# IMPLEMENTACIÓN DE TÉCNICAS DE DATA MINING, PARA LA PREDICCIÓN DE LA DESERCIÓN DE LOS ESTUDIANTES DEL PROGRAMA DE INGENIERIA INDUSTRIAL DE LA UNIVERSIDAD ICESI

**EDUARDO JOSÉ POLO SAA**
**SEBASTIAN REYES BAUER**

**UNIVERSIDAD ICESI**
**FACULTAD DE INGENIERÍA**
**PROGRAMA DE INGENIERÍA INDUSTRIAL**
**CALI**
**MAYO 2016**

# IMPLEMENTACIÓN DE TÉCNICAS DE DATA MINING, PARA LA PREDICCIÓN DE LA DESERCIÓN DE LOS ESTUDIANTES DEL PROGRAMA DE INGENIERIA INDUSTRIAL DE LA UNIVERSIDAD ICESI

**EDUARDO JOSÉ POLO SAA**
**SEBASTIAN REYES BAUER**

**Proyecto de Grado para optar el título de Ingeniero Industrial**

**Director proyecto**
**FERNANDO QUINTERO MORENO**

**UNIVERSIDAD ICESI**
**FACULTAD DE INGENIERÍA**
**PROGRAMA DE INGENIERÍA INDUSTRIAL**
**CALI**
**MAYO 2016**

# Contenido

pág.

# Lista de Figuras

# GLOSARIO

**Educational Data Mining:** Es una nueva disciplina, basada en la minería de datos, que se encarga de explorar información del sistema educacional, para posteriormente, con la ayuda de algoritmos, técnicas, modelos, entre otras herramientas, encontrar patrones descriptivos, e incluso realizar predicciones, con el fin de encontrar respuestas a preocupaciones educacionales.

**Data Mining:** Es un proceso orientado a la extracción del conocimiento útil y comprensible, a partir de gran cantidad de información almacenada en bases de datos.

**Modelo descriptivo:** Es una representación simplificada de una determinada realidad. En "*machine learning*", se suele aplicar en el aprendizaje no supervisado, al reproducir los patrones dentro de la información, que permitan evidenciar la relación entre los datos estudiados.

**Modelo predictivo:** Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos. En machine learning, se suele aplicar en el aprendizaje supervisado, para estimar futuros valores de variables dependientes, siendo estas variables continuas o discretas.

**Modelo Cognitivo:** Se caracteriza por estudiar cómo el ser humano conoce, piensa y recuerda, centra su atención en el papel como elabora, crea e interpreta la información el hombre como sujeto pensante. Resalta su preocupación por el desarrollo de habilidades mentales y su representación en el aprendizaje.

***Supervised Learning* (aprendizaje supervisado):** En aprendizaje automático y minería de datos, el aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento.

**Unsupervised Learning (aprendizaje no supervisado):** Aprendizaje no supervisado es un método de aprendizaje automático donde un modelo es ajustado a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. Así, el aprendizaje no supervisado típicamente trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos.

**Red neuronal:** Una red neuronal es un sistema compuesto de muchos elementos procesadores simples operando en paralelo, cuya función es determinada por la estructura de la red, fuerza en las conexiones y el procesamiento realizado por los elementos computacionales en los nodos.

**Cluster:** Consiste en la agrupación de los datos acorde a la similitud entre los mismos.

**Algoritmo:** Conjunto ordenado y finito de operaciones que permite hallar la solución de un problema. El algoritmo consiste en un método para resolver un problema mediante una secuencia de pasos a seguir.

**Método:** Proceso o camino sistemático establecido para realizar una tarea o trabajo con el fin de alcanzar un objetivo predeterminado.

**Técnica:** Se refiere a los procedimientos y recursos que se emplean para lograr un resultado específico, detallando la manera en que han de realizarse dichos procedimientos.

**Deserción Universitaria:** Puede entenderse como el abandono del sistema educativo por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno.

**Deserción precoz:** Aquella presentada cuando un estudiante es admitido en un programa educativo, pero por diferentes motivos decide no matricularse.

**Deserción temprana:** Aquella deserción que se presenta en los primeros cuatro semestres de un programa educativo, en el caso de las universidades.

**Folds:** se refiere a la cantidad de divisiones que se realizan sobre el total de datos, para obtener aleatoriamente varios grupos, que para este proyecto serán grupos de estudiantes, con el objetivo de llevar a cabo la técnica de validación cruzada.

# 1. CAPITULO I. Definición del Problema

## 1.1 Contexto del Problema

Durante mucho tiempo se concibió el rendimiento académico como indicador del nivel de exigencia y calidad de las instituciones educativas. Sin embargo, el actual enfoque educativo apunta a la medición del aprendizaje en términos de competencias y habilidades, no solo disciplinares, sino también en relación a la formación personal, pensamiento crítico, creatividad, trabajo en equipo y aprendizaje para el resto de la vida. En este sentido, los sistemas evaluativos vienen siendo reestructurados para incorporar evidencias de aprendizajes en términos de realidades prácticas del conocimiento, en la habilidad para aprender, y no centrado en la consolidación del conocimiento aplicado o inerte.

De otro lado pero en el mismo contexto, las realidades y expectativas de las nuevas generaciones encuentran dificultades frente a los estilos tradicionales de enseñanza y evaluación, ocasionando un fenómeno de alta deserción en las carreras universitarias, al cual se le han ido estudiado sus causas. En Colombia, se ubica entre el 40% y 50% de los estudiantes universitarios (Dinero, 2015), y en el programa de pregrado de Ingeniería Industrial de la Universidad Icesi el panorama no es diferente, pues se presenta una deserción mayor al 25% durante los primeros cuatro semestres cursados. En una noticia reciente del periódico El Espectador titulada “Deserción, ¿qué estamos haciendo mal?”, se plantea como esta problemática surge al tiempo que se ha logrado un aumento en la cobertura de la educación, y agravada al interior de las instituciones de alta calidad. Son cifras equivalentes de deserción respecto a las que se presentan en otros países latinoamericanos pero muy altas comparadas con países desarrollados, estas por debajo del 25%.

El abanico causal de la deserción incluye asuntos financieros (inestabilidad económica), adaptación al medio universitario, la tradición cultural en el entorno

familiar y social, ausencia de mejores bases desde la educación secundaria (especialmente en matemáticas y lecto-escritura), didácticas de clase o enfoques de enseñanza-aprendizaje que desmotivan, priorización en contenidos y no en habilidades, restricciones al acceso o costos de los créditos educativos, programas escogidos por intereses y no por vocación, están entre tantas posibilidades.

De otro lado, el desarrollo de la minería de datos (*Data Mining*) ha abarcado el sector educativo para estructurar lo que se conoce como EDM (*Educational Data Mining*) que busca detectar patrones en los comportamientos entre datos almacenados, con el fin de encontrar los factores de mayor incidencia en múltiples procesos educativos. Muchos de estos estudios son asociados a la predicción del rendimiento académico o a la posibilidad de desertar o no. En efecto, apelar a las técnicas implícitas en la minería de datos ha determinado un campo de estudio que busca predecir desempeños académicos o posibilidades de finalización de estudios, teniendo como punto de partida los atributos diferenciados de la psicología del estudiante, una vasta información de su entorno familiar y social, sus antecedentes académicos y extracurriculares, caracterización de sus instituciones previas, el estilo de aprendizaje-enseñanza identificado y otras variables.

En el contexto de la Universidad Icesi hay una apuesta por una política incluyente a tener jóvenes de diversos orígenes socio-económicos, a través de la apertura de créditos, becas y el reciente subsidio a la demanda que ha otorgado el gobierno nacional en el programa “Ser pilo paga”. Esta diversidad amplía aún más el reto para detectar variables causales y de incidencia en la deserción y en el rendimiento académico.

## 1.2 Análisis y Justificación

En la Universidad Icesi se cuenta con un amplio registro de información que incluye aspectos básicos de información personal, escolar, familiar y socio-económica del estudiante, y abre posibilidades para incluir un mayor número de atributos o variables, que enriquezcan la implementación de técnicas del EDM. Establecer los niveles de incidencia de múltiples entradas en los resultados de desempeño académico puede servir para diseñar estrategias en los procesos de selección de estudiantes, y políticas de apoyo y acompañamiento al estudiante en su vida universitaria, soporte o flexibilización crediticia y hasta replanteamiento de los niveles de control académico que aseguren la permanencia del estudiante y la respectiva finalización de sus estudios.

En el panorama actual, el proceso de admisión se ha centrado en la ejecución de filtros para el acceso directo, luego, el paso por entrevista para quienes no cumplan el primer filtro, y finalmente una tercera instancia para definir los admitidos en un semestre académico. La preocupación inicial durante el proceso de admisión es elegir los candidatos(as) que aseguren un buen rendimiento académico y por ende, su permanencia hasta la graduación. Ante esta situación, cabe preguntar si los actuales criterios de filtro, tienen un sustento basado en los antecedentes de la población, o por si el contrario existe la posibilidad de detectar atributos relevantes que marquen patrones de deserción. De igual manera, se podrían anticipar alarmas justificadas sobre recién admitidos que pudiesen estar en un cierto estado inicial propenso a la deserción, y sobre quienes se pudiesen dirigir estrategias de acompañamiento y mayor apoyo.

## 1.3 Formulación del Problema

El programa de pregrado de Ingeniería Industrial de la Universidad Icesi trabaja en la disminución de la tasa de deserción de sus estudiantes matriculados en los primeros cuatro semestres y si bien hay una experticia que identifica eventos causantes de la deserción, no cuenta con estudios de evaluación de incidencia y potencial predicción basados en la caracterización de los candidatos, según antecedentes poblacionales recientes.

# 2. CAPITULO II- Objetivos

## 2.1 Objetivo General

Contribuir con información y relaciones relevantes, al objetivo de reducción de las tasas de deserción en el programa de pregrado de Ingeniería Industrial de la Universidad Icesi.

## 2.2 Objetivo del Proyecto

Proporcionar una herramienta basada en minería de datos que contribuya a determinar la propensión de un perfil de estudiante a desertar del programa de Ingeniería Industrial de la Universidad Icesi.

## 2.3 Objetivos Específicos

1. Clasificar la información de la población estudiantil, en perspectiva de su adecuación, para posterior implementación de técnicas estadísticas y de Data Mining.
2. Experimentar y validar modelos de predicción de la deserción a partir de la base de datos construida.
3. Construir una herramienta de apoyo a los procesos de admisión y al diseño de estrategias que sustentan el acompañamiento al estudiante en su vida universitaria.

# 3. CAPÍTULO III. Marco de Referencia

## 3.1 Antecedentes o Estudios Previos

Retornando a una visión global, la educación viene siendo repensada en términos del contexto social, económico y ambiental, para desarrollar formas de pensamiento, creatividad y aplicación del conocimiento y generación del nuevo, que ayude a solucionar problemas. Por mucho tiempo hubo una fijación casi exclusiva a supeditar todo al rendimiento académico, sin considerar competencias y habilidades para la vida y la profesión. Si bien hoy este enfoque aún predomina, se empiezan a generar procesos de mayor y mejor aseguramiento de las competencias de formación disciplinar y complementarias en el desarrollo humano. Para tal propósito, se plantea un dilema al querer sostener el marco evaluativo que ayude a identificar asociaciones cuantitativas del aprendizaje, al tiempo que se evalúan factores incidentes en el rendimiento académico, en lo posible, más alineado con las habilidades de formación.

Debido a lo anterior, varias investigaciones se han desarrollado con el fin de relacionar este fenómeno a través de diferentes aspectos, que según (Peña-Ayala, 2014) en su encuesta sobre el desarrollo de *Educational Data Mining* (EDM), hasta el año 2014 se podían contabilizar 240 publicaciones. En este mismo informe, se hace énfasis en los tipos de modelos de EDM que existen, siendo estos de dos tipos: descriptivos y predictivos.

Los descriptivos, aplican generalmente sistemas de aprendizaje no supervisados de *Machine Learning*, caracterizados por ser algoritmos que permiten a la máquina la posibilidad de aprender y encontrar patrones en bases de datos grandes, sin necesidad de proporcionar parámetros o categorías previas (valores asignados). Entre algunos de sus objetivos está explicar o generalizar algún comportamiento, encontrar relaciones e interconexiones entre los datos minados y buscar estructuras intrínsecas en los datos proporcionados. Por su parte, los predictivos aplican funciones de aprendizaje

supervisado, donde se le brinda al algoritmo información necesaria sobre las variables de entrada, incluyendo instrucciones sobre el tipo de dato (categoría) y sus posibles valores (comportamiento). Estos, se enfocan generalmente en la estimación de valores desconocidos o futuros de variables dependientes, en este caso lo que se desea predecir, basados en atributos de las variables independientes, en este caso las de entrada.

La aplicación de estos modelos en el EDM está dirigida a la búsqueda de seis funcionalidades que se pudieron identificar en el informe según (Peña-Ayala, 2014). Estas funcionalidades son: *Student Behavior Modeling, Student Performance Modeling, Assesment, Student Modeling, Student Support & Feedback y Curriculum, Domain Knowldedge, sequencing & teacher support*.

| Functionalities | Counting | Percentage(%) | Accumulative counting | Acummulative percentaje (%) |
|---|---|---|---|---|
| Student behavior modeling | 48 | 21.62% | 48 | 21.62% |
| Student performance modeling | 46 | 20.72% | 94 | 42.34% |
| Assessment | 45 | 20.27% | 139 | 62.61% |
| Student Modeling | 43 | 19.37% | 182 | 81.98% |
| Student support and feedback | 21 | 9.46% | 203 | 91.44% |
| Curriculum, domain knowledge, sequencing, teacher support | 19 | 8.56% | 222 | 100.00% |
| **Total** | 222 | 100.00% | | |

*Figura 1. Estudios de EDM según sus funcionalidades. (Peña-Ayala, 2014)*

*Student behavior modeling* es un tipo de funcionalidad orientada a dar forma a los aspectos que caracterizan a un estudiante; en las investigaciones o trabajos desarrollados bajo esta perspectiva se tienen en cuenta factores, entre los cuales están las emociones, conocimientos, habilidades, la forma en que aprende, etc. Lo que busca normalmente este tipo de trabajos, es dar bases sólidas que permitan adaptar la enseñanza al tipo de aprendizaje de los estudiantes, con el fin de cumplir con las expectativas de conocimiento que ellos tienen.

*Student performance modeling* es la segunda funcionalidad de la cual se han hecho más investigaciones o trabajos. Esta es orientada a representar y anticipar el rendimiento de los estudiantes. La modelación desde esta perspectiva es usada para

identificar y calcular factores como la eficiencia, hacer evaluación de logros, buscar competencias adquiridas en los estudiantes y asociar el rendimiento a atributos como el tiempo, los recursos consumidos, las deficiencias, etc. El objetivo final perseguido desde este punto de vista es estimar que tan capaz es el estudiante de lograr o aprender un nuevo conocimiento.

*Student support & feedback* es la tercera en importancia para nosotros, pues las otras funcionalidades se alejan de lo requerido para este proyecto. Este tipo de funcionalidad busca, por medio de la modelación, obtener fundamentos que permitan dar sugerencias, quejas, hacer solicitudes y evaluaciones a los estudiantes. El fin último de este tipo de trabajos es mejorar el rendimiento del estudiante por medio del cumplimiento de los logros.

Ya conociendo las funcionalidades más importantes que pueden tener los proyectos o trabajos, es necesario conocer la forma en que estos se pueden implementar. Como característica principal, cada tipo de implementación está asociada a un tipo de modelo (Peña-Ayala, 2014).

En modelos descriptivos, las técnicas de implementación que se conocen, según este informe, son *clustering, reglas de asociación* y *análisis de correlación.* En los modelos predictivos se conocen trabajos e investigaciones que aplican *clasificación, regresión* y *categorización.* Estas técnicas de implementación deben existir en cualquier estudio de EDM y sólo una es aplicada normalmente en el mismo. De acuerdo con (Peña-Ayala, 2014), la técnica más utilizada en los estudios hasta ahora es la de *clasificación*, seguida por la técnica de *clustering*.

Además de tener técnicas de implementación, estos estudios se han caracterizado por tener métodos y/o algoritmos. En el informe se aclara que se han encontrado trabajos que puedan aplicar uno, varios o ninguno.

La siguiente información refleja los resultados de la creación de la base de datos de trabajos sobre EDM en el informe de (Peña-Ayala, 2014). Los métodos más utilizados

son el teorema de bayes y los árboles de decisión. Los algoritmos que más han usado son *los K- Means, Expectaction Maximization* y el *J48.* Por último, las técnicas más utilizadas son la regresión logística y la regresión lineal.

| Metodo | items | Counting | Percentage(%) | Accumulative counting | Acummulative percentaje (%) |
|---|---|---|---|---|---|
| Teorema de Bayes | 1 | 48 | 19.67% | 48 | 19.67% |
| Árboles de decisión | 1 | 44 | 18.03% | 92 | 37.70% |
| Instance-based learning | 1 | 22 | 9.02% | 114 | 46.72% |
| Hidden Markov Model | 1 | 20 | 8.20% | 134 | 54.92% |
| others from 5 to 13 | 5 | 54 | 22.13% | 188 | 77.05% |
| others from 2 to 4 | 11 | 25 | 10.25% | 213 | 87.30% |
| others with 1 | 32 | 31 | 12.70% | 244 | 100% |
| **Total** | 52 | 244 | 100% | | |

*Figura 2.Algoritmos aplicados en estudios de EDM. (Peña- Ayala, 2014)*


| Metodo | items | Counting | Percentage(%) | Accumulative counting | Acummulative percentaje (%) |
|---|---|---|---|---|---|
| Teorema de Bayes | 1 | 48 | 19.67% | 48 | 19.67% |
| Árboles de decisión | 1 | 44 | 18.03% | 92 | 37.70% |
| Instance-based learning | 1 | 22 | 9.02% | 114 | 46.72% |
| Hidden Markov Model | 1 | 20 | 8.20% | 134 | 54.92% |
| others from 5 to 13 | 5 | 54 | 22.13% | 188 | 77.05% |
| others from 2 to 4 | 11 | 25 | 10.25% | 213 | 87.30% |
| others with 1 | 32 | 31 | 12.70% | 244 | 100% |
| **Total** | 52 | 244 | 100% | | |


*Figura 3.Algoritmos aplicados en estudios de EDM. (Peña- Ayala, 2014)*

Entrando en lo particular, en una investigación se puede encontrar información útil que ayude a definir la metodología que luego se implementará. Se han podido identificar diversos factores que han servido como objeto de estudio para este tema, ya que la predicción del rendimiento académico de los estudiantes puede ser influenciada por una amplia variedad de aspectos, lo que ocasiona que haya necesidad de crear límites de estudio para conseguir un modelo adecuado y bien definido. En los siguientes

párrafos se ilustran informes que sirven de ejemplo y permiten obtener información sobre los objetivos que se han planteado, las variables medidas, las técnicas de implementación, algoritmos y métodos empleados, además de la representación o el nivel de ajuste logrado con la creación del modelo.

(Thai-nghe, Horváth, & Schmidt-Thieme, 2011) es una investigación de la Universidad de Hildesheim en Alemania, la cual enfocó su búsqueda en la predicción del rendimiento académico por estudiante, para así recomendar al mismo las áreas de estudio más adecuadas de acuerdo con su desempeño. En este trabajo se utiliza una técnica de implementación de clasificación, abordada desde la disciplina probabilística, utilizando como método principal un modelo de factorización, en conjunto con un algoritmo llamado *Tensor Factorization Forecasting*, el cual se puede describir brevemente como una mezcla de matrices de factorización, cadenas de Markov y métodos de pronósticos, parecido a los sistemas de recomendación.

Las variables principales tenidas en cuenta para el desarrollo del trabajo fueron el rendimiento histórico del estudiante, el rendimiento promedio de la materia y el tiempo transcurrido desde la obtención de los conocimientos previos. El nivel de representación o ajuste se hizo por medio de la medición del RMSE, obteniendo un valor de 0.30159.

(Elbadrawy, Studham, & Karypis, 2014) en su trabajo para la Universidad de Minnesota tuvo como objetivo principal predecir el desempeño de un estudiante en una materia, para saber el riesgo que tenía de perderla y los aspectos más influyentes en la nota. El tipo de implementación es de regresión, utilizando la regresión múltiple como método principal y la ayuda de ecuaciones para la determinación de la relación entre los atributos y los resultados finales de los estudiantes. Para el desarrollo se tuvo en cuenta la historia académica del estudiante (promedios de materias y promedio acumulado), el área del conocimiento del curso (dificultad del curso, departamento al que pertenece, tipo de evaluaciones) y la Interacción con e-learning (frecuencia de entrada, participación en el curso, notas).

El nivel de representación o ajuste también se evaluó por medio del RMSE, con un valor de 0.145. A las conclusiones que se llegaron fueron que entre más regresiones se utilicen, menor va a ser el RMSE y en cuanto a los estudiantes, se determinó que los mayores contribuyentes a la predicción del rendimiento para este caso fueron los sesgos tanto del estudiante como del curso.

(Harwati, Alfiani, & Wulandari, 2015) es un caso de estudio hecho en la Universidad Islámica de Indonesia. Su objetivo fue hacer un mapeo de los estudiantes con el fin de encontrar patrones escondidos y clasificarlos basados en su entorno demográfico, con el fin de tener bases para el diseño de programas con opción de mejora del rendimiento académico. La técnica de implementación es *clustering*, utilizando un algoritmo conocido como K-Mean. Las variables tenidas en cuenta para el desarrollo del caso fueron el perfil académico del estudiante, tomando como base el promedio acumulado, los trabajos importantes del estudiante en la universidad, las notas de laboratorio, además del perfil de actividades de los estudiantes, tales como las tasas de asistencia, la participación en la organización y la zona de origen.

Como resultado final se logró el agrupamiento de los estudiantes en tres categorías, siendo estas estudiantes inteligentes y activos (45.75% de la población estudiada), estudiantes con capacidad promedio de aprendizaje (33.3%) y estudiantes con menos capacidad de aprendizaje (20.91%).

(Natek & Zwilling, 2014) es basado en la creación de un sistema administrativo sobre el conocimiento estudiantil en instituciones universitarias hecho en la escuela internacional de estudios sociales y de negocios en Eslovenia. Su meta más importante fue crear un modelo que permitiera hacer predicción de la nota definitiva en un curso informático a estudiantes con pequeños grupos de datos. El estudio implementa la clasificación como técnica, con métodos que involucran arboles de decisión y algoritmos como el *J48* y el *REPTree*. Las variables que se tuvieron en cuenta fueron en gran parte información sacada de cursos informáticos universitarios, el año de estudios,

género del estudiante, su año de nacimiento y si actualmente trabaja. Estos datos fueron procesados con la herramienta *MS Excel table tools*, utilizada para el DM.

Esta información posteriormente sirvió para la creación de un modelo en un software diseñado para este objetivo llamado WEKA, el cual fue analizado con varios algoritmos con el fin de observar cual tenía mayor porcentaje de representación. Como resultado final se logró un porcentaje de predicción del 65%, con el algoritmo REPTree, además se concluyó que no siempre es necesario tener grandes cantidades de datos para lograr una predicción aceptable.

(Kabakchieva, Stefanova, & Kisimov, 2009) se enfoca en los resultados de un caso de estudio hecho entre la Universidad Nacional y la Universidad de Sofía, ambas de Bulgaria, el cual buscaba encontrar patrones interesantes en la información disponible de las universidades para predecir el rendimiento basado en las características personales y pre universitarias de los estudiantes. Las variables tenidas en cuenta fueron el género, año de nacimiento, lugar de nacimiento, ubicación de la vivienda (Rural o Urbana), el perfil, el colegio y la puntuación total conseguida en este, además del año de admisión universitaria, puntuación en exámenes de admisión y las calificaciones de la universidad.

El caso anterior, se llevó a cabo bajo el uso de un Software llamado WEKA, en el cual se crearon los parámetros o categorías que se tuvieron en cuenta para la variable clasificatoria o de salida, siendo estas cinco de acuerdo al puntaje total universitario (Excelente, muy bueno, bueno, promedio, bajo y muy bajo). Este software permite el uso de varios algoritmos clasificadores, lo que ayuda a comparar cual se adapta y predice mejor de acuerdo a los datos de entrada; los que se usaron en este caso fueron el árbol de decisión J48, clasificadores de Bayes como *NaiveBayes* y *BayesNet*, el algoritmo k del vecino más cercano (kNN), (IBk) y dos algoritmos de estudio de reglas (OneR y JRip).

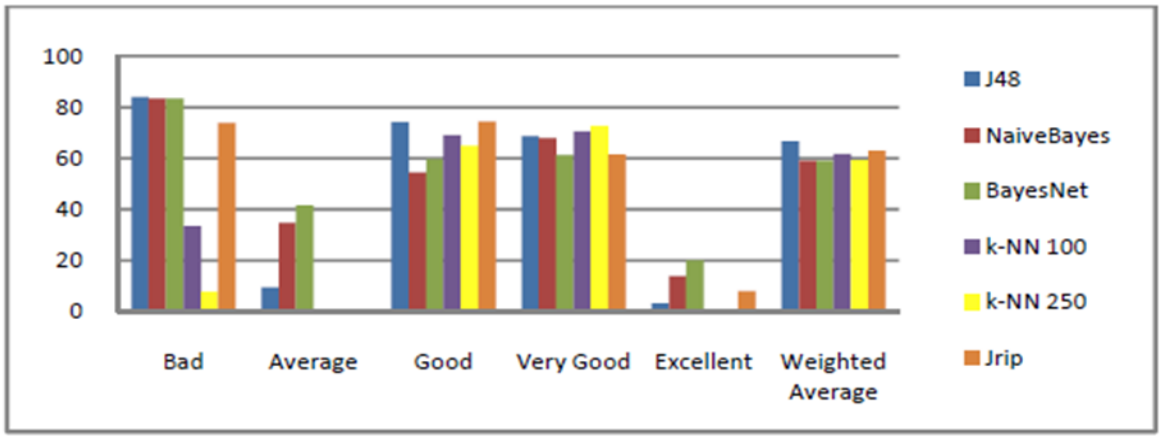



*Figura 4. Porcentajes de predicción alcanzados por los algoritmos. (Kabakchieva, Stefanova, & Kisimov, 2009)*

Los resultados del test mostraron que el árbol de decisión J48 tuvo un mejor desempeño que los demás, seguido por el algoritmo de estudio de regla JRip, sin embargo la predicción promedio de las clases fue inferior al 70%, en parte por el bajo porcentaje de predicción en las clases excelente y promedio.

(Marquez-Vera, Romero, & Ventura, 2011) muestra un estudio hecho para la Universidad de Zacatecas en México el cual buscaba encontrar los aspectos más relevantes en la predicción del rendimiento académico de los estudiantes. Los datos que se tuvieron en cuenta para el desarrollo del modelado de información fueron información de la misma universidad que databa un periodo de quince años, sacada principalmente de tres encuestas que tenían, entre otras, indagaciones sobre características personales y familiares de los estudiantes, entorno socio económico de los mismos y las notas que sacaron en las diferentes materias a lo largo del curso, para un total de setenta y siete atributos. La variable de salida que desearon predecir fue rendimiento final, cuyos resultados posibles fueron pasar o no pasar.

En total se obtuvo información de seiscientos setenta estudiantes, de los cuales seiscientos pasaron y setenta no. Para la creación del modelo también se utilizó el software WEKA, en el cual simularon los datos para diez tipos de algoritmos clasificadores los cuales fueron cinco algoritmos inductores (JRip, NNge, OneR, Prism y Ridor) y cinco tipos te arboles de decisión (J48, SimpleCart, ADTree, RandomTree y

REPTree). De los setenta y siete atributos, se logró reducir a los quince más importantes de acuerdo a la frecuencia con que el software los utilizó para la clasificación, sin dañar la representación de los mismos.

Después de esto, se observó que existía un problema por la clasificación de los datos la cual era no balanceada (600 y 70), lo que hacía que los resultados no fueran del todo representativos. Para esto se utilizó un algoritmo re balanceador de datos que tiene el software, llamado SMOTE (Synthetic Minority Over-sampling Technique), que lograba tener un 50-50 en las dos clases para compararlas equitativamente y como comparación alternativa, un sistema de clasificación costo-sensitivo que otorgaba una importancia cuatro veces mayor a los estudiantes que no pasaban. Los resultados muestran que el algoritmo de árbol de decisión J48 obtuvo el mejor desempeño clasificando a los estudiantes que no pasaron (93.3%), con media geométrica del 94.6%.

Como conclusiones, se logró identificar que los quince aspectos más importantes en la predicción del rendimiento, para estos datos, fueron entre otros: las notas de física, humanidades, matemática e inglés, además de atributos como la edad, tener más de un hermano, estudiar en la noche, el salón y los estudiantes con los que ve la materia, y el nivel de motivación hacia la carrera.

De acuerdo con el artículo (Ramesh, Parkavi, & Ramar, 2013), que plantea estudio con modelos de predicción, se encuentra que la ocupación de los padres, juega un rol importante en la predicción de las notas académicas para los estudiantes en la escuela secundaria. Por otro lado, se identifica que el tipo de escuela no influye en dicho rendimiento. La técnica más acertada, con que lograron un nivel de 72.38% de exactitud en la predicción, fue *Multi Layer Perception Classifier*. Para dicho modelo, se consideraron varios aspectos sobre el contexto en que se encuentra estudiante, incluyendo sus hábitos alimenticios, zona donde vive, número de integrantes de la familia, intereses deportivos, entre otros, para determinar que tanto influyen estos factores a nivel psicológico y emocional, y recomiendan incluir una lista más extensa, puesto que esto se ve reflejado en su rendimiento académico.

Siguiendo en la misma dirección del tipo de estudio, (AL-Malaise, Malibari, & Alkhozae, 2012) considera modelos de predicción, basado en la técnica *multi agent data mining*. Se aplica la metodología de estudio por internet, e-learning, y se obtienen datos directamente de la plataforma Moodle, tomando como atributos la cantidad de exámenes realizados, cantidad de preguntas buenas sobre el total, nota final, entre otros, son datos directamente relacionados con el rendimiento del estudiante. Luego, a través de técnicas de clasificación de datos y de reconocimiento de patrones, logran obtener modelos de predicción. Se utilizan herramientas como árboles de decisión, matriz de confusión y *ensemble of clasiffiers*. Posteriormente, emplean herramientas como *Logitboost, SAMME, AdaBoost* para mejorar los modelos obtenidos, al aumentar considerablemente el porcentaje de aproximación o exactitud de los datos obtenidos, que fue del 80%.

Por otro lado, se encontró un artículo que considera el modelo descriptivo como fuente de patrones, que permiten evidenciar las correlaciones entre el orden en que el estudiante presenta sus materias y respectivos exámenes, frente a las notas obtenidas finalmente. Se trata de (Campagni, Merlini, Sprugnoli, & Verri, 2015), que incorpora técnicas de *clustering*, como *K-means*, que indica la cercanía de los datos obtenidos, dado sus atributos*. SPAM y CloSpam*, fueron los algoritmos para reconocer patrones, que luego, junto con *bubblesort*, fueron incorporados en ciertos grupos, para ir definiendo sus respectivos posibles resultados. Finalmente, encuentran que los estudiantes de mayor rendimiento académico, fueron aquellos que siguieron el plan curricular tal y como este fue definido, y sin presentar ningún tipo de atraso en la presentación de los respectivos exámenes para cada materia. Por otro lado, se propone indagar un poco más acerca de aspectos socio-culturales, ya que se evidenció que los estudiantes presentan las materias acorde a sus gustos y habilidades, lo cual podría permitir identificar la carrera ideal para los estudiantes, quienes a veces evitan ciertos estudios, debido a sus propósitos como profesionales.

Continuando con los múltiples factores que se consideran al momento de encontrar patrones para determinar el rendimiento académico del estudiante, se logra evidenciar que, otro aspecto importante a considerar es el nivel de conocimiento y habilidades que tiene el estudiante, e incluso ver la facilidad con que las puede aprender. Es por ello, que el artículo (Thai-nghe & Drumond, 2011) se concentra en definir un nivel de aprendizaje y adaptación del estudiante para con el nuevo conocimiento. Se pretende cuantificar la capacidad del estudiante para responder acertadamente, sea porque sabe del tema, o porque su misma personalidad y actitud lo conlleva a hacerlo.

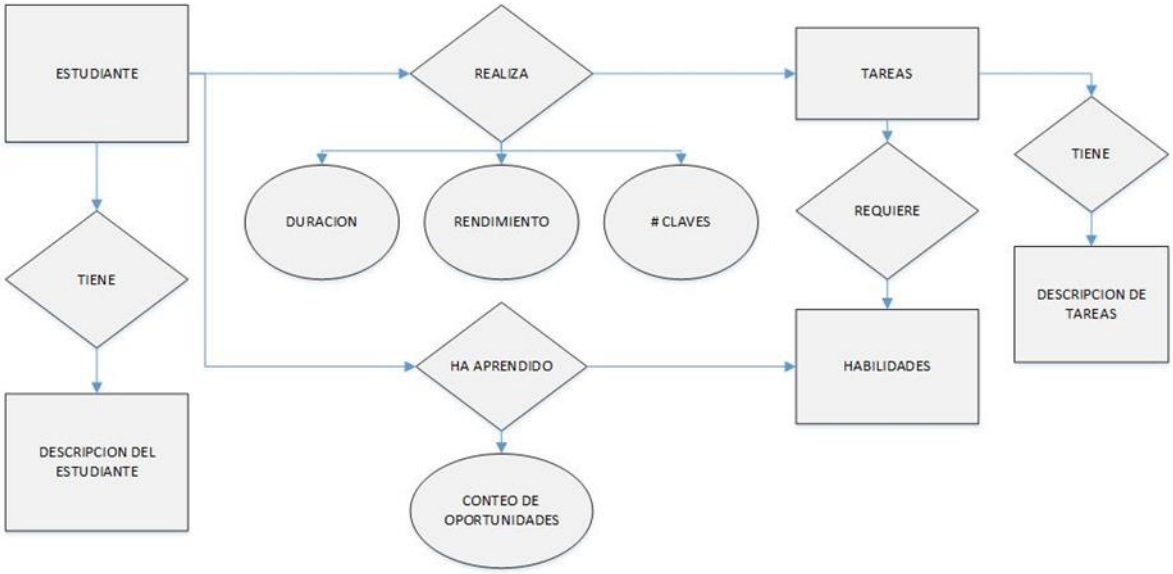



*Figura 5.Proceso (Thai-nghe & Drumond, 2011)*

Esta imagen muestra las conexiones que se consideran entre las habilidades y el estudiante, donde el estudiante requiere ciertas destrezas para cada tarea, que pueden hacer parte de una o varias materias, pero que finalmente se deben poner a prueba. Utilizando *multi-relational matrix factorization*, logran dimensionar las relaciones entre estudiante y tareas y habilidades, creando un modelo cognitivo que incorpora los

factores de aprendizaje, acorde a las características del estudiante y a su respectiva situación que incluye oportunidad de aprendizaje.

En relación a la Universidad Icesi, el análisis de la deserción en los diferentes programas ha sido estudiado mayormente en el orden cualitativo, y complementado recientemente con un estudio de correlación de factores incidentes en el rendimiento académico. Se intentó conocer de cerca la metodología y los resultados de dicho estudio pero no hubo respuesta del principal responsable.

## 3.2 Marco Teórico

Los motivos que llevan a la deserción, en el caso de la Universidad Icesi, son variados y no solamente dependen del rendimiento académico, por eso se considera conveniente analizar todos los aspectos posibles que influyan.

El aprendizaje supervisado se elige cuando se quiere predecir el comportamiento de los datos y descubrir qué algoritmo o método se acomoda más, con el fin de que esta predicción sea lo más acertada posible. El aprendizaje no supervisado, por su parte, es recomendable cuando se quieren encontrar relaciones o patrones en la información suministrada, con el fin de establecer los diferentes grupos de datos que se pueden formar a partir de la misma. Como el objetivo del proyecto es predecir la deserción, el aprendizaje supervisado es adecuado. Sin embargo, la sola elección del tipo de aprendizaje que se pretende establecer, no es suficiente para empezar la implementación, pues existen dos tipos principales de técnicas para implementar este tipo de aprendizaje: clasificación y regresión.

La clasificación se utiliza cuando el atributo principal que se pretende analizar tiene valores discretos (Ejemplo: 1 o 0) o puntuales (ejemplo: sí/no), es más efectivo cuando los demás atributos o variables presentan estos mismos tipos de datos y

permite que los porcentajes de predicción del modelo puedan ser elevados, en comparación con la técnica de regresión.

La regresión, por su parte, se utiliza cuando el atributo que se desea analizar tiene valores continuos (Ejemplo: promedios académicos) y busca encontrar cual función se puede acomodar más al comportamiento de los datos.

El porcentaje de predicción usado para evaluar los datos, luego de escoger el tipo de aprendizaje y la técnica a usar, es obtenido por medio de los clasificadores (Classifiers), los cuales fueron presentados como métodos y algoritmos en el capítulo anterior. El objetivo de su uso, es el aporte que le dan a la toma de decisión, para la elección del que mejor represente el comportamiento de los datos.

En casos donde no haya un porcentaje de predicción elevado, estos también pueden informar sobre atributos que influyen mayormente en la predicción, con el fin de eliminar aquellos otros que generan ruido en el modelo y no permiten un análisis adecuado, además de informar sobre la calidad de los datos que se proporcionan.

Con *Data Mining* se busca entender, entre otras cosas, los datos estadísticos y encontrar los factores que influyeron en dichos datos.

# 4. CAPITULO IV – Metodología

A continuación se presentará la metodología que se pretende seguir para el cumplimiento de cada uno de los objetivos específicos planteados en el segundo punto del proyecto de grado, que a su vez contribuyen a que se desarrolle el objetivo final de este Proyecto de Grado.

## 4.1 Clasificar la información de la población estudiantil, en perspectiva de su adecuación, para posterior implementación de técnicas estadísticas y de Data Mining

Por medio de una petición hecha al Programa de Ingeniería Industrial de la Universidad Icesi, se pretende obtener información completa sobre la población estudiantil que ha ingresado al programa durante los últimos años. Los contextos académico, social, económico, emocional y psicológico de los estudiantes, reúnen atributos que pueden influir significativamente en la deserción. Organizar la información tiene gran relevancia, pues permite conocer la situación actual de deserción en el programa de Ingeniería Industrial y colabora a la elección adecuada de la técnica a usar para el análisis y construcción del modelo. En este caso, la técnica de implementación escogida es la de clasificación, con el enfoque predictivo que se busca en el proyecto y se acomoda perfectamente al tipo de resultado que se desea analizar (Deserción: Sí o No), una variable discreta.

Por lo anterior, se deben crear categorías y definir los atributos que conformarán una base de datos clasificada y categorizada, indicando el tipo de variable como los valores que puede tomar. Se debe buscar que los atributos escogidos para hacer todo el procedimiento sean lo suficientemente claros y contribuyan a que se logre el objetivo principal. En este sentido, se organiza la información en tablas de Excel, para luego reconocer y descartar atributos o variables que sean irrelevantes para el modelo

(Ejemplo: nombres y códigos), además de la discretización de otros, que generalmente tienen valores continuos (Ejemplo: promedios).

La base de datos debe ser dividida en dos grupos de: el de entrenamiento y validación. El primero busca implementar aprendizaje de máquina por medio de la aplicación de algoritmos a un subgrupo de los datos, el cual es prudente que sea mayor o igual al 60% de la muestra, y lo segundo busca validar si la información encontrada tiene la suficiente precisión para ser considerada una buena aproximación al comportamiento de los datos. Esta es la razón de la división de los datos y su respectivo uso. Además, es necesario que el tamaño de cada grupo sea lo suficientemente grande para que el modelo obtenga buenos resultados, especialmente en el test de entrenamiento, que toma como parámetro de comparación, lo sucedido y lo simulado.

Para este trabajo de grado se implementará una división de:

- 70% - datos de entrenamiento
- 30% - validación de los datos

Para esta división, se debe buscar que la proporción de personas que desertaron versus las que no, sean respetadas a la generada de la base de datos general, es decir que si en el total de datos la proporción es 50-50, esta debe ser la misma tanto para el test de entrenamiento como para el de validación. Esto tiene como fin último que haya un equilibrio que permita comparar, con mayor precisión los datos reales con el modelo creado, y conocer el porcentaje de predicción acertado y su variación, indicadores sobre los cuales se escogerá el modelo que mejor se adapte.

## 4.2 Experimentar y validar modelos de predicción de la deserción a partir de la base de datos construida.

El modelo como tal, consistirá en la incorporación de una serie de algoritmos y técnicas de *Educational Data Mining*, que permita asimilar la relación entre aquellos atributos de mayor importancia, con su respectivo impacto en la posible deserción del estudiante, proporcionando información relevante sobre aquel individuo que se esté evaluando. La información otorgada por el modelo, será con variables discretas, donde se identificarán categorías para comprender el panorama de deserción.

Con respecto a los algoritmos y técnicas a incorporar, se utilizará WEKA (Waikato Environment for Knowledge Analysis), una herramienta computacional de uso libre que implementa *machine learning*, y que tiene como objetivo, brindar algoritmos de minería de datos. Dentro de esta herramienta, se realizarán los siguientes pasos:

1. Pre-procesar la información, para eliminar datos atípicos, valores extremos, entre otros, logrando identificar la información adecuada previamente.
2. Identificar atributos de mayor impacto mediante técnicas de *ranking*, para trabajar con los atributos más relevantes.
3. Implementar algoritmos de clasificación, que acorde con su nivel de precisión frente a los datos históricos, se escogen los más efectivos.
4. Elegir y validar el modelo que mejor represente la función de los datos.

Durante el proceso de validación de modelos, dentro del total de datos, se decide seleccionar un 30% para esta etapa. Con este porcentaje se intenta hacer una mejor aproximación a los parámetros de la investigación, al verificar el nivel de aproximación presentado entre el testeo y entrenamiento de los datos. Esta primera etapa, permite comparar los resultados arrojados por el 70% de los datos, con el 30% restante, lo cual

hace referencia a una primera validación, con datos reales y que tiene como objetivo, definir oportunamente el modelo.

Una de las técnicas a implementar es validación cruzada, ya que define el modelo final, como el promedio de varios modelos obtenidos a partir de diferentes evaluaciones. Esta consiste en definir una serie de posibles combinaciones que puedan presentarse dadas las proporciones de datos de entrenamiento y testeo, y pueden definirse ya sea por subconjuntos, o tomas aleatorias de carácter individual. También se pueden asignar tan solo unos cuantos datos a ser validados, con el resto de información como entrenamiento.

A continuación, se presenta un gráfico que ejemplifica la técnica de validación cruzada de K iteraciones, que permite reconocer como se van definiendo y asignando los conjuntos a validar.

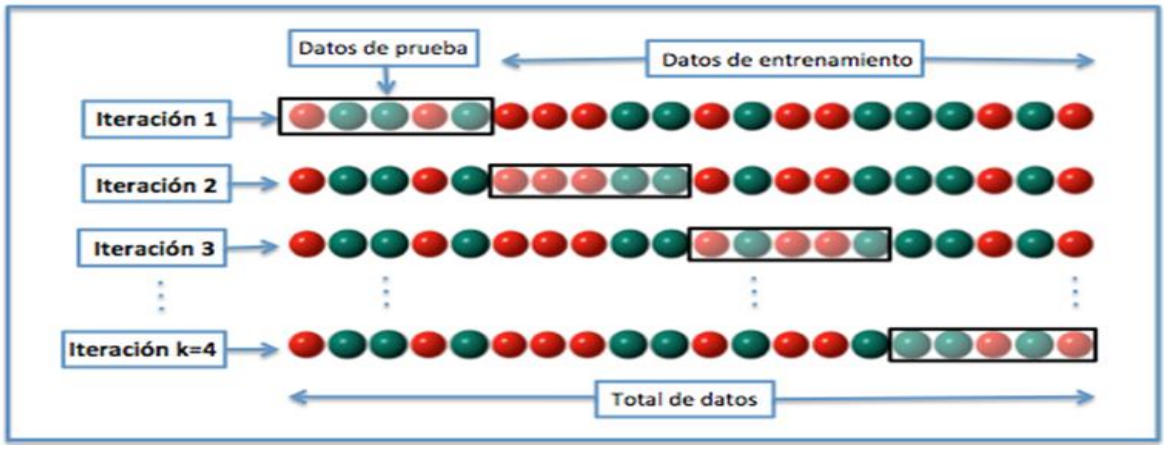



*Figura 6.Validación Cruzada (Wikipedia)*

Nota: La imagen fue extraída de Wikipedia, dado que es la que mejor ilustra el concepto de validación cruzada, que es expuesto a lo largo de este proyecto, bajo otras fuentes más confiables.

## 4.3 Construir una herramienta de apoyo a los procesos de admisión y al diseño de estrategias que sustentan el acompañamiento al estudiante en su vida universitaria.

Con el modelo de predicción elegido, se construirá una herramienta de predicción de probabilidad de deserción, y así apoyar los procesos de admisión y acompañamiento al estudiante. Esta etapa consiste en demostrar la efectividad del modelo, que se define como la cantidad de predicciones acertadas sobre el total de predicciones. Como técnica de validación, se ejecuta una matriz de confusión, la cual consiste en establecer las categorías de predicción dentro de una matriz, teniendo tanto vertical como horizontalmente las mismas categorías, pero unas haciendo referencia a datos reales y las otras a la predicción realizada.

| | | REALIDAD | |
|---|---|---|---|
| | | ESTUDIANTE DESERTA | ESTUDIANTE NO DESERTA |
| PREDICCION | ESTUDIANTE DESERTA | X | Y |
| | ESTUDIANTE NO DESERTA | Y | X |

*Figura 7, Matriz de confusión*

En la Figura 7 registra una matriz con dos categorías que indican si el estudiante deserta o no. Los valores que toman X corresponden a predicciones correctas, y los valores en Y 2 tipos de errores. Un error, es predecir que un estudiante va a desertar, cuando en realidad no lo hace, el otro error es decir que el estudiante no va a desertar cuando en cambio, si lo hace. Con una matriz similar, se espera realizar la validación final del modelo, midiendo su efectividad en la predicción, al comparar las predicciones correctas frente a las incorrectas. Esto demuestra el nivel de confianza del modelo.

# 5. CAPITULO V - Objetivos

## 5.1 Clasificar la información de la población estudiantil, en perspectiva de su adecuación, para posterior implementación de técnicas estadísticas y de Data Mining

### 5.1.1 Obtención de la Información de los estudiantes (Base de Datos)

Antes de empezar con el desarrollo de este primer objetivo, fue necesario solicitar, al programa de Ingeniería Industrial, el histórico poblacional de estudiantes matriculados en el programa entre los años 2005 y 2014, sumando un total de 20 semestres, advirtiendo que los estudiantes del cohorte 14-2 se encuentran cursando su cuarto semestre a la fecha presente. A partir de ahora esta información se le identificará como base de datos (BD).

La base de datos es una mezcla de tipo secundaria y primaria, gran parte de ésta proviene de información directa entregada por los mismos estudiantes cuando se inscribieron a los procesos de admisión y otra ha sido calculada y registrada por parte de la Universidad. Por motivos de confidencialidad, la BD tiene los códigos de los estudiantes trocados, respetando entonces la identidad y privacidad de cada estudiante y los atributos son asignados a un código aleatorio anónimo.

### 5.1.2 Unión de las Bases de datos

Es importante mencionar que la información fue entregada en dos fases. Primero se recibió información directamente de aquella disponible por el programa y la segunda

fue extraída desde Simbiosis, el sistema de información institucional de la universidad. A continuación se detalla el tipo de información.

La primera base de datos consta de un total de mil ochocientos dieciséis (1816) registros de estudiantes ingresados desde la cohorte 51, hasta la cohorte 142, cuenta con registro único para cada estudiante, es decir que no existen códigos repetidos entre los mismos.

La segunda parte de la información consta de un total de once mil seiscientos noventa y cuatro (11694) registros, desde la cohorte 972 hasta la cohorte 142 y posee registros múltiples para cada código, ya que detalla la evolución de los atributos de cada estudiante por cada semestre cursado.

Por lo descrito, es necesario que quede un registro por código del estudiante en la segunda BD, única manera posible de unir las dos bases de datos. Con múltiples registros, la función BUSCARV de Excel falla, y ante este inconveniente, se procedió a automatizar la creación de registros únicos con la programación de una macro. Primero, la cantidad de códigos obtenidos a través de filtros avanzados en esta BD fue de mil novecientos noventa (1990), cifra más cercana a la primera base de datos, y luego se creó la interfaz para que funcionará la macro, la cual por medio del código debía encontrar todos los registros posibles y ordenarlos del último al primero. Esto con el fin de copiar los datos del registro más actualizado y pegarlos en una hoja nueva que tenía los mismos campos de la original. Al ser estos 1990 códigos, se decidió crear una función repetitiva que permitiera obtenerlos solo con correr una vez la macro.

Definida la segunda BD con un total de 1990 registros, se efectuó el proceso de unión, sin embargo hubo quinientos setenta y siete (577) registros de estudiantes que tienen estado indeterminado o estado de estudiantes admitidos pero que no se han matriculado. En la Figura 8 se muestra la proporción de registros existentes no útiles

donde el 32% corresponden a aquellos estudiantes que pasan el filtro creado por la universidad y son admitidos, pero por diferentes motivos no ingresan al programa.

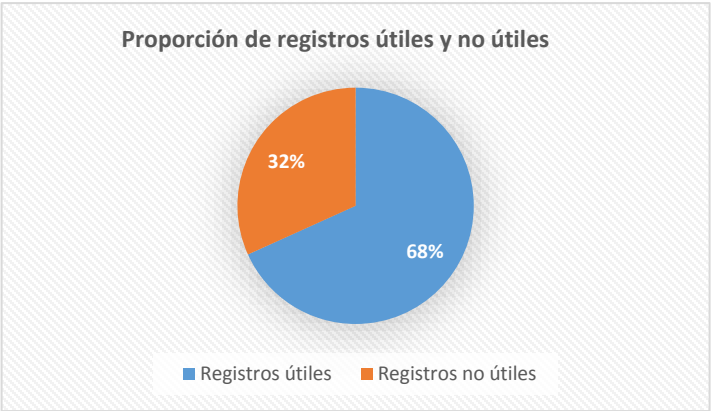



*Figura 8. Total de registros entre útiles y no útiles*

De acuerdo con lo anterior, se eliminan estos registros no útiles en cada base de datos, dejando 1240 registros en la primera BD y 1413 en la segunda. Se procede a la unión de las dos bases de datos por medio del complemento de Microsoft Excel llamado Power Query, el cual a través de consultas permite ejecutar diferentes opciones, como combinar tablas o bases de datos con un atributo en común, en este caso, el código real del estudiante para luego si trocar los registros y conservar el anonimato estipulado. Los resultados finales de dimensionamiento de la BD única es un total de 1240 registros.

### 5.1.3 Descripción de la Base de Datos

La BD única consta de mil doscientos cuarenta (1240) registros de estudiantes ya graduados, retirados, en estado de prueba y cursando normalmente. Las categorías o atributos almacenados para cada estudiante fueron treinta y ocho (38). A continuación, una breve descripción de algunas categorías.

| ATRIBUTOS PARA CADA ESTUDIANTE | | |
|---|---|---|
| Código | PRO_ACTUAL | MATRICULADAS_PERIODO_ACTUAL |
| Período de ingreso | NO_PROGRAMAS | CARGO_PADRE |
| puesto_icfes | FECHA_ICFES | CARGO_MADRE |
| Estado | BECA | DIMENSION_01 |
| Fecha de nacimiento | OTRA_BECA | DIMENSION_01_VALOR |
| Género | BECA_GOBIERNO | DIMENSION_02 |
| Estado civil | PROMEDIO_PERIODO | DIMENSION_02_VALOR |
| Ciudad de nacimiento | PROMEDIO_ACUM | DIMENSION_04 |
| Colegio | NUMERO_PRUEBA | DIMENSION_04_VALOR |
| Categoría del colegio | NUMERO_RETIRO | Puntaje ICFES |
| Estrato | CANCELACIONES | CURSADAS |
| Admitido | PRO_INICIAL | PERDIDAS |
| DIMENSION_03_VALOR | DIMENSION_03 | |

*Figura 9. Atributos por cada estudiante*

**Código:** Se refiere al código de estudiante asignado por la universidad Icesi, el cual por temas de confidencialidad fue trocado. La BD antes de trocar se halla en la oficina de la dirección programa permitiendo una futura adición de información.

**Período de ingreso**: Corresponde al año y semestre en que el estudiante ingresa. Se rige por un digito para el año y otro para el semestre académico, por ejemplo, 71, corresponde al año 2007 y primer semestre, o 122 corresponde al segundo semestre del año 2012.

**Programa**: Este atributo es siempre 'IND' (ingeniería industrial).

**Semestre**: Son valores enteros del 0 al 10, donde el valor 0 corresponde a aquellos que no entraron finalmente al programa y el resto de valores al semestre como tal. Cualquier valor diferente al 0 o 10, representa el último semestre cursado antes del retiro.

**Icfes**: Puntaje obtenido en la prueba del estado Icfes, obligatoria al final de los estudios de secundaria. Su valor es la suma de los puntajes obtenidos en cada uno de los componentes de la prueba. Se aclara que durante el periodo a analizar (2005 – 2014), estas pruebas mantuvieron el esquema de evaluación, basado en competencias y la misma escala en los puntajes, siendo datos comparables.

**Puesto Icfes**: Puesto otorgado al estudiante dentro de un grupo aleatorio de 1000 estudiantes, dado su puntaje del Icfes, siendo el 1er puesto el de mayor puntaje.

**Cohorte**: Corresponde a los mismos valores del periodo de ingreso.

**Estado**: Se refiere al estado que otorga Icesi a sus estudiantes, donde indica si el estudiante se encuentra en prueba (P), graduado (G), normal (N) o retirado (R).

**Categoría del colegio**: Esta es una categoría que el mismo gobierno asigna a los colegios dado el rendimiento o puntaje de sus estudiantes en las pruebas Icfes. Más adelante se indican las posibles categorías.

**Estado de admisión**: Indica si el estudiante fue admitido (S) o no fue admitido (N) a la Universidad Icesi.

**Categorías de Dimensión**: Las dimensiones se refieren al estilo de aprendizaje del estudiante, utilizadas en la universidad para darle una descripción a los profesores del perfil de aprendizaje de cada estudiante, para elaborar didácticas que faciliten el aprendizaje. Son 4 dimensiones, cada una con dos posibles valores, y se sintetizan a partir de la definición dada por (Lamamie de Clairac Palarea, 2015) y (Ventura, 2000).

La dimensión 1 tiene como posibles valores activo y reflexivo. El valor *Activo* describe a un estudiante que encuentra el aprendizaje en experiencias nuevas y se caracteriza por ser improvisador y espontáneo. El valor *Reflexivo* se refiere a un estudiante observador, analítico, receptivo y concienzudo.

La dimensión 2 cuenta con los valores sensorial e intuitivo. Esta clasifica el tipo de percepción de los estudiantes, siendo sensoriales cuando manifiestan preferencia por

hechos reales y detalles, caracterizados por tener un tipo de pensamiento procedimental. Los intuitivos, por su parte, aprenden encontrando relaciones y significados subyacentes, y poseen habilidades innovadoras y creativas.

La dimensión 3 clasifica a los estudiantes según su preferencia para el aprendizaje mediante explicaciones, siendo estas visual o verbal. Los verbales prefieren explicaciones orales o escritas, los visuales prefieren representaciones gráficas, diagramas y demostraciones.

La dimensión 4 se asocia al tipo de comprensión de los estudiantes, siendo posibles lo secuencial y lo global. El primero se orienta al entendimiento analítico, siguiendo procesos lineales y predeterminados, y el estilo global a un entendimiento holístico, los conceptos en un sentido amplio.

Los valores de cada dimensión representan el grado de cada característica con el que se alinea el estudiante en su aprendizaje, siendo 1 la calificación más baja y 11 la más alta. En cada dimensión, a uno de los valores posibles se le cuantifica. Por ejemplo, si un estudiante registra 11 en el valor secuencial de la cuarta dimensión, demuestra que es radicalmente secuencial y con nula comprensión holística.

**Atributos de Becas**: contiene el nombre de la beca otorgada al estudiante. Puede ser institucional, empresarial o gubernamental.

**NÚMERO_PRUEBA**: número de veces que un estudiante ha quedado en prueba dado su último registro.

**NÚMERO_RETIROS**: número de veces que un estudiante se ha (o ha sido) retirado hasta su último registro.

**Cursadas**: Es el total de materias o asignaturas cursadas por el estudiante hasta su último registro.

**Perdidas**: total de materias reprobadas hasta el último registro de cada estudiante

### 5.1.4 Adecuación de la Base de Datos

#### 5.1.4.1 Eliminación de registros con falta de información

Se encontraron y eliminaron catorce (14) registros de estudiantes que no tenían información completa a través de todos los atributos, dejando la base de datos con mil doscientos veintiséis (1226) registros útiles.

#### 5.1.4.2 Recuperación de información relevante

Se encontró que faltaba el puntaje Icfes de casi todos los estudiantes (150) que ingresaron durante el año 2006. Siendo este dato el referente actual del primer filtro de admisión (con puestos por debajo de 100 son admitidos directamente), se considera relevante recuperar dicha información. Para esto, se encontró que a través de la página web del ministerio de educación, con el "código Icfes" se puede relacionar los puntajes obtenidos. Con macros en Excel se automatizó el procedimiento de identificar los puntajes de cada área y sumarlos para los estudiantes que carecían de dicho valor en la BD.

#### 5.1.4.3 Adición y caracterización de atributos

A la BD se le adicionan atributos importantes para el desarrollo del modelo tales como la fecha de ingreso a la universidad, obtenida de acuerdo al cohorte de ingreso; la edad al ingreso, obtenida restando la fecha de ingreso con la fecha de nacimiento y por último, el atributo deserción, el cual se obtuvo por medio del análisis del atributo Estado, cuya explicación se hará posteriormente.

Además, se decidió hacer modificación a las categorías 'Estado Civil', 'Categoría del Colegio', 'Dimensión 1', 'Dimensión 2', 'Dimensión 3' y 'Dimensión 4', lo cual consistió

en cambiar los valores que estos atributos podían tomar, por medio de abreviaciones que los representaran correctamente. Todo esto con el fin de tener información clara y concreta, evitando además, posibles errores de código.

Este fue el cambio hecho en cada una de las categorías:

| **Dimension 1** | |
|---|---|
| ***Valores*** | ***Abreviatura*** |
| Activo | A |
| Reflexivo | R |

*Figura 10 Dimensión 1*

| **Dimension 2** | |
|---|---|
| ***Valores*** | ***Abreviatura*** |
| Sensorial | S |
| Intuitivo | I |

*Figura 12 Dimensión 2*

| **Dimension 3** | |
|---|---|
| ***Valores*** | ***Abreviatura*** |
| Visual | VS |
| Verbal | VB |

*Figura 11 Dimensión 3*

| **Categoría del colegio** | |
|---|---|
| *Valores* | *Abreviatura* |
| Inferior | **I** |
| Baja | **B** |
| Media | **M** |
| Alta | **A** |
| Superior | **S** |
| Muy Superior | **MS** |

*Figura 15 Categoría del Colegio*

| **Dimension 4** | |
|---|---|
| ***Valores*** | ***Abreviatura*** |
| Global | G |
| Secuencial | S |

*Figura 14 Dimensión 4*

| **Estado Civil** | |
|---|---|
| *Valores* | *Abreviatura* |
| Soltero | **S** |
| Casado | **C** |
| Unión Libre | **UL** |

*Figura 13 Estado Civil*

#### 5.1.4.4 Eliminación de atributos no representativos

Las primeros atributos que se eliminaron, son aquellos que tenían información no relevante sobre las pruebas saber 11 o ICFES. Estos fueron el código ICFES y Fecha ICFES, los cuales hacen referencia al registro del estudiante en la prueba y la fecha de presentación de la misma. Igualmente, se eliminaron los atributos de los nombres de los padres y el nombre del colegio. También atributos como el código del estudiante, la cohorte, el programa, la fecha de ingreso, la fecha de nacimiento, ciudad de nacimiento y admisión de cada estudiante fueron eliminados. Esto, teniendo en cuenta que aportaron a la creación de atributos más importantes para el análisis, siendo el caso de las fechas y el atributo admisión, o porque no aportaban tanto en el análisis que se quería hacer, en este caso el código y el programa.

### 5.1.5 Incorporación del atributo 'Deserción'

A continuación, por medio del atributo Estado, el cual hace referencia al estado actual del estudiante en la carrera y tiene como valores posibles 'graduado' (G), 'en prueba' (P), 'retirado' (R) y 'cursando normal' (N), se creó el atributo **Deserción**, atributo sobre el cual se va a construir el modelo, el cual tuvo dos posibles valores:

- 'SI', para cuando el estudiante presenta el estado 'R'.
- 'NO' para cuando el estudiante presenta el estado 'P', 'N' o 'G'.

### 5.1.6 Análisis de la Base de Datos

Teniendo en cuenta lo anteriormente desarrollado, se creó un gráfico de comparación entre la cantidad de estudiantes ingresados por cohorte y la cifra de aquellos que desertaron por cualquier causa en algún momento. El comportamiento de la deserción durante estos años se muestra por medio del siguiente gráfico:

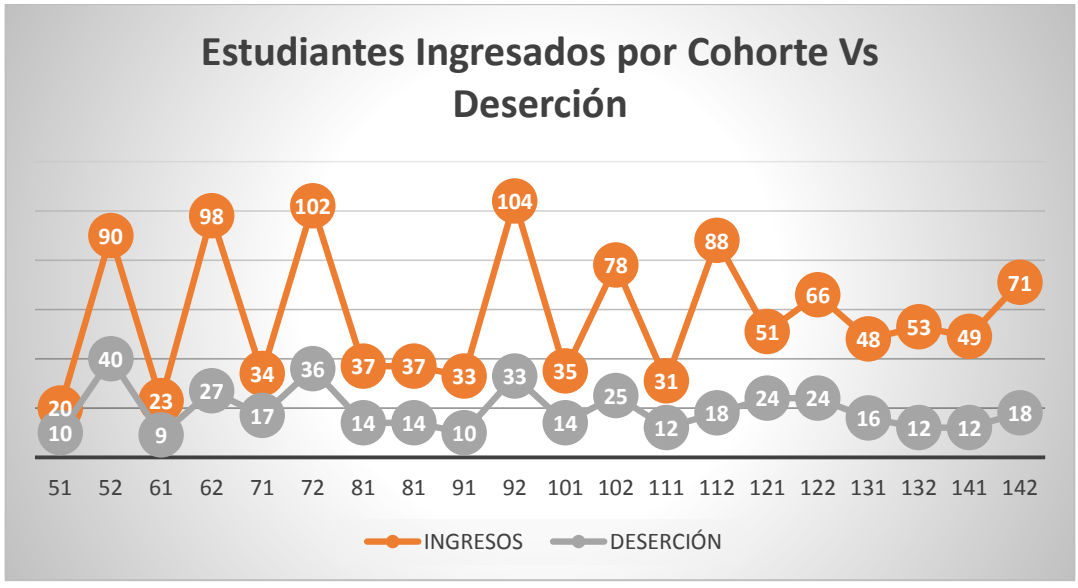



*Figura 16: Ingresados frente a Deserción*

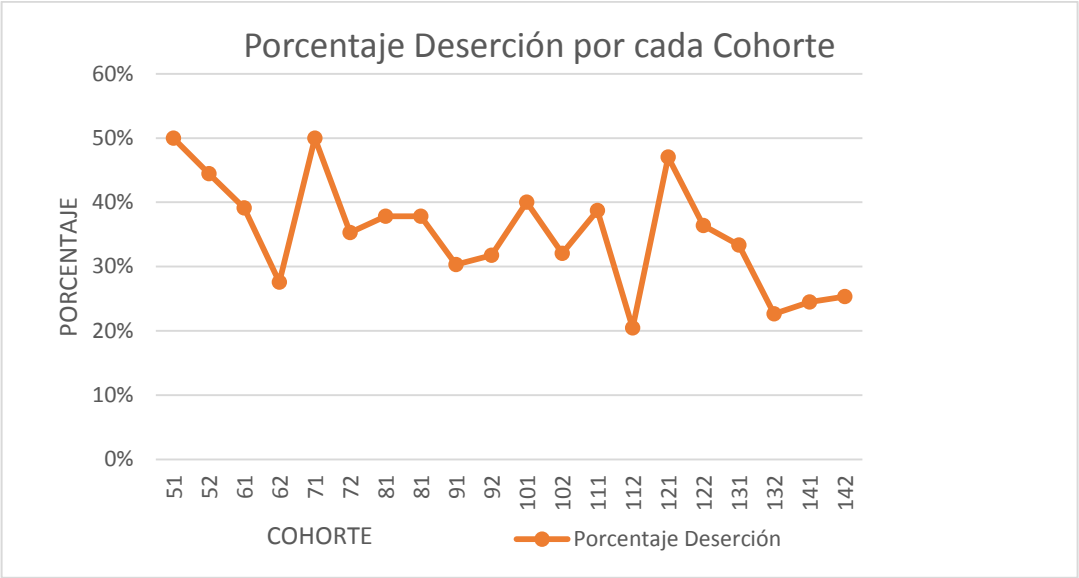
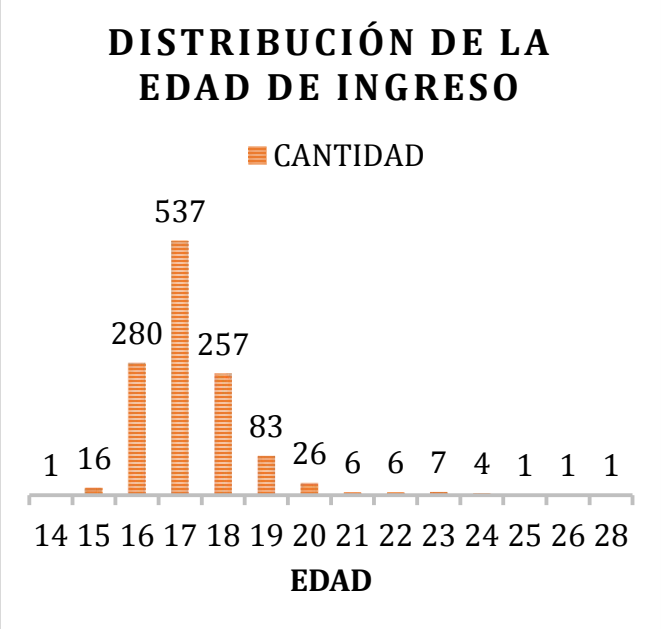



*Figura 17: Porcentaje de Deserción dado el Cohorte*

Como se observa, la deserción siempre ha estado por encima del 20%, llegando incluso a valores cercanos al 50%, como sucedió en los cohortes 5-1, 7-1 y 12-1. Siguiendo con el análisis, se puede ver que para los dos últimos cohortes estudiados, los cuales todavía pueden entrar a ser analizados en el contexto de deserción temprana, hay probabilidad que aumente un 16,3% para el cohorte 14-1 y 21,1% para el cohorte 14-2, debido a los estudiantes que se encuentran actualmente en prueba académica, sin tener en cuenta aquellos que por cualquier otro motivo puedan desertar. Estas cifras sin duda son preocupantes y por eso este proyecto se vuelve importante y necesario con el fin de obtener información relevante que ayude a explicar las posibilidades de deserción, la cual en promedio es del 35.2% según los datos suministrados.

Otro aspecto importante a analizar, fueron las edades de ingreso al programa, ya que podían influir de algún modo en la deserción de los estudiantes. En la base de datos actualizada, el rango de edad de ingreso al programa fue desde los 15 años hasta los 28 años y este fue su comportamiento.




| *Estadistica Descriptiva* | |
|---|---|
| Media | 17,2756933 |
| Error típico | 0,03688088 |
| Mediana | 17 |
| Moda | 17 |
| Desviación estándar | 1,29135767 |
| Varianza de la muestra | 1,66760462 |
| Curtosis | 10,5454326 |
| Coeficiente de asimetría | 2,30198304 |
| Rango | 14 |
| Mínimo | 14 |
| Máximo | 28 |
| Suma | 21180 |
| Cuenta | 1226 |
| Nivel de confianza(95,0%) | 0,07235669 |

*Figura 18: Edad con que ingresan los estudiantes al programa de Ingeniería Industrial*

En este gráfico se puede ver claramente la distribución de los datos, la media de ingreso cercana a la moda, cuyo valor es 17 años. Además, se puede observar un sesgo que principalmente es dado por diferentes datos atípicos como las edades de ingreso de 26 y 28 años que se presentaron con la frecuencia mínima.

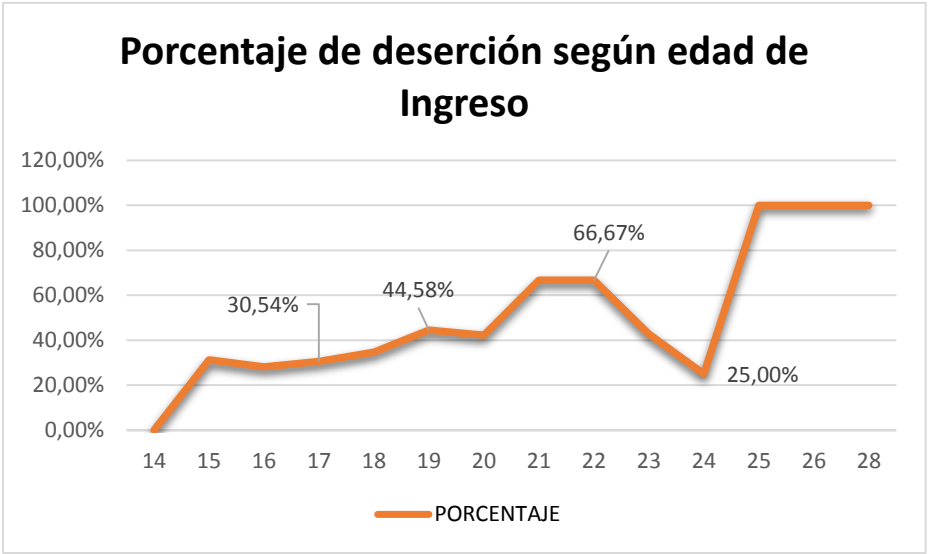



*Figura 19. Porcentaje de deserción por edad*

Entrando a analizar este atributo, se puede que a medida que las edades de ingreso van aumentando, las probabilidades de deserción también aumentan. Sin embargo, entrar a concluir algo sobre este aspecto es precipitado, pues la cantidad de ingresados con edades altas representan valores atípicos lo que los convierte en datos no representativos. De todas maneras, esta información puede llegar a ser importante en el futuro, pues a pesar de que pocas veces se presentan, las probabilidades de deserción podrían ser altas.

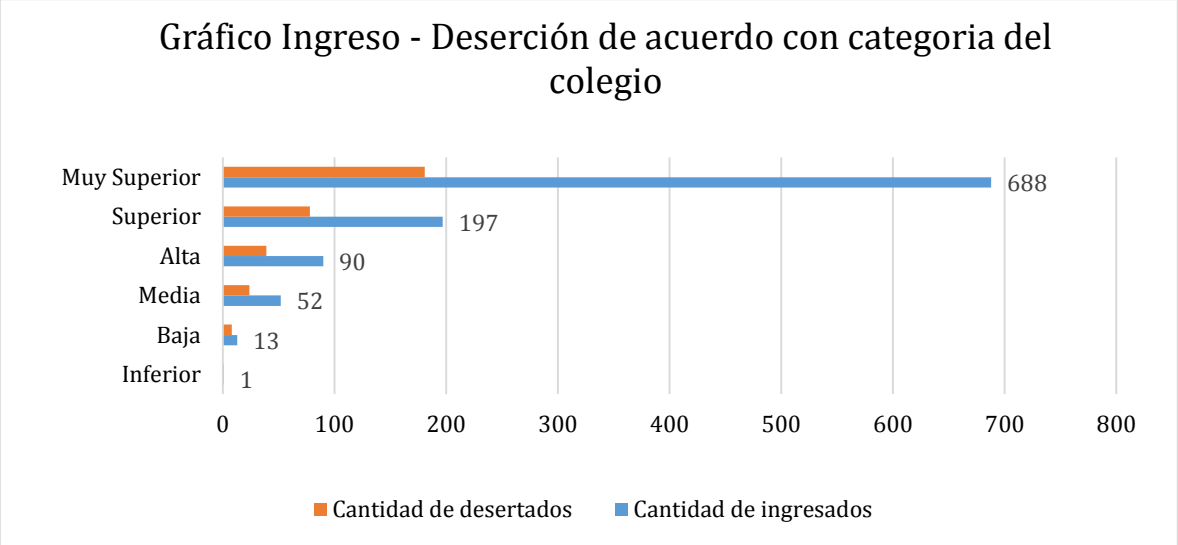



*Figura 20.Nivel de deserción dada la categoría del colegio de los estudiantes*

En la Figura 20 se presenta el nivel de deserción dada la categoría del colegio de los estudiantes. En el gráfico se puede observar que el Programa de Ingeniería Industrial de la Universidad Icesi, tiende históricamente a tener en cuenta la categoría del colegio en la admisión de estudiantes, pues se ve la diferencia de admitidos en cada una de las categorías, siendo mayores en colegios de calidad muy superior y menores conforme a la disminución de categoría. En el siguiente gráfico se puede observar que conforme la categoría del colegio es menor, el porcentaje de deserción es mayor. Esto indica, que a lo largo de los semestres estudiados, el porcentaje de deserción de los estudiantes provenientes de colegios de categoría muy superior es mucho menor al de los de otras

categorías, no importando si en porcentaje son más los estudiantes admitidos de esta categoría a los de otras categorías. Esto convierte al atributo categoría del colegio, uno importante a analizar en este proyecto.

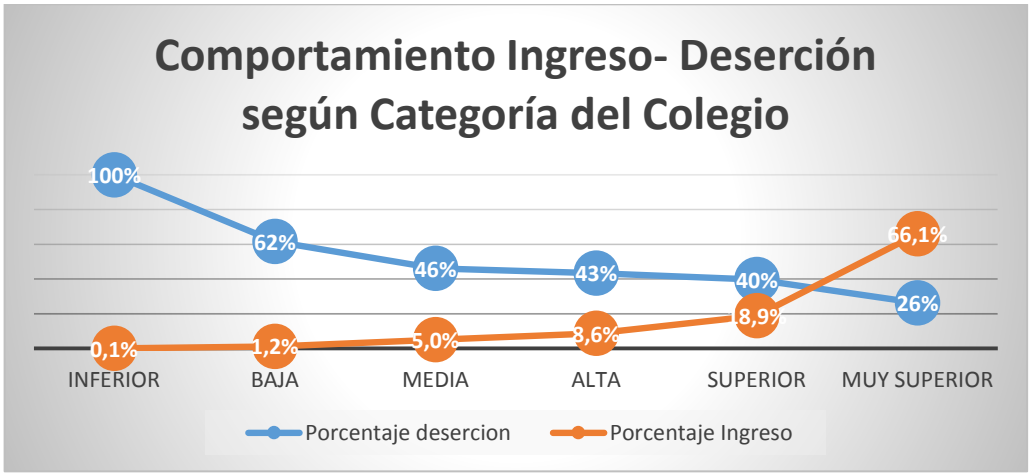



*Figura 21: Comportamiento Ingreso-Deserción, según Categoría del Colegio*

Posteriormente, se hizo un análisis de los atributos, puntaje ICFES y puesto ICFES, ya que un atributo puede ser causal del otro y ocasionar un error en el análisis de los resultados del proceso de *data mining.* Se decidió hacer un análisis de correlación de los dos atributos y los resultados se presentan a continuación:

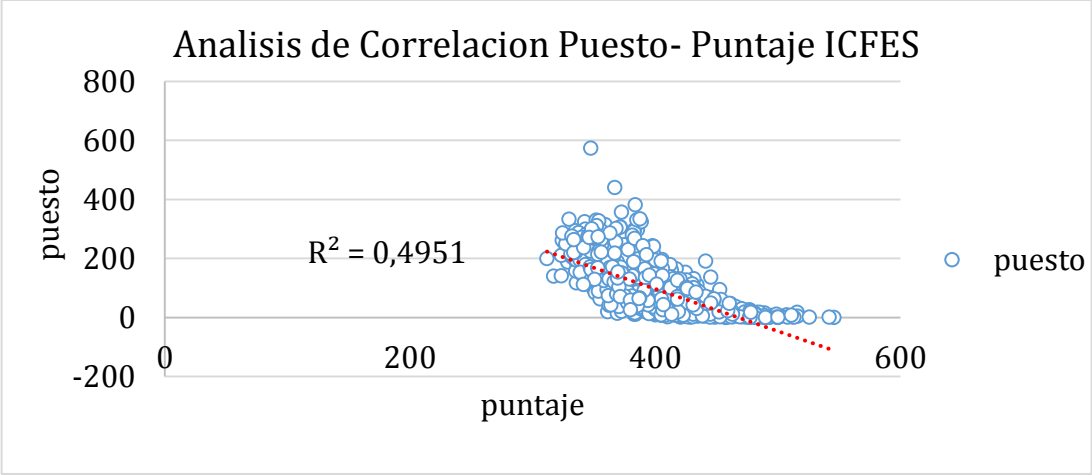



*Figura 22: Correlación entre el puesto y el puntaje obtenido en la prueba de saber (ICFES)*

Con los 1041 estudiantes de la base de datos, se puede ver claramente que la correlación no es significativa, lo que hace dudar que un atributo sea causal del otro. Ha habido una discusión de la manera como el ICFES hace la clasificación de los puestos, dividiendo a los estudiantes en grupos de mil arbitrariamente, permitiendo que en ocasiones puntajes similares tengan puestos totalmente diferentes. Esto conduce a que no haya homogeneidad entre grupos y se presenten comparaciones indebidas de puestos entre ellos. Por lo tanto, se considera que el atributo "Puesto ICFES" podría deteriorar la realidad presentada en el puntaje y se decide eliminarlo, y el puntaje se convierte en el único atributo que evalúa el rendimiento del estudiante en la prueba.

### 5.1.7 Creación del "*Training set*" y el *"test set"*

Se hace la división de la totalidad de los registros de los estudiantes en dos grupos como son el "training set" y el "test set", cuya funcionalidad radica en tener un grupo de datos al cual se apliquen los algoritmos de *data mining* (Training), del cual se deriva un modelo, y por otra parte, este modelo debe ser validado en otro grupo de datos (Test) para conocer el grado de representación.

Los porcentajes definidos previamente fueron del 70% para el training set y del 30% para el test set. Además, se debe tener en cuenta la relación entre los valores posibles del atributo deserción, para mantenerse en la misma proporción en los dos grupos, con el fin de evitar posibles errores de incidencia de alguno de los atributos en los resultados de predicción. Es decir, que un valor del atributo DESERCIÓN que se presente en mayor proporción en uno de los dos grupos puede ocasionar que el modelo se dirija solo hacia ese resultado y genere ruido en el modelo. Como un 32.6% de los registros el atributo deserción tiene un valor de 'SÍ' y 67.4% un valor "NO", son 400 estudiantes para el valor 'SI' y 826 para el valor 'NO'. Al separar la base de datos en dos, quedando un grupo de solo valores SI, y otro de solo valores NO, se parte la base en cada grupo con los porcentajes del 70% y 30%, para el *training* y el *test*, respectivamente, aplicando la función aleatorio de Excel.

A continuación se muestra el resultado de la división de la base de datos en dos grupos.

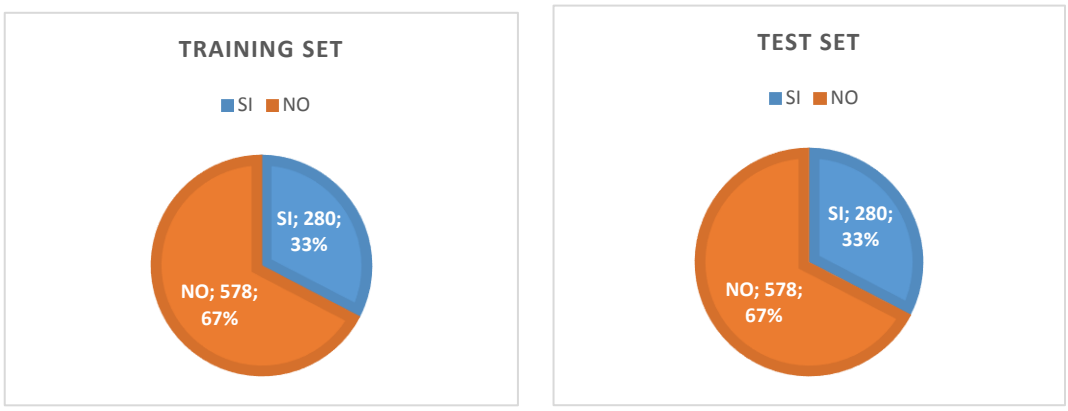



*Figura 23. Test Set*

Si se analizan los valores de la deserción, en cada grupo se obtiene una distribución similar a la existente en la tabla de datos original. Además, se respetan los porcentajes del 70% para el *training set* y del 30% para el *test set*. Con esta división, los datos quedan listos para ser analizados con técnicas de *Data Mining*, el siguiente objetivo del proyecto.

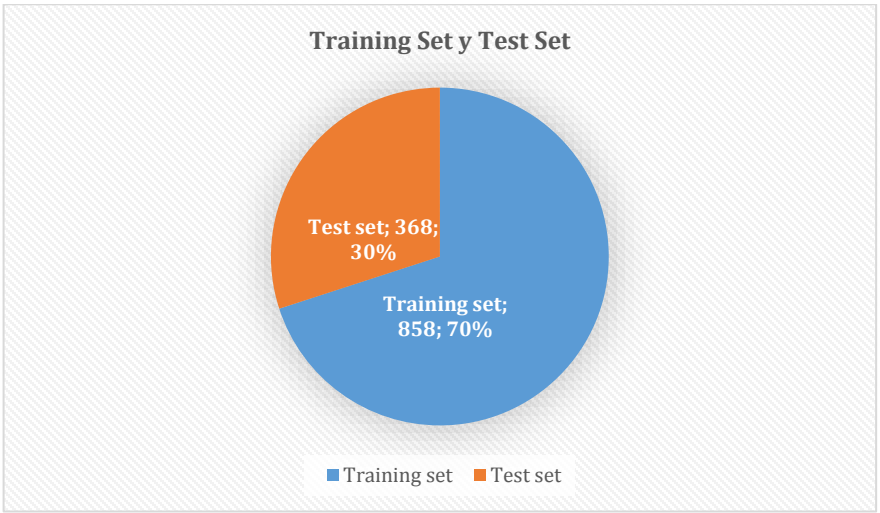



*Figura 25. Training Set y Test set*

## 5.2 Experimentar y validar modelos de predicción de la deserción, a partir de la base de datos construida.

Ya con la base de datos adecuada, se procede a implementar técnicas estadísticas y de data mining, para generar diferentes modelos de predicción de la probabilidad de deserción de los estudiantes del programa de ingeniería industrial de la universidad Icesi. Estos modelos de predicción se van reestructurando acorde a sus necesidades, pues incluso nuevas adecuaciones de los datos pueden ser sugeridas a partir del análisis de los resultados obtenidos en cada uno de los modelos, por lo cual se deben experimentar diversos modelos de predicción, con diferentes algoritmos y técnicas, para luego evaluar y validar su nivel de rendimiento y definir el que mejor se adapte al objetivo del estudio.

### 5.2.1 Cambio de formato del archivo de Excel a arff

Para empezar con este objetivo, lo primero que se debe tener en cuenta es que el software utilizado para analizar, experimentar y validar la base de datos construida (WEKA), no recibe archivos con el formato (.xslx) usado normalmente en los archivos de Excel y, por ende, es el tipo de archivo de la base de datos construida en el objetivo anterior. Debido a esto, se efectúa un proceso de adaptación de la BD al formato (.arff) aceptado por WEKA, lo que implica separar la información perteneciente a cada registro y pegarla en un libro nuevo que contenga solo la hoja donde está ubicada la BD, y guardar la información bajo la opción de CSV (delimitado por comas). Este tipo de archivo permite que la hoja de cálculo pueda ser abierta con bloc de notas y guardar en el tipo de formato requerido por WEKA.

El tipo de archivo relacional (arff) tiene una estructura que debe ser respetada para que pueda ser leído, estructura que no tiene en el momento el archivo CSV. A continuación se muestra el flujo de pasos requeridos para convertir el archivo al formato relacional.

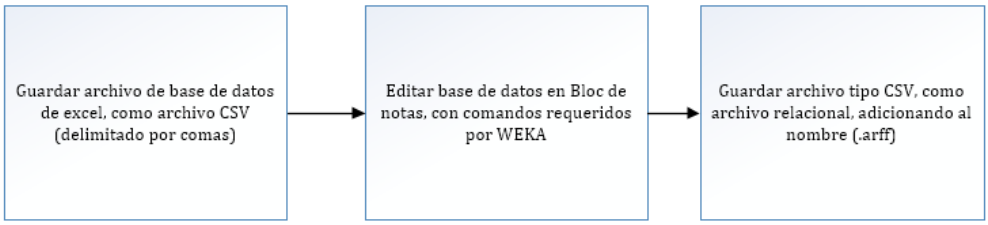



*Figura 26. Pasos para convertir archivo .xslx a .arff*

Ya con el archivo tipo CSV, se puede abrir la base de datos con el Bloc de Notas, el cual contiene la información del archivo de Excel en forma de celdas, donde cada una de estas es delimitada por punto y coma (;). Este tipo de distribución no sirve como archivo relacional, por lo que se procede a editarlo con los parámetros característicos del tipo de archivo relacional. Lo primero es separar los nombres de los atributos del conjunto de datos, dejando cada uno en una columna diferente, respetando el orden en el que aparecen los valores en el conjunto de datos. Lo segundo es dar un nombre al archivo por medio del código @*relation*, el cual permite asignarle el nombre al tipo de relación que se analiza, en este caso el nombre fue DESERCION.

Por motivos de compilación, WEKA necesita conocer qué nombres llevarán los atributos que se presentan, y es por eso que antes del inicio de cada nombre se debe ingresar el código @*attribute,* y después del nombre, declarar que tipo de atributo es, o que tipo de valores toma. Para los atributos con números enteros se introduce la palabra *Integer*, para los números reales se introduce *Real*, para variables discretas se introducen llaves { },  donde se ingresan los valores específicos que puede tomar cada atributo. A continuación se muestra el formato de archivo creado con estos pasos:

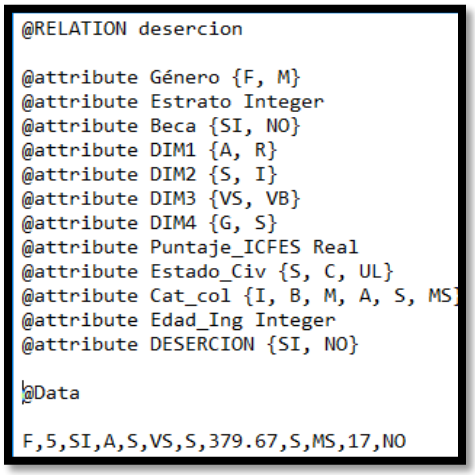



*Figura 27 Formato útil para WEKA*

Ya como último paso, se guarda el archivo introduciendo la extensión (.arff) seguida del nombre, lo que convierte el archivo CSV al formato deseado, permitiendo entonces que la información pueda ser leída por el software WEKA.

### 5.2.2 Análisis de atributos con WEKA

#### 5.2.2.1 Análisis de las Dimensiones de los estudiantes

Para empezar, se interpretan las dimensiones del aprendizaje de los estudiantes en relación a la estrategia del aprendizaje activo que emplea la universidad en todos sus programas. Este atributo en primera instancia se considera importante, dado que en el caso de no darse la adaptación, el estudiante se puede tentar a desertar. Se procede a analizar dicho atributo mediante técnicas estadísticas y de minería de datos, y determinar su impacto en el objetivo de deserción.

Inicialmente, no se considera el valor numérico que acompaña el nivel de dichos atributos, por ejemplo “Activo (5) o Reflexivo (11)”, para esta primera etapa tan solo

se estudia la incidencia de poseer alguna de las características de cada dimensión, no su magnitud.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         826               67.3736 %
Incorrectly Classified Instances       400               32.6264 %
Kappa statistic                          0
Mean absolute error                      0.4396
Root mean squared error                  0.4688
Relative absolute error                 99.9752 %
Root relative squared error            100      %
Total Number of Instances             1226
```

*Figura 28. Resultados interfaz Explorer de WEKA*

```
=== Confusion Matrix ===

   a   b   <-- classified as
   0 400 |   a = SI
   0 826 |   b = NO
```

*Figura 29. Matriz de confusión entregada por WEKA*

Se corre entonces un algoritmo que intente predecir los valores de la deserción, considerando únicamente las dimensiones del estudiante. Se obtiene como resultado, tan solo la aplicación del algoritmo ZeroR, el cual consiste en predecir siempre aquel valor que mayormente se presenta, que en este caso sería predecir que ningún estudiante va a desertar, lo cual no aporta ninguna información. Con estos datos, se evidencia que las dimensiones de aprendizaje del estudiante, por si solas, no presentan información contundente para el algoritmo y se procede entonces a especificar la magnitud de estas dimensiones, en busca de seguir evaluando resultados.

Se asumen los valores 1, 3 y 5 como de poca magnitud y 7, 9 y 11 como mucha magnitud. Para organizar los datos con esta última consideración, se realiza un condicional en Excel, que permita clasificar el estudiante en su respectiva dimensión y su peso

respectivo, dándole la connotación “poco o muy”. La dimensión “neutro” no se tendrá en cuenta, debido a que esta se establece solo cuando ninguna dimensión del estudiante ha sido evaluada, probablemente por algún error en los datos suministrados o, por la ausencia de la evaluación a dichos estudiantes.

A continuación se presenta la modificación planteada para una serie de casos.

| DIM 1 | DIM 2 | DIM 3 | DIM 4 |
|---|---|---|---|
| Activo (1) | Sensorial (1) | Visual (7) | Global (3) |
| Activo (0) | Sensorial (0) | Visual (0) | Secuencial (0) |
| Activo (5) | Sensorial (7) | Visual (9) | Secuencial (3) |
| Activo (5) | Intuitivo (1) | Visual (1) | Secuencial (3) |
| Activo (7) | Intuitivo (1) | Verbal (5) | Secuencial (3) |

| DIM 1 | DIM 2 | DIM 3 | DIM 4 |
|---|---|---|---|
| poco activo | poco Sensorial | muy Visual | poco Global |
| neutro | neutro | neutro | neutro |
| poco activo | muy Sensorial | muy Visual | poco Secuencial |
| poco activo | poco Intuitivo | poco Visual | poco Secuencial |
| muy activo | poco Intuitivo | poco Verbal | poco Secuencial |

*Figura 30. Atributos de Dimensión de aprendizaje*

Luego, con la creación de tablas dinámicas en Excel, teniendo como referencia el nivel de deserción promedio (32,626%), se podrán alertar los perfiles de estudiantes que se alejen de dicho porcentaje, con un margen considerable (10%).

**DIMENSION 1**

| Count of DESERCION | Co | | | |
|---|---|---|---|---|
| Row Labels | NO | SI | Grand Total | |
| muy activo | 160 | 62 | 222 | 27,928% |
| muy reflexivo | 23 | 6 | 29 | 20,690% |
| poco activo | 385 | 178 | 563 | 31,616% |
| poco reflexivo | 184 | 61 | 245 | 24,898% |
| **Grand Total** | **752** | **307** | **1059** | |

**DIMENSION 3**

| Count of DESERCION | Co | | | |
|---|---|---|---|---|
| Row Labels | NO | SI | Grand Total | |
| muy Verbal | 8 | 3 | 11 | 27,273% |
| muy Visual | 321 | 109 | 430 | 25,349% |
| poco Verbal | 59 | 38 | 97 | 39,175% |
| poco Visual | 364 | 157 | 521 | 30,134% |
| **Grand Total** | **752** | **307** | **1059** | |

**DIMENSION 2**

| Count of DESERCION | Co | | | |
|---|---|---|---|---|
| Row Labels | NO | SI | Grand Total | |
| muy Intuitivo | 18 | 13 | 31 | 41,935% |
| muy Sensorial | 240 | 69 | 309 | 22,330% |
| poco Intuitivo | 103 | 68 | 171 | 39,766% |
| poco Sensorial | 391 | 157 | 548 | 28,650% |
| **Grand Total** | **752** | **307** | **1059** | |

**DIMENSION 4**

| Count of DESERCION | Co | | | |
|---|---|---|---|---|
| Row Labels | NO | SI | Grand Total | |
| muy Global | 37 | 6 | 43 | 13,953% |
| muy Secuencial | 76 | 28 | 104 | 26,923% |
| poco Global | 255 | 111 | 366 | 30,328% |
| poco Secuencial | 384 | 162 | 546 | 29,670% |
| **Grand Total** | **752** | **307** | **1059** | |

*Figura 31. Análisis atributos de Dimensiones de aprendizaje*

En la Figura 31 se identifica en amarillo aquellas probabilidades de deserción que se encuentran moderadamente por debajo del valor referente, y ello indica que los estudiantes "muy reflexivos, muy sensoriales y muy globales" presentan una considerable menor tasa de deserción. En cuanto a los datos sombreados de color naranja, representan los valores que están más por encima de la tasa de deserción media, siendo los "muy verbales, y en general la dimensión intuitivo".

De otro lado, se identifica que la "Dimensión 2" es la que cuenta con la mayor desviación en el porcentaje de deserción, por lo cual será información que se plasmará como datos interesantes a tener en cuenta, con el fin de que se pueda hacer seguimiento a esta dimensión del estudiante en un futuro.

Ya finalizando este punto, se decide generar modelos de predicción de la deserción, incluyendo las dimensiones pero considerando otros atributos con los que se cuentan hasta el momento. Se obtienen mejores niveles de predicción cuando se asumen las dimensiones a nivel general, o sea sin incluir las magnitudes de las mismas. Esto se presenta debido a que de por si las magnitudes no terminan distanciando las tasas de deserción asociadas a la dimensión como tal, por lo cual resulta más útil para el modelo, contar con mayor representatividad en una dimensión, que asumir varias magnitudes que el no logro agrupar. Se decide entonces considerar las dimensiones del estudiante a nivel general, y por ello se clasifican todas las magnitudes dentro de una dimensión.

#### 5.2.2.2 Análisis de las becas

Dentro de la base de datos, los estudiantes cuentan con diferentes tipos de becas, que terminan siendo más o menos 15. Las becas pasan a ser clasificadas dentro de ciertas categorías, para facilitar el entendimiento de la información. A continuación, se muestran dichas categorías y sus respectivos niveles de deserción.

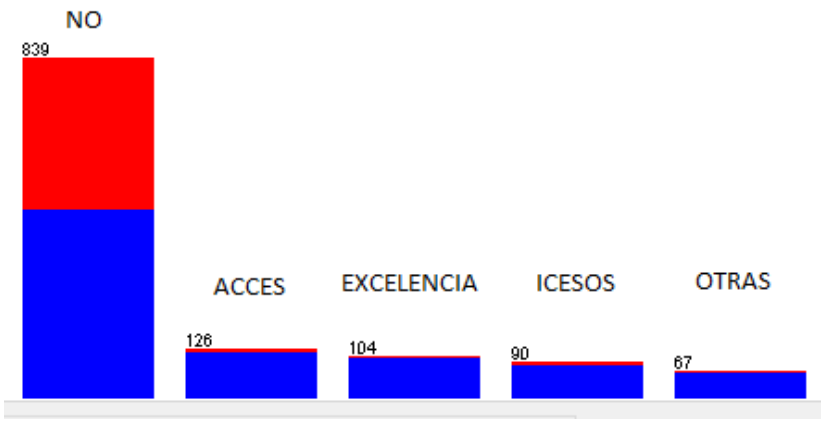



*Figura 32. Análisis del atributo Beca y deserción*

A partir de los resultados obtenidos, se elabora la Figura 33 incluyendo el tipo de beca, que indica en azul aquellos estudiantes que no desertan y en rojo los que sí. La primera columna representa aquellos estudiantes sin beca y las demás columnas se refieren a aquellos que obtuvieron algún tipo de beca. Se concluye que aquellos estudiantes con algún tipo de beca presentan un nivel mínimo de deserción, en todos los tipos de beca.

Teniendo en cuenta todo lo anterior, en el modelo de predicción de deserción se incluirá el atributo "BECA" pero tan solo considerando si el estudiante posee algún tipo de beca (independientemente de cual sea) y por lo tanto los valores de dicho atributo serán "SI" o "NO".

#### 5.2.2.3 Análisis de estrato y categoría de colegio

Con el objetivo de evitar atributos que inciden directamente en otros y generen distorsión, se analiza también el caso del estrato socio económico – categoría del colegio (otorgada por el ICFES), donde dado el sistema de educación en Colombia, se presume que estudiantes de estratos altos provienen mayoritariamente de colegios de alta calidad.

Se acondicionan entonces los datos para evaluar de manera independiente con la herramienta WEKA y lograr establecer el grado de predicción de la "categoría del colegio" en los algoritmos de clasificación, a partir del atributo "estrato". En la siguiente gráfica, se muestra el resultado de 5 algoritmos y el valor asociado del porcentaje de predicción obtenido.

| Dataset | (1) rules.Ze | (2) bayes | (3) trees | (4) lazy. | (5) rules |
|---|---|---|---|---|---|
| Estrato | (100) 26.29 | 36.38 v | 36.16 v | 36.16 v | 36.08 v |
| | (v/ /*) | (1/0/0) | (1/0/0) | (1/0/0) | (1/0/0) |

*Figura 33. Análisis de correlación entre categoría de colegio y estrato*

Los niveles de predicción son bastante bajos, lo cual estaría indicando que un atributo no influye en el otro, pues se emplearon algoritmos de Bayes, de árboles de decisión, de reglas de asociación, entre otros, y ninguno evidenció un dato influyente. Sin embargo se decide mostrar un poco más acerca de este comportamiento, buscando mayor claridad para tomar la decisión de incluir o no ambos atributos y para ello se presenta la siguiente gráfica.

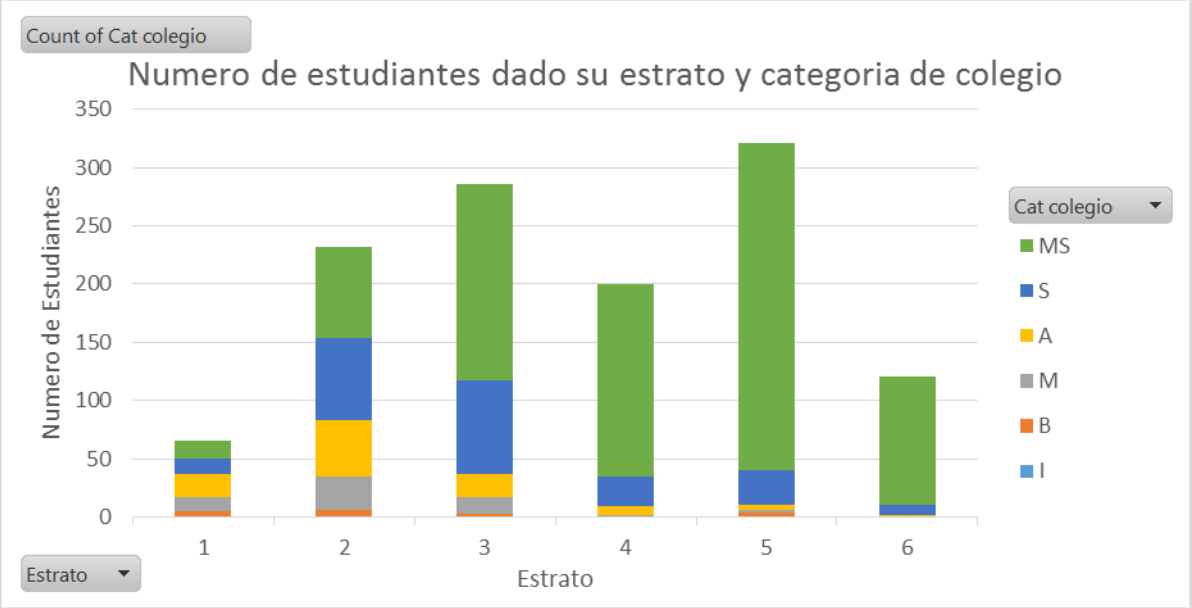



*Figura 34. Análisis para atributos estrato y categoría de colegio*

Al emplear algoritmos de Bayes, de árboles de decisión, reglas de asociación, entre otros, se evidencia que los niveles de predicción son relativamente bajos y que un atributo no influye en el otro. Sin embargo, se decide ahondar en el análisis. En la Figura 36, las columnas representan los estratos y van del "estrato 1" a "estrato 6" de derecha a izquierda, mientras que los colores son las diferentes categorías del colegio, siendo evidentes las siguientes:

Verde – Categoría Muy Superior

Azul – Categoría Superior

Amarillo – Categoría Alta

Gris – Categoría Media

En los primeros estratos (bajos) hay mayor heterogeneidad con individuos que provienen de colegios de diferente categoría, y en los estratos altos hay mayoría de egresados de colegios con calidad superior y minoría de menor categorización de sus

instituciones. Esto explica la baja predicción de los algoritmos, dada la composición mixta de la población estrato-categoría colegio. Por lo tanto, ambos atributos han de considerarse en la construcción del modelo de predicción.

#### 5.2.2.4 Análisis de la ciudad de origen (Residencia)

Hay un total de 76 diferentes municipios de origen en la BD y con mucha variabilidad en la participación de cada uno, teniendo municipios con escasa población representativa, lo que genera ruido en los árboles de decisión al obtener ramas cuya decisión no aporta valor. En este punto se analizan dos aspectos:

- Identificar la mejor manera de agrupación de los datos, acorde a su nivel de impacto sobre el objetivo "Deserción". La idea es evitar casos en que un atributo puede no aportar al modelo si no se filtran los datos realmente relevantes.
- Identificar aquellos datos representativos e incluso "datos curiosos" por reconocer a lo largo del análisis.

Se determinan tres criterios de agrupación: 1) estudiantes que pertenezcan a la ciudad de Cali, 2) municipios diferentes a Cali y 3) ciudades fuera del país. Se presenta la siguiente información, donde los valores de las columnas para [SI] o [NO] hacen referencia a la cantidad de estudiantes que desertan o no, conforme a su ciudad de proveniencia.

| | Cali | Dif. Cali | Fuera del País |
|---|---|---|---|
| SI | 270 | 129 | 1 |
| NO | 612 | 213 | 1 |
| % DESERCION | 30,61% | 37,72% | 50,00% |

*Figura 35. Análisis de Ciudad frente al atributo deserción*

No se presenta sesgo alguno en las probabilidades de deserción en relación a la ciudad si se sostiene la diferencia básica de "Cali" y "No Cali". En la categoría de "No Cali" se

evalúa internamente y se deciden las ciudades que logran ser más representativas según tamaño de población perteneciente (mínimo 16 estudiantes).

| | Desercion | | |
|---|---|---|---|
| Ciudad | SI | NO | % Desercion |
| Palmira | 11 | 39 | 22,00% |
| Cali | 270 | 612 | 30,61% |
| Tuluá | 5 | 11 | 31,25% |
| Buga | 6 | 10 | 37,50% |
| Bogotá | 18 | 24 | 42,86% |
| Popayán | 13 | 17 | 43,33% |

*Figura 36. Análisis de ciudades diferentes a Cali*

A nivel general, se tiene una deserción promedio del 34,42%, dato aproximado a la tasa de deserción promedio, por lo cual se puede asumir que no se aporta información relevante dado que los datos que se distancian de la media son muy pocos, y no habría patrones significativos por ciudad u origen.

En conclusión, el atributo "Ciudad de origen" no será incluido dentro de la BD final utilizada para la búsqueda del modelo de predicción. Sin embargo, algunos datos deben servir de referencia en la herramienta de predicción propuesta, con el fin de hacer seguimiento a los estudiantes de fuera de la ciudad de Cali.

#### 5.2.2.5 Análisis de Perfil laboral de los padres

En la BD, la columna de "Cargos" se refiere a la ocupación laboral o actividad económica de los padres. En este punto, se contabilizan 661 valores diferentes debido a que no hay un patrón definido y especifico cuando los prospectos diligencian el formulario. Es notorio que 306 estudiantes tienen en blanco este atributo y solo 441 estudiantes y se decide agrupar todos aquellos cargos que conservaban un término en común y que fueron listados un amplio número de veces.

| TIPO CARGO PADRE | NO DESERTA | SI DESERTA | % Desercion |
|---|---|---|---|
| MEDICO | 11 | 2 | 15,38% |
| DIRECTOR | 23 | 5 | 17,86% |
| INDEPENDIENTE | 37 | 9 | 19,57% |
| ASESOR | 14 | 4 | 22,22% |
| JEFE | 15 | 5 | 25,00% |
| COMERCIANTE | 11 | 4 | 26,67% |
| INGENIERO | 11 | 4 | 26,67% |
| VENDEDOR | 10 | 4 | 28,57% |
| JUBILADO | 9 | 4 | 30,77% |
| ADMINISTRADOR | 15 | 7 | 31,82% |
| GERENTE | 115 | 56 | 32,75% |
| ABOGADO | 6 | 3 | 33,33% |
| DOCENTE | 21 | 11 | 34,38% |
| PROPIETARIO | 16 | 9 | 36,00% |
| | | | |
| **TOTAL** | **314** | **127** | **27,213%** |

*Ilustración 37. Análisis de cargo de padres*

Se tiene el porcentaje de deserción presentado ante cada tipo de cargo, con lo cual por ejemplo, quienes tienen un padre "abogado", presentan una tasa de deserción del 33%. En la tabla, el color rojo indica valores altos de probabilidad de deserción, mientras que en amarillo dicho valor va mermando, hasta quedar en verde se simboliza una baja tasa de deserción. Lo remarcable es que de los se encuentran en tono rojizo, constituyen cerca del 60% de estudiantes que presentan deserción.

| *Descripcion de datos* | |
|---|---|
| | |
| Mean | 27,2127 |
| Standard Error | 1,73623 |
| Median | 27,619 |
| Mode | 26,6667 |
| Standard Deviation | 6,49639 |
| Range | 20,6154 |

*Figura 38. Estadística descriptiva de los cargos de los padres*

Finalmente, se decide no considerar este atributo para la elaboración del modelo de predicción, pues son muchos los datos faltantes y muchos otros que no logran ser

representativos dado el nivel de especificidad del atributo, donde además, se puede estar evaluando tan solo un sesgo durante el proceso de la agrupación.

Debido a lo anterior, y a pesar que en múltiples estudios indican que el nivel de educación de los padres es influyente al tema de la deserción, se sugiere a la Universidad Icesi que solicite no solamente la ocupación de los padres bajo ciertos estándares para evitar tantos valores, sino también su nivel máximo de estudios alcanzados, pues podrá aportar a futuros estudios de *data mining*.

### 5.2.3 Aplicación de técnicas de data mining a la base de datos, para la generación y posterior validación de los modelos de predicción

#### 5.2.3.1 Eliminación de datos atípicos y extremos

Se realiza un análisis final de base de datos, encaminado a obtener datos atípicos o valores extremos que puedan afectar la calidad de la información. En este sentido, se aplica el filtro *interquartilrange*, el cual hace un análisis de la distribución de los datos en cada atributo y, de acuerdo a su relación con la clase a analizar o atributo a predecir (DESERCIÓN), clasifica los datos entre atípicos y no atípicos, seguida por una columna que señala valores extremos o no extremos. Dicho esto, a continuación se muestran los resultados obtenidos con el filtro.

Una vez aplicado el algoritmo, se encuentran 17 datos atípicos, que aunque son muy pocos, se toma la decisión de eliminarlos, pues no representan información útil para el testeo de diferentes algoritmos y para la posterior construcción de la herramienta de predicción.

Continuando con este proceso, se revisa la cantidad de valores considerados extremos por el algoritmo aplicado, donde se encuentran otros 3 datos y, aunque no presentan un riesgo de ruido, también se toma la decisión de eliminarlos del total de registros.

#### *5.2.3.2 Análisis de la predicción planteada por diferentes algoritmos*

Con los datos atípicos y valores extremos eliminados, se lleva cabo un análisis preliminar de la predicción con diferentes algoritmos, aspecto mencionado anteriormente en la metodología. El objetivo es conocer cuál de estos se ajusta mejor al comportamiento de la información introducida, pues el que mejor desempeño presente para cada uno de los escenarios del proceso de admisión y de apoyo al estudiante, va a ser escogido como base para la construcción de la herramienta de predicción de deserción. Para este proceso se utiliza la interfaz *experimenter* para analizar exhaustivamente la BD con diferentes algoritmos, e identificar cuál presenta mejor porcentaje de predicción, teniendo en cuenta las curvas ROC y error absoluto de cada uno. A continuación se presenta una imagen basada en información obtenida por el software donde muestra el porcentaje de datos correctos predichos para cada algoritmo.

| Rules | | | Trees | | | Bayes | Lazy | Meta | |
|---|---|---|---|---|---|---|---|---|---|
| ZeroR | Jrip | OneR | J48 | DesicionStump | RepTree | NaiveBayes | IBK | AdaBoost | Vote |
| **67,74** | **71,18** | **61,35** | **72,47** | **67,74** | **71,76** | **73,57** | **68,6** | **73,76** | **72,77** |

*Figura 39. Muestra de resultados obtenidos con el Experimenter de WEKA*

Como se observa en la imagen anterior, la base de datos sin atípicos fue testeada con un total de 10 algoritmos, 3 de ellos son algoritmos de reglas de decisión, 3 algoritmos que permiten crear arboles de decisión, 2 algoritmos meta heurísticos, una función y por último, un algoritmo de Bayes. Ahora bien, con el fin de generar una idea del tipo de clasificación que aplican a los modelos de datos, se muestra una descripción breve de estos:

**JRip:** este algoritmo implementa un aprendizaje basado en reglas proposicionales llamado RIPPER, que consiste en eliminar progresivamente ramas que generan ruido en el modelo de datos.

**OneR:** Es un clasificador que tiene en cuenta el error mínimo de cada atributo del modelo de datos, para generar el resultado. Como característica, tiende a discretizar las categorías con valores numéricos.

**Decision Stump:** algoritmo normalmente usado en conjunto con algoritmos aceleradores de predicción, el cual para modelos de clasificación basa su decisión en la entropía. Como característica, convierte la información faltante en valores independientes.

**REPTree:** algoritmo de árbol de aprendizaje rápido. Construye el árbol de decisión basado en la varianza de cada atributo y utiliza el error para la eliminación de ramas.

**IBk:** algoritmo basado en la teoría del vecino más cercano. Se utiliza en conjunto con el método de validación cruzada.

**AdaBoost M1:** algoritmo de aprendizaje acelerado. Se utiliza principalmente en problemas de clases nominales, mejorando el rendimiento de los modelos, pero generando algunas ocasiones un problema efectuando ajustes excesivos al modelo.

**Vote:** algoritmo que permite combinar diferentes clasificadores, en este caso se usa para combinar el algoritmo de árbol J48 y el algoritmo de NaiveBayes.

A continuación se presenta una tabla con resultados resumidos de este primer análisis con un 95% de confianza, el cual tiene como objetivo mostrar el desempeño en cuatro indicadores.

| Algoritmo | Porcentaje de predicción | ROC Area | F measure | Error Absoluto |
|---|---|---|---|---|
| ZeroR | 67,74% | 50% | 0% | 44% |
| Jrip | 71,18% | 65% | 49% | 39% |
| OneR | 61,35% | 50% | 23% | 39% |
| **J48** | **72,47%** | **73%** | **53%** | **34%** |
| DecisionStur | 67,74% | 69% | 0% | 38% |
| REPTree | 71,76% | 74% | 49% | 35% |
| **NaiveBayes** | **73,57%** | **78%** | **51%** | **34%** |
| Ibk | 68,60% | 63% | 50% | 31% |
| AdaBoostM1 | 73,76% | 78% | 50% | 35% |
| Vote | 72,77% | 66% | 53% | 27% |
| **MEJOR** | **73,76%** | **78%** | **53%** | **27%** |

*Figura 40.Desempeño por algoritmo en primer análisis*

Como se puede observar, se resaltan los mejores desempeños, de acuerdo a cada aspecto analizado en la lista de algoritmos. Lo anterior se hace con el objetivo de generar una clasificación de algoritmos, que permita ir reduciendo la cantidad final de ellos a solo uno. Entendido esto, con letra en negrilla se resalta aquellos valores que obtuvieron un desempeño alto en los cuatro aspectos analizados, ya con estos se pasa a una siguiente etapa de análisis más profunda para posterior elección del modelo que mejor representa la distribución de los datos.

Entrando a explicar en detalle en que consiste cada indicador, se empieza por el *% de predicción*, indicador más importante en la toma de decisiones, el cual permite conocer que parte del total de datos fue correctamente clasificada, es decir aquellas instancias en las que el algoritmo, y el resultado previo de la base de datos, coincidieron.

*ROC area*, hace referencia a la curva de aprendizaje de cada algoritmo. Dicha curva indica el porcentaje o cantidad de datos correctos que se logran obtener dada una cantidad de datos considerados, que para este proyecto podría ser el número de desertores encontrados, dado un cierto número de estudiantes considerados.

Para ejemplificar lo anterior, suponga que hay una población de 1000 estudiantes, de los cuales 200 desertan y 800 no desertan. Suponga además, que se tiene un modelo de

predicción que indica que 300 estudiantes van a desertar, de los cuales tan solo la mitad, o sea 150 estudiantes, realmente son desertores. Ahora bien, habrán estudiantes con mayor probabilidad de deserción que otros, y ahí es donde empieza a jugar un rol importante el indicador *ROC Area*, ya que si por ejemplo, 30 estudiantes tienen una probabilidad de deserción del 90%, seguramente la curva de aprendizaje logrará encontrar cerca del 10% del total de desertores, con tan solo evaluar esos 30 estudiantes de 1000, que corresponden al 3%. Así entonces, al principio la curva tiene a tener una pendiente bastante pronunciada hacia arriba, pero a medida que la probabilidad de deserción del estudiante va decreciendo, también lo ira haciendo la pendiente, creando un comportamiento logarítmico, pues tendrá que incorporar mayor número de estudiantes e ir obteniendo menor porcentaje de desertores reales.

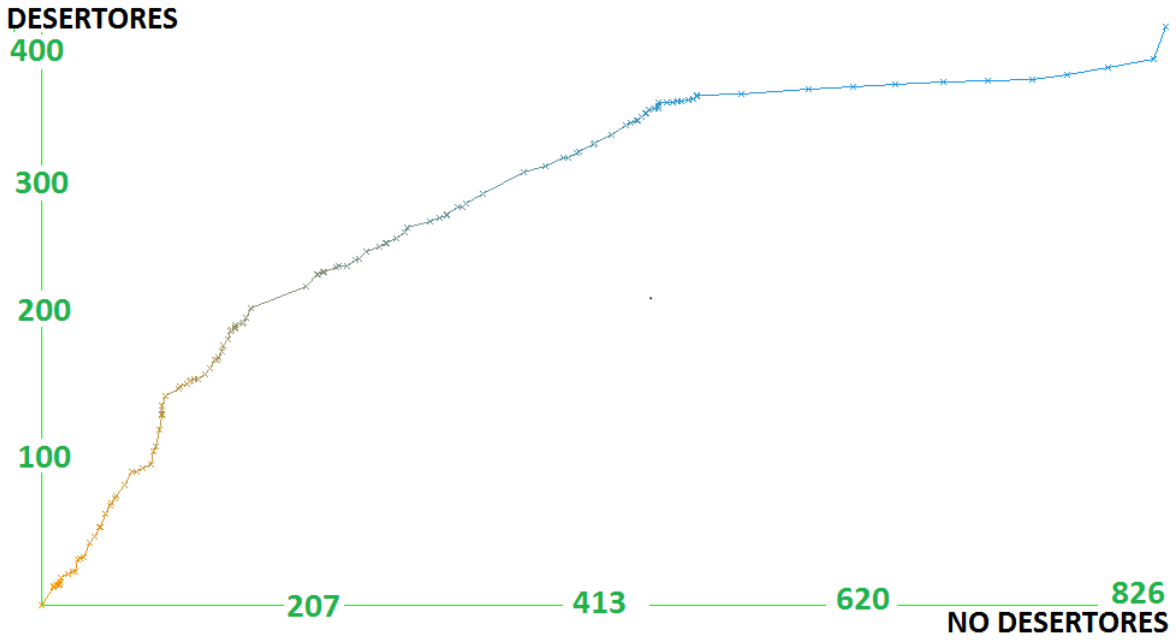



*Figura 43. ROC Área para algoritmo J48*

La figura 43 muestra el comportamiento de un modelo de predicción generado con el algoritmo J48 y a partir de la base de datos construida. Se puede evidenciar que la curva empieza con un rápido crecimiento en el eje 'y', el cual indica los estudiantes desertores que se van encontrando al tiempo que se van encontrando también estudiantes no

desertores, que corresponden al eje 'x'. Cerca del 50% de los desertores se encuentran con tan solo incorporar un 15% de los no desertores.

La medida F (F measure), funciona de acuerdo a una ecuación que tiene en cuenta la precisión de las predicciones y el total de errores encontrados. Esta medida se toma por cada posible valor, es decir para el SI y NO, del atributo deserción, y que al final permite hacer un promedio del resultado para cada valor, el cual es mostrado en la Figura anterior.

A continuación se demuestra cómo se obtienen algunos elementos.

La precisión viene dada por:

$$Precisión = \frac{Predicciones\ acertadas}{predicciones\ acertadas + predicciones\ no\ acertadas}$$

Recall, segundo componente de la ecuación representa la fracción de todas las predicciones acertadas sobre el total de predicciones a acertar.

$$Recall = \frac{Predicciones\ acertadas\ para\ un\ valor\ determinado}{Total\ de\ registros\ con\ un\ valor\ predeterminado}$$

La medida F, entonces viene representada por la siguiente ecuación

$$F\ Measure = \frac{2 * Precisión * Recall}{Precisión + Recall}$$

Como se puede observar, es una medida completa que ayuda también a visualizar el comportamiento del algoritmo dado el conjunto de datos.

Por último, el error absoluto indica que tan acertadas fueron las predicciones respecto al total de datos, este valor entre más pequeño es mejor, indicando que la predicción y el valor real coinciden.

#### 5.2.3.3 Balanceo de la clase Deserción

Antes de definir el algoritmo con mejor porcentaje de predicción para cada herramienta a construir, se considera importante analizar los datos con un balanceo de la clase deserción. Este balanceo genera nuevos registros, para entonces considerar una nueva proporción para el atributo de deserción. Lo anterior se hace debido a que puede presentarse una tendencia en la predicción, al considerar mayormente la "NO deserción", que representa el 67.75% del total de dicha variable, proporción que puede incidir en la posterior ejecución del algoritmo de predicción. En otras palabras, este valor puede incidir en los resultados, haciendo que los algoritmos tiendan a predecir, en la mayoría de oportunidades, registros con el valor NO, generando curva de aprendizaje de alta precisión para este valor, pero a su vez, afectando negativamente la curva de aprendizaje del resultado SI. Tomar decisiones empleando el algoritmo con mejor desempeño en estas condiciones, puede sesgar el resultado.

Se presenta el porcentaje correcto de predicción correcto para cada valor en los algoritmos J48 y NaiveBayes, mejores en cuanto a indicadores de desempeño escogidos.

| Algoritmo | % correcto valor SI | % correcto valor NO | PROMEDIO |
|---|---|---|---|
| J48 | 47,50% | 83,10% | 71,50% |
| NaiveBayes | 37% | 89,10% | 72,10% |

*Figura 41. Nivel de predicción con cada algoritmo*

En la figura 41, se observa que el porcentaje de predicción es alto, sin embargo, hay un desbalance en la predicción para cada valor que toma la deserción, logrando un nivel alto para cuando es [NO] pero bajo cuando es [SI], pues este último se ubica por debajo del 50%. Considerando que el objetivo del proyecto es analizar la probabilidad de deserción de los estudiantes, estos porcentajes pueden no terminar siendo los más adecuados, y se propone continuar evaluando nuevos resultados. Se decide entonces validar los resultados obtenidos cuando se aplica una técnica de balanceo, para ver el

impacto que se tiene sobre diferentes indicadores, en especial sobre los últimos mencionados.

Para hacer este balanceo, los registros con valores [NO] en la deserción, se dejan intactos, y se crean registros a partir de la información perteneciente a aquellos estudiantes con valor de deserción [SI], por medio de un algoritmo que se basa en los registros reales, para crear instancias a partir de los patrones encontrados. Este algoritmo es llamado SMOTE (Syntethic Minority Oversampling Technique), el cual fue usado para duplicar la cantidad de registros con valor de deserción igual a [SI]. Es importante aclarar que este algoritmo altera la cantidad de datos, más no la realidad encontrada en los patrones de estudiantes con valor de deserción igual a SI, pues estos nuevos registros poseerán características similares a aquellos que se encuentran en la base de datos.

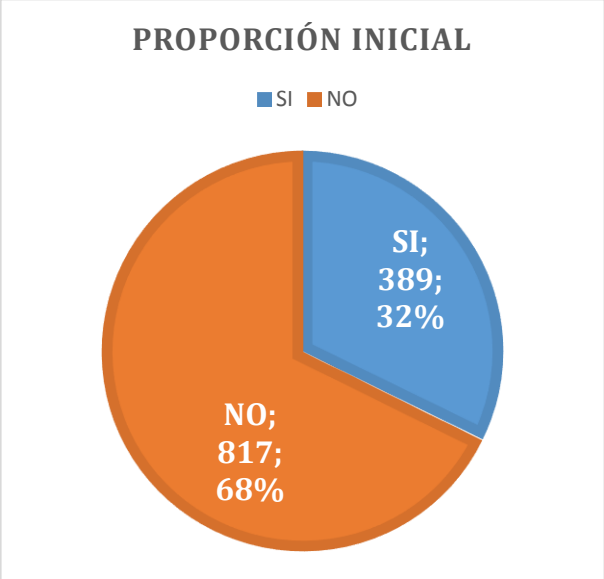



*Figura 42. Proporción de valores de Deserción iniciales*

Este proceso fue efectuado en varios proyectos de Educational Data Mining (EDM) mejorando significativamente los indicadores de desempeño de la clase a pesar de volúmenes menores de número de registros. Sin embargo, no existe un sustento teórico que indique hasta qué punto pueden considerarse adecuados estos registros sintéticos. Lo expresado anteriormente tiene como propósito aclarar que la obtención de un mejor desempeño, pero debido a que se sale del alcance del proyecto obtener nuevos datos que mejoren el nivel de predicción, se aplica este algoritmo a los registros. El resultado final de la proporción de los dos valores de la deserción, se presenta en la Figura 44.

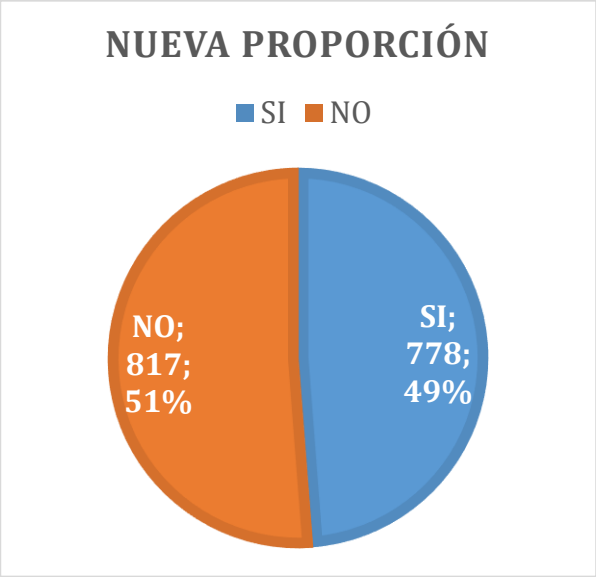



*Figura 43. Proporción con SMOTE realizado*

Esta nueva proporción va a evitar posibles sesgos en la predicción de los algoritmos y se evidencia la mejora en los algoritmos J48 y NaiveBayes, exhibiendo el antes y después de este procedimiento.

| Algoritmo | % correcto valor SI | | % correcto valor NO | | PROMEDIO |
|---|---|---|---|---|---|
| | INICIAL | NUEVA | INICIAL | NUEVA | |
| *J48* | 47,50% | 74,70% | 83,10% | 70,50% | **71,50%** |
| *NaiveBayes* | 37% | 81,40% | 89,10% | 64,60% | **72,80%** |

*Figura 44. Comparación de predicciones con SMOTE aplicado*

Se constata una mejoría en los porcentajes de predicción del valor SI, además de una equiparación, vista más que todo en el algoritmo J48, de los niveles de predicción, y contribuye a obtener modelos con mejores indicadores de predicción.

Ahora se procede a analizar los dos algoritmos resaltados en la figura anterior, tanto con los datos balanceados como con los reales, con el fin de obtener un punto de comparación de desempeño final que permita generar bases para la posterior elección de mejor algoritmo que se ajuste a los datos.

#### *5.2.3.4.1 Análisis para los mejores algoritmos con SMOTE*

Con SMOTE aplicado, se procede a hacer un análisis de los algoritmos NaiveBayes y J48. En las siguientes tablas se muestran los resultados generados por cada algoritmo y sus respectivos indicadores con un 95% de confianza.

| J48 | | | | | |
|---|---|---|---|---|---|
| **Seed** | **Predicción** | **Error Absoluto de la Media** | **F measure** | **ROC Area** | **Precisión** |
| 1 | 72,54% | 33,58% | 72,50% | 76,90% | 72,60% |
| 2 | 73,04% | 33,30% | 73,00% | 77,50% | 73,20% |
| 3 | 73,98% | 33,09% | 74,00% | 77,40% | 74,20% |
| 4 | 74,16% | 33,03% | 74,10% | 77,00% | 74,40% |
| 5 | 72,92% | 33,46% | 72,90% | 77,10% | 73,20% |
| 6 | 73,23% | 33,43% | 73,20% | 77,10% | 73,40% |
| 7 | 72,92% | 33,38% | 72,90% | 76,50% | 73,20% |
| 8 | 72,85% | 33,52% | 72,80% | 76,40% | 73,00% |
| 9 | 73,29% | 33,13% | 73,30% | 77,90% | 73,50% |
| 10 | 72,66% | 33,65% | 72,16% | 77,00% | 72,90% |
| **Promedio** | 73,16% | 33,36% | 73,09% | 77,08% | 73,36% |

*Figura 45. Resultados J48 con SMOTE*

| NaiveBayes | | | | | |
|---|---|---|---|---|---|
| **Seed** | **Predicción** | **Error Absoluto de la Media** | **F measure** | **ROC Area** | **Precisión** |
| 1 | 72,79% | 34,15% | 72,60% | 81,60% | 73,70% |
| 2 | 72,54% | 34,20% | 72,40% | 81,50% | 73,50% |
| 3 | 72,48% | 34,18% | 72,30% | 81,60% | 73,40% |
| 4 | 72,48% | 34,25% | 72,30% | 81,50% | 73,40% |
| 5 | 72,62% | 34,22% | 72,50% | 81,50% | 73,40% |
| 6 | 72,65% | 34,21% | 72,50% | 81,50% | 73,60% |
| 7 | 72,48% | 34,24% | 72,30% | 81,40% | 73,30% |
| 8 | 72,29% | 34,16% | 72,10% | 81,60% | 73,20% |
| 9 | 72,23% | 34,21% | 72,10% | 81,60% | 73,10% |
| 10 | 72,66% | 34,18% | 72,50% | 81,60% | 73,50% |
| **Promedio** | 72,52% | 34,20% | 72,36% | 81,54% | 73,41% |

*Figura 46.Resultados NaiveBayes con SMOTE*

#### *5.2.3.4.2 Análisis de los mejores algoritmos sin SMOTE*

Aquí solo se consideran los registros reales. Para el algoritmo de árbol J48 y el algoritmo NaiveBayes, los resultados se muestran a continuación:

| J48 | | | | | |
|---|---|---|---|---|---|
| **Seed** | **Predicción** | **Error Absoluto de la Media** | **F measure** | **ROC Area** | **Precisión** |
| 1 | 72,14% | 34,09% | 71,20% | 71,00% | 70,90% |
| 2 | 72,13% | 34,30% | 71,40% | 72,50% | 71,10% |
| 3 | 73,13% | 33,72% | 72,60% | 72,70% | 72,40% |
| 4 | 72,22% | 34,30% | 72,20% | 71,50% | 71,30% |
| 5 | 72,63% | 33,68% | 72,20% | 72,20% | 72,00% |
| 6 | 73,47% | 33,45% | 73,10% | 74,20% | 72,90% |
| 7 | 72,97% | 33,67% | 72,50% | 73,40% | 72,30% |
| 8 | 71,60% | 34,71% | 70,50% | 70,80% | 70,30% |
| 9 | 72,30% | 34,28% | 71,70% | 72,50% | 71,50% |
| 10 | 72,10% | 34,17% | 71,20% | 71,80% | 70,90% |
| **Promedio** | 72,47% | 34,04% | 71,86% | 72,26% | 71,56% |

*Figura 47. Resultados J48 sin SMOTE*

| NaiveBayes | | | | | |
|---|---|---|---|---|---|
| **Seed** | **Predicción** | **Error Absoluto de la Media** | **F measure** | **ROC Area** | **Precisión** |
| 1 | 74,20% | 34,35% | 72,60% | 77,80% | 73,00% |
| 2 | 73,71% | 34,16% | 72,10% | 78,30% | 72,40% |
| 3 | 73,54% | 34,31% | 72,00% | 77,80% | 72,20% |
| 4 | 73,60% | 34,24% | 72,00% | 78,10% | 72,30% |
| 5 | 73,21% | 34,28% | 71,60% | 78,00% | 71,80% |
| 6 | 73,30% | 34,32% | 71,70% | 77,80% | 71,90% |
| 7 | 73,63% | 34,41% | 72,00% | 77,60% | 72,30% |
| 8 | 73,30% | 34,38% | 71,60% | 77,70% | 71,90% |
| 9 | 73,55% | 34,26% | 72,00% | 78,00% | 72,20% |
| 10 | 73,50% | 34,23% | 72,10% | 78,10% | 72,20% |
| **Promedio** | 73,55% | 34,29% | 71,97% | 77,92% | 72,22% |

*Figura 48. Resultados NaiveBayes sin SMOTE*

En estos cuadros, el número semilla o "Seed", permite la generación de números pseudoaleatorios, los cuales generan dispersión en las secuencias escogidas en la generación de *folds* (Ver Glosario) en el proceso de validación cruzada, con el fin de que el algoritmo pueda ser evaluado con distintas secuencias de escogencia de datos, y se generen diferentes resultados de acuerdo al cambio de las mismas. Esto se efectúa con el único fin de tener suficiente evidencia que certifique los valores entregados por cada algoritmo, obteniendo suficiente información para determinar la escogencia del mejor desempeño.

La Figura 49 presenta la el promedio de las medidas de cada indicador. Para la evaluación de los algoritmos, se considera la precisión como el aspecto con más peso, pues de este depende la construcción de la herramienta, seguido de los otros indicadores, los cuales tienen como función principal soportar la decisión tomada.

| PROMEDIO GENERAL | | | | | |
|---|---|---|---|---|---|
| **Algoritmo** | **Predicción** | **Error Absoluto de la Media** | **F measure** | **ROC Area** | **Precisión** |
| J48 | 72,81% | 33,70% | 72,47% | 74,67% | 72,46% |
| NaiveBayes | 73,04% | 34,25% | 72,17% | 79,73% | 72,82% |

*Figura 49. Desempeño general de cada algoritmo*

En esta misma figura, la cual tiene como resultado el promedio de las medidas obtenidas en los algoritmos, el desempeño es muy similar. Se puede ver que el algoritmo con mejores indicadores hasta el momento es el NaiveBayes, el cual tiene tres de cinco indicadores con mejor desempeño. Respecto al algoritmo J48, se puede observar que entrega mejores resultados en error absoluto y F measure, aspectos que también se deben tener en cuenta en la toma de decisión.

Por este motivo, se decide hacer una etapa final de predicción, donde además de tener en cuenta los registros, deben pasar por una restricción de número mínimo de estudiantes en cada agrupación, la cual es de gran importancia para la representatividad del modelo, y la obtención de información de alta fidelidad. A continuación, se describe el proceso efectuado en la última etapa de análisis, previo a la construcción de las herramientas.

### 5.2.4 Resultados finales del modelo de predicción

Como último paso dentro de este objetivo, se procede a la generación de modelos de predicción con el algoritmo J48 y NaiveBayes, ya con todas las consideraciones detalladas hasta el momento, que serán finalmente los que conlleven a la creación de las herramientas que aporten al tema de la deserción. Dichas herramientas contendrán tanto datos de referencia encontrados a lo largo del proyecto, como un modelo de predicción de la deserción que pueda ser aplicado tanto para la etapa de admisión, como de seguimiento o apoyo al estudiante durante su vida universitaria.

Se procede entonces a generar 2 modelos de predicción con los atributos ya identificados y bajo las condiciones mencionadas, pero además, jugando un poco con las numero de estudiantes mínimo que obtendría cada rama del árbol J48 y cada probabilidad causal del NaiveBayes. Estos números mínimos, darán mayor representatividad conforme al número aumenta, pues tener una rama de decisión con tan solo unos 5 casos seguramente será muy poco representativo frente al total de

datos, mientras que ramas con mayor cantidad de casos podrán ser más representativas.

El número mínimo de estudiantes finalmente asumido es de 35, lo cual permitiría tener un modelo con un nivel de predicción alto, que al observar los resultados del algoritmo J48 con un nivel de predicción 74.225%, se confirma. En adicción a esto, la cifra de 35 estudiantes como mínimo, permite obtener un buen indicador de representatividad, pues evita el surgimiento de ramas sin representatividad y permite la creación de caminos con un total de estudiantes aún superiores al mínimo establecido, lo cual termina siendo entonces un gran aporte, pues ilustra la realidad, con patrones realmente representativos.

Con respecto a este número de estudiantes, se podría creer que iría deteriorando el modelo y su predicción conforme aumenten en número, más sin embargo no es así, puesto que bajo varios casos, el modelo tiende a predecir erróneamente cuando existen ramas poco representativas, que en caso de ser eliminadas, conllevan a la obtención de una predicción mejor del algoritmo.

Se logra entonces generar un modelo con indicadores altos, que además logra un nivel de precisión del 67.4% para aquellos desertores y del 75.9% para los no desertores. En cuanto al indicador de falsos positivos, es tan solo del 9.6%, que indica que el modelo tiende a considerar a muy pocos no desertores dentro de aquellos que clasifica como desertores, siendo esto algo importante a mencionar, pues el impacto negativo, en este caso, termina siendo bastante reducido. Finalmente se presentan los datos obtenidos con WEKA para dicho modelo planteado.

| Resultados del modelo de prediccion con los datos historicos | | | | | |
|---|---|---|---|---|---|
| CLASS | TP RATE | FP RATE | PRESICION | F - MESURE | ROC AREA |
| SI | 0,408 | 0,096 | 0,674 | 0,508 | 0,745 |
| NO | 0,904 | 0,593 | 0,759 | 0,825 | 0,745 |
| AVERAGE | 0,742 | 0,43 | 0,731 | 0,722 | 0,745 |

| MATRIZ DE CONFUSION | | |
|---|---|---|
| | PREDICCIONES | |
| | DESERTA | NO DESERTA |
| DESERTORES | 163 | 237 |
| NO DESERTORES | 79 | 747 |

*Figura 50. Resultados modelo de predicción J48 con datos históricos*

Como conclusión, los resultados obtenidos con el algoritmo J48, para construir un modelo de predicción de la deserción, en este caso son bastante buenos, ya que se tiene un buen nivel de predicción, que es del 74.225%, una buena cantidad de datos para sustentar la predicción, pues es de mínimo 35 estudiantes, y además, un adecuado nivel de precisión, del 73.1% en promedio, que logra resultados eficientes para ambos posibles valores de deserción.

Estos datos terminan siendo incluso muy superiores en casi todos los aspectos cuando se compara con el algoritmo de NaiveBayes, el cual al incrementar el número de estudiantes mínimo, tiende a deteriorarse, siendo el nivel de predicción para 35 estudiantes como mínimo, de tan solo 72.1%, muy por debajo del obtenido con J48.

| RESULTADOS DE DESEMPEÑO PARA NAIVEBAIYES CON NUMERO MINIMO DE 35 ESTUDIANTES | | | | | | |
|---|---|---|---|---|---|---|
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **Fmeasure** | **ROC Area** | **Class** |
| 37.0% | 10.9% | 62.2% | 37.0% | 46.4% | 77.4% | SI |
| 89.1% | 63.0% | 74.5% | 89.1% | 81.1% | 77.4% | NO |
| 72.1% | 46.0% | 70.5% | 72.1% | 69.8% | 77.4% | |

| **Class** | **MATRIZ DE CONFUSION** | |
|---|---|---|
| SI | **148** | 252 |
| NO | 90 | **736** |
| **Class** | SI | NO |

*Figura 51. Resultados de desempeño de NaiveBayes para número mínimo de 35 estudiantes*

Debido a estos resultados, y teniendo en cuenta que el desempeño general de los dos algoritmos es relativamente similar, por motivos de representatividad del modelo se decide escoger el algoritmo J48 como el indicado para ser base de la construcción de las herramientas.

Teniendo en cuenta esta elección, se parte de la premisa de que el modelo de predicción generado mediante la implementación de la técnica SMOTE, parece ser bastante adecuado, pues logra una predicción correcta en el 73.98% de los casos, y además, tiene niveles de precisión por encima del 70% para ambas posibles predicciones de deserción o no deserción. Esto supone algo bastante positivo, pues se logra un muy buen cubrimiento sobre aquellos posibles desertores, y además el modelo termina siendo eficiente, pues de los que señala como desertores, finalmente lo son cerca del 72.7% de los casos.

También, es importante tener en cuenta un modelo sin este algoritmo de creación sintética de registros, pues esta técnica altera la realidad en el sentido de que aumenta los registros de una clase en particular, lo cual puede ser perjudicial para un modelo que realmente posea fundamentos que soporten la generación de probabilidades de deserción.

## 5.3 Construir una herramienta de apoyo a los procesos de admisión y al diseño de estrategias que sustentan el acompañamiento al estudiante en su vida universitaria.

Como objetivo final de este proyecto, se pretende materializar y dar utilidad a todo el conocimiento que fue extraído de la base de datos creada y de los modelos de predicción generados. Es por ello que se propone la creación de una herramienta, que pueda alimentarse de dicho conocimiento y que permita identificar la probabilidad de deserción de los estudiantes de Ingeniería Industrial de la Universidad Icesi, para entonces apoyar tanto los procesos de admisión, como el proceso de seguimiento y apoyo a los estudiantes.

Mediante la implementación de data mining y de técnicas estadísticas, se logró obtener información relevante sobre el impacto que tienen ciertos atributos del estudiante, frente al tema de deserción presentado en la carrera de Ingeniería Industrial de la Universidad Icesi.

Los modelos que se lograron obtener a partir de todo el estudio, y que puntualmente se muestran en el objetivo 2 de este proyecto, fueron todo el tiempo encaminados hacia el aporte que podrían tener para la construcción de la herramienta, pues se requería destacar 2 modelos de predicción con diferentes dimensiones, ya que para la etapa de admisión se puede contar con cierta cantidad de atributos, sin incorporar otros atributos que, para la etapa de apoyo al estudiante, si estarían presentes.

### 5.3.1 Descripción del proceso de admisión para el programa de Ingeniería Industrial

Para tener un mayor entendimiento sobre los procesos de admisión que se llevan a cabo en la Universidad Icesi, y así mismo argumentar e identificar el tipo de intervención que

podría llevarse a cabo con la herramienta que se va a construir, se presenta el siguiente gráfico.

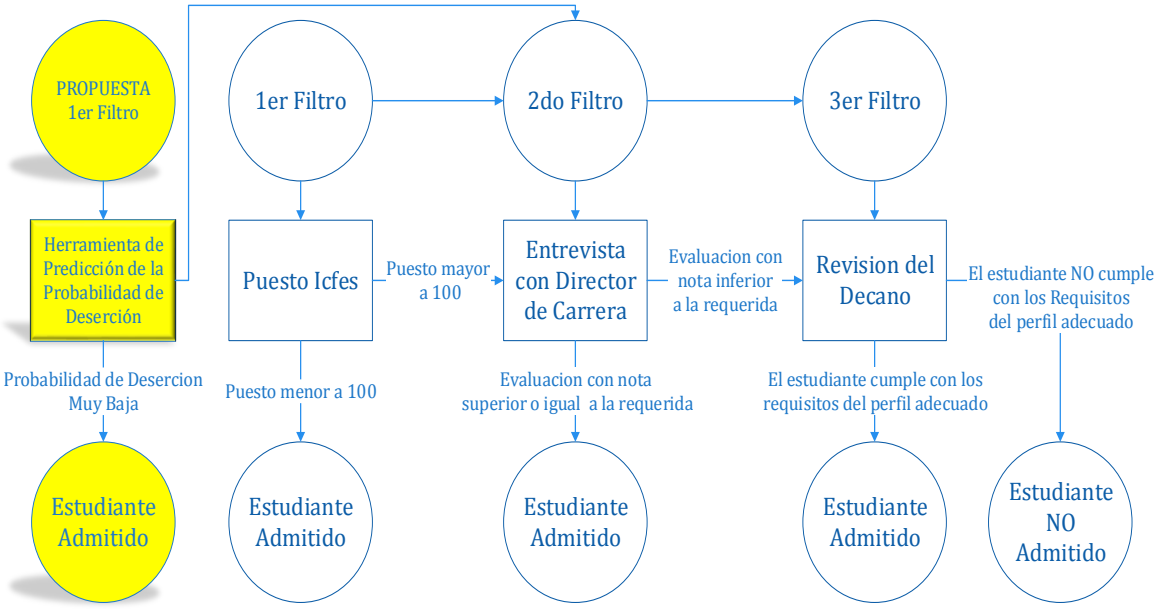



*Figura 52. Etapas de filtro de la Universidad y la propuesta planteada*

En el grafico se pueden ver los 3 filtros actualmente existentes para el proceso de admisión. Para el momento en que se realiza el primer filtro, el estudiante proporciona cierta información que solicita Icesi, sin embargo solo se tiene en cuenta el puesto del estudiante en el Icfes, donde un puesto menor o igual a 100, le permite ser admitido directamente. Aquí se identifica una falencia, pues se cuenta con más información del estudiante, pero no entra a ser considerada para determinar la admisión o no del estudiante, en ese instante.

Es por ello que se considera que la herramienta de predicción de la probabilidad de deserción, que se propone en este proyecto, debería ser el primer filtro, pues incorpora no solo el puesto del Icfes que obtuvo el estudiante, sino muchos más atributos que podrían dar mayor contundencia al momento de identificar aquel perfil de estudiante ideal, que entre otras cosas, debería NO desertar, puesto que se ha evidenciado que es un problema importante y que la Universidad intenta reducir, o incluso eliminar.

### 5.3.2 Creación de una herramienta como primer filtro

Una vez identificadas las partes críticas en donde puede ser utilizada la información obtenida anteriormente, se procede a crear la primera herramienta de predicción de la probabilidad de deserción. Como se pudo constatar en el primer objetivo, no existe una correlación clara entre el puesto Icfes y la deserción, donde el puntaje obtenido en las pruebas, tampoco presenta una fuerte correlación con el puesto ocupado en las mismas. Se implementa el algoritmo J48 eliminando aquellos atributos que aún no se conocen en esa instancia, como son las dimensiones de aprendizaje y el uso de beca.

En el objetivo anterior se pudo observar que hubo dos maneras de trabajar los algoritmos de predicción, una de estas fue utilizando netamente los datos reales, los cuales pueden dar como resultado una predicción adecuada para la instancia NO, y una inferior para la instancia SI de la deserción. La otra fue con un algoritmo generador de instancias ficticias llamado SMOTE, el cual mejoró significativamente la predicción de la instancia SÍ y balanceó la proporción de registros.

Con esta información clara, se procede a crear dos versiones de la herramienta para primer filtro, teniendo como único fin ver el desempeño de cada una al ser testeada con un conjunto de datos del modelo real, que al final ilustrará cuál de las dos maneras logra mejor desempeño la hora de predecir la deserción.

#### 5.3.2.1 Creación de Herramienta para primer filtro sin SMOTE

Cuando se aplica el algoritmo J48 a los datos, sin los atributos de beca y dimensiones de aprendizaje, el porcentaje de predicción baja ostensiblemente. Esto, lo que demuestra es que los atributos eliminados juegan un papel importante en la predicción de la probabilidad de deserción.

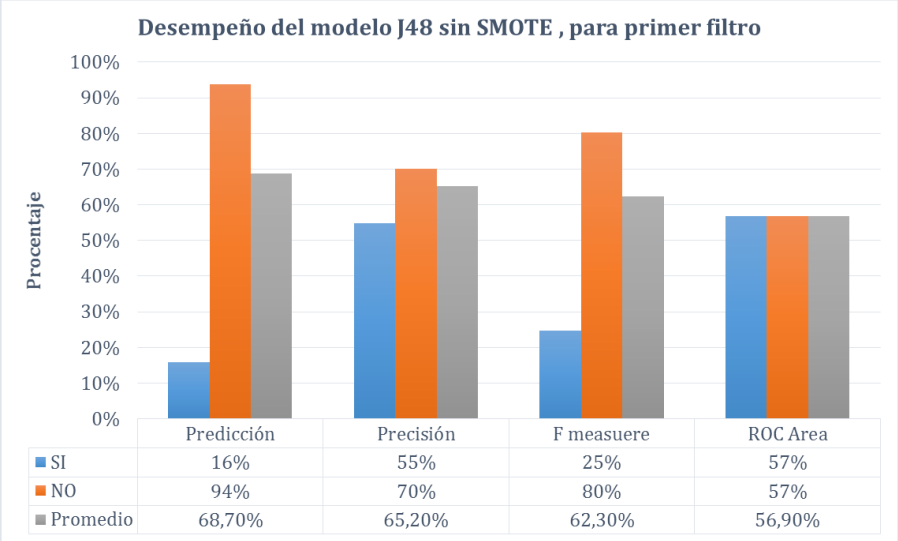



*Figura 53.Desempeño del algoritmo J48 sin SMOTE, para primer filtro*

Como se observa en la figura 53, este modelo tiene un comportamiento similar al obtenido testeando con todos los atributos. Sin embargo, sus indicadores son más bajos como lo muestra el promedio, además la diferencia entre las instancias SI y NO es aún más marcada.

En el anexo (1) encontrarán el árbol de decisión creado con estos registros y a continuación se presenta la importancia que le dio el algoritmo a cada atributo, para la posterior creación del árbol de decisión.

| **Ranquin** | |
|---|---|
| ***Atributo*** | ***Importancia*** |
| Puntaje ICFES | 42,8% |
| Categoria Colegio | 39,9% |
| Género | 17,3% |
| Estrato | 0,0% |
| Edad Ingreso | 0,0% |

*Figura 54. Resultados Ranquin para primer filtro sin SMOTE*

Se puede observar que al no estar el atributo beca, ni las dimensiones de aprendizaje, el atributo Puntaje ICFES pasa a ser el principal decisor del modelo de predicción. Respecto a los resultados de predicción para la clase SI, se puede concluir que son muy bajos, por lo que se decide no tener en cuenta aquellos caminos, puesto a que no

representan información fundamentada, en la creación de la herramienta para primer filtro sin SMOTE.

Además de esto, se decide eliminar los caminos no representativos, es decir aquellos en los cuales el algoritmo toma la decisión con muy pocos registros, pues la toma de decisión sobre la probabilidad de deserción debe ser bien fundamentada y no obtenida a partir de un patrón encontrado en pocos registros.

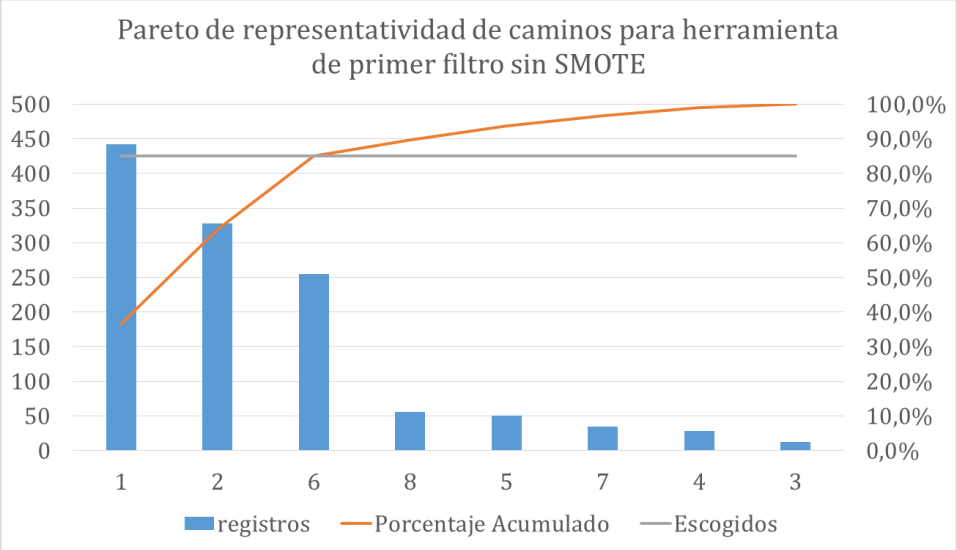



*Figura 55. Pareto de representatividad de caminos para herramienta de primer filtro sin SMOTE*

En la figura 55 se puede observar que el algoritmo genera un total de 8 caminos (para ver descripción de caminos (ver Anexos 1 y 2), de los cuales tres representan el 85% del total de registros (regla de Pareto). Estos tres caminos, presentan diferentes maneras de saber cuándo un estudiante tiene pocas probabilidades de desertar, los cuales tienen soporte en los indicadores de desempeño ilustrados anteriormente para la instancia NO, la cual proporciona un 94% de predicciones correctas.

Con todo lo anterior, se crea la herramienta que determina que estudiante debe pasar a una segunda instancia, debido a que no cumple ninguno de los criterios establecidos en alguno de los tres caminos de decisión. Como esta herramienta podría ser utilizada por las personas de la dependencia de registro académico, quienes pueden o no tener un conocimiento intermedio o avanzado de las hojas de Excel, se decide replicar el árbol, con los caminos que son representativos, por medio de una macro. De esta

manera, la herramienta va recorriendo los caminos de decisión y tiene como resultado final si el estudiante debe pasar a entrevista con el Director de Programa o no. Esta decisión se fundamenta en el no cumplimiento de ninguno de los atributos que tiene aquellos estudiantes con menores probabilidades de deserción.

Esta herramienta se encuentra adjunta en el CD perteneciente a este proyecto de grado, donde se pueden observar las macros creadas, que le otorgan la practicidad en el momento de uso.

#### 5.3.2.2 Creación de herramienta de primer filtro con SMOTE

Los resultados obtenidos por el algoritmo de árbol J48, con SMOTE aplicado, se presentan en los anexo 3, 4 y 5. Este árbol ilustra los diferentes caminos de decisión sobre el atributo deserción y la probabilidad de que estos resultados se cumplan.

Respecto a los indicadores de desempeño, para este caso se obtienen peores resultados a los obtenidos sólo con datos reales, lo que puede indicar que el modelo no es susceptible de mejora con un balanceo de clases, marcando un indicio que la heterogeneidad de los registros sobresale como característica principal del modelo.

Se puede observar que se logra un balanceo en la predicción de ambas clases, característica que no posee el modelo de datos reales. En la figura 59 se muestra la importancia que el algoritmo dio a cada atributo, para la posterior construcción del árbol de decisión.

Como se puede observar tanto en la figura como en el anexo mencionado anteriormente, el puntaje ICFES también es el principal atributo decisor, y otros atributos como la edad de ingreso, la categoría del colegio y el género aportan información relevante en el análisis de la probabilidad de deserción final. Por este motivo, se procede a crear una herramienta con SMOTE, que tenga en cuenta todos estos atributos, para la primera etapa de filtro.

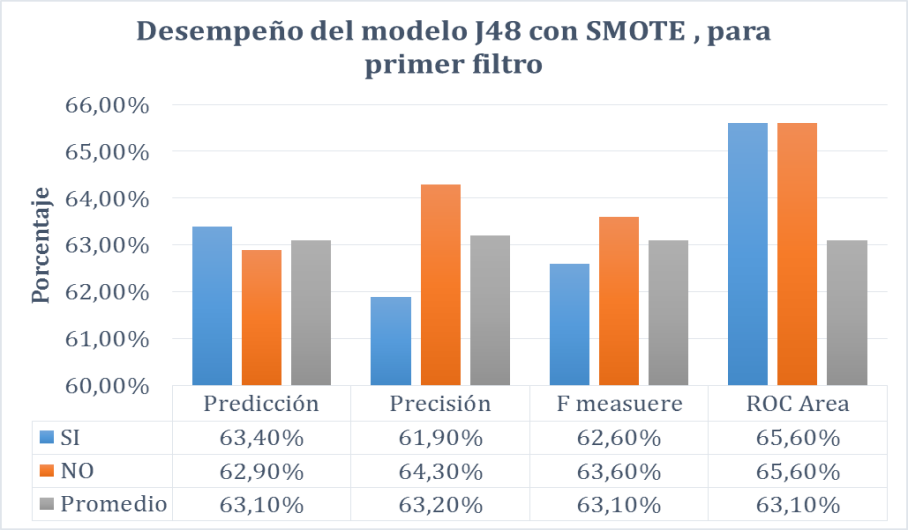



*Figura 56. Desempeño del modelo J48 con SMOTE, para primer filtro*

| Puesto | Atributo | Porcentaje |
|---|---|---|
| **1** | *Puntaje ICFES* | 34,7% |
| **2** | *Edad Ingreso* | 28,8% |
| **3** | *Categoría del Colegio* | 19,5% |
| **4** | *Género* | 15,2% |
| **5** | *Estado Civil* | 1,8% |
| **6** | *Estrato* | 0,0% |

*Figura 57. Resultados Ranquin para primer filtro con SMOTE*

Antes de crear esta herramienta, se hace un procedimiento similar al efectuado para la primera herramienta creada, en el cual es obtenido un Pareto, donde se tienen en cuenta la cantidad de estudiantes al final de cada rama de árbol. Ahora bien, con el objetivo de tener la mayor cantidad de caminos posibles, se decide eliminar aquellos, donde el algoritmo toma la decisión con un total de 30 estudiantes o menos, pues su grado de representación, frente a los 1700 registros estudiados, es muy bajo.

A continuación, se muestra el resultado del Pareto y en el anexo 3, se muestran las etapas de decisión detalladas de cada camino.

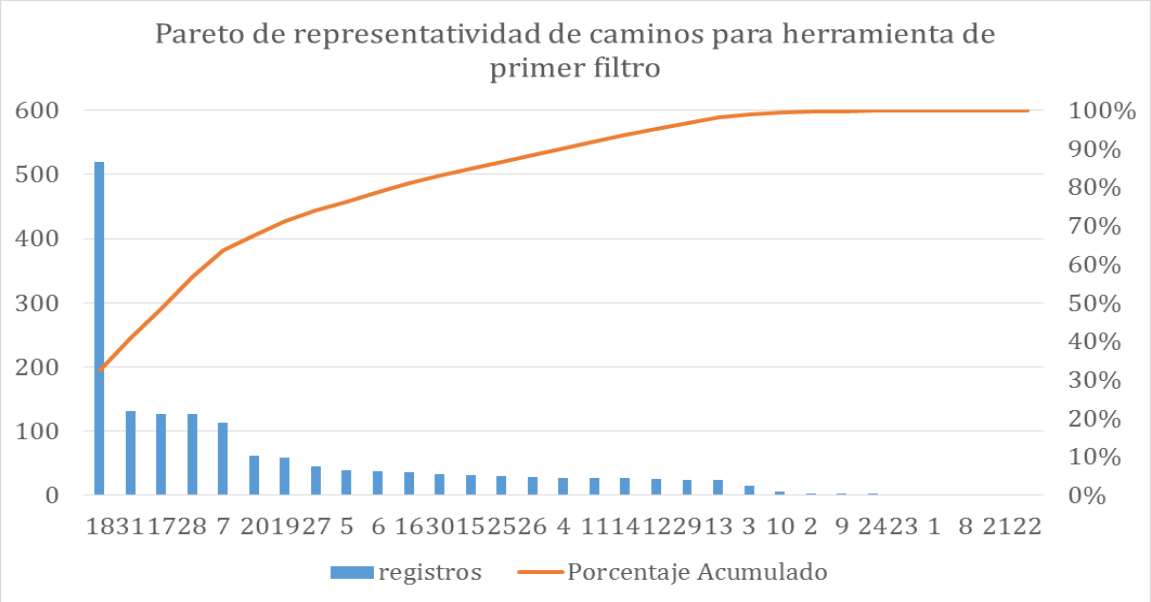



*Figura 58. Análisis de Pareto a caminos para herramienta de primer filtro*

Una vez escogidos aquellos caminos representativos, se procede a crear la herramienta para el filtro de admisión. Esta herramienta se crea por medio de una macro en Excel, que permite recorrer todos los caminos escogidos por medio de condicionales anidados, totalmente interactiva y obteniendo como resultado final, la probabilidad de deserción del estudiante y una instrucción posterior de si el estudiante queda admitido inmediatamente o debe pasar a entrevista con el director de programa.

La anterior decisión se basa en el porcentaje de deserción obtenido, el cual si es menor de 50%, significa que el estudiante tiene pocas posibilidades de desertar de acuerdo con los datos suministrados, por lo cual puede ser admitido y pasar a la etapa de matrículas.

Es importante dejar claro que esta herramienta sólo es útil en esta primera etapa de filtro, pues es claro que la información faltante, es decir el atributo beca y los estilos de aprendizaje, desempeña un papel importante en la etapa posterior a la matricula,

cuando el estudiante empieza a cursar la carrera, donde los procesos de apoyo pueden ser de gran utilidad en la disminución del fenómeno de la deserción.

Esta herramienta (Ver anexo 6) también se encuentra adjunta en el CD de este proyecto de grado en el mismo archivo de Excel de la primera herramienta, esto con el fin de que se pueda observar el proceso en cada una de estas, y se puedan comparar los resultados obtenidos en cada una de las etapas de construcción.

#### 5.4.3.3 Testeo de verificación de herramientas con los datos reales

Después de la creación de las herramientas, se hace un test de verificación que pretende ver el desempeño de cada modelo, cuando se le da a predecir los mismos datos reales con los que se construyó. A la vista de los proyectos de *Data Mining*, esto es un procedimiento erróneo, pues la verificación de un modelo debe hacerse con datos diferentes a los que se utilizaron, ya que, si se efectúa este procedimiento con la misma información, se espera que los niveles de predicción sean elevados y no den posibilidad de ver el desempeño real del modelo con información nueva.

Al no contar con la posibilidad de utilizar información nueva, debido a que los semestres no utilizados en este modelo no entran en el alcance de este proyecto, el cual abarca aquellas cohortes que están en su cuarto semestre o más adelante, se decide efectuar el procedimiento con la misma información obtenida para construirlo. Se espera que los modelos tengan niveles de predicción mayores o iguales a los obtenidos al momento de aplicar los algoritmos al conjunto de datos, por lo cual, cualquier resultado que no presente este comportamiento en los indicadores, puede conducir a que el modelo no tiene una precisión adecuada y que la heterogeneidad de los datos no permite que se logre un buen indicador.

Curiosamente, esta primera evaluación muestra algo bastante desviado de la teoría para el modelo con SMOTE aplicado, aunque también comprensible dada la intervención y adecuación de los datos. Este modelo de predicción demostró tener un

nivel general de predicción bastante similar al inicial, siendo este del 68.3%, sin embargo los niveles de predicción ahora son bastante inferiores, siendo del 57.5% para las predicciones de quienes desertan, lo cual demuestra nuevamente un deterioro de este nivel de eficiencia del modelo, que seguramente surge de aquel balanceo, en donde ahora el algoritmo tiende a clasificar como desertores a aquellos que antes no lo eran.

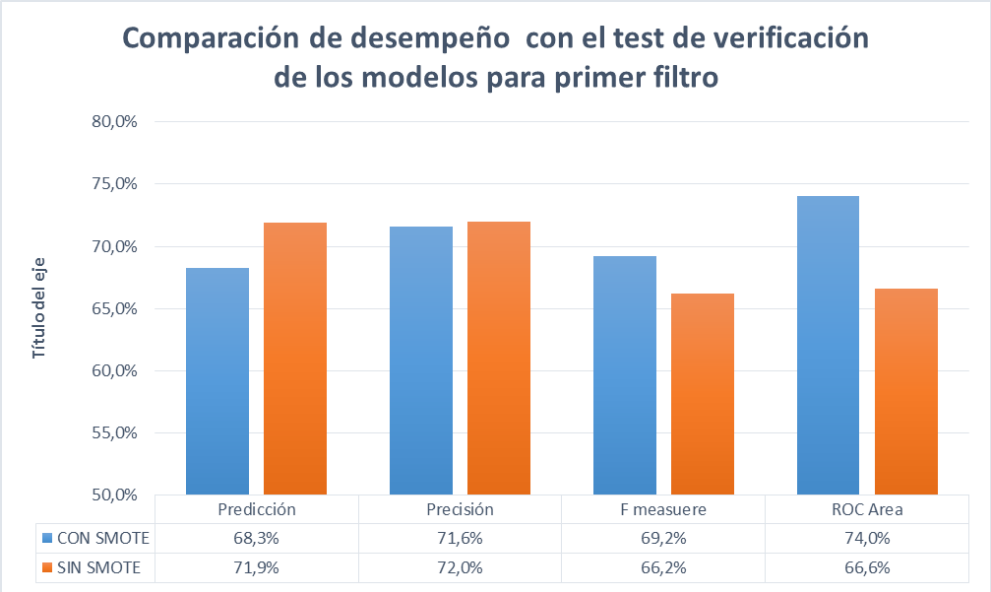



*Figura 59.Comparación de desempeño de herramientas para primer filtro*

Los indicadores obtenidos luego de efectuar este test de verificación, ilustran que efectivamente hubo un desempeño comparativamente alto de los indicadores, respecto a los obtenidos en los modelos originales, tanto para el modelo sin SMOTE, como para el creado con SMOTE, a pesar del deterioro visto en su predicción.

En este caso, se puede ver que el modelo con los datos reales tiene mejores indicadores de precisión y predicción, también se observa que el balancear las clases, para este caso, no se obtienen resultados significativamente mejores a los obtenidos con la información real.

Ahora bien, entrando a analizar el modelo creado con SMOTE, los resultados e indicadores mencionados anteriormente, fueron obtenidos producto de un proceso una

validación cruzada, lo que podría traer conflictos frente a la realidad, pues el modelo se evalúa con varios folds, pero todos ellos considerando siempre la adaptación que de primer momento realizo el algoritmo SMOTE, al crear datos ficticios para balancear el nivel de deserción.

Debido a lo anterior, se procede a realizar otro tipo de evaluaciones, justamente intentando evitar que la validación pueda verse influenciada por el algoritmo SMOTE y así mismo evitar que sus resultados se ajusten de acuerdo a la información alterada que este mismo modelo obtiene. Se plantea entonces una evaluación más para este modelo en específico, que parte del manejo del tema de data mining y que surge como una propuesta dentro de este proyecto, dado el caso que se presenta.

Lo que se propone es hacer un nuevo tipo de testeo, esta vez dividendo la base de datos real en 65% para el entrenamiento del modelo "training" y del 35% para el testeo "test". Dicha división de los datos se realiza de manera aleatoria en Excel, y posteriormente se hace una comprobación que permita que el balanceo de desertores siga presente y se respete la proporción existente en la realidad, donde finalmente se obtuvo una deserción entre 34% y 31% de deserción, lo cual se puede considerar dentro de la realidad de deserción actual, que es del 32.6% aproximadamente.

Se tiene entonces una base de datos de unos 800 estudiantes que permite crear el modelo mediante el algoritmo J48 y además aplicando el SMOTE. Con ello se obtiene nuevamente un modelo tal y como se había obtenido previamente, pero ahora si se podrá poner aprueba con datos reales y que no se han considerado para la construcción del mismo, evitando entonces que el testeo pueda acomodarse a sus datos de entrada y que el mismo pueda afectarse por el algoritmo de SMOTE, pues los otros 400 estudiantes para la etapa del testeo se dejarán intactos.

Los resultados nuevamente muestran un deterioro considerable, pues ahora la predicción baja al 61.3% y los niveles de precisión se ubican en promedio en un 60%, pero donde es tan solo del 40.2% para cuando predice que el estudiante será desertor.

Estos testeos realizados, demuestran que la aplicación del algoritmo SMOTE no conlleva a nada seguro, y que en cambio podría estar deteriorando la predicción para este caso. Por este motivo, y al no encontrar sustento teórico que valide si es correcta la aplicación de SMOTE en modelos de predicción, se decide dar mayor validez a la herramienta creada a partir de los datos reales. Sin embargo, las dos herramientas quedan disponibles para ser testeadas con nueva información, donde se pretende que posteriormente puedan ser implementadas en los procesos de primer filtro.

### 5.3.3 Creación de una herramienta para identificar a aquellos estudiantes con alta probabilidad de deserción, que sugiera apoyo a dichos estudiantes

Esta herramienta se crea utilizando el modelo de predicción obtenido con todos los atributos considerados, pues ya para la etapa de seguimiento al estudiante, la Universidad Icesi contará con la información de las dimensiones o estilos de aprendizaje de sus estudiantes y también se sabrá que estudiantes están becados.

Esta herramienta es presentada en un documento de Excel y consiste en una matriz que solicita la información del estudiante y posteriormente muestra la probabilidad de deserción del mismo y también indica la población o muestra con que se argumenta dicha predicción. Se encontrará una escala de colores del verde al rojo, donde aquellos valores en rojo indican la alta probabilidad de desertar y por ende sugieren una especial atención (alerta primaria) en dichos estudiantes. El continuo seguimiento al estudiante puede alinearse con estrategias que den asistencia en temas académicos, o incluso económicos o sociales, que mitiguen las causas de la deserción

| CODIGO | BECA | ESTRATO | CAT COLEGIO | GENERO | DIM 2 | DIM 4 | DESERCION | PROB. DESERCION | MUESTRA |
|---|---|---|---|---|---|---|---|---|---|
| 12457845 | Si | 3 | MUY SUPERIOR | MASCULINO | SENSORIAL | GLOBAL | NO | 6,977% | 387 |
| 56894512 | NO | 6 | MUY SUPERIOR | FEMENINO | INTUITIVO | SECUENCIAL | NO | 37,143% | 35 |
| 45154826 | SI | 4 | INFERIOR | FEMENINO | INTUITIVO | SECUENCIAL | NO | 6,977% | 387 |
| 26954878 | NO | 5 | BAJA | FEMENINO | SENSORIAL | GLOBAL | SI | 61,905% | 21 |
| 12569132 | NO | 2 | MEDIA | MASCULINO | SENSORIAL | SECUENCIAL | SI | 70,635% | 126 |
| 23675489 | NO | 5 | ALTA | MASCULINO | INTUITIVO | SECUENCIAL | SI | 65,517% | 29 |
| 45175829 | NO | 6 | SUPERIOR | FEMENINO | INTUITIVO | GLOBAL | NO | 42,553% | 47 |
| 26758416 | NO | 4 | MUY SUPERIOR | MASCULINO | SENSORIAL | GLOBAL | NO | 32,500% | 480 |

*Figura 60. Vista a Herramienta de procesos de apoyo*

### 5.3.4 Identificación de datos de referencia, que sirvan como material de apoyo frente al tema de deserción.

Es un material programado en un documento en Excel, que registra datos de referencia identificados a lo largo del proyecto, e incorpora información obtenida por el modelo de predicción final, conformada por ramas de predicción con algunas particularidades interesantes y objeto de ser analizadas. Todo está sustentado por datos numéricos, a partir de inferencias estadísticas con gráficos de apoyo para dimensionar fácilmente los escenarios, y derivar conclusiones rápidas. La herramienta o documento se divide en varias pestañas, cada una haciendo alusión a un tema en particular, y se intenta incorporar toda la información pertinente que demuestre el valor de realizar este tipo de proyectos de minería de datos.

# 6. CAPITULO VI – Conclusiones y Recomendaciones

## 6.1 Conclusiones

- El algoritmo J48, implementado para generar modelos de predicción de la probabilidad de la deserción, logró los mejores resultados para este proyecto, alcanzando un nivel de predicción del 74.23%. Esto demuestra el valor de la minería de datos para este tipo de estudios, que pretenden obtener información relevante sobre el campo educacional, y en este caso sobre la deserción.

- En un modelo de predicción, incrementar el número de estudiantes con que se toman las decisiones, puede terminar mejorando el nivel de predicción, contrario a lo que se podría pensar, pues ahora el modelo será más representativo y así mismo no se verá afectado por la ausencia de datos muy particulares o atípicos, que durante la etapa de entrenamiento se presentaban.

- En la Universidad Icesi, para la carrera de Ingeniería Industrial, la tasa de deserción es del 32.6% a nivel general, mientras que para los estudiantes becados es tan solo de 9.3%. Por otra parte, aquellos estudiantes que no están becados y cuyos estratos son 1 o 2, presentan una tasa de deserción del 70%. Esto demuestra que obtener una beca, reduce significativamente la probabilidad de deserción. Por otra parte, se concluye que el factor económico influye considerablemente sobre la deserción.

- La aplicación de técnicas estadísticas sobre la base de datos estudiada, permite obtener un panorama claro de lo que se está analizando, dando paso a que la aplicación de algoritmos de minería de datos pueda lograr mejores resultados, ya que estructurar la información adecuadamente, es vital para generar un buen modelo de predicción.

- Mediante diferentes datos estadísticos obtenidos a lo largo del proyecto, se encontró una variedad de atributos con posibles patrones frente a la deserción, por lo cual se

infiere que, incluir datos como ciudad de nacimiento o cargo del padre del estudiante, entre otros, puede mejorar futuros modelos de predicción.

- Los atributos más relevantes para predecir la deserción, en orden de importancia son, Beca, Estrato, Categoría del Colegio, Género, Dimensión 1 y Dimensión 2. Finalmente fueron estos 6 atributos, los que permitieron predecir la probabilidad de deserción de los estudiantes, para el Programa de Ingeniería Industrial de la Universidad Icesi.

## 6.2 Recomendaciones

- Existen amplias posibilidades de plantear estudios más exhaustivos con información más detallada y ampliada, a partir de los datos almacenados y recolectados por parte de la Universidad Icesi, respecto a la población candidata a ingresar al programa de Ingeniería Industrial, y la matriculada con atributos de desempeño y estilos de aprendizaje.

- Se sugiere crear formularios web de inscripción con respuestas estándar, que contribuyan a registrar datos referentes al número de hermanos, ocupación específica y nivel de estudios de los padres, historia académica del candidato(a), e incluso, intereses extracurriculares. Estos aspectos enriquecen la capacidad de asertividad de los algoritmos de minería de datos, y por ende, las conclusiones que apoyen el proceso de selección y acompañamiento al estudiante.

- Con el análisis de correlación implementado para los atributos puntaje ICFES y puesto ICFES, se puede concluir que no existe una fuerte correlación entre estos. Así mismo, se infiere que el puesto ICFES, termina distorsionando la realidad que pretende evaluar esta prueba, que es el rendimiento del estudiante. Por lo tanto, se sugiere plantear la posibilidad de cambiar el procedimiento efectuado actualmente en la etapa de admisión de estudiantes, para el primer filtro, donde una posibilidad de cambio podría ser el uso de la herramienta desarrollada en este proyecto, creada a partir de la aplicación de las técnicas de Data Mining, que a pesar de contar con un porcentaje de

predicción relativamente bajo del 65%, con un 95% de confianza, serviría como un primer paso, para acercarse más a las características que inciden en que un estudiante tenga mayor probabilidad de deserción.

- De acuerdo al análisis efectuado con la totalidad de atributos, se puede evidenciar que el algoritmo con mejor desempeño, en este caso el J48, una vez aplicado el algoritmo de Ranquin, establece que uno de los atributos con menor importancia a la hora de predecir la deserción es el puntaje ICFES.
  Lo anterior, confirma una vez más que el proceso de admisión puede estar efectuándose con falta de información para evitar la deserción, que en este caso podría ser motivada por diversos aspectos, incluso aquellos diferentes al académico. Es por esto, que la segunda herramienta, creada para la etapa posterior a la matricula, abre las puertas para un análisis más profundo de los aspectos más comunes que provocan que un estudiante deserte, por lo cual se podría utilizar como guía en los procesos de apoyo específicos para cada estudiante.

- En el mediano plazo, sería ideal contar con formularios que evalúen los estilos de aprendizaje en el proceso de inscripción, pues como se puede observar en los datos obtenidos una vez aplicado el Data Mining, esta información puede aportar suficientes bases para tomar la decisión de admisión de cada estudiante.

## BIBLIOGRAFÍA


Dinero, R. (2 de Febrero de 2015). *Revista Dinero*. Obtenido de Revista Dinero: http://www.dinero.com/pais/articulo/niveles-desercion-universitaria-los-andes-colombia/205330

Lamamie de Clairac Palarea, F. (5 de Marzo de 2015). *LinkedIn.* Obtenido de Pulse: https://www.linkedin.com/pulse/activo-reflexivo-te%C3%B3rico-y-pragm%C3%A1tico-cu%C3%A1l-es-tu-de-francisco

Silva Numa, S., & Baena, M. P. (17 de Octubre de 2015). *El Espectador*. Obtenido de El Espectador: http://www.elespectador.com/noticias/educacion/desercion-estamos-haciendo-mal-articulo-593308

Ventura, A. C. (2000). *PERFIL DE ESTILOS DE APRENDIZAJE DE ESTUDIANTES.* Santa Fe. Rosario: Uniroja.

AL-Malaise, A., Malibari, A., & Alkhozae, M. (2012). STUDENTS' PERFORMANCE PREDICTION SYSTEM USING MULTI AGENT DATA MINING TECHNIQUE, *2*(5), 1–8.

Campagni, R., Merlini, D., Sprugnoli, R., & Verri, M. C. (2015). Data mining models for student careers. *Expert Systems with Applications*, *42*(13), 5508–5521. http://doi.org/10.1016/j.eswa.2015.02.052

Dinero, R. (2 de Febrero de 2015). *Revista Dinero*. Obtenido de Revista Dinero: http://www.dinero.com/pais/articulo/niveles-desercion-universitaria-los-andes-colombia/205330

Elbadrawy, A., Studham, S., & Karypis, G. (2014). Personalized Multi-Regression Models for Predicting Students Performance in Course Activities, 10.

Harwati, Alfiani, A. P., & Wulandari, F. A. (2015). Mapping Student's Performance Based on Data Mining Approach (A Case Study). *Agriculture and Agricultural Science*

*Procedia*, *3*, 173–177. http://doi.org/10.1016/j.aaspro.2015.01.034

Kabakchieva, D., STEFANOVA, K., & KISIMOV, V. (2009). Analyzing University Data for Determining Student Profiles and Predicting Performance. *Proceedings of the 4th International Cenference on Educational Data Mining, EindhovenKabakchieva, D., & Kl, S. (2009). Analyzing University Data for Determining Student Profiles and Predicting Performance. Proceedings of the 4th International Cenference O*.

Marquez-Vera, C., Romero, C., & Ventura, S. (2011). Predicting School Failure Using Data Mining. *... Data Mining, ...*, (December). Obtenido de http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_paper11_short_Marquez-Vera.pdf

Natek, S., & Zwilling, M. (2014). Student data mining solution-knowledge management system related to higher education institutions. *Expert Systems with Applications*, *41*(14), 6400–6407. http://doi.org/10.1016/j.eswa.2014.04.024

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4 PART 1), 1432–1462. http://doi.org/10.1016/j.eswa.2013.08.042

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting Student Performance : A Statistical and Data Mining Approach. *International Journal of Computer Applications*, *63*(8), 35–39.

Silva Numa, S., & Baena, M. P. (17 de Octubre de 2015). *El Espectador*. Obtenido de El Espectador: http://www.elespectador.com/noticias/educacion/desercion-estamos-haciendo-mal-articulo-593308

Thai-nghe, N., & Drumond, L. (2011). Multi-Relational Factorization Models for Predicting Student Performance, (September 2015).

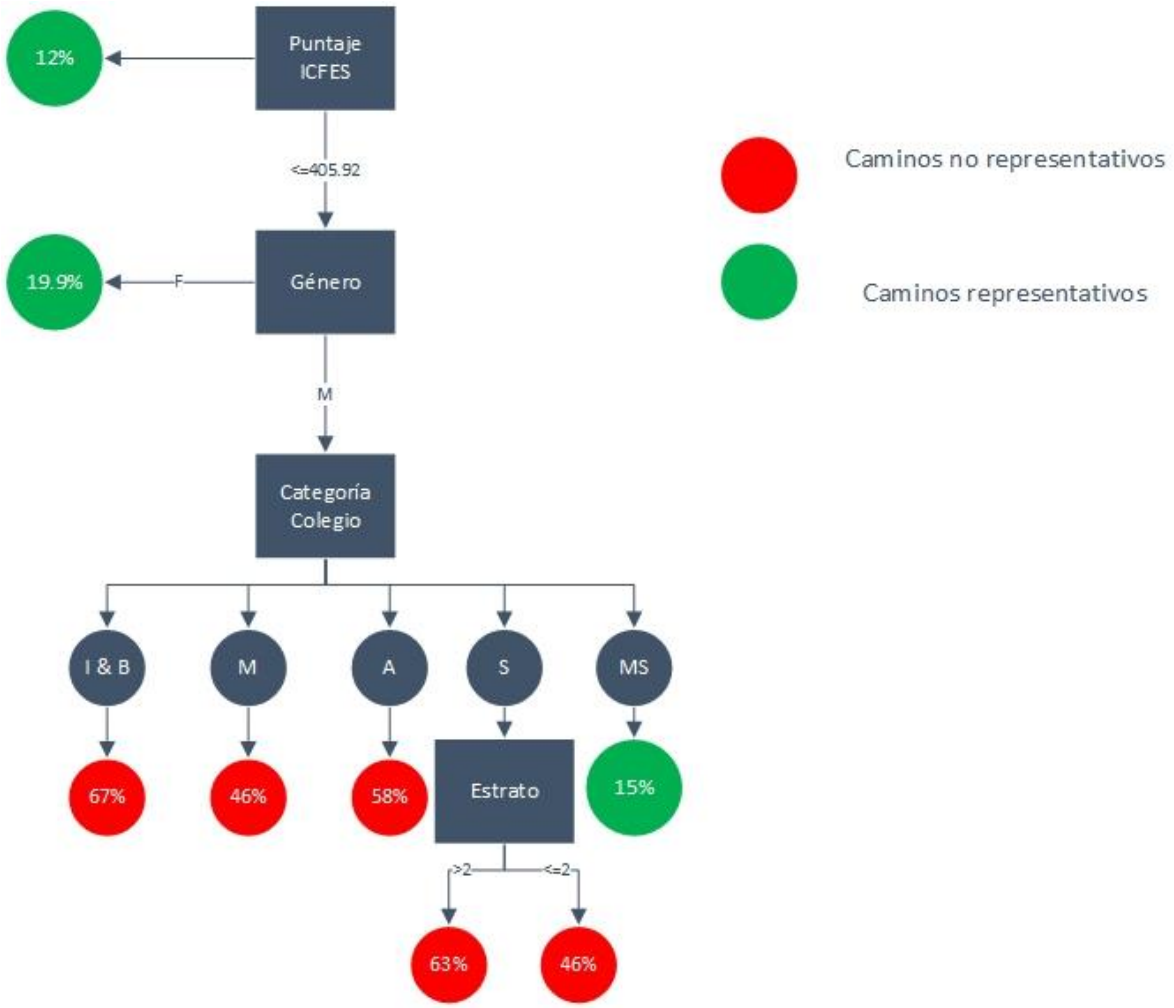
Thai-nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization Models for Forecasting Student Performance.

# ANEXOS

**ANEXO 1:** Total registros por camino en modelo sin SMOTE, para primer filtro

| Indice | Camino | | | | | | | registros |
|---|---|---|---|---|---|---|---|---|
| 1 | Picf>405,92 | | | | | | | 442 |
| 2 | Picf<=405,92 | > | Gén=F | | | | | 328 |
| 6 | Picf<=405,92 | > | Gén=M | > | CatCol=MS | | | 255 |
| 8 | Picf<=405,92 | > | Gén=M | > | CatCol=S | > | Estr>2 | 56 |
| 5 | Picf<=405,92 | > | Gén=M | > | CatCol=A | | | 50 |
| 7 | Picf<=405,92 | > | Gén=M | > | CatCol=S | > | Estr<=2 | 35 |
| 4 | Picf<=405,92 | > | Gén=M | > | CatCol=M | | | 28 |
| 3 | Picf<=405,92 | > | Gén=M | > | CatCol=I&B | | | 12 |

**ANEXO 2:** Árbol de caminos en modelo sin SMOTE, para primer filtro

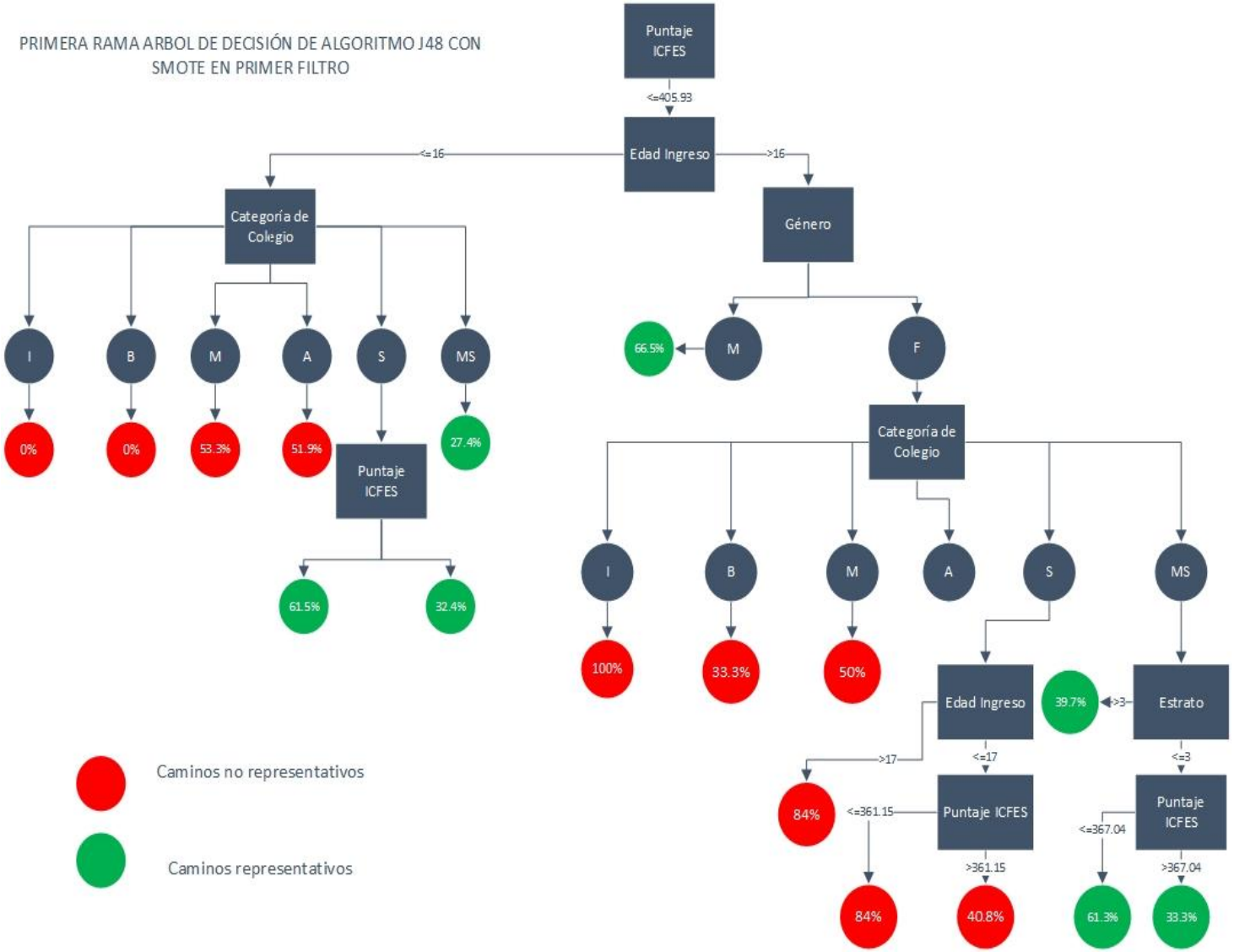



**ANEXO 3:** Total registros por camino en modelo sin SMOTE, para primer filtro

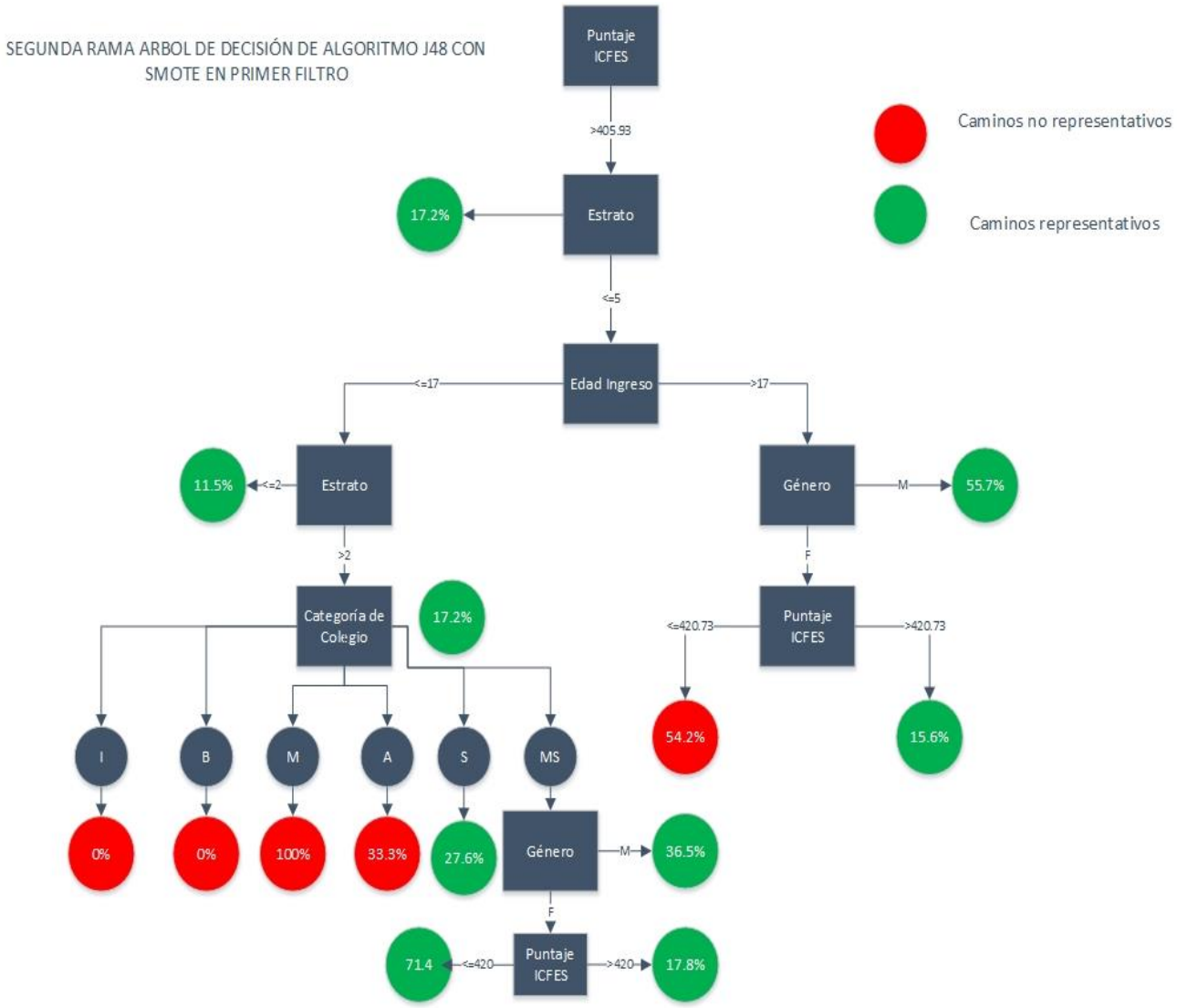
| Indice | Camino | | | | | | | | | | | | | | | | | | | | | registros |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Plcf<=405,9368 | > | Eing<=16 | > | Ctcol=I | | | | | | | | | | | | | | | | | 0 |
| 2 | Plcf<=405,9369 | > | Eing<=16 | > | Ctcol=B | | | | | | | | | | | | | | | | | 3 |
| 3 | Plcf<=405,9370 | > | Eing<=16 | > | Ctcol=M | | | | | | | | | | | | | | | | | 15 |
| 4 | Plcf<=405,9371 | > | Eing<=16 | > | Ctcol=A | | | | | | | | | | | | | | | | | 27 |
| 5 | Plcf<=405,9372 | > | Eing<=16 | > | Ctcol=S | > | Plcf<=366,59 | | | | | | | | | | | | | | | 39 |
| 6 | Plcf<=405,9373 | > | Eing<=16 | > | Ctcol=S | > | Plcf>366,59 | | | | | | | | | | | | | | | 37 |
| 7 | Plcf<=405,9373 | > | Eing<=16 | > | Ctcol=MS | | | | | | | | | | | | | | | | | 113 |
| 8 | Plcf<=405,9368 | > | Eing>16 | > | Gen= F | > | Ctcol=I | | | | | | | | | | | | | | | 0 |
| 9 | Plcf<=405,9369 | > | Eing>16 | > | Gen= F | > | Ctcol=B | | | | | | | | | | | | | | | 3 |
| 10 | Plcf<=405,9370 | > | Eing>16 | > | Gen= F | > | Ctcol=M | | | | | | | | | | | | | | | 6 |
| 11 | Plcf<=405,9371 | > | Eing>16 | > | Gen= F | > | Ctcol=A | | | | | | | | | | | | | | | 27 |
| 12 | Plcf<=405,9372 | > | Eing>16 | > | Gen= F | > | Ctcol=S | > | Eing>17 | | | | | | | | | | | | | 25 |
| 13 | Plcf<=405,9373 | > | Eing>16 | > | Gen= F | > | Ctcol=S | > | Eing<=17 | > | Plcf<=361,15 | | | | | | | | | | | 23 |
| 14 | Plcf<=405,9374 | > | Eing>16 | > | Gen= F | > | Ctcol=S | > | Eing<=17 | > | Plcf>361,15 | | | | | | | | | | | 27 |
| 15 | Plcf<=405,9373 | > | Eing>16 | > | Gen= F | > | Ctcol=MS | > | Estr<=3 | > | Plcf<=367,04 | | | | | | | | | | | 31 |
| 16 | Plcf<=405,9374 | > | Eing>16 | > | Gen= F | > | Ctcol=MS | > | Estr<=3 | > | Plcf>367,04 | | | | | | | | | | | 36 |
| 17 | Plcf<=405,9373 | > | Eing>16 | > | Gen= F | > | Ctcol=MS | > | Estr>3 | | | | | | | | | | | | | 126 |
| 18 | Plcf<=405,9373 | > | Eing>16 | > | Gen= M | | | | | | | | | | | | | | | | | 519 |
| 19 | Plcf>405,9374 | > | Estr>5 | | | | | | | | | | | | | | | | | | | 58 |
| 20 | Plcf>405,9375 | > | Estr<=5 | > | Eing<=17 | > | Estr<=2 | | | | | | | | | | | | | | | 61 |
| 21 | Plcf>405,9376 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=I | | | | | | | | | | | | | 0 |
| 22 | Plcf>405,9377 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=B | | | | | | | | | | | | | 0 |
| 23 | Plcf>405,9378 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=M | | | | | | | | | | | | | 1 |
| 24 | Plcf>405,9379 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=A | | | | | | | | | | | | | 3 |
| 25 | Plcf>405,9380 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=S | | | | | | | | | | | | | 29 |
| 26 | Plcf>405,9381 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=MS | > | Gen=F | > | Plcf<=420 | | | | | | | | | 28 |
| 27 | Plcf>405,9382 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=MS | > | Gen=F | > | Plcf>420 | | | | | | | | | 45 |
| 28 | Plcf>405,9383 | > | Estr<=5 | > | Eing<=17 | > | Estr>2 | > | Ctcol=MS | > | Gen=M | | | | | | | | | | | 126 |
| 29 | Plcf>405,9384 | > | Estr<=5 | > | Eing>17 | > | Gen=F | > | Plcf<=420,73 | | | | | | | | | | | | | 24 |
| 30 | Plcf>405,9385 | > | Estr<=5 | > | Eing>17 | > | Gen=F | > | Plcf>420,73 | | | | | | | | | | | | | 32 |
| 31 | Plcf>405,9386 | > | Estr<=5 | > | Eing>17 | > | Gen=M | > | | | | | | | | | | | | | | 131 |

**ANEXO 4:** Primera rama de Árbol de caminos en modelo con SMOTE, para primer filtro

PRIMERA RAMA ARBOL DE DECISIÓN DE ALGORITMO J48 CON SMOTE EN PRIMER FILTRO



**ANEXO 5:** Segunda rama de Árbol de caminos en modelo con SMOTE, para primer filtro



**ANEXO 6.** Imagen de herramienta para primer filtro