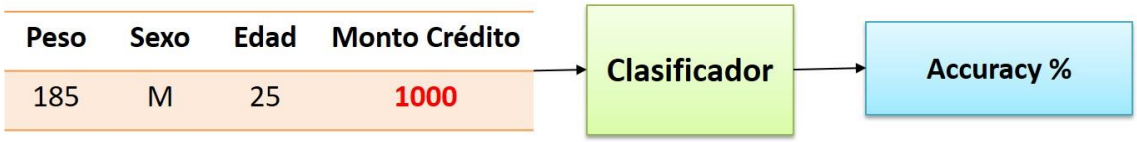



*Figura 25 Valores sin incertidumbre*

### Rangos de Incertidumbre

En este proceso como vemos el atributo numérico muestra grados de incertidumbre en base al porcentaje I que se establecieron. Es aquí donde se aplicarán las fases presentadas en el capítulo 3. Donde ira variando el grado de incertidumbre en el atributo y procediendo a muestrear N valores dentro de ellos, y se obtendrá un valor de accuracy por cada combinación de parámetro. En la Figura 26 se puede observar la combinación de los parámetros establecidos.

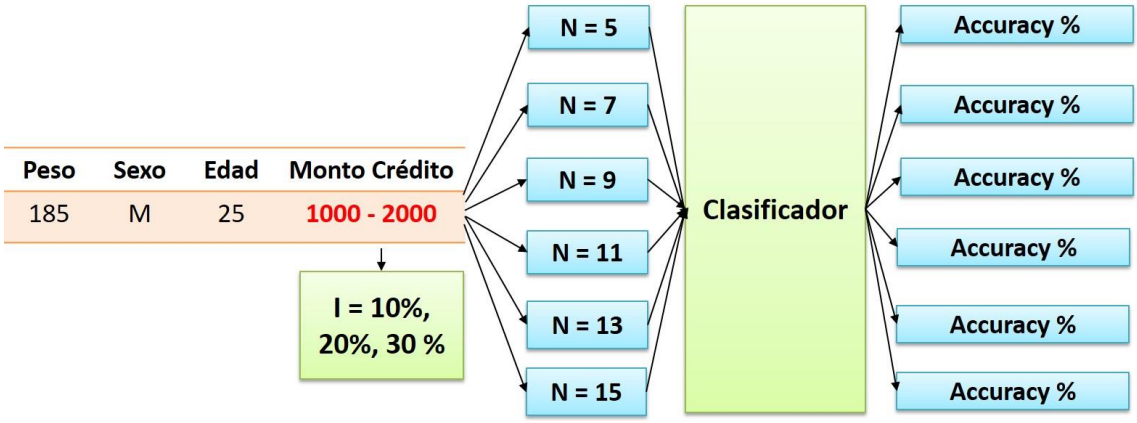



*Figura 26 Rangos de Incertidumbre*

### Incertidumbre máxima

Este escenario se presenta cuando no se conoce el valor del atributo que vendrá a representar cuando se tiene incertidumbre máxima.

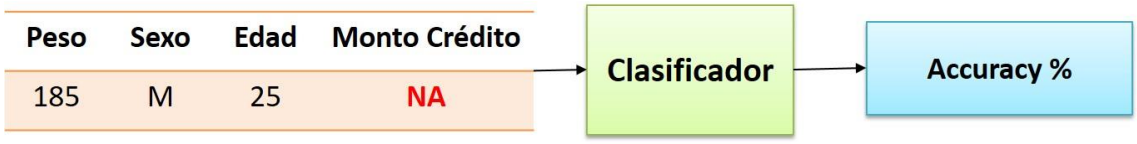



*Figura 27 Incertidumbre máxima*

Existen tres escenarios establecidos para la ejecución del experimento presentado, donde el modelo se ejecutará cuando no se tiene incertidumbre presente en sus datos; es decir que se conoce su valor, cuando se tiene un cierto grado de incertidumbre, teniendo un

conocimiento aproximado del mismo, y finalmente cuando no conocemos su valor y se tiene una incertidumbre máxima.

Para cada uno de los escenarios planteados sus ejecuciones se realizarán 20 veces, aplicando los procesos descritos anteriormente, y se calcula la media de sus resultados, evitando con esto la variabilidad, obteniendo resultados más estables para su análisis. En la figura 28 se muestran un ejemplo de los resultados obtenidos.

| Parámetro | Accuracy |
|---|---|
| I = 0 (Valor Conocido) | 84.04 |
| I = 10% N = 5 | 83.07 |
| I = 20% N = 5 | 82.95 |
| I = 30 % N = 5 | 82.55 |
| I = Maxima | 78.45 |

Probabilidad de Pertenencia

| Parámetro | Accuracy |
|---|---|
| I = 0 (Valor Conocido) | 84.04 |
| I = 10% N = 5 | 83.10 |
| I = 20% N = 5 | 82.96 |
| I = 30% N = 5 | 82.55 |
| I = Maxima | 78.45 |

Voto Mayoritario

*Figura 28 Resultado obtenido con las aproximaciones*

## 4.2. Recursos computacionales.

### 4.2.1. Características de los equipos

La descripción de las características de los equipos utilizados para realizar la ejecución de la experimentación se resume en la tabla 1.

| Equipo | Sistema Operativo | Memoria RAM | Procesador | Disco Duro |
|---|---|---|---|---|
| **Ordenador Portátil** | Windows 10 | 4 GB | Pentium Dual Core | 500 GB |
| **Máquina Virtual (DSIIC)** | Windows Server Estándar 2012 | 12 GB | Intel Xeon | 500 GB |

*Tabla 1 Equipos utilizados para la experimentación.*

### 4.2.2. Lenguaje de programación R

Es un lenguaje de programación pensado para el análisis estadístico y representación gráfica de datos, siendo un software libre que se distribuye bajo licencia GNU GPL, además de ser un conjunto de herramientas integradas.

R integra variables, datos, resultados y funciones en un área de trabajo mediante objetos que llevan un nombre definido. Su información se encuentra estructurada en paquetes y librerías.
Adicionalmente existe RStudio, que es la interfaz gráfica para desarrollo en R. RStudio incluye una consola, editor de sintaxis que permite la ejecución de código, así como herramientas para creación de gráficos, depuración y la gestión del espacio de trabajo [27]. Para el presente trabajo se han utilizado las siguientes librerías definidas en la Tabla 2.

| **Paquete** | **Descripción** |
|---|---|
| **Rpart** | Usado para aplicar el modelo de árboles de decisión. |
| **RandomForest** | Usado para aplicar el modelo Random Forest |
| **e1071 (naiveBayes)** | A través de esta librería se ha realizado la aplicación de modelo Naive Bayes. |
| **Ggplot2** | Es un sistema de trazado para R, basado en la gramática de gráficos. |
| **Runif** | Esta función genera valores de distribución uniforme en un intervalo |

*Tabla 2 Librerías utilizas en R*

Podremos encontrar en [28] información sobre el uso y la descripción de manera detallada de las librerías que se muestran en la tabla.

## 4.3. Descripción de los Datasets

Para realizar la aplicación de modelos de clasificación se han utilizado tres datasets, los cuales fueron obtenidos del repositorio de la UCI Machine Learning Repository (*University California Irvine*) y del Repositorio de Kaggle. Las características de los *datasets* se resumen en la Tabla 3 y sus descripciones completas pueden ser encontradas en [29] [30] [31].

| Descripción | German Credit | Credit Approval | IBM HR Analytics Employee Attrition & Performance |
|---|---|---|---|
| Número de Instancias | 1000 | 690 | 1470 |
| Número de atributos | 20 | 15 | 34 |
| Clases (Número de grupo) | 2 | 2 | 2 |
| Repositorio | UCI | UCI | Kaggle |
| Abreviatura | G. Credit | C. Approval | IBM HR |

*Tabla 3 Metadatos de los datasets G. Credit, C. Approval, IBM HR.*

## 4.4. Resultados

Después de haber realizado las diferentes ejecuciones para cada uno de los conjuntos de datos, se procedió a unificar sus resultados para cada técnica de clasificación empleada.

### 4.4.1. Análisis de los resultados.

Los resultados de las Figuras 29,31,33 muestran que a medida aumenta la incertidumbre en los datos, los clasificadores van perdiendo precisión, viéndose afectados en su accuracy, teniendo un comportamiento similar para las tres técnicas empleadas. También se ha podido determinar que Random Forest es el clasificador que mayor accuracy ha obtenido dentro de la fase experimental, seguido por Arboles de decisión, y por último Naive Bayes cuando se tiene un escenario normal donde los datos no presentan incertidumbre.

Las figuras 30,32,34 presentan resultados de accuracy de acuerdo con el grado de incertidumbre y los N valores generados, los resultados obtenidos han variado según la técnica empleada al muestrear diferentes N, pero la mayoría de casos las mejores predicciones de los modelos se han obtenido cuando se genera un mayor número de instancias dentro del intervalo por lo general entre 11 y 15.

La Figura 35 muestra los resultados de la degradación de la clasificación, pudiendo notar que al tener una máxima incertidumbre presente en los datos la técnica de Random Forest es la que menor degradación tiene en su precisión, también se puede ver que a un menor rango de incertidumbre por lo general Naive Bayes es una de las técnicas que menor degradación tiene con respecto a los valores de accuracy de cada modelo empleado.

Los resultados obtenidos cuando el rango de incertidumbre es mínimo, no tienen mayor perdida con respecto cuando se conoce su valor. Por lo que la aplicación del presente método puede ser de gran ayuda para enfrentar este tipo de escenario. A continuación, se realiza un análisis en base a los parámetros generados de manera visual, junto con sus descripciones.

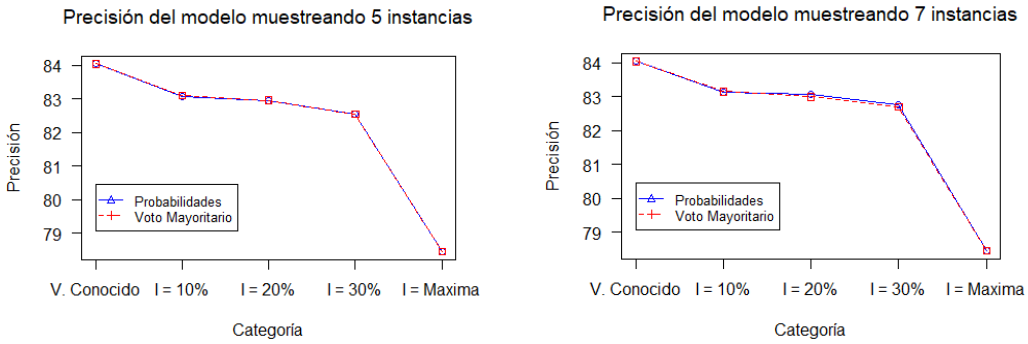
### 4.4.2. Análisis de la predicción de los modelos.

La tabla 4 muestra los resultados obtenidos de accuracy al aplicar los diferentes parámetros establecidos para cada modelo de clasificación empleado.

| Algoritmo | | Árbol de Decisión | | Random Forest | | Naive Bayes | |
|---|---|---|---|---|---|---|---|
| Método | | Probabilidad | Voto Mayoritario | Probabilidad | Voto Mayoritario | Probabilidad | Voto Mayoritario |
| **I = 0 (Valor Conocido)** | | **84.04** | **84.04** | **84.95** | **84.95** | **80.53** | **80.53** |
| **I = 10 %** | **N = 5** | 83.07 | 83.10 | 84.20 | 84.20 | 79.81 | 79.81 |
| | **N = 7** | 83.12 | 83.15 | 84.33 | 84.38 | 79.84 | 79.84 |
| | **N = 9** | 83.13 | 83.15 | 84.31 | 84.33 | 79.79 | 79.79 |
| | **N = 11** | 83.07 | 83.17 | 84.31 | 84.33 | 79.86 | 79.86 |
| | **N = 13** | 83.20 | 83.12 | 84.39 | 84.45 | 79.84 | 79.84 |
| | **N = 15** | 83.28 | 83.24 | 84.36 | 84.54 | 79.83 | 79.83 |
| **I = 20 %** | **N = 5** | 82.95 | 82.96 | 84.01 | 84.01 | 79.65 | 79.65 |
| | **N = 7** | 83.06 | 83.00 | 84.09 | 84.06 | 79.67 | 79.67 |
| | **N = 9** | 82.96 | 82.90 | 83.97 | 83.99 | 79.71 | 79.70 |
| | **N = 11** | 83.01 | 82.91 | 83.98 | 83.97 | 79.70 | 79.69 |
| | **N = 13** | 83.06 | 83.00 | 83.97 | 84.06 | 79.74 | 79.73 |
| | **N = 15** | 83.13 | 83.05 | 84.11 | 84.13 | 79.70 | 79.70 |
| **I = 30 %** | **N = 5** | 82.55 | 82.55 | 83.07 | 83.71 | 79.63 | 79.63 |
| | **N = 7** | 82.76 | 82.70 | 83.91 | 83.89 | 79.60 | 79.60 |
| | **N = 9** | 82.87 | 82.89 | 83.90 | 83.94 | 79.61 | 79.61 |
| | **N = 11** | 82.79 | 82.87 | 83.87 | 83.97 | 79.71 | 79.70 |
| | **N = 13** | 82.87 | 82.87 | 83.95 | 83.95 | 79.64 | 79.64 |
| | **N = 15** | 82.75 | 82.74 | 83.84 | 83.84 | 79.63 | 79.62 |
| **I = Máxima** | | **78.45** | **78.45** | **81.35** | **81.35** | **70.93** | **70.93** |

*Tabla 4 Resultados globales del accuracy de cada modelo de clasificación.*

La Figura 29 muestra los resultados obtenidos de accuracy del modelo de Árboles de Decisión, en la cual se muestra los resultados cuando se conoce su valor, y los diferentes rangos de incertidumbre generados hasta tener una incertidumbre máxima, para cada aproximación realizada, se puede observar de manera visual como el clasificador se ve afectado a medida aumenta el nivel de incertidumbre.



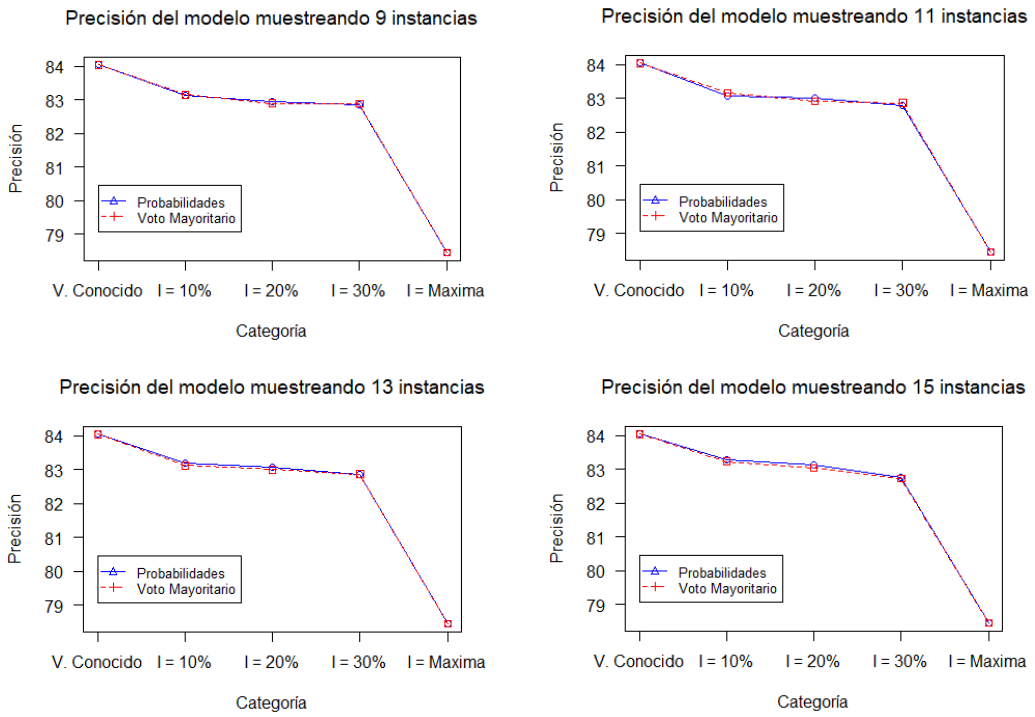
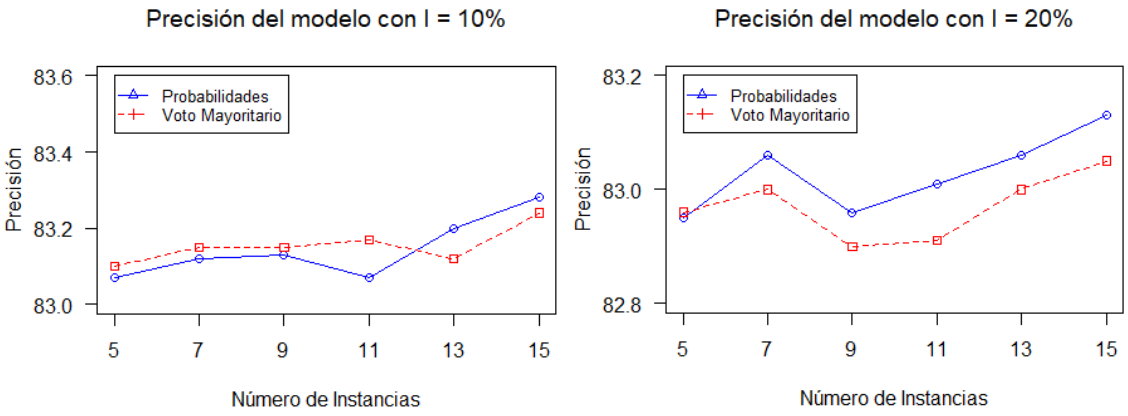



*Figura 29 Resultados globales del modelo AD muestreando N instancias*

La figura 30 muestra los resultados desde otra perspectiva del modelo de árbol de clasificación, se puede observar que cuando I=10% el punto más alto de accuracy se presenta cuando N = 15 con un 83.28% para la aproximación de probabilidad de pertenencia, y para voto mayoritario al igual cuando N = 15 con 83.24. Cuando I=20% se puede identificar que la aproximación de probabilidad de pertenencia obtiene mejores resultados con respecto al voto mayoritario teniendo su punto más alto cuando N = 15 con 83.13% y cuando I=30% los resultados se muestran más similares entre sus aproximaciones habiendo una mejoría cuando N = 9 para voto mayoritario con el 82.89%, analizando los resultados desde una visión general se puede decir que se obtiene mejores resultados con el método de probabilidades de pertenencia.



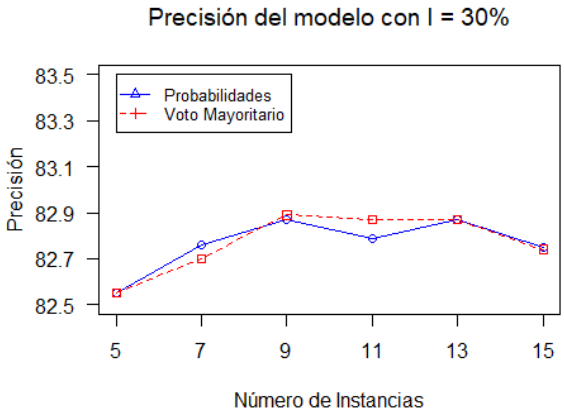
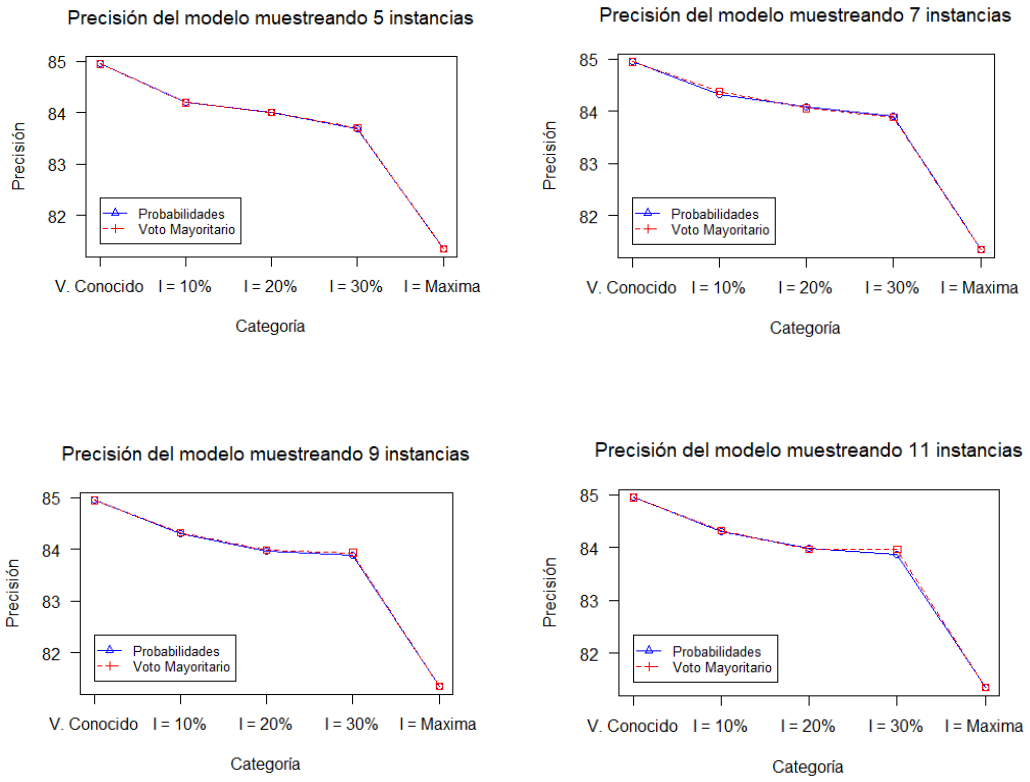



*Figura 30 Resultados globales del modelo AD variando su incertidumbre I.*

La Figura 31 muestra los resultados obtenidos de accuracy del modelo random forest, en la cual se muestra los resultados cuando se conoce su valor, y los diferentes rangos de incertidumbre generados hasta tener una incertidumbre máxima, para cada aproximación realizada, se puede observar de manera visual como el clasificador se ve afectado a medida aumenta el nivel de incertidumbre.



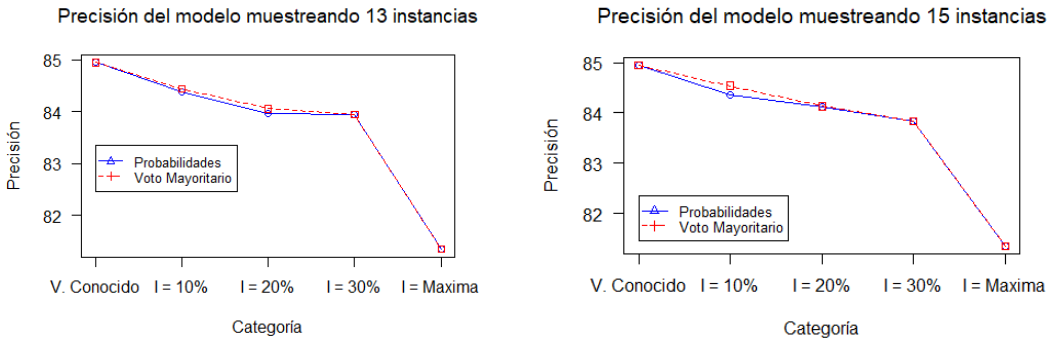



*Figura 31 Resultados globales del modelo RF muestreando N instancias.*

La figura 32 muestra los resultados desde otra perspectiva del modelo random forest, donde se puede observar que cuando I=10% se obtienen mejores resultados con la aproximación de voto mayoritario teniendo su punto más alto cuando N = 15 con un 84.54%. Cuando I=20% se observa que las aproximaciones muestran alguna variación en sus predicciones, pero la que presenta un mejor accuracy es cuando N =15 para voto mayoritario con un 84.13% y cuando I= 30% los resultados se muestran más similares entre sus aproximaciones habiendo una mejoría cuando N = 11 para voto mayoritario con un 83.97%. Teniendo una perspectiva general de los resultados se puede ver que Random Forest obtiene mejores resultados con el método de voto mayoritario.

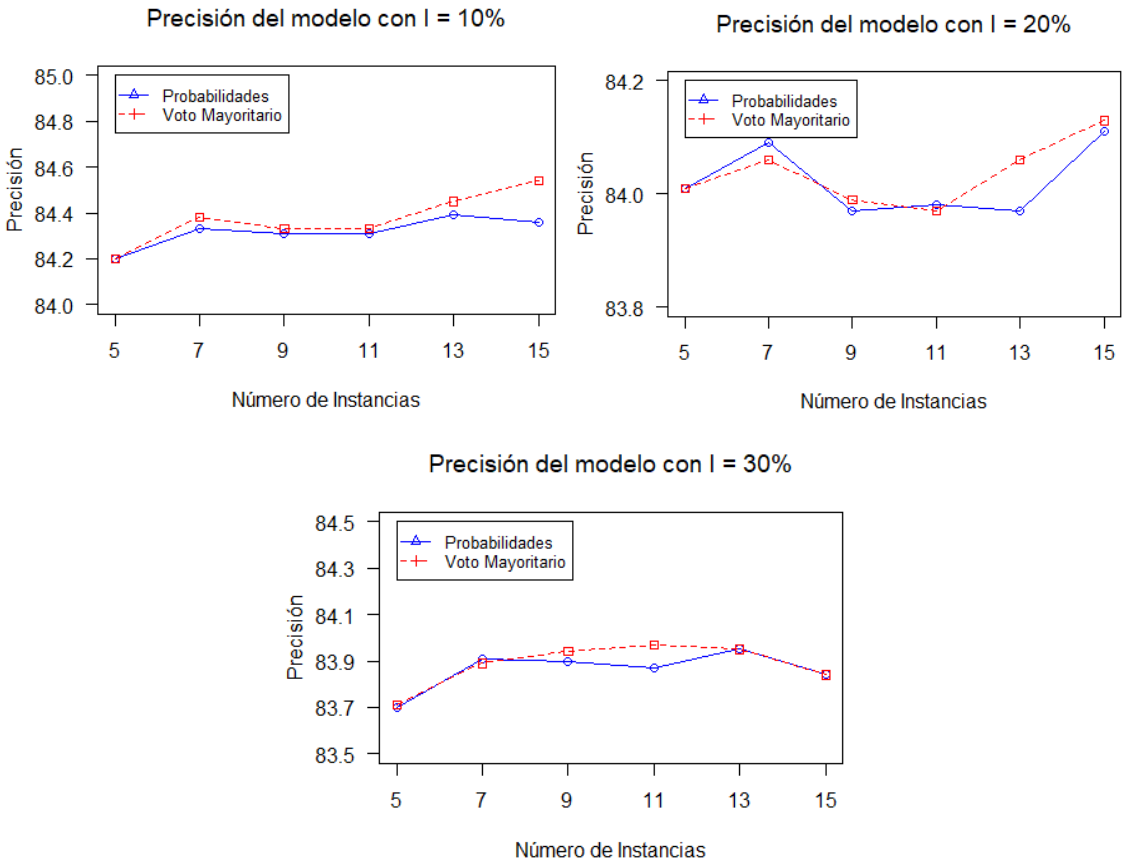



*Figura 32 Resultados globales del modelo RF variando su incertidumbre I.*

La Figura 33 muestra los resultados obtenidos de accuracy del modelo naive bayes, en la cual se muestra los valores cuando se conoce su valor, y los diferentes rangos de incertidumbre generados hasta tener una incertidumbre máxima en su atributo, se

puede observar de manera visual como el clasificador se ve afectado a medida aumenta el nivel de incertidumbre. los resultados también muestran que la perdida de accuracy cuando la incertidumbre es mínima es muy reducida pero cuando se tiene una máxima incertidumbre esta cae considerablemente.

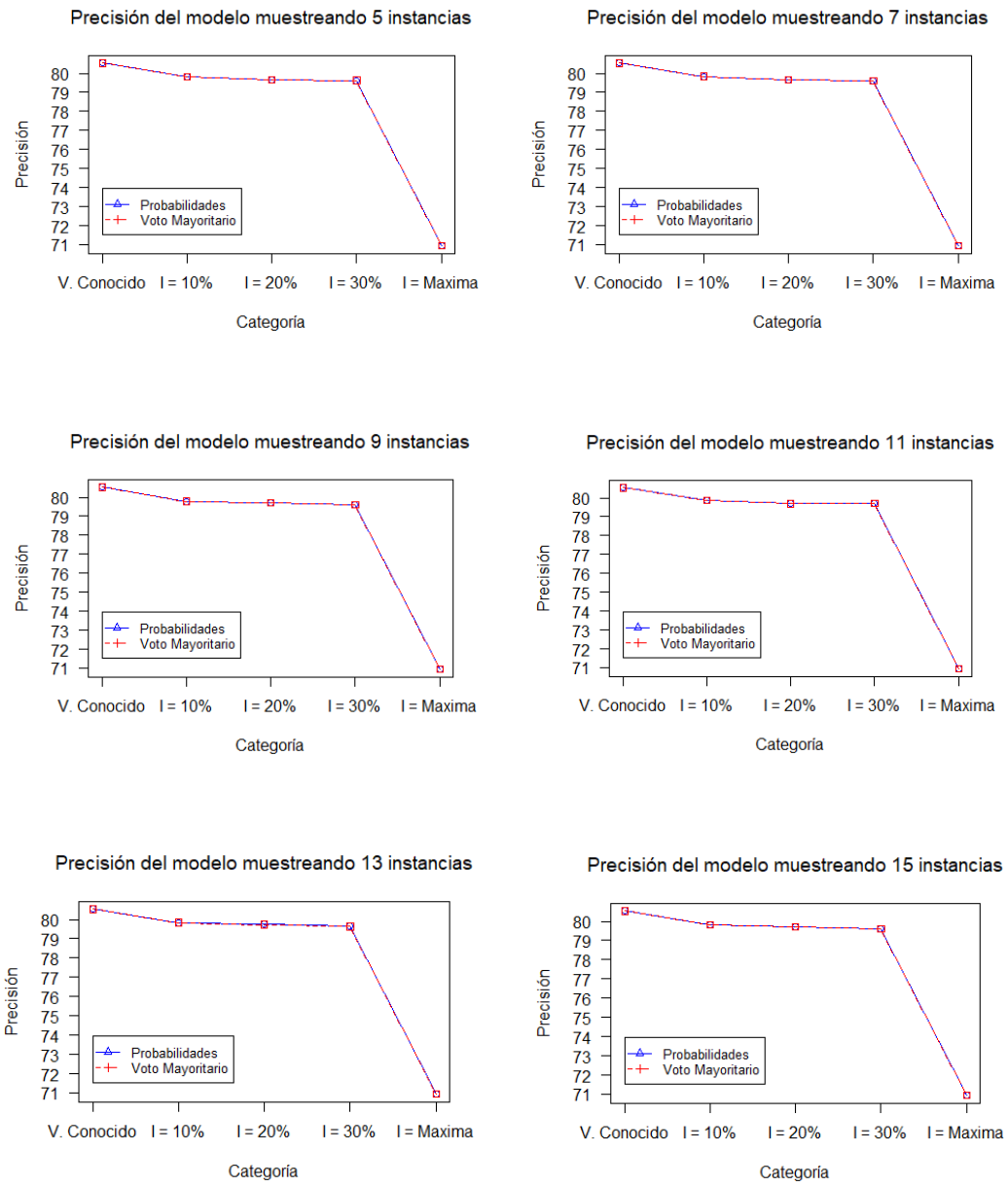



*Figura 33 Resultados globales del modelo NB muestreando N instancias.*

La figura 34 muestra los resultados desde otra perspectiva del modelo naive bayes, se puede observar que los resultados en sus aproximaciones no muestran mayor variación siendo similar en su gran mayoría. cuando I=10%. Sus predicciones son iguales tanto para voto mayoritario y probabilidad de pertenencia teniendo su punto más alto cuando N = 11 con un 79.84%. Cuando I=20% existe una pequeña diferencia entre los resultaos de sus aproximaciones, pero obtienen su mejor rendimiento cuando N = 13 para probabilidad de pertenencia con un 79.74%, y cuando I= 30% de igual manera sus resultados en las aproximaciones son muy similares, pero sobresale la probabilidad de

pertenencia cuando N = 11 con un 79.71%. Al ser un método probabilístico la variación de sus aproximaciones se ven de manera muy similar.

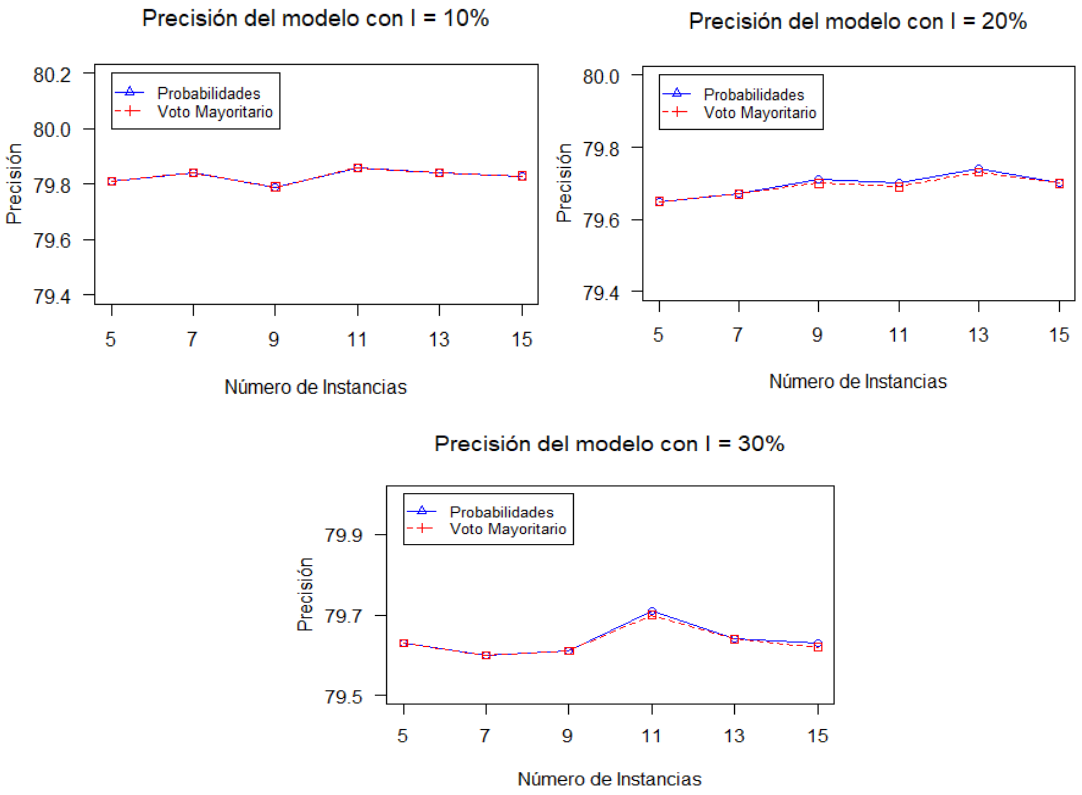



*Figura 34 Resultados globales del modelo NB variando su incertidumbre I.*

### 4.4.3. Análisis de la degradación de la clasificación.

La tabla 5 muestra los resultados de la degradación de accuracy de los modelos aplicados a medida aumenta la incertidumbre para cada N valores generados. se puede observar que de las técnicas aplicadas la que mejor controla la incertidumbre cuando hay una máxima incertidumbre es Random Forest degradando la clasificación en un 3.60% con respecto cuando se conoce su valor. Los Arboles de Decisión presenta una pérdida del 5.59%, mientras que Naive Bayes es la que mayor degradación presenta con un 9.60%.

| **Algoritmo** | | **Árbol de Decisión** | | **Random Forest** | | **Naive Bayes** | |
|---|---|---|---|---|---|---|---|
| **Método** | | Probabilidad | Voto Mayoritario | Probabilidad | Voto Mayoritario | Probabilidad | Voto Mayoritario |
| **I = Máxima** | | **5.59** | **5.59** | **3.60** | **3.60** | **9.60** | **9.60** |
| N = 5 | I = 10 % | 0.97 | 0.94 | 0.75 | 0.74 | 0.71 | 0.71 |
| | I = 20 % | 1.09 | 1.08 | 0.93 | 0.93 | 0.88 | 0.88 |
| | I = 30 % | 1.49 | 1.49 | 1.25 | 1.23 | 0.9 | 0.9 |
| N = 7 | I = 10 % | 0.92 | 0.89 | 0.61 | 0.56 | 0.69 | 0.69 |
| | I = 20 % | 0.98 | 1.04 | 0.85 | 0.89 | 0.86 | 0.86 |
| | I = 30 % | 1.28 | 1.34 | 1.03 | 1.06 | 0.93 | 0.93 |
| N = 9 | I = 10 % | 0.91 | 0.89 | 0.64 | 0.61 | 0.74 | 0.74 |
| | I = 20 % | 1.08 | 1.14 | 0.98 | 0.95 | 0.82 | 0.83 |
| | I = 30 % | 1.17 | 1.15 | 1.05 | 1.00 | 0.92 | 0.92 |
| N = 11 | I = 10 % | 0.97 | 0.87 | 0.64 | 0.62 | 0.67 | 0.67 |
| | I = 20 % | 1.03 | 1.13 | 0.97 | 0.97 | 0.82 | 0.83 |
| | I = 30 % | 1.25 | 1.17 | 1.07 | 0.97 | 0.82 | 0.83 |
| N = 13 | I = 10 % | 0.84 | 0.92 | 0.56 | 0.49 | 0.68 | 0.68 |
| | I = 20 % | 0.98 | 1.04 | 0.98 | 0.88 | 0.79 | 0.80 |
| | I = 30 % | 1.17 | 1.17 | 0.99 | 0.99 | 0.88 | 0.88 |
| N = 15 | I = 10 % | 0.76 | 0.80 | 0.59 | 0.40 | 0.70 | 0.70 |
| | I = 20 % | 0.91 | 0.99 | 0.83 | 0.81 | 0.82 | 0.82 |
| | I = 30 % | 1.29 | 1.30 | 1.10 | 1.10 | 0.89 | 0.90 |

*Tabla 5 Resultados de la degradación de los clasificadores.*

La Figura 35 muestra la degradación de los modelos de clasificación en base a las N instancias generadas. Se puede ver que Naive Bayes es la técnica que mayor degradación sufre cuando se tiene incertidumbre máxima, pero se puede notar que en la mayoría de casos cuando I=10%, I=20% I=30% es una de las técnicas que menos degradación sufre en su predicción. Controlado mejor la incertidumbre cuando esta es mínima. Pero también se debe tener en cuenta que a nivel de accuracy es la técnica que menores resultados obtuvo.

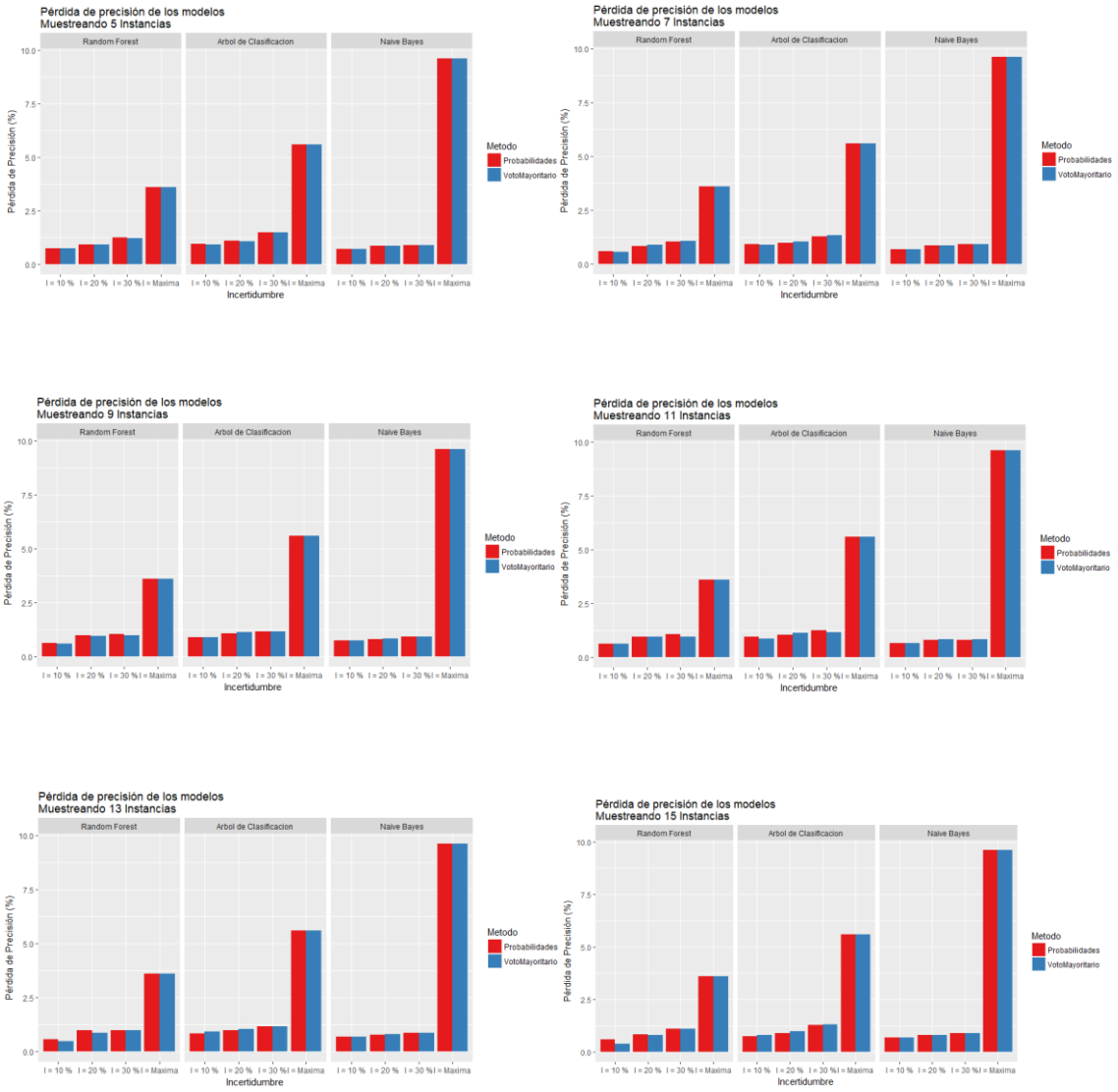



*Figura 35 Resultados de la degradación de los modelos de clasificación.*

# 5. Conclusiones y Trabajos Futuros

Este trabajo presenta un estudio que permite en base a dos aproximaciones poder abordar problemas de clasificación, donde se tiene un modelo aprendido con datos limpios y procesados, y este se aplica a datos que presentan incertidumbre en un atributo numérico. Con los experimentos realizados se ha podido comprobar que la precisión de los clasificadores se ve afectada a medida que aumenta el grado de incertidumbre presente en los datos.

Se ha detectado un comportamiento muy similar para los tres algoritmos utilizados en los diferentes conjuntos de datos, siendo Random Forest el algoritmo que mejores resultados obtiene cuando tiene incertidumbre máxima (con respecto a la medida de evaluación usada). Y Naive Bayes destaca con mejores resultados en general cuando la incertidumbre es mínima en base a los parámetros generados.

Las dos aproximaciones empleadas, voto mayoritario y probabilidad de pertenencia, han presentado comportamientos diferentes dependiendo de la técnica de aprendizaje empleada para aprender el modelo: Random Forest ha obtenido mejores resultados aplicando voto mayoritario, los Árboles de Decisión muestran resultados similares sobresaliendo no por mucho la aproximación de probabilidades de pertenencia y los resultados de Naive Bayes son muy similares sin notar mayor variación entre las dos aproximaciones por lo que ambas son igualmente válidas en este escenario. Mirando resultados globales (no por técnica), la aproximación de voto mayoritario se puede decir que obtiene mejores resultados.

Con respecto al número de instancias muestreadas por intervalo, las mejores predicciones obtenidas se obtuvieron cuando se generaron entre 11 y 15 puntos/instancias.

Los resultados experimentales que se obtuvieron con las aproximaciones empleadas son muy prometedores ya que muestran que la precisión alcanzada por nuestros métodos, aunque está por debajo de cuando no existe incertidumbre en los datos, pero mejora notablemente al caso de tener una incertidumbre máxima (la situación habitual de tratar este problema hoy en día)., Por ello, pensamos que sería interesante continuar con los trabajos futuros que se proponen a continuación.

**Trabajos Futuros.**

- Realizar un estudio que permita comparar la técnica empleada con respecto a otros métodos donde los datos son completados por regresión, medias, entre otros, y no tratan la incertidumbre que estos presentan.

- Realizar el experimento con un número mayor de dataset, en el presente trabajo se realizaron con 3 dataset ya que los datos obtenidos son prometedores y contribuirán en la veracidad de los resultados obtenido.

- Realizar un estudio en el que se pueda validar el porcentaje de datos que presentan incertidumbre dentro del conjunto de aplicación y a partir de que punto sería recomendable aplicar algún método para tratar esta incertidumbre.

- Generar un modelo que permita abordar temas de clasificación donde la incertidumbre se encuentre presente en valores nominales en el conjunto de aplicación.

# 6. Referencias


[1] S. Swapna, P. Niranjan, B. Srinivas y R. Swapna, «Data cleaning for data quality,» de *International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016.

[2] S.-F. Wu, C.-Y. Chang y S.-J. Lee, «Time series forecasting with missing values,» de *International Conference on Industrial Networks and Intelligent Systems (INISCom)*, 2015.

[3] I. Berkan Aydilek y A. Ahmet, «A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm,» de *Department of Computer Engineering, Selçuk University, Konya, Turkey*, 2013.

[4] I. Pratama, A. Erna Permanasar, I. Ardiyanto y R. Indrayani, «A Review of Missing Values Handling Methods on Time-Series Data,» de *International Conference on Information Technology Systems and Innovation (ICITSI)*, 2016.

[5] A. dhevi, «Imputing missing values using Inverse Distance Weighted Interpolation for time series data,» de *International Conference on Advanced Computing (ICoAC)* , 2014.

[6] K. L. Marie D. Sobrevilla, A. G. Quiñones, K. V. S. Lopez y V. T. Azaña , «Daily weather forecast in Tiwi, Albay, Philippines using Artificial Neural Network with missing values Imputation,» de *IEEE Region 10 Conference (TENCON)*, 2016.

[7] M. Mediavilla, «Método de imputación de los valores no observados.,» 2012. [En línea]. Available: http://2012.economicsofeducation.com/user/pdfsesiones/035.pdf.

[8] J. Hernandez Orralo, M. J. Ramírez Quintana y C. Ferri Ramírez, Introduccion a la Mineria de Datos, PEARSON Prentice Hall, 2004.

[9] D. S. G. César Pérez López, Data Mining Soluciones con Enterprise Miner, RA-MA Editorial, 2006.

[10] L. C. M. Félix, «Data mining: torturando a los datos hasta que confiesen,» [En línea]. Available: http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html.

[11] Z. Yun, L. Weihua y C. Yang, «Applying Balanced ScordCard Strategic Performance Management to CRISP-DM,» de *International Conference on Information Science, Electronics and Electrical Engineering*, 2014.

[12] A. Nadali, E. Naghizadeh kakhky y H. Eslami Nosratabadi, «Evaluating the Success Level of Data Mining Projects Based on CRISP-DM Methodology by a Fuzzy Expert System,» de *International Conference on Electronics Computer Technology*, 2011.

[13] C. C. Lorena, «Aplicación de algoritmos de clasificación supervisada usando weka,» de *Congresos Labsis*, 2009.

[14] Q. A. Al-Radaideh y E. Al Nagi, «Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance,» de *International Journal of Advanced Computer Science and Applications*, 2012.

[15] ORACLE, «Data Mining Concepts,» [En línea]. Available: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#DMCON005.

[16] S. Sayad, «An Introduction to Data Mining,» [En línea]. Available: http://www.saedsayad.com/data_mining_map.htm.

[17] S. Segrera Francia y M. N. Moreno García , «Multiclasificadores: Métodos y Arquitecturas,» 2006. [En línea]. Available: https://www.researchgate.net/publication/39698361_Multiclasificadores_metodos_y_arquitecturas.

[18] N. Sutton-Charani, S. Destercke y T. Denoeux, «Learning Decision Trees from Uncertain Data with an Evidential EM Approach,» de *International Conference on Machine Learning and Applications*, 2013.

[19] R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent y L. R. Zúnica, «Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos,» 2007.

[20] QuanDare, «ARTIFICIAL INTELLIGENCE Random forest vs Simple tree,» [En línea]. Available: https://quantdare.com/random-forest-vs-simple-tree/.

[21] L. Breiman, «University of California, Berkeley - RANDOM FORESTS,» [En línea]. Available: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

[22] A. Yunjing, S. Sun y S. Wang, «Naive Bayes Classifiers for Music Emotion Classification Based on Lyrics,» de *International Conference on Computer and Information Science (ICIS)*, 2017.

[23] K. M. Leung, «Naive Bayesian Classifier,» 28 11 2007. [En línea]. Available: http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf. [Último acceso: 23 08 2017].

[24] G. Zheng y Y. Tian, «Chinese Web Text Classification System Model Based on Naive Bayes,» de *International Conference on E-Product E-Service and E-Entertainment*, 2010.

[25] J. Hernández Orallo y C. Hervás Martínez, «Evaluacion sensible a la distribucion y el coste,» [En línea]. Available: http://slideplayer.es/slide/2312433/.

[26] T. Song, L. Shao y X. Han, «A quick Otsu-Kmeans algorithm for the internal pipeline detection,» de *International Conference on Mechatronics and Automation (ICMA)*, 2017.

[27] «RStudio,» [En línea]. Available: https://www.rstudio.com/products/rstudio/. [Último acceso: 22 08 2017].

[28] RStudio, «RDodumentation,» [En línea]. Available: https://www.rdocumentation.org/. [Último acceso: 15 08 2017].

[29] H. Hofmann, «UCI Machine Learning Repository,» [En línea]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data). [Último acceso: 27 05 2017].

[30] «UCI Machine Learning Repository,» [En línea]. Available: http://archive.ics.uci.edu/ml/datasets/credit+approval. [Último acceso: 05 05 2017].

[31] International Business Machines. (IBM), «kaggle,» [En línea]. Available: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset. [Último acceso: 27 05 2017].

[32] Y. Holtz, «The R Grahp Gallery,» [En línea]. Available: http://www.r-graph-gallery.com/about/. [Último acceso: 05 08 2017].

[33] F. McCown, «Producing Simple Graphs with R,» 2016. [En línea]. Available: https://www.harding.edu/fmccown/r/.

[34] A. Baradi y K. Enders, «An introduction to modern missinug data analysis,» *Journal of School Psychology,* vol. 48, pp. 5-37, 2018.

[35] R. Little y D. Rubin, Statistical Analysis with Missing Data, Hoboken, NJ, 2002.

# 7. Anexos

| Datos | | | G. Credit | | | Credit Approved | | | IBM HR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Algoritmo | | | Árbol de Clasificación | Random Forest | Naive Bayes | Árbol de Clasificación | Random Forest | Naive Bayes | Árbol de Clasificación | Random Forest | Naive Bayes |
| **I = 0 (Valor Conocido)** | | | **78.27** | **78.70** | **75.40** | **88.02** | **87.08** | **82.84** | **85.83** | **89.06** | **83.34** |
| **I = 10 %** | **N = 5** | **Probabilidad** | 77.66 | 77.58 | 75.20 | 86.62 | 86.59 | 81.65 | 84.93 | 88.42 | 82.60 |
| | | **Voto Mayoritario** | 77.66 | 77.54 | 75.20 | 86.71 | 86.65 | 81.65 | 84.93 | 88.42 | 82.60 |
| | **N = 7** | **Probabilidad** | 77.58 | 77.79 | 75.24 | 86.59 | 86.62 | 81.65 | 85.19 | 88.59 | 82.64 |
| | | **Voto Mayoritario** | 77.61 | 77.85 | 75.24 | 86.65 | 86.71 | 81.65 | 85.20 | 88.59 | 82.64 |
| | **N = 9** | **Probabilidad** | 77.55 | 77.54 | 75.22 | 86.59 | 86.80 | 81.56 | 85.24 | 88.59 | 82.58 |
| | | **Voto Mayoritario** | 77.70 | 77.59 | 75.22 | 86.59 | 86.83 | 81.56 | 85.19 | 88.57 | 82.58 |
| | **N = 11** | **Probabilidad** | 77.52 | 77.73 | 75.24 | 86.65 | 86.65 | 81.71 | 85.03 | 88.55 | 82.64 |
| | | **Voto Mayoritario** | 77.68 | 77.71 | 75.24 | 86.77 | 86.74 | 81.71 | 85.07 | 88.53 | 82.64 |
| | **N = 13** | **Probabilidad** | 77.77 | 77.76 | 75.26 | 86.71 | 86.83 | 81.65 | 85.12 | 88.59 | 82.62 |
| | | **Voto Mayoritario** | 77.75 | 77.88 | 75.26 | 86.53 | 86.89 | 81.65 | 85.08 | 88.59 | 82.62 |
| | **N = 15** | **Probabilidad** | 77.89 | 77.73 | 75.32 | 86.68 | 86.74 | 81.53 | 85.27 | 88.61 | 82.64 |
| | | **Voto Mayoritario** | 77.95 | 78.12 | 75.32 | 86.62 | 86.89 | 81.53 | 85.16 | 88.63 | 82.64 |
| **I = 20 %** | **N = 5** | **Probabilidad** | 77.52 | 77.24 | 75.02 | 86.44 | 86.56 | 81.35 | 84.88 | 88.25 | 82.57 |
| | | **Voto Mayoritario** | 77.52 | 77.28 | 75.02 | 86.47 | 86.56 | 81.35 | 84.88 | 88.21 | 82.57 |
| | **N = 7** | **Probabilidad** | 77.65 | 77.38 | 75.00 | 86.41 | 68.65 | 81.44 | 85.11 | 88.25 | 82.57 |
| | | **Voto Mayoritario** | 77.58 | 77.32 | 75.00 | 86.41 | 86.62 | 81.44 | 85.01 | 88.23 | 82.57 |
| | **N = 9** | **Probabilidad** | 77.46 | 77.28 | 75.10 | 86.44 | 86.32 | 81.47 | 84.99 | 88.30 | 82.55 |
| | | **Voto Mayoritario** | 77.44 | 77.32 | 75.10 | 86.44 | 86.35 | 81.44 | 84.81 | 88.32 | 82.55 |
| | **N = 11** | **Probabilidad** | 77.55 | 77.12 | 75.10 | 86.53 | 86.47 | 81.44 | 84.96 | 88.34 | 82.57 |
| | | **Voto Mayoritario** | 77.65 | 77.19 | 75.10 | 86.35 | 86.41 | 81.41 | 84.73 | 88.33 | 82.57 |
| | **N = 13** | **Probabilidad** | 77.63 | 77.11 | 75.18 | 86.68 | 86.47 | 81.47 | 84.88 | 88.33 | 82.57 |
| | | **Voto Mayoritario** | 77.57 | 77.33 | 75.18 | 86.65 | 86.50 | 81.44 | 84.77 | 88.36 | 82.57 |
| | **N = 15** | **Probabilidad** | 77.74 | 77.28 | 75.16 | 86.65 | 86.65 | 81.41 | 85.00 | 88.41 | 82.54 |
| | | **Voto Mayoritario** | 77.69 | 77.29 | 75.16 | 86.47 | 86.74 | 81.41 | 84.99 | 88.37 | 82.54 |
| **I = 30 %** | **N = 5** | **Probabilidad** | 77.21 | 76.93 | 74.84 | 85.96 | 86.02 | 81.56 | 84.47 | 88.14 | 82.49 |
| | | **Voto Mayoritario** | 77.21 | 76.94 | 74.84 | 85.96 | 86.08 | 81.56 | 84.47 | 88.13 | 82.49 |
| | **N = 7** | **Probabilidad** | 77.43 | 77.17 | 74.89 | 86.05 | 86.38 | 81.38 | 84.81 | 88.19 | 82.53 |
| | | **Voto Mayoritario** | 77.33 | 77.09 | 74.89 | 86.05 | 86.35 | 81.38 | 84.71 | 88.22 | 82.53 |
| | **N = 9** | **Probabilidad** | 77.52 | 77.15 | 74.92 | 86.08 | 86.29 | 81.44 | 85.00 | 88.26 | 82.47 |
| | | **Voto Mayoritario** | 77.57 | 77.16 | 74.92 | 86.08 | 86.35 | 81.44 | 85.01 | 88.32 | 82.47 |
| | **N = 11** | **Probabilidad** | 78.65 | 77.01 | 75.06 | 85.78 | 86.35 | 81.59 | 84.95 | 88.26 | 82.49 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Voto Mayoritario** | 77.64 | 77.24 | 75.06 | 86.05 | 86.38 | 81.56 | 84.93 | 88.30 | 82.49 |
| | **N = 13** | **Probabilidad** | 77.51 | 77.28 | 75.02 | 86.05 | 86.38 | 81.47 | 85.04 | 88.19 | 82.45 |
| | | **Voto Mayoritario** | 77.53 | 77.28 | 75.02 | 86.05 | 86.38 | 81.47 | 85.04 | 88.19 | 82.45 |
| | **N = 15** | **Probabilidad** | 77.43 | 76.94 | 75.06 | 85.81 | 86.32 | 81.44 | 85.01 | 88.27 | 82.40 |
| | | **Voto Mayoritario** | 77.48 | 77.00 | 75.06 | 85.87 | 86.26 | 81.41 | 84.88 | 88.27 | 82.40 |
| **I = Máxima** | | | **71.53** | **73.67** | **66.48** | **83.23** | **83.94** | **71.27** | **80.59** | **86.45** | **75.05** |

*Tabla 6 Resultados de las ejecuciones de los dataset y cada modelo de clasificación.*

## 7.1. Anexo 1

La tabla 7 muestra los resultados obtenidos de los conjuntos de datos, aplicando los diferentes métodos de clasificación, a continuación, se describen de manera gráfica sus resultados.

### Dataset 1: (G. Credit)

### Modelo de clasificación: Árbol de Clasificación.

### Resultados variando el número de instancias N

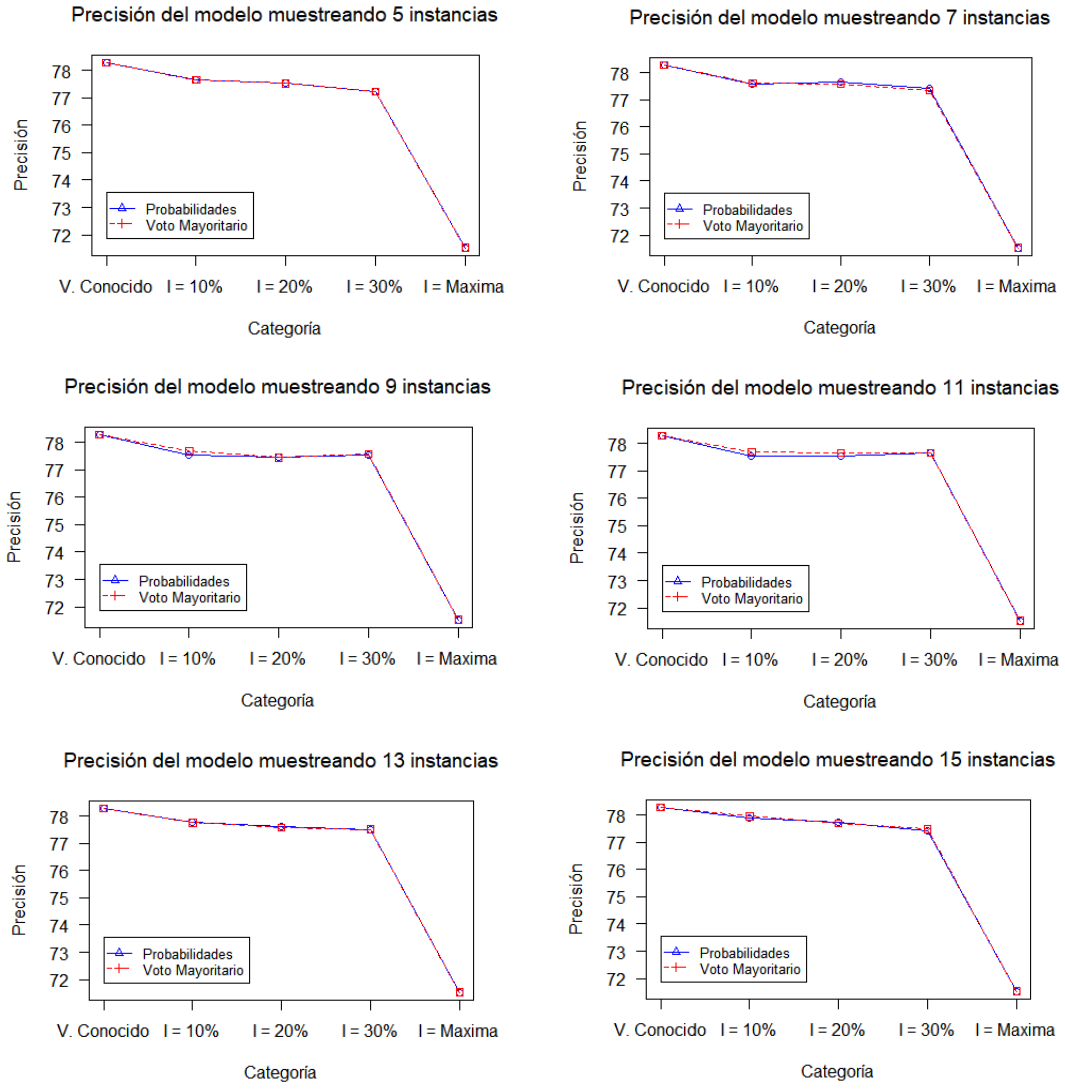



*Figura 36. Resultados (G. Credit) del modelo AD muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.

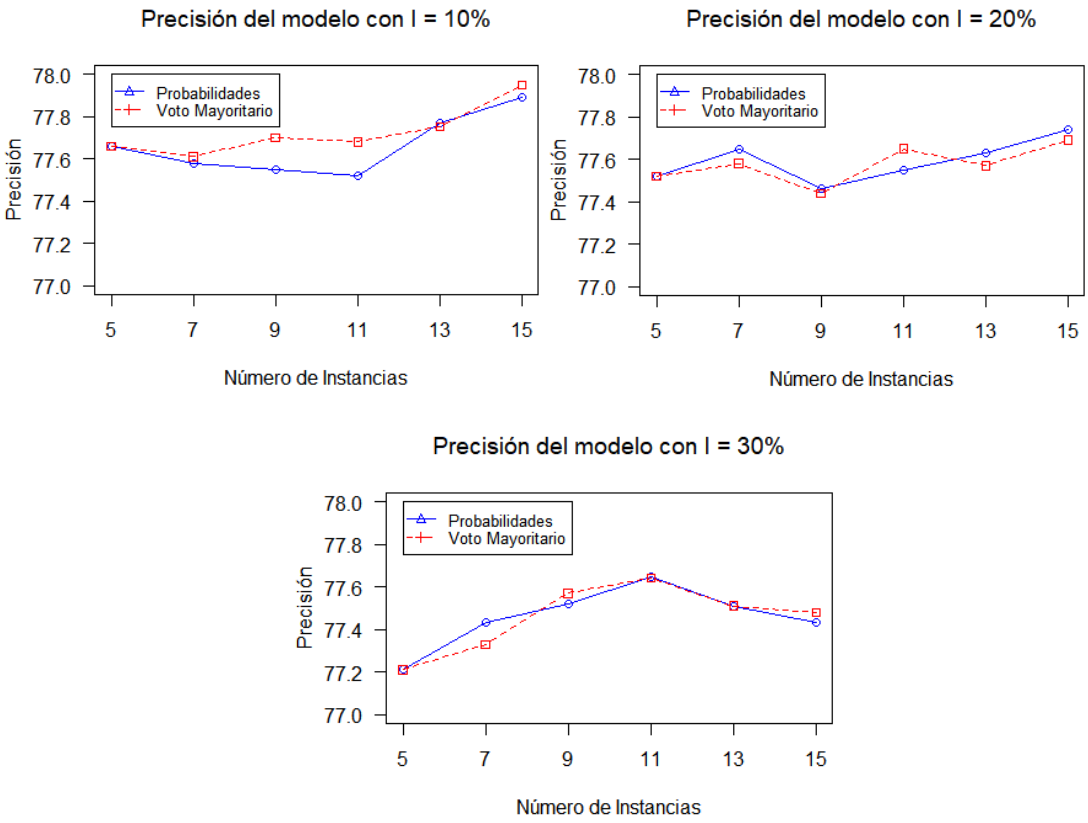
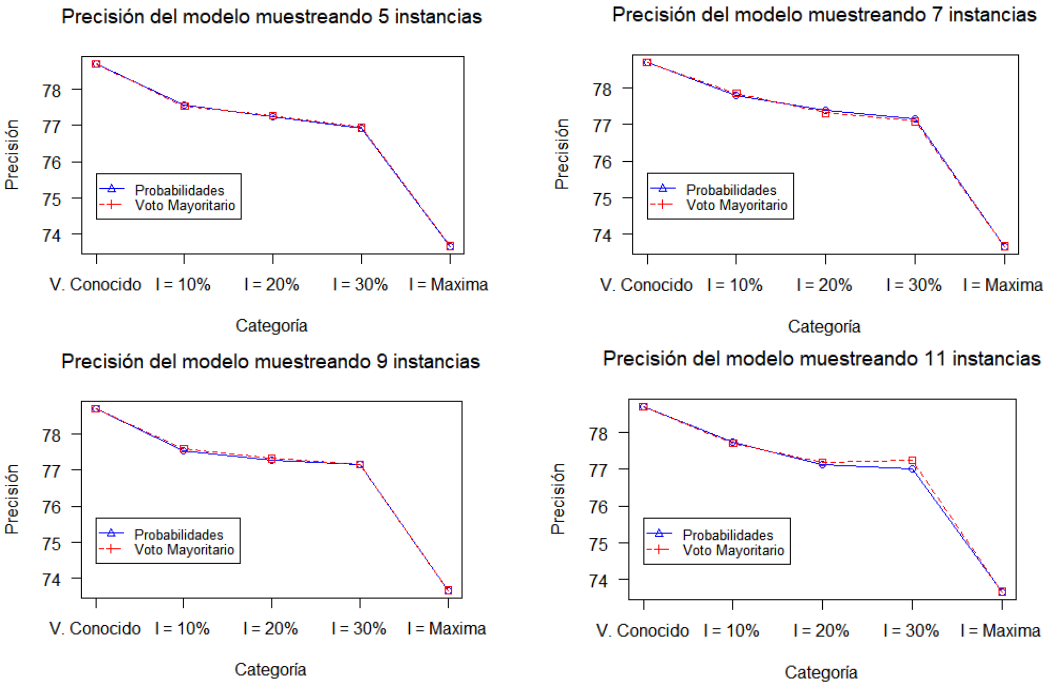



*Figura 37 Resultados (G. Credit) del modelo AD variando su incertidumbre I.*

## Modelo de clasificación: Random Forest

## Resultados variando el número de instancias N



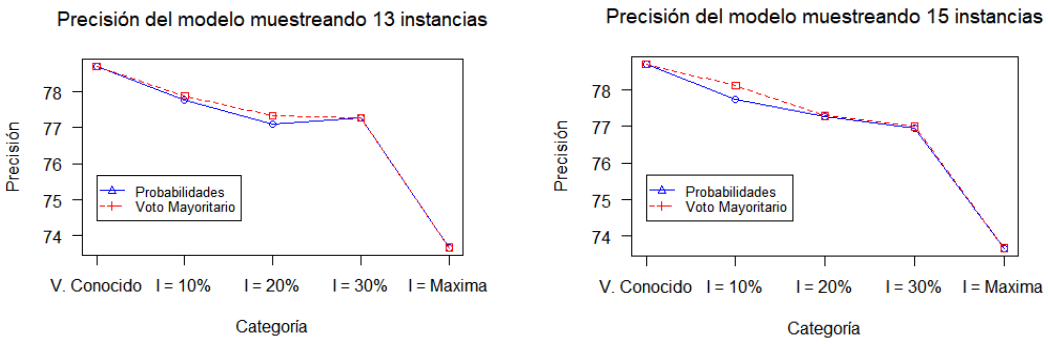



*Figura 38 Resultados (G. Credit) del modelo RF muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.

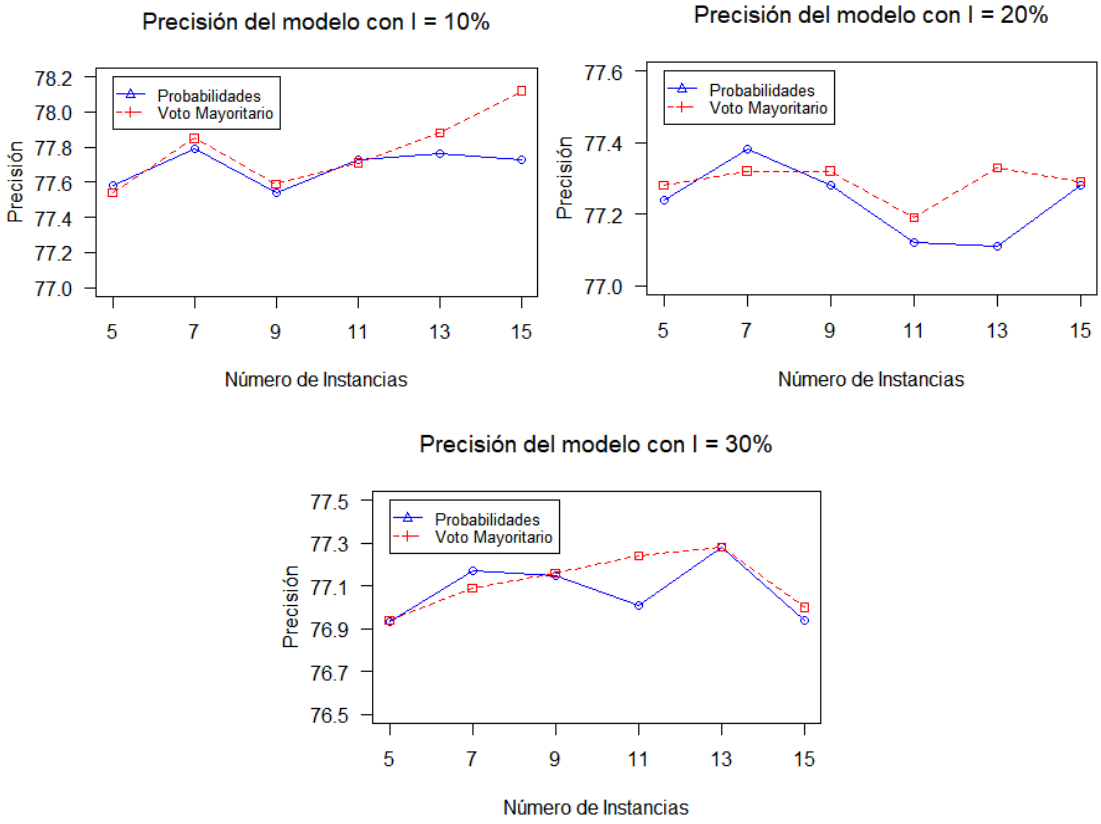



*Figura 39 Resultados (G. Credit) del modelo RF variando su incertidumbre I.*

## Modelo de clasificación: Naive Bayes

## Resultados variando el número de instancias N

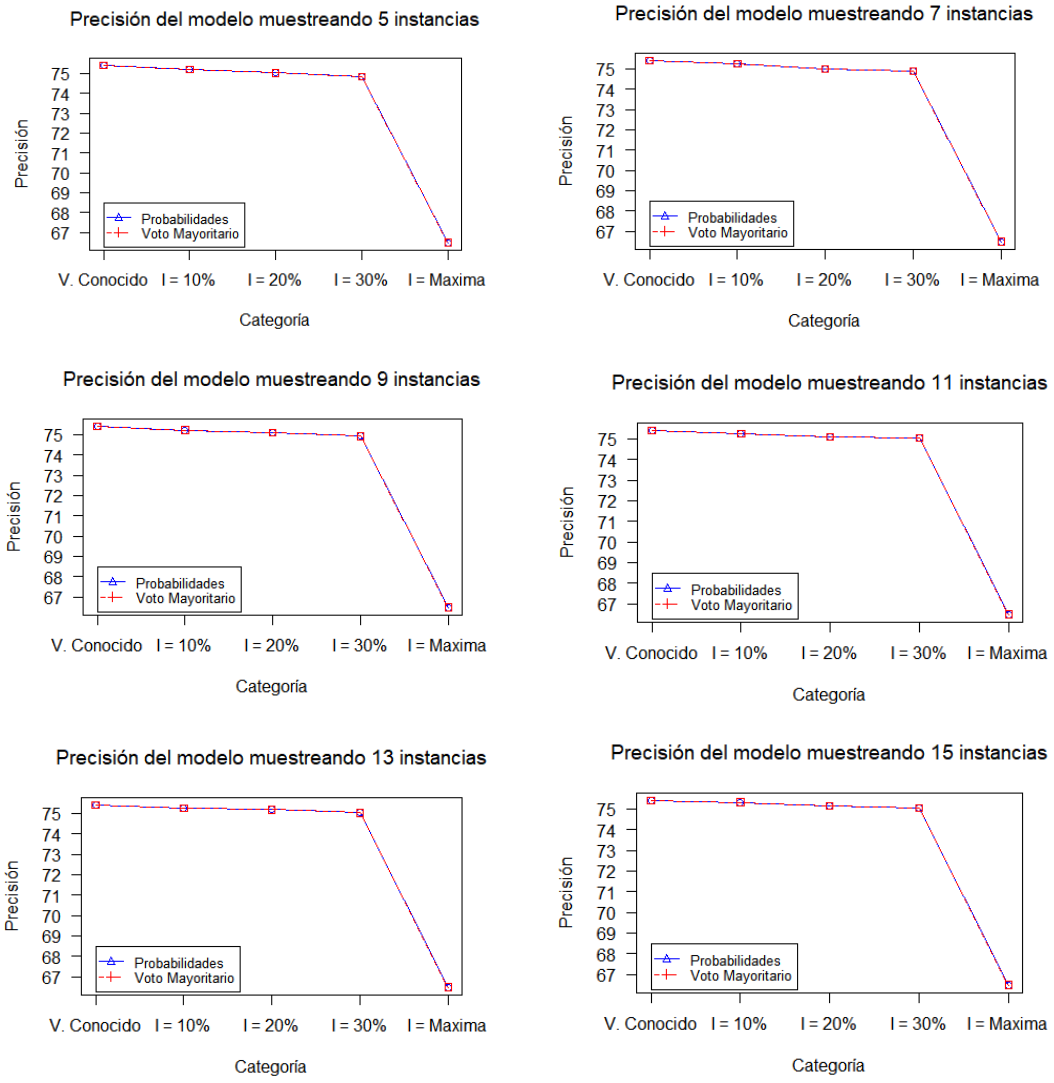
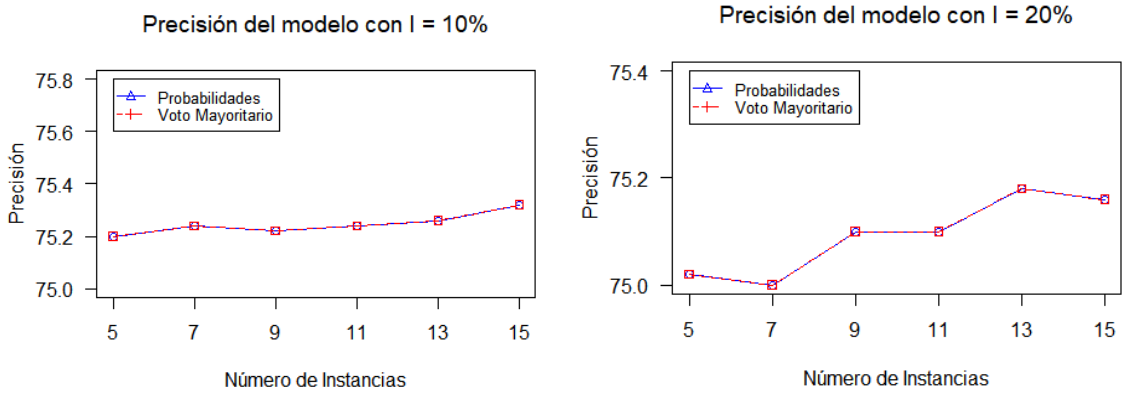



*Figura 40 Resultados (G. Credit) del modelo NB muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.



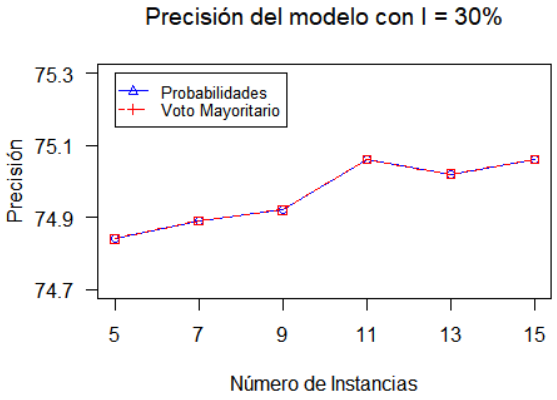


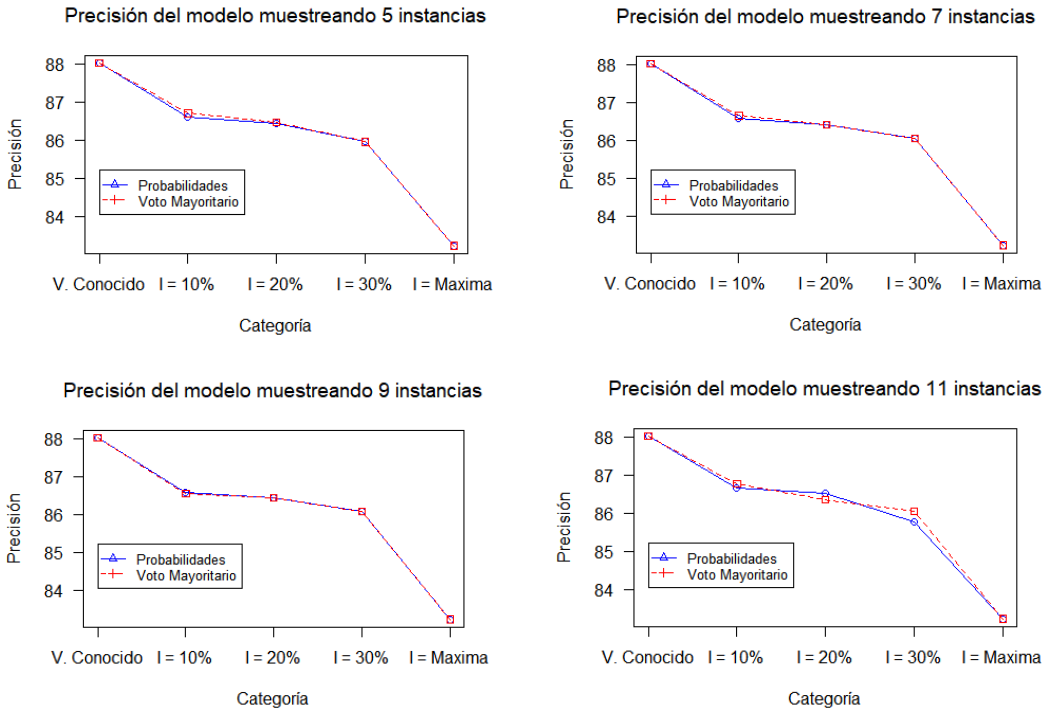

*Figura 41 Resultados (G. Credit) del modelo NB variando su incertidumbre I.*

## 7.2. Anexo 2

### Dataset 2: (Credit Approved)

### Modelo de clasificación: Árbol de Clasificación

### Resultados variando el número de instancias N



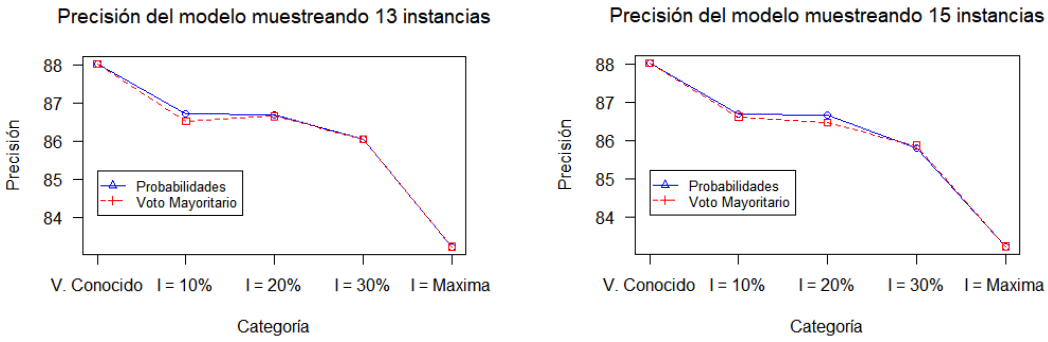



*Figura 42 Resultados (Credit Approved) del modelo AD muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.

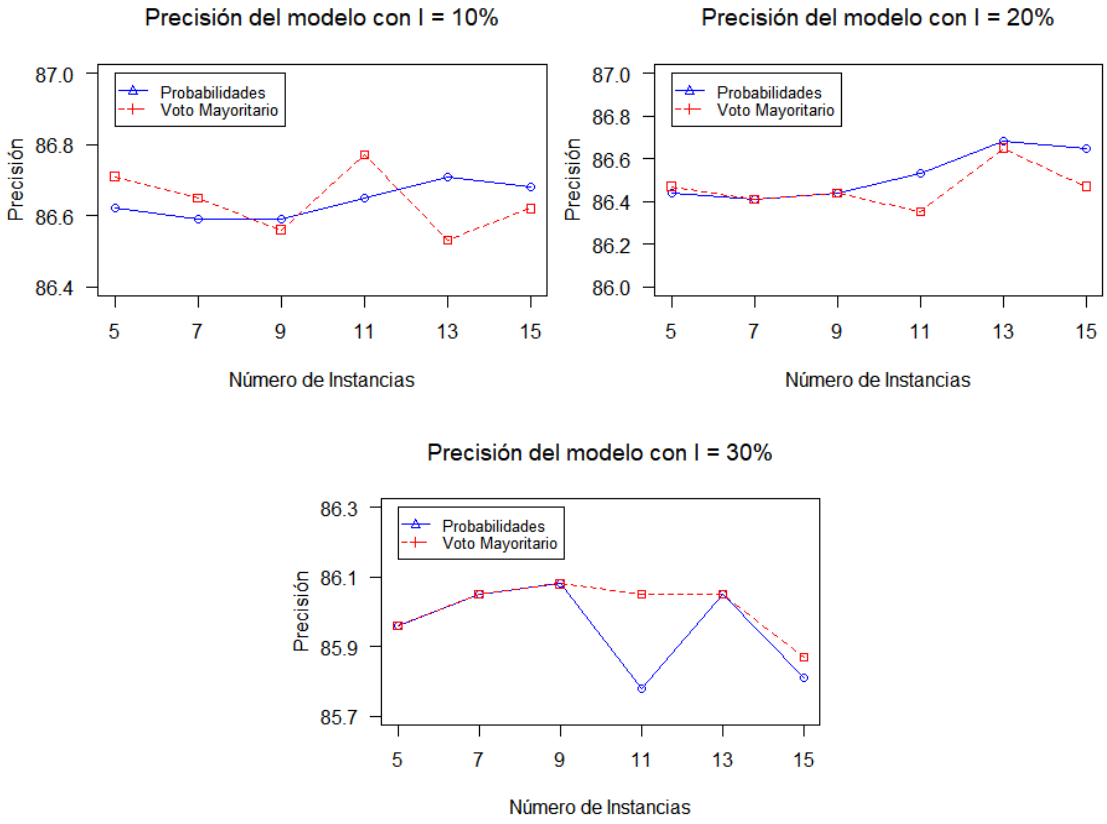



*Figura 43 Resultados (Credit Approved) del modelo AD variando su incertidumbre I.*

## Modelo de clasificación: Random Forest

## Resultados variando el número de instancias N

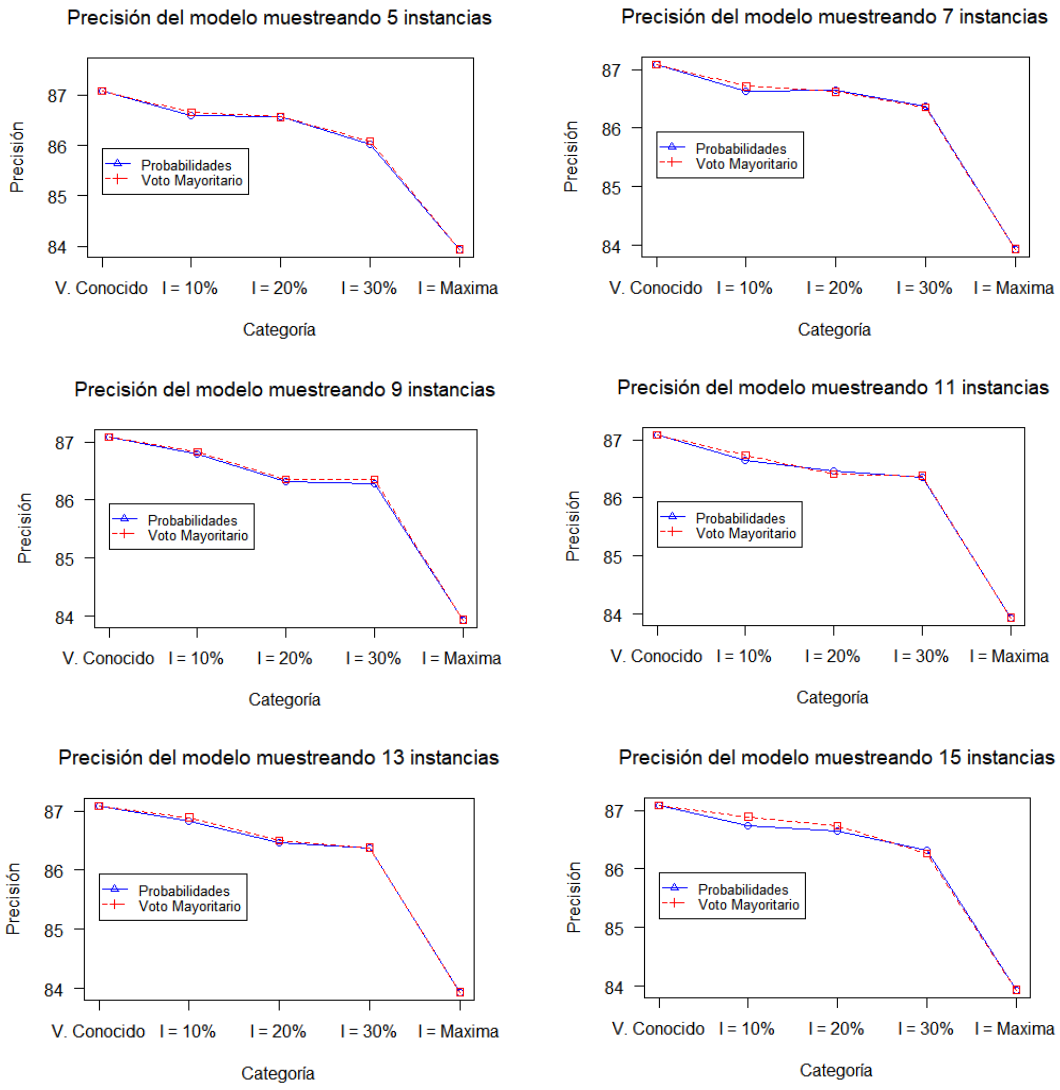
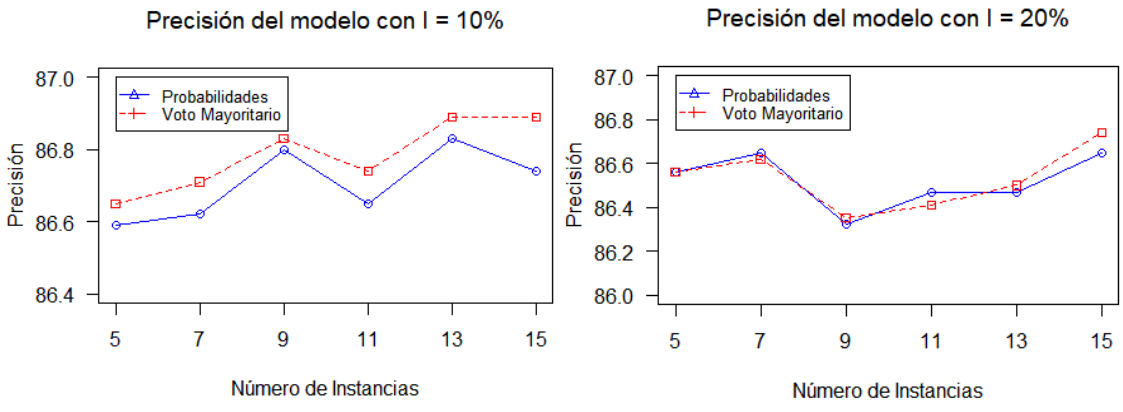



*Figura 44 Resultados (Credit Approved) del modelo RF muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.



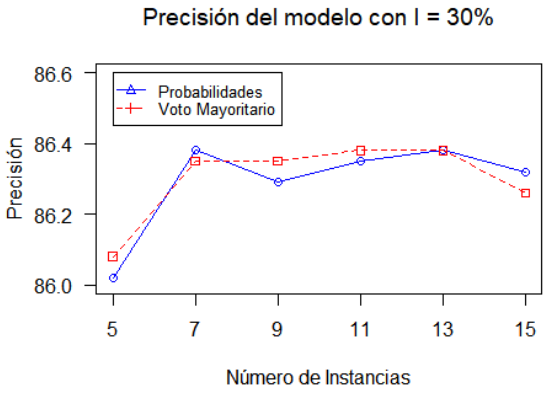



*Figura 45 Resultados (Credit Approved) del modelo RF variando su incertidumbre I.*

## Modelo de clasificación: Naive Bayes

## Resultados variando el número de instancias N

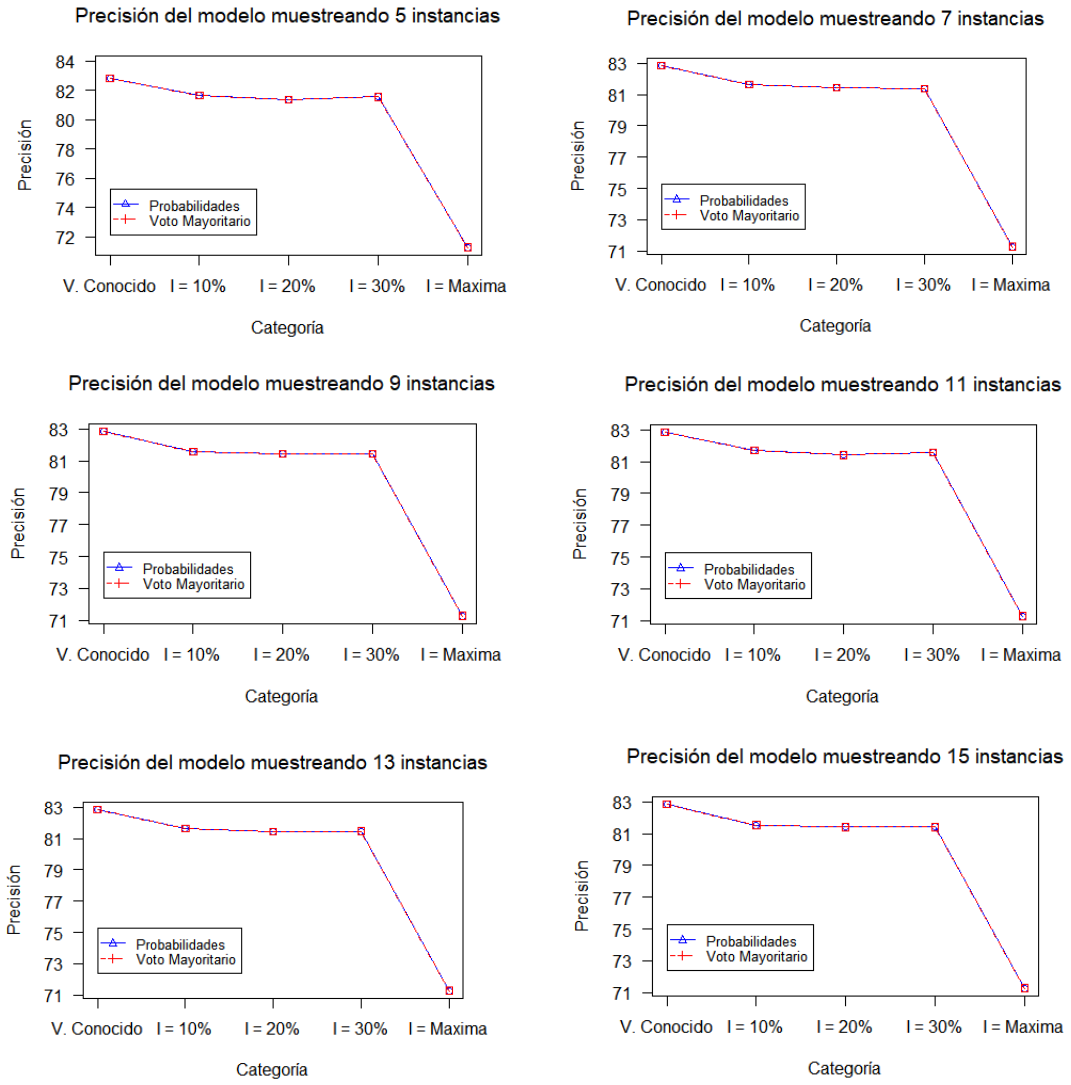



*Figura 46 Resultados (Credit Approved) del modelo NB muestreando N instancias.*

### Resultados variando el porcentaje de incertidumbre I.

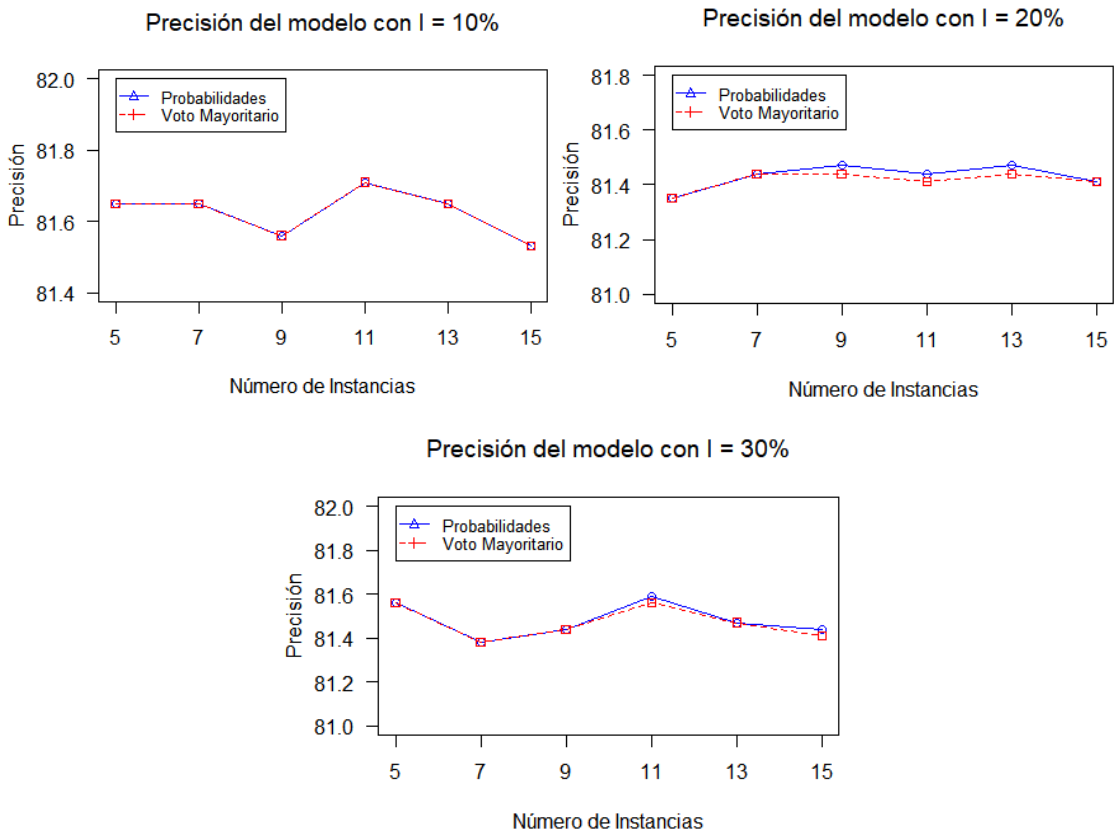


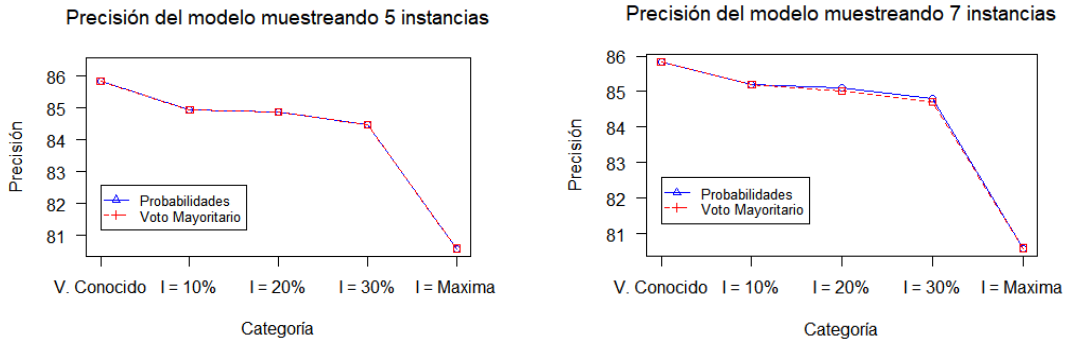

*Figura 47 Resultados (Credit Approved) del modelo NB variando su incertidumbre I.*

## 7.3. Anexo 3

### Dataset 3: (IBM HR)

### Modelo de clasificación: Árbol de Clasificación

### Resultados variando el número de instancias N



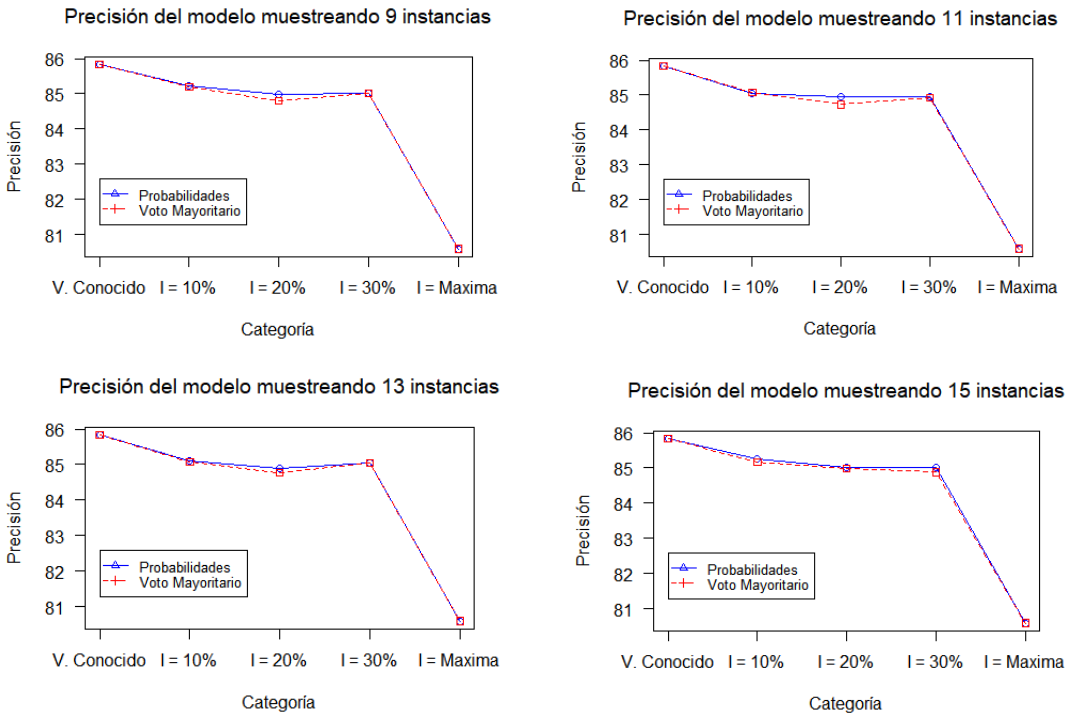



*Figura 48 Resultados (IBM HR) del modelo AD muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.

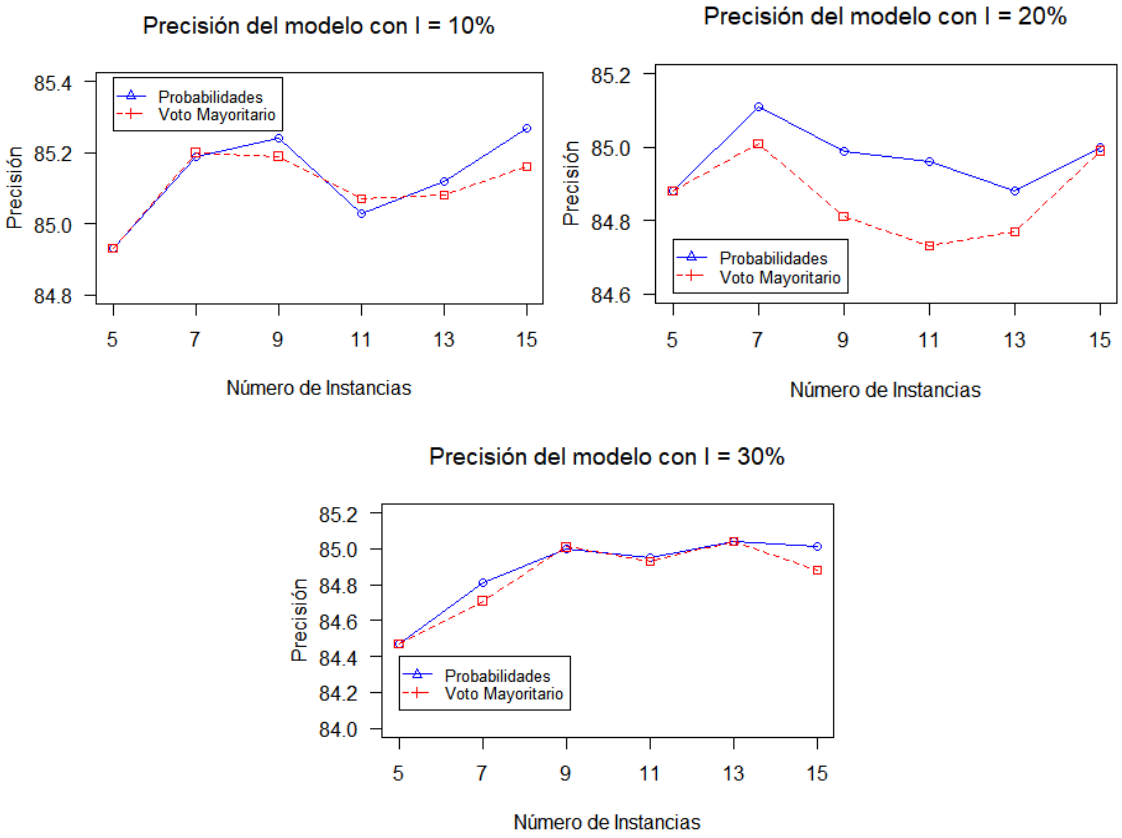



*Figura 49 Resultados (IBM HR) del modelo AD variando su incertidumbre I.*

## Modelo de clasificación: Random Forest

## Resultados variando el número de instancias N

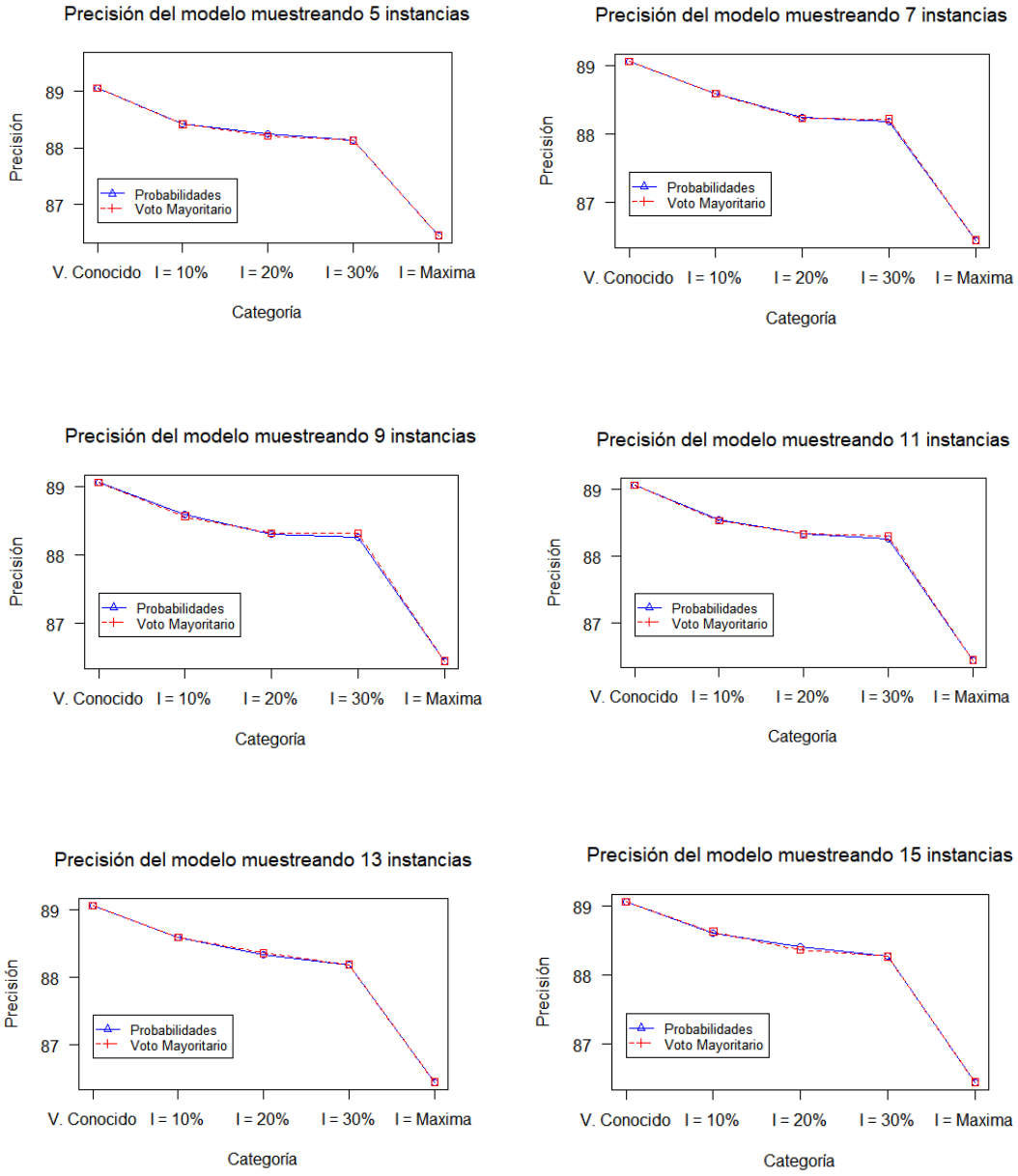



*Figura 50 Resultados (IBM HR) del modelo RF muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.

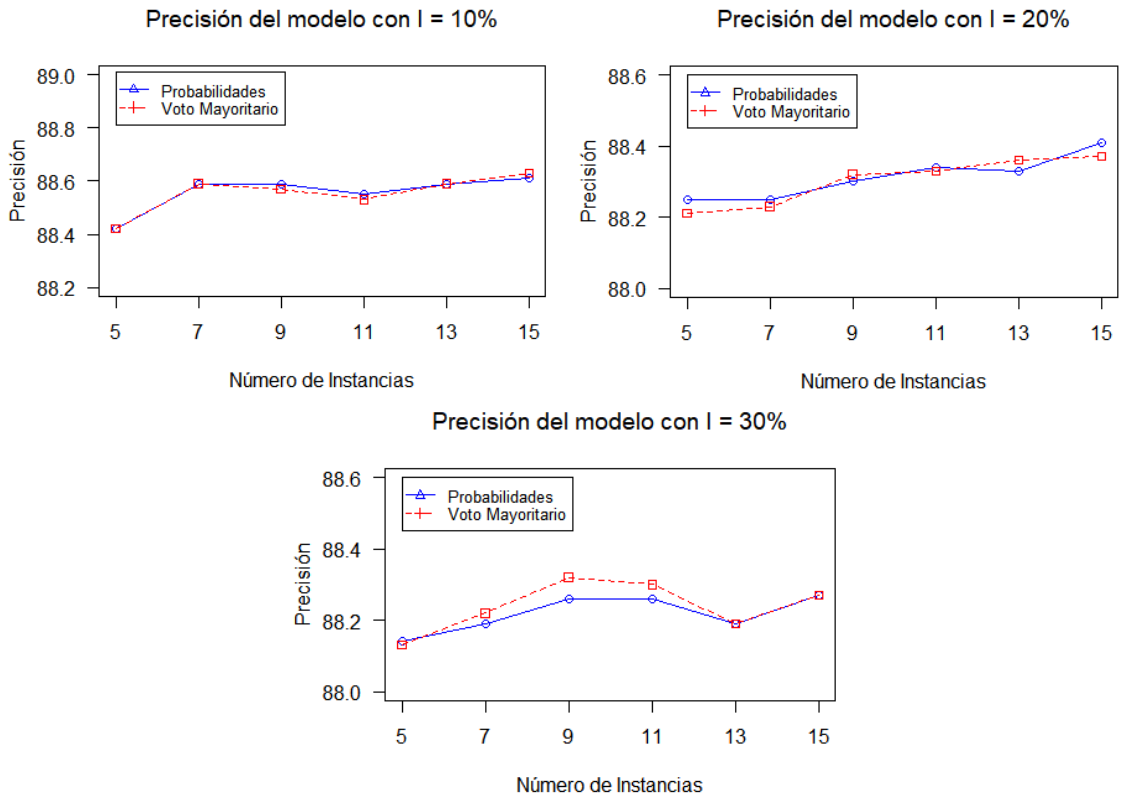
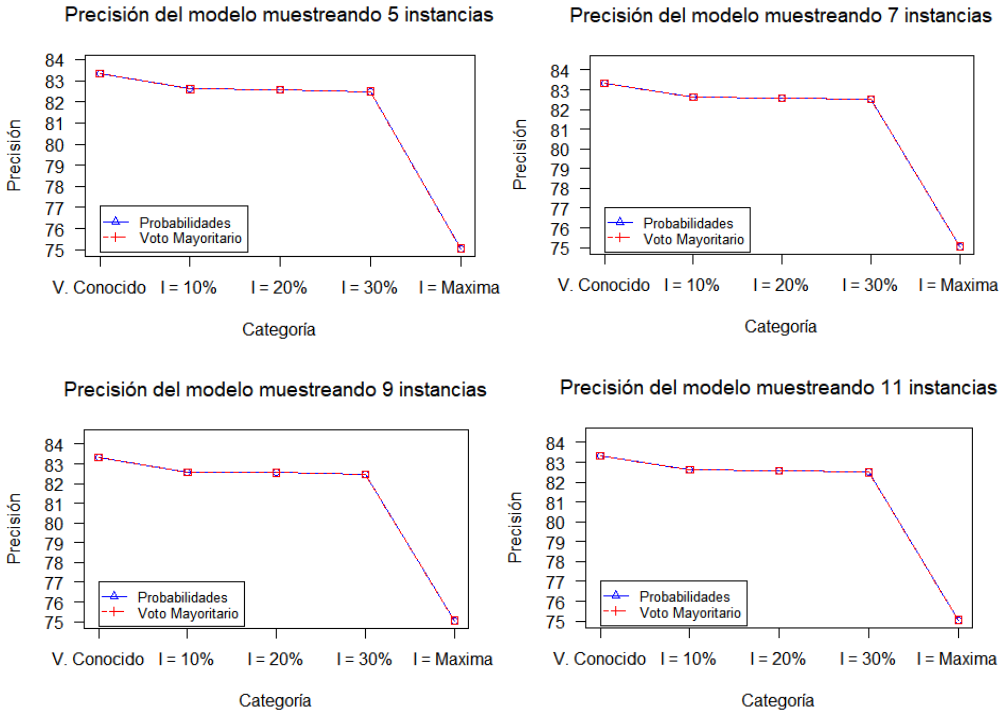



*Figura 51 Resultados (IBM HR) del modelo RF variando su incertidumbre I.*

## Modelo de clasificación: Naive Bayes

## Resultados variando el número de instancias N



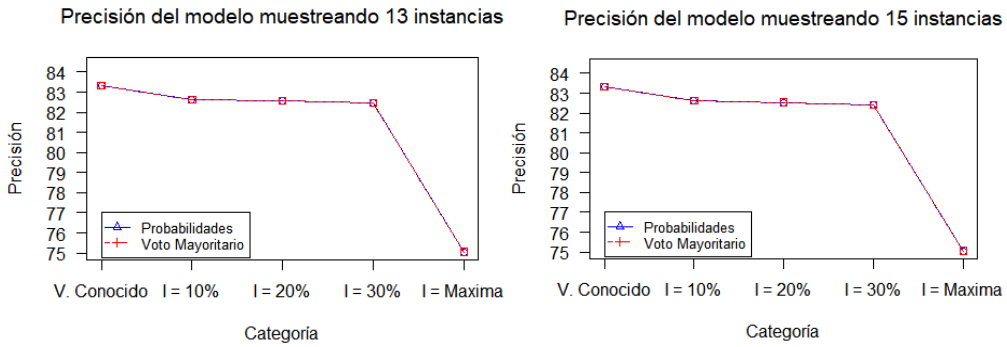



*Figura 52 Resultados (IBM HR) del modelo NB muestreando N instancias.*

## Resultados variando el porcentaje de incertidumbre I.

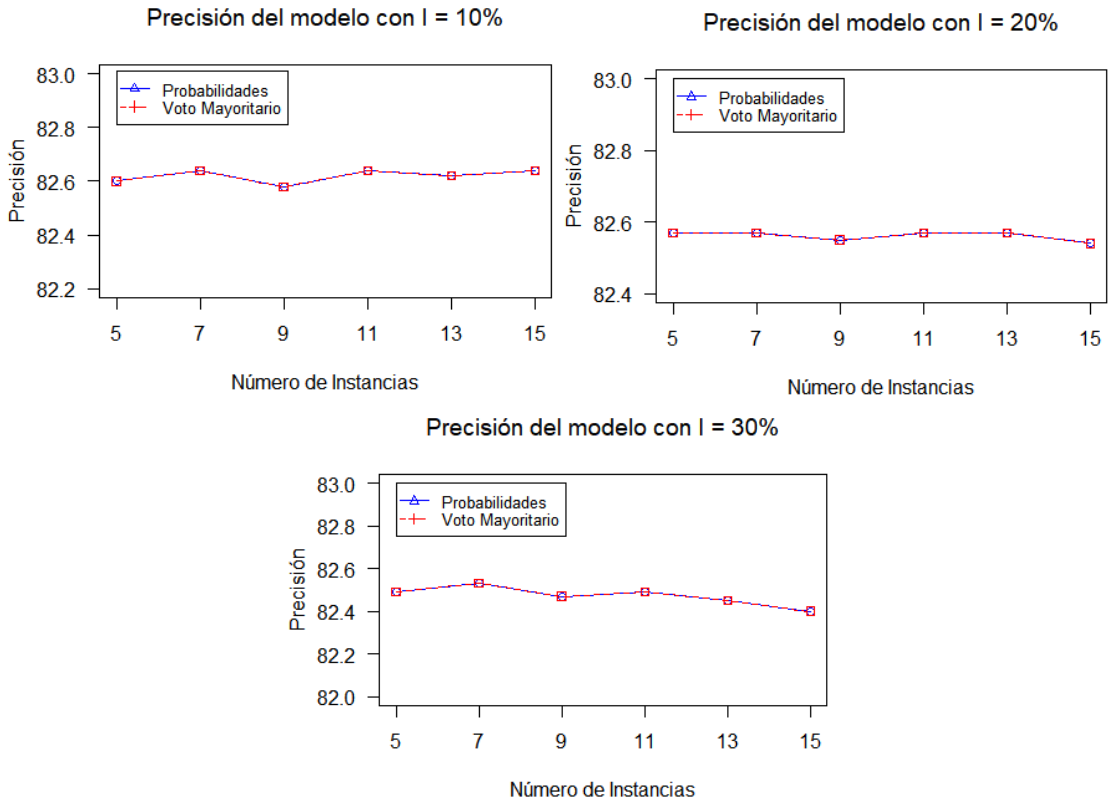



*Figura 53 Resultados (IBM HR) del modelo NB variando su incertidumbre I.*