DESARROLLANDO MODELOS DE ANÁLISIS PREDICTIVO PARA MEJORAR LA PROBABILIDAD DE GRADUACIÓN DE ESTUDIANTES SUBGRADUADOS

por

Jesús Daniel Hernández Londoño

Proyecto sometido en cumplimiento parcial de los requisitos para el grado de

MAESTRÍA EN CIENCIAS
en
MATEMÁTICA ESTADÍSTICA

UNIVERSIDAD DE PUERTO RICO
RECINTO MAYAGÜEZ
2022

Aprobado por:

Roberto Rivera Santiago, Ph.D. Presidente, Comité Graduado	Fecha
Wolfgang A. Rolke, Ph.D. Miembro, Comité Graduado	Fecha
Edgardo Lorenzo González, Ph.D. Miembro, Comité Graduado	Fecha
Omar Colon Reyes, Ph.D. Director del Departamento	Fecha

Dissertation Abstract Presented to Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

DEVELOPING PREDICTIVE ANALYTICS MODELS TO IMPROVE CHANCES OF GRADUATION FOR UNDERGRADUATE STUDENTS

By

Jesús Daniel Hernández Londoño

2022

Chair: Roberto Rivera Santiago, Ph.D.
Major Department: Mathematical Sciences

University graduation rates in Puerto Rico are alarmingly low. On average, around 45 % of undergraduate students obtain their degree at 150 % program length [43]. This percentage becomes even more dramatic when viewed in conjunction with the school dropout rate. The 40 % of students don't finish high school and of those whether who do, not all go to college, but suppose they do [34]. Of that 60 % who go to college after high school only 45 % graduate at 150 % program length. Although some colleges offer workshops and mentoring to students to help them graduate, there is no objective way to predict whether a particular student will complete her degree. For such reasons, this research project aims to implement predictive models using machine learning methods to predict whether a student will complete their university degree at UPRM 150 % program length. These models were applied to 24,432 data of students admitted to the Mayaguez University Campus (RUM) between 1999 and 2010; which includes variables such as the student's study program, standardized entrance test scores, parent's level of education, among others. 6 machine learning methods including Stochastic Gradient Boosting and TabNet (a new artificial neural network architecture for tabular data) were used to predict whether an undergraduate student graduates at 150 % program length from UPRM. Stochastic Gradient Boosting and TabNet showed the best performance of all the methods when predicting whether a student graduates at 150 % program length from UPRM. Said tool allows university officials to detect specific students and develop intervention strategies to increase their chances of graduation.

Resumen de Disertación Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

DESARROLLANDO MODELOS DE ANÁLISIS PREDICTIVO PARA MEJORAR LA PROBABILIDAD DE GRADUACIÓN DE ESTUDIANTES SUBGRADUADOS

por

Jesús Daniel Hernández Londoño

2022

Consejero: Roberto Rivera Santiago, Ph.D.
Departamento: Matemáticas

Las tasas de graduación universitarias en Puerto Rico son alarmantemente bajas. En promedio, cerca del 45 % de los estudiantes subgraduados obtienen su grado al 150 % de la duración de su programa académico [43]. Este porcentaje se hace aún más dramático si se mira en conjunto con la tasa de deserción escolar. El 40 % de los estudiantes no terminan la escuela superior y de los que si la terminan, no todos van a la universidad, pero supongamos que sí [34]. De ese 60 % que ingresan a la universidad después de la escuela superior solo el 45 % se gradúa al 150 % de la duración de su programa académico. Aunque algunas universidades ofrecen talleres y consejería a estudiantes para facilitar que se gradúen, no existe una manera objetiva de predecir si una estudiante en particular completará su grado. Por tales motivos, este proyecto de investigación tiene por objetivo implementar modelos predictivos utilizando métodos de aprendizaje automático para predecir si un estudiante completará su grado universitario en el RUM al 150 % de la duración de su programa académico. Esos modelos se aplicaron a 24,432 datos de estudiantes admitidos al Recinto Univesitario de Mayaguez (RUM) entre el 1999 hasta el 2010; el cual incluye variables como programa de estudio del estudiante, puntuación en pruebas estandarizadas de entrada, nivel de educación de los padres, entre otras. Se usaron 6 métodos de aprendizaje automatizado incluyendo Stochastic Gradient Boosting y TabNet (una nueva arquitectura de red neuronal artificial para datos tabulares) para predecir si un estudiante subgraduado se gradúa al 150 % de la duración de su programa académico del RUM. Stochastic Gradient Boosting y TabNet mostraron el mejor desempeño de todos los métodos al predecir si un estudiante se gradúa al 150 % de la duración de su programa académico del RUM. Dicha herramienta le permite a oficiales de la universidad detectar estudiantes específicos y desarrollar estrategias de intervención para aumentar sus posibilidades de graduación.

"Hay una ley de vida, cruel y exacta,
que afirma que uno debe crecer o, en caso contrario,
pagar más por seguir siendo el mismo"
Norman Mailer

Agradecimientos

En primera instancia, quiero agradecer a Dios por su cuidado, gracia y bondad, por ubicarme en este camino a personas que me enseñaron cosas maravillosas sobre la vida.

A mis padres Jesús Aderh Hernández Cardenas e Isabel Cristina Londoño Sierra, por ser mi fuente de motivación, por su amor incondicional, su paciencia, su acompañamiento y cada gota de conocimiento.

A mis hermanos Jesús A Hernández Londoño, Bella C Hernández Londoño y mi melliza Laura D Hernández Londoño, por aconsejarme, motivarme y acompañarme en mis momentos más difíciles.

Al Dr. Roberto Rivera Santiago por su dedicación y compromiso con mi trabajo, por su paciencia, y todo lo que me permitió aprender de él bajo su dirección.

A mis amigos Nathalia Penagos, Carlos Murillo, Adriana Gaitán y Danier Ruiz por el apoyo, sus palabras y acompañamiento en esta aventura del conocimiento.

Al Dr. Omar Colon por su ayuda y apoyo con herramientas tecnológicas para el desarrollo de este proyecto.

A la Oficina de Planificación, Investigación y Mejoramiento Institucional (OPIMI) de la Universidad de Puerto Rico Recinto Mayagüez por facilitarme las bases de datos.

A la maravillosa Universidad de Puerto Rico Recinto Mayagüez, el Departamento de matemáticas y a nuestra Isla del encanto, por acogerme, enseñarme, y educarme en estos 3 años.

Índice general

Índice de figuras

Índice de cuadros

Objetivo

Predecir si un estudiante subgraduado se gradúa al 150 % de la duración de su programa académico de la Universidad de Puerto Rico Recinto Mayagüez, usando métodos de aprendizaje automático según varias variables del estudiante.

Capítulo 1

Introducción

El ser humano empieza a transmitir el conocimiento de manera constante a través del tiempo. Lo hace con el fin de poder desarrollar su economía, el acceso a la riqueza, la ciencia y la tecnología. Todo empezó con la transmisión y asimilación del conocimiento hace miles de años en Asia y Europa principalmente, luego en América del Norte y mucho después en el Sur. La explosión demográfica que se da justo en el momento de la industrialización de las grandes economías acelera descomunalmente la demanda de todo tipo de productos. En consecuencia, es indispensable tener obreros y empleados capacitados para desempeñar funciones nuevas cada día. Así pues, la educación formal se impone como un medio efectivo para adquirir herramientas y conocimientos esenciales para la vida cotidiana y así, acceder y ascender a los diferentes puestos de trabajo, y diferentes bienes y servicios. Seres humanos educados brinda a la nación una sociedad capaz de luchar, generar cambio, crear, decidir con criterio propio, crecer económicamente y comunicativamente para expresar sentimientos de afecto y solidaridad. En palabras del expresidente colombiano Andrés Pastrana, una nación educada es la mejor garantía del progreso de una nación.

Dicho lo anterior, ¿cómo está la educación universitaria en Puerto Rico? Los datos del consejo de educación de Puerto Rico revelan que la tasa de graduación universitaria es alarmantemente baja. En promedio, cerca del 45 % de los estudiantes subgraduados obtienen su grado al 150 % de la duración de su programa académico [43]. Aunque la tasa de graduación universitaria al 150 % de la duración de su programa académico (G150) pueda ser tan baja como del 10 % (el caso de la Universidad Central Del Caribe) y tan alta como del 82 % (el caso de la Universidad Dewey-Manatí), el 45 % (promedio tasa de graduación universitaria en Puerto Rico) se hace aún más dramático si se mira en conjunto con la tasa de deserción escolar. El 40 % de los estudiantes no terminan la escuela superior y de los que si la terminan, no todos van a la universidad, pero supongamos que sí [34]. De ese 60 % que ingresan a la universidad después de la escuela superior solo el 45 % se gradúa al 150 % de la duración de su programa académico.

Las bajas tasas de graduación universitaria son un problema que afecta a todos en general. Los estudiantes pierden ingresos anuales por cada año adicional que pasen en la universidad, en promedio $\$43,000$ [8]. Si algún estudiante no se gradúa y decide conseguir empleo, perdería alrededor de $\$17,880$ anuales, dado que, [12] según las estadísticas

laborales de Estados Unidos, la tasa de ingresos promedio anual para empleados con un grado universitario sin posgrado es $\$55,510$ y para empleados sin grado universitario es $\$37,630$. Además, tendría menor probabilidad de acceder a planes de jubilación [16] y tres veces más probable de vivir en la pobreza [15]. De igual forma, limita el acceso a la educación post-secundaria, afectando los cupos académicos y llevando a las universidades a perder ingresos adicionales por matrícula y cuotas [1], "Georgia State calcula que cada aumento de un punto porcentual en la tasa de graduación tiene un valor de $3 millones de dólares al año en ingresos adicionales por matrícula y cuotas". Asimismo, se ve afectado el presupuesto federal y estatal, estudiantes no graduados pagan menos impuestos.

Lo descrito anteriormente exige al país y a las instituciones de educación post-secundaria prestar atención a las bajas tasas de graduación universitarias. Con los datos de estudiantes que tienen las universidades a su disposición se deberían desarrollar estrategias y modelos predictivos que permitan aumentar la cifra de estudiantes que completan sus grados exitosamente.
El propósito de este proyecto de investigación es implementar un modelo predictivo utilizando métodos de aprendizaje automático con el objetivo de predecir si un estudiante subgraduado se gradúa al 150 % de la duración de su programa académico de la Universidad de Puerto Rico Recinto Mayagüez. Para alcanzar nuestro objetivo se compararán Árboles de Clasificación, Bosque Aleatorio, Stochastic Gradient Boosting, Naïve Bayes, Regresión Logística y TabNet, con datos imputados y datos no imputados. Para imputar los datos faltantes se usará la imputación multivariante por ecuaciones encadenadas (MICE) y para datos no imputados se eliminarán los valores faltantes. Cabe resaltar que este proyecto no propone una herramienta para decidir si se admite un estudiante o no. No sería ético si se usa para admisiones, dado que, se estaría sesgando a que personas con mejores beneficios sean admitidas (ver sección 3.2 y capitulo 4).

1.1. Descripción del problema y motivación

El Sistema Integrado de Datos de Educación Post-Secundaria (IPEDS) maneja información general sobre colegios, universidades e instituciones técnicas y vocacionales de Estados Unidos. IPEDS entre sus diversos reportes informa las tasas de G150 de los programas. Para programas de 4 años el tiempo de finalización al 150 % sería 6 años y para programas de 5 años sería 7.5 años pero se consideran 8 años. A continuación se describen las tasas de G150 de algunas universidades en Puerto Rico usando IPEDS para detallar a grandes rasgos la alarmante situación de las bajas tasas de graduación. [9] Para los programas de 4 años, la Pontificia Universidad Católica de Puerto Rico tiene una tasa de graduación de 35.3 % (Comienzo otoño 2011) y 39 % (Comienzo otoño 2013). La Universidad Interamericana de Puerto Rico de 33 % (Comienzo otoño 2011) y 38.2 % (Comienzo otoño 2013). La Universidad de Puerto Rico Recinto de Río Piedras de 58 % (Comienzo otoño 2011) y 56 % (Comienzo otoño 2013). La Universidad de Puerto Rico Recinto de Mayagüez de 47 % (Comienzo otoño 2011) y 49 % (Comienzo otoño 2013). Por otro lado, para programas de 5 años (comienzo otoño 2011) la tasa de graduación para la Pontificia Universidad Católica de Puerto Rico es de 40.6 %. La Universidad Interamericana de Puerto Rico es 37.3 %. La Universidad de Puerto Rico Recinto Río Piedras de 65 % y

la Universidad de Puerto Rico Recinto Mayagüez de 52 %. Al comparar a Estados Unidos con Puerto Rico para programas de 4 años (al 150 % de la duración de su programa académico) se tiene a Estados unidos con mayor variabilidad en sus tasas de graduación aunque en promedio tiene una tasa de graduación para esos mismos programas alrededor del 63.4 % [11]. Por ejemplo, [9] la Universidad de Dakota del Norte tiene una tasa de graduación de 54 % (Comienzo otoño 2011) y 61 % (Comienzo otoño 2013). La Universidad de Minnesota-Rochester de 61 % (Comienzo otoño 2011) y 59 % (Comienzo otoño 2013). La Universidad de Texas en Austin de 83 % (Comienzo otoño 2011) y 86 % (Comienzo otoño 2013). El Instituto de Tecnología de Massachusetts de 94 % (Comienzo otoño 2011) y 95 % (Comienzo otoño 2013).

Algunos factores que afectan las tasas de graduación universitaria en Estados Unidos, [6] son la formación académica previa, falta de interés y enfoque, selección incorrecta de clases durante el proceso de matrícula, falta de empatía de los profesores a la hora de enseñar [4], la consejería académica insuficiente o inadecuada [3], los grupos grandes en los cursos, el ausentismo frecuente del estudiante, pocos cursos, [1] factores económicos, entre otras. Adicionalmente, se ha sugerido [3] establecer un sistema de alerta temprana y consejería académica compuesta por estudiantes para aumentar la tasa de graduación. Por otra parte, en Puerto Rico no existen estudios de factores que afectan las tasas de graduación, por ende, no se puede asegurar si son los mismos factores de Estados Unidos los que afecte a Puerto Rico, debido a que son culturas socialmente diferentes, aunque estén bajo un mismo gobierno nacional.

La Universidad de Puerto Rico Recinto Mayagüez (RUM) ha puesto en marcha planes estratégicos para aumentar sus tasas de graduación. Por ejemplo, el Centro de Redacción en Español (CRE) y el English Writing Center (EWC), que hacen hincapié en el desarrollo de la escritura como habilidad profesional integral. También se ofrecen tutorías para apoyar a todos los estudiantes de la universidad que cursan asignaturas entorno a las ciencias matemáticas, éste denominado “Centro de Apoyo”, dirigido por el Departamento de Matemáticas – Facultad de Artes y Ciencias. También se cuenta con conferencias orientadas a cómo organizar el tiempo y el manejo del estrés académico, dirigido por el Departamento de Psicología. Se cuenta con tres programas PEARLS, R2DEEP y CAHSI. [28] La Alianza Informática de Instituciones al Servicio de los Hispanos (CAHSI) se formó en el 2004 pero se estableció en el 2017 en la universidad de Puerto Rico como CAHSI Circuito Caribe con el fin de ayudar a los estudiantes adquirir habilidades de investigación en el pregrado y posgrado. [29] El Programa para el acceso, la retención y el éxito de LIATS en ingeniería (PEARLS) dirigido a estudiantes de licenciatura y maestría en todas las disciplinas de la ingeniería se creó con el fin de complementar el trayecto académico de los estudiantes a través de tutoría profesional individualizada, oportunidades de investigación, pasantías y actividades de tutoría entre pares, trabajando por un objetivo común: el éxito de los estudiantes.[30] El programa de pre-ingeniería Reclutamiento, Retención y Educación en Ingeniería a Distancia (R2DEEP) se creó con el fin de mejorar la cantidad y calidad de estudiantes que ingresan a programas ofrecidos en el Colegio de Ingeniería. Sin embargo, en el RUM estos planes estratégicos están disponibles a la comunidad en general y por ende no existen estrategias que se enfoquen en estudiantes específicos.

Las estrategias que se enfocan en estudiantes específicos permiten intervenir al estudiante para orientarlo de acuerdo a su problemática particular y en consecuencia proveer un servicio apropiado. Por ejemplo, según el Informe Hechinger [1], la Universidad del Sur de Florida, la Universidad Estatal de Arizona y la Universidad Estatal de Georgia han visto aumentar sus tasas de graduación después de recurrir al análisis predictivo para identificar y brindar apoyo oportuno a los estudiantes que podrían tener dificultades académicas. En su informe, presentan el caso de la Universidad Estatal de Georgia en el que el asesor académico del estudiante Keenan Robinson en su primer año lo orientó a elegir otro programa, dado que su plan de elegir enfermería era arriesgado porque era probable que abandonara la universidad si elegía dicho programa, ya que aunque su promedio de calificaciones (Grade Point Average, GPA) era sólido el análisis predictivo detectó problemas. Por otra parte, sus análisis predictivos también buscaban estudiantes que tenían buenas calificaciones pero que estaban pagando tarde las facturas de matrícula, indicando que el estudiante podría abandonar la escuela por razones financieras. Similarmente, el University College de la Universidad de Maryland usa un programa que permite identificar en tiempo real el éxito de un estudiante en un curso en particular. El programa facilita a los miembros de la facultad y los especialistas en apoyo a los estudiantes a ver señales de advertencia que afectan el éxito de los estudiantes e intervenir en consecuencia. Su programa identifica con precisión entre el 80 y 90 por ciento de las veces. De igual manera, ayuda a los profesores a identificar estudiantes que pueden reprobar su curso y estudiantes que podrían necesitar un poco de ayuda adicional para pasar de un B a un A. La Universidad Estatal de Arizona usa eAdvisor un programa que monitorea el progreso del estudiante para brindar la intervención necesaria en momentos estratégicos hacia la finalización de su título, por ejemplo, si un estudiante no aprueba un curso eAdvisor lo programaría para reunirse con un asesor y si la conversación va bien se podría conducir al estudiante a inscribirse en otro programa [31].

Por lo tanto, predecir la graduación de los estudiantes e intervenir a tiempo es de gran importancia, dado que, brinda mayor oportunidades al estudiante, permite aumentar los ingresos de la universidad [1], medir la eficacia de la institución ante los organismos de acreditación y el gobierno [4], y brinda una sociedad capaz de luchar, generar cambio, crear y decidir con criterio propio.

El presente trabajo está organizado de la siguiente manera. En el Capítulo 2 se establece la metodología que permite el desarrollo de los resultados estudiados. Para ello, se presenta una introducción sobre el aprendizaje automático y sus desafíos, el método de imputación de datos usada, los métodos de aprendizaje automático y las métricas de evaluación consideradas para el desarrollo de este proyecto. En el Capítulo 3 se expone la población de estudio (los estudiantes subgraduados de la Universidad de Puerto Rico Recinto Mayagüez), el análisis exploratorio de la información obtenida (usando los datos de entrenamiento al eliminar los valores faltantes), un análisis de los valores faltantes, un análisis del rendimiento predictivo de los métodos aplicados y la elección final de los métodos. En el Capítulo 4 se plantean las conclusiones y sugerencias futuras.

Capítulo 2

Metodología

2.1. Aprendizaje Automático

El Aprendizaje Automático (ML) es "la ciencia de hacer que las computadoras actúen sin estar programadas explícitamente" - Andrew Ng. Es decir, el ML permite entrenar a las computadoras para que aprendan a descubrir por si solas las relaciones subyacentes que hay en los datos. En la actualidad, con el crecimiento continuo de la información las compañías están utilizando el aprendizaje automático con todos los datos que puedan recopilar para mejorar su negocio, de hecho, han tomado mejores decisiones y han desarrollado mejores servicios, lo que conduce a un negocio más sólido y viable. Compañías como Tesla, Google, Amazon y Apple son un ejemplo de negocios solidos que han desarrollado automóviles sin conductor, asistentes de voz (siri, alexa, google) y búsqueda web efectiva, entre otros.

El ML se divide en tres tipos de aprendizaje (figura 2.1). El aprendizaje supervisado, en el que se conoce o se tiene una variable objetivo que puede ser cualitativa o cuantitativa. Si es cuantitativa los métodos o algoritmos se conocen como métodos de regresión, en caso contrario, métodos de clasificación. El aprendizaje no supervisado, opuesto al anterior, no se conoce una variable objetivo y generalmente está dedicado a tareas de agrupamiento. Por ultimo, el aprendizaje por reforzamiento, aquí los métodos emplean prueba y error para encontrar una solución al problema. El método obtiene recompensas o penalizaciones por las acciones que realiza. Este proyecto utiliza el aprendizaje supervisado y métodos de clasificación dada la naturaleza de lo datos, puesto que, se tiene la variable objetivo Y_i que representa si un estudiante i se graduó ($Y = 1$) o no ($Y = 0$).

2.1.1. Hiperparámetros

Los hiperparámetros de los métodos de aprendizaje automático son similares a los parámetros de un modelo estadístico. La diferencia radica, en que los hiperparámetros se encuentran de manera iterativa. Se ajustan o se afinan mientras el modelo está aprendiendo. [6] Es decir, se entrenan varios modelos con distintas combinaciones de hiperparámetros y se eligen los que consigan un mejor trabajo con respecto a una métrica de evaluación. En este proyecto se usan varias combinaciones de algunos hiperparámetros partiendo de los valores predeterminados.

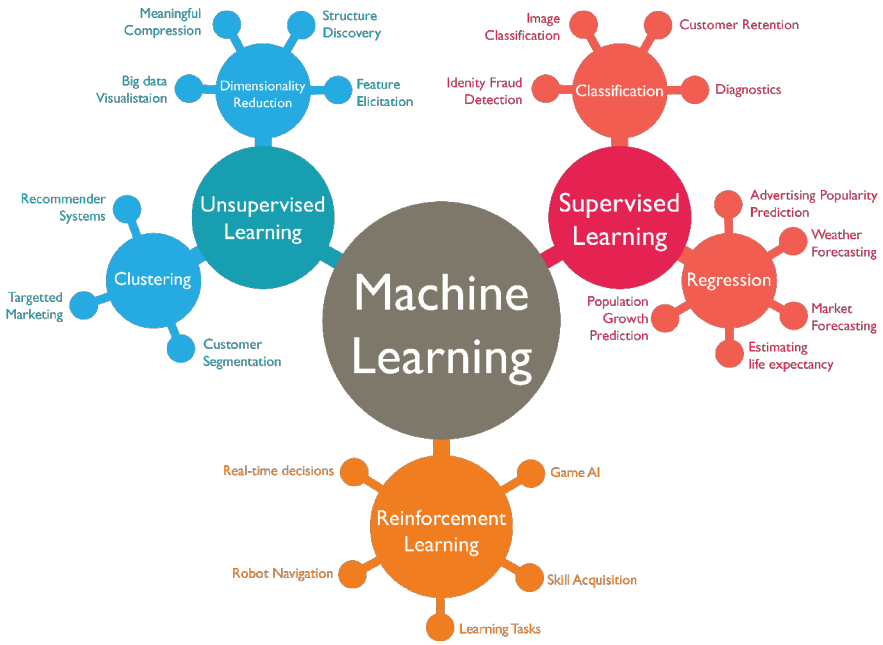

Figura 2.1: Tipos de aprendizaje en el aprendizaje automático. (Extraída de [25])

2.1.2. Desafíos

Un desafío central del aprendizaje automático es generalizar sus métodos o algoritmos más allá de la base de datos de entrenamiento. Un método que pueda generalizarse quiere decir, que tiene un buen desempeño en datos que nunca ha visto el modelo (datos previamente no observados) [32]. Para obtener una buena generalización se deben elegir métodos con hiperparámetros que reduzcan la brecha entre el error de entrenamiento y el error de prueba. Entre los factores que impiden la generalización de un método de aprendizaje automático se encuentran el Ajuste Insuficiente (underfitting), Sobreajuste (overfitting) y compensación sesgo-varianza o Bias-Variance Trade-Off (ver figura 2.2).

El underfitting se presenta cuando el modelo no logra un error suficientemente bajo en los datos de entrenamiento. El overfitting se presenta cuando la brecha entre el error de entrenamiento y el error de prueba es enorme [32]. Ambos problemas producen predicciones incorrectas. En la compensación sesgo-varianza, el sesgo hace referencia al error de aproximar un problema a la vida real. Es decir, que tan cerca esta mi modelo del que realmente es. La varianza se refiere a cuan cambiante es mi modelo al realizar pequeños cambios en los datos de entrenamiento. A continuación se describe un ejemplo de alta varianza y alto sesgo. Supongamos que queremos ajustar una función no lineal con un modelo de regresión lineal, debido a que es imposible ajustar una linea recta a una curva se genera una estimación imprecisa del modelo. En consecuencia, la regresión lineal da como resultado alto sesgo [13]. Por otro lado, para entrenar un modelo se usa una base de datos de entrenamiento y se elijen los hiperparámetros que permitan un buen ajuste. Ahora se hacen pequeños cambios en la misma base de datos de entrenamiento y se vuelve a entrenar el mismo modelo pero se obtienen hiperparámetros totalmente diferentes. En consecuencia, mi modelo tendría alta varianza.

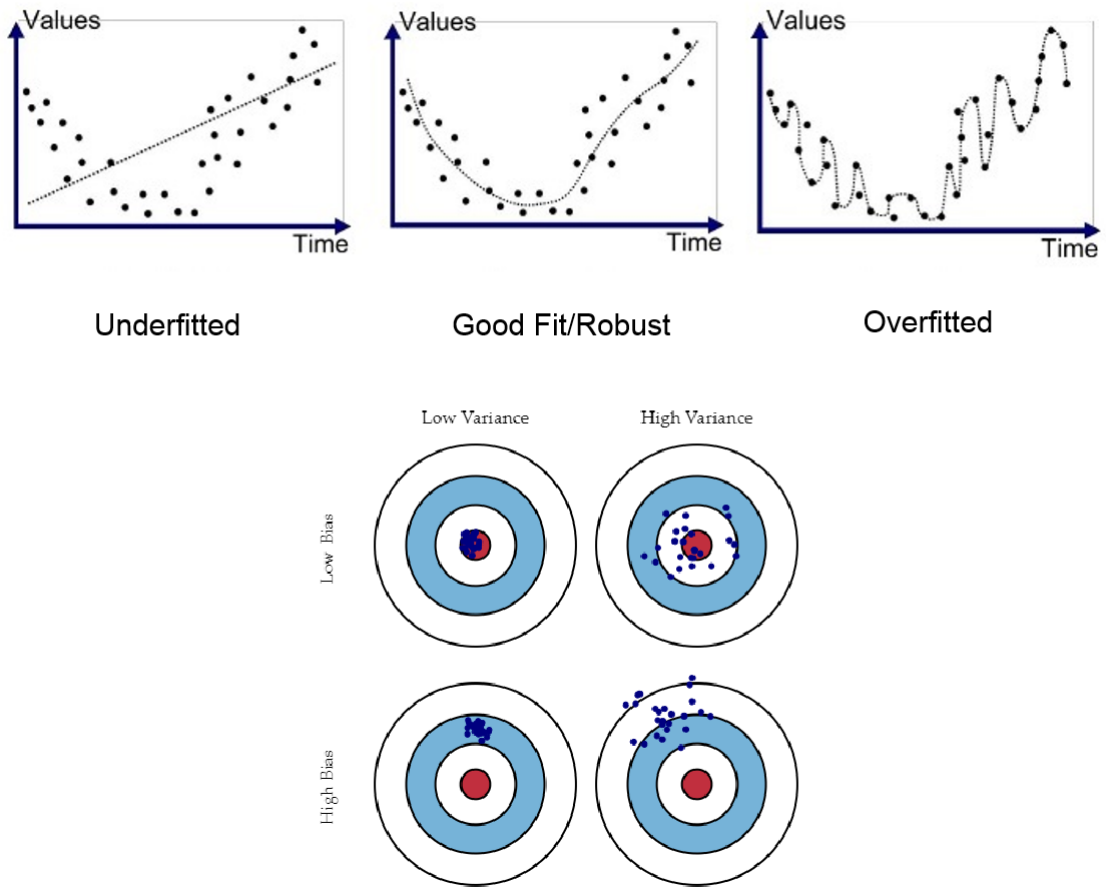

Figura 2.2: Ilustración factores que impiden la generalización de un método de aprendizaje automático. (Extraída de [19] y [33])

2.2. Datos Faltantes

En muchos casos los valores faltantes son inherentes en las bases de datos. Manejarlos es una tarea importante en el preprocesamiento de datos en aprendizaje automático. Hay tres mecanismos que tratan de explicar los valores faltantes es una base de datos.

- Valores faltantes completamente al azar (MCAR).
- Valores faltantes al azar (MAR).
- Valores faltantes no al azar o no ignorables (NMAR).

Los datos faltantes se denominan MCAR (Missing Completely At Random) cuando es probable que ese valor no dependa de otras predictoras, ni de los valores faltantes. Por ejemplo, cuando se agota la batería de un sensor, el sensor deja de enviar datos a los servidores [26]. Un valor faltante es de tipo MAR (Missing at Random) cuando es probable que ese valor dependa de otras predictoras (se puede explicar con otras predictoras dentro de la base de datos) pero no depende de los valores faltantes. Por ejemplo, las mujeres suelen evitar revelar su edad en las encuestas (el género está relacionado con la falta de la

predictora edad) [26]. NMAR (Missing Not At Random) ocurre cuando es probable que ese valor faltante dependa de los valores faltantes. Sucede cuando las personas no quieren revelar algo muy personal acerca de ellas. Por ejemplo, las personas con salarios altos evitan revelar sus ingresos en las encuestas [26].

Se dice que si los valores faltantes representan entre un 5 % a un 20 % del total de datos entonces se requiere de métodos sofisticados. Muchas veces la estrategia para manejar los datos faltantes es usar técnicas básicas o fundamentales como la media, moda, mediana, una constante o incluso descartar esas filas a la que pertenecen, sin embargo, usar esas estrategias provocan sesgos en los análisis posteriores de los conjuntos de datos [26]. La imputación multivariante por ecuaciones encadenadas (MICE) es un método sofisticado que asume los valores faltantes de tipo MAR. Este método se puede usar como alternativa para imputar los valores faltantes y evitar sesgo [26]. MICE predice el valor faltante usando un modelo donde la predictora a la que pertenece dicho valor faltante es la variable dependiente y las demás predictoras son independientes. Por ejemplo, las variables binarias se pueden modelar mediante regresión logística y las variables continuas mediante regresión lineal [26]. El cuadro 2.1 describe el funcionamiento de MICE. Se adaptó con fines explicativos sencillos.

Algoritmo. *MICE*

1. Haga una imputación básica (media, moda u otra imputación) para cada valor faltante.
2. Tome una variable que contiene valores faltantes y vuelva a poner como NA sus valores faltantes originales. Será la variable dependiente a modelar.
3. Ajuste un modelo en el que la variable del paso 2 es la variable dependiente y las restantes las independientes.
4. Haga predicciones con el modelo del paso 3 y reemplace los valores faltantes con esas predicciones.
5. Repita del paso 2 al 4 hasta tomar todas las variables que tenían valores faltantes.

Cuadro 2.1: Algoritmo de MICE.(Adaptado de [26])

2.3. Métodos de Aprendizaje Automatizado

Los métodos de aprendizaje automático empleados son métodos para problemas de clasificación. Se describen tres métodos basados en árboles, Árbol de Clasificación, Bosque Aleatorio y Boosting. Dos métodos basados en probabilidad, Naïve Bayes y Regresión Logística. Por último, un método basado en redes neuronales artificiales para datos tabulares, Tabnet.

2.3.1. Basado en Árboles

1. Árbol de Clasificación

El origen de los árboles aparentemente se da en el análisis de datos de encuestas en 1964 por Sonquist y Morgan, debido a que en la mayoría de los análisis cuando giran en torno a la existencia de efectos de interacción no se han manejado razonablemente bien. Sonquist y Morgan proponen un nuevo método aplicado a problemas con variable respuesta cuantitativa (enfoque de "árbol de regresión") que no impone restricciones sobre los efectos de interacción en los datos de encuestas, centrándose en la importancia de reducir el error predictivo. Más tarde en 1971 se extendió a problemas con variable respuesta cualitativa (enfoque de "árbol de clasificación") y en 1984 Breiman, Friedman, Olshen y Stone desarrollan una metodología llamada CART (Classification and Regression Trees) para la construcción de árboles en problemas de regresión y clasificación [38]. Este proyecto utiliza la metodología CART y se describe a continuación.

CART es una metodología no paramétrica (esto significa que no requiere suposiciones) donde los árboles son gráficos con estructura de árbol en el que se ilustran reglas de decisión. La estructura del árbol consta de un solo nodo raíz (ubicado en la parte superior; es el primer nodo), nodos internos y nodos hoja (ubicados en la base del árbol), enlazados con ramas que contiene reglas de decisión desde la parte superior hasta la base del árbol. Los nodos representan una predictora a excepción del nodo hoja que tiene un valor específico utilizado como predicción. Las reglas de decisión en las ramas son puntos de división utilizados para construir el árbol de manera recursiva. CART incluye el desarrollo para dos tipos de árboles, el árbol de regresión y el árbol de clasificación, el uso de esos dos tipos de árboles depende de la variable respuesta. Si la variable respuesta es cuantitativa se usa el de regresión y si la variable respuesta es cualitativa se usa el de clasificación. Dada la naturaleza de los datos de este proyecto (ver sección 2.1) se usa el árbol de clasificación. El procedimiento para desarrollar un árbol de clasificación usando CART consisten en tres pasos [37]

- Construcción del árbol.
- Poda del árbol.
- Elección del árbol óptimo.

A medida que se construye el árbol los puntos de división se eligen de tal forma que se obtenga la mejor división en los datos y para ello se usan las medidas de impureza. Después de construirlo se poda y finalmente se elige el árbol óptimo. Estos procedimientos se especifican en la siguientes secciones. Para el caso de un árbol de regresión es lo mismo, excepto que la mejor división en los datos será el punto y predictora que minimice la suma de los cuadrados del error.

Un problema importante con los árboles es que pueden ser inestables, debido a que pequeños cambios en los datos puede alterar potencialmente un punto de división y, por ende, se puede obtener un árbol muy diferente [17]. Esta problemática de inestabilidad hace referencia a la alta varianza que tienen los árboles y, en consecuencia, pueden ser

vulnerables al sobreajuste [17], porque puede seguir la tendencia de los datos de entrenamiento (bajo sesgo) y al seguir esa tendencia tiene un error de entrenamiento mucho menor que el error de prueba (alta varianza). Bagging y Boosting reducen la problemática de alta varianza. Bagging utiliza diferentes muestras aleatorias del tamaño del conjunto de datos de entrenamiento con reemplazamiento para entrenar múltiples versiones del mismo modelo (que pueden ser árboles) y Boosting ajusta secuencialmente múltiples modelos (que pueden ser árboles) para que aprendan de los errores del modelo anterior. Otro inconveniente es con las predictoras cualitativas, porque tienden a imponer un sesgo en las medidas de impureza, debido a la gran cantidad de divisiones potenciales que puede tener dicha predictora [17], si el número de categorías es grande esto puede conducir al sobreajuste [20].

La visualización del árbol se hace difícil en presencia de una gran cantidad de datos, sin embargo, ha sido ampliamente utilizado en diferentes áreas de investigación, como la medicina, la biología, la agricultura, la economía entre otros [35]. Su amplio uso es debido a la versatilidad al construir un modelo con distintos tipos de predictoras (cuantitativas o cualitativas) [35] y valores faltantes. Si hay un valor faltante en alguna predictora X solo se consideran las observaciones que no faltan en X para ajustar un árbol [20], pero a medida que se ajusta ese árbol en cada nodo se deben identificar divisiones sustitutas o predictoras sustitutas y puntos de división que se usarán en caso de que falte un valor en X. Se debe formar una lista ordenada de predictoras sustitutas y puntos de división, debido a que en el momento de entrenar o predecir puede faltar un valor en la primera predictora sustituta, entonces se usa la segunda sustituta, y así sucesivamente. Para elegir la primera división sustituta en algún nodo se debe elegir la predictora y punto de división que mejor imite la división de los datos de la predictora que divide en dicho nodo [20]. La segunda división sustituta es el predictora y el punto de división correspondiente al segundo mejor resultado, y así sucesivamente. Con el fin de ejemplificar la división sustituta, supongamos que se ajusta un árbol de un solo nodo con dos predictoras A1, A2 y una variable respuesta Y. Los datos se muestran en el cuadro 2.2.

Y	A1	A2
1	10	10
1	9	9
1	8	8
1	7	7
1	6	5
0	5	6
0	NA	4
0	3	3
0	2	2
0	1	1

Cuadro 2.2: Ejemplo de división sustituta de un árbol de un solo nodo. (Adaptado de [36])

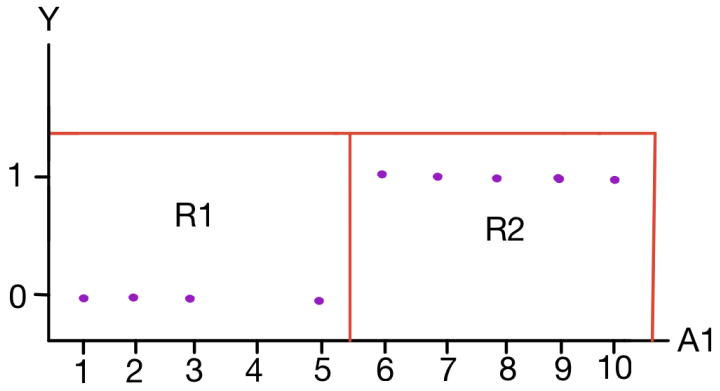

Figura 2.3: Ejemplo Regiones o Hiper-Rectángulos en un Plano Bidimensional usando solo A1>5.5.

El punto de división A1 $>$ 5.5 proporciona una división perfecta para la variable respuesta Y, dividiendo a A1 en dos regiones rectangulares en un espacio bidimensional como se muestra en la figura 2.3, entonces, el árbol ajustado tendrá el único nodo con esa división como se muestra en la figura 2.5, pero como hace falta una observación en A1 se debe buscar una división sustituta que imite la división que se obtuvo con A1 en Y. Para ello no se tiene en cuenta la observación A2 $=$ 4 y en consecuencia la división A2 $>$ 4.5 pues en esa fila se obtiene un valor faltante para A1. Ignorando ese valor faltante se obtiene que A2 $>$ 3.5 imita la división de A1 en Y, porque el punto de división A2$>$ 3.5 separa a Y en dos regiones rectangulares donde al menos una región tiene todas las observaciones de una sola categoría en Y, que para este caso es la primera región R1 con $Y = 0$ (ver figura 2.4).

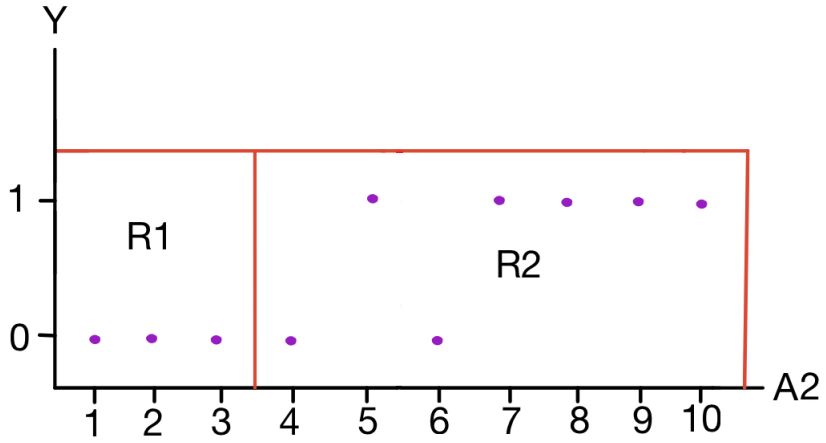

Figura 2.4: Ejemplo Regiones o Rectángulos en un Plano Bidimensional usando solo A2>3.5.

Ahora se desea predecir la variable respuesta Y con A1 $=$ 3 y A2 $=$ 3, observe que como el árbol ajustado es de un solo nodo solo debe usar el valor A1 $=$ 3. Al enviar esta observación a lo largo del árbol se obtiene $Y = 0$ porque A1 $=$ 3$<$5.5 (ver figura 2.5). Ahora si A1 $=$ NA y A2 $=$ 4 se debe usar la división sustituta con predictora A2 y punto de división A2 $>$ 3.5 puesto que A1 tiene un valor faltante, como A2 $=$ 4$>$3.5 entonces $Y = 1$ (ver figura 2.4).

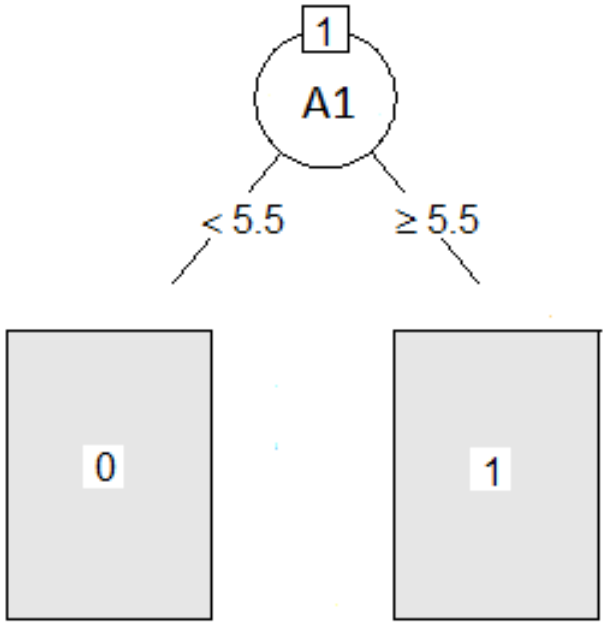

Figura 2.5: Ejemplo Árbol de Decisión.

Construcción del Árbol

La construcción de un árbol de clasificación usando la metodología CART se lleva acabo mediante un procedimiento de división o partición binaria de manera recursiva. La división recursiva divide los datos de las predictoras en una serie de regiones rectangulares en dos dimensiones, cubos en tres dimensiones o hipercubos en dimensiones superiores, la dimensión depende de la cantidad de predictoras [17]. La división recursiva empieza por el nodo raíz que será el primer nodo, luego los nodos hijos o nodos internos recursivamente que serán las siguientes divisiones.

En un conjunto de datos hay diferentes opciones de división por cada predictora, debido a que, en predictoras cuantitativas las opciones serán los puntos medios de cada observación después de organizar las observaciones y en las predictoras cualitativas serán las agrupaciones por la misma categoría. Como resultan varios puntos de división se debe elegir la predictora y punto de división que permita minimizar alguna función de error, intuitivamente sería la precisión de la clasificación o la tasa de mala clasificación, sin embargo, esa función resulta ser menos sensible a los cambios en los nodos [20]. Dicho de otra forma, se busca una medida de impureza que califique a los nodos en función de si contienen datos que pertenecen principalmente a una de las clases de la variable respuesta Y [17], y entre mas mínima sea la impureza en cada nodo indica que más datos en los nodos pertenecen a una sola clase en Y. Es decir, las medidas de impureza permiten cuantificar que predictor a medida que crece el árbol es capaz de distinguir mejor las clases de la variable objetivo o respuesta Y. Aunque existen alrededor de 16 medidas de impureza, en la practica son preferibles el Índice Gini y la Entropía Cruzada [13]. Para esta proyecto se tuvo en cuenta el Índice Gini dado que la Entropía Cruzada computacionalmente puede ser un poco más costosa.

- Índice Gini es definido por

$$G(m) = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.3.1}$$

  Donde $\hat{p}_{mk}$ representa la proporción de observaciones de entrenamiento en el m-ésimo hipercubo que pertenecen a la K-ésima clase [13]. Para dos clases $G(m) = 2\hat{p}(1-\hat{p})$, donde $\hat{p}$ es la proporción de instancias que pertenecen a una de las clases.

- Entropía Cruzada es definida por

$$D(m) = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk}) \tag{2.3.2}$$

  Donde $\hat{p}_{mk}$ representa la proporción de observaciones de entrenamiento en el m-ésimo hipercubo que pertenecen a la K-ésima clase [13]. Para dos clases $D(m) = -\hat{p}\log(\hat{p}) - (1-\hat{p})\log(1-\hat{p})$, donde $\hat{p}$ es la proporción de instancias que pertenecen a una de las clases.

Note que ambas medidas de impureza dependen del m-ésimo hipercubo que hace referencia al m-ésimo nodo. Esto significa, que la medida de impureza se debe calcular en cada m-ésimo nodo. Supongamos que se desea dividir un m-ésimo nodo que se llamará nodo madre, como la división o partición es binaria al intentar dividir este nodo resultan dos posibles nodos hijos, uno izquierdo m_L y uno derecho m_R para cada predictora y punto de división en ese nodo madre, entonces, para cuantificar la reducción de la impureza con cada predictora y punto de división en los nodos hijos se usa la siguiente expresión

$$I(m) = i(m) - P[m_L]i(m_L) - P[m_R]i(m_R) \tag{2.3.3}$$

Donde $i(m)$ es la medida de impureza del nodo madre, $P[m_L]$ y $P[m_R]$ es la probabilidad de que una observación caiga dentro del nodo izquierdo m_L y derecho m_R, y $i(m_L)$ y $i(m_R)$ es la medida de impureza del nodo hijo izquierdo y derecho, respectivamente. La predictora y punto de división que mejor reduzca la impureza en el nodos hijos debe maximizar la expresión 2.3.3 y en consecuencia, será elegida como la división de ese m-ésimo nodo "madre". El cuadro 2.3 resume la idea de la construcción de un árbol de clasificación.

Poda del Árbol y Elección del Árbol Óptimo

Una forma de prevenir el sobreajuste en los árboles de clasificación es eliminar algunos nodos, este proceso se conoce como podar el árbol. Observe que después de construir un árbol y podarlo, resultan una serie de diferentes árboles de tamaños decrecientes llamados subárboles, la pregunta que precede a esto es ¿cómo decido que subárbol usar? Para responder a esa pregunta la metodología CART usa una función conocida como costo-complejidad $R_\alpha(T)$ que permite comparar los subárboles, quien minimice $R_\alpha(T)$ será la elección más óptima. La función $R_\alpha(T)$ es una versión penalizada de la tasa de error de clasificación de un árbol T [17] y se expresa a continuación

Algoritmo. *Construyendo un Árbol de Clasificación*

1. Divida los datos de las predictoras en un subconjunto de hipercubos.
2. Calcule la medida de impureza.
3. Calcule la expresión 2.3.3.
4. Elige la predictora y punto de división que maximice el paso anterior para asignarla como un nodo.
5. Repite el paso del 1 al 4 en las particiones resultantes después del paso 3.

Cuadro 2.3: Construcción de un Árbol de Clasificación. (Adaptado de [13])

$$R_\alpha(T) = Resub(T) + \alpha|T| \tag{2.3.4}$$

La expresión 2.3.4 tiene tres componentes $Resub(T)$, $|T|$ y α. $Resub(T) = 1 - \underset{k}{max}(\hat{p}_{mk})$ es la tasa de error de clasificación del árbol T (en el caso de un árbol de regresión se usa la suma de los cuadrados del error del árbol T), $|T|$ es el número de nodos del árbol T y el parámetro de complejidad α es un peso mayor o igual a cero que compensa el número de nodos de un árbol. Para elegir el árbol óptimo se pueden tener en cuenta los siguientes pasos

1. Aplique la función $R_\alpha(T)$ con distintos valores de α por cada subárbol.
2. Compare los resultados obtenidos en 1 y elija los subárboles con menor resultado de $R_\alpha(T)$.
3. Del paso anterior resultarán varios subárboles con su respectivo α, note que, encontrar un valor de α nos llevará directamente al tamaño del árbol, si $\alpha = 0$ se obtiene el árbol más grande, si α se incrementa reduce el tamaño del árbol y si α disminuye se incrementa el tamaño del árbol. Por lo anterior, Breiman [38] propone obtener la tasa de error de clasificación de los subárboles del paso dos mediante validación cruzada sobre los datos de entrenamiento.
4. El árbol que obtenga en promedio un error de clasificación más bajo por validación cruzada en 3 será el árbol óptimo.

La metodología CART usa la validación cruzada con k particiones [39], dado que, reduce el sobreajuste, generalmente k es igual a 10. El procedimiento consiste en dividir la base de datos de entrenamiento en k particiones mutuamente excluyentes y aproximadamente de igual tamaño, de esas k particiones tome k-1 para entrenar el árbol de clasificación y la k-ésima para probarlo, repita ese proceso k-veces, al final promedie el error de clasificación. El árbol que tiene la menor tasa de error de clasificación se selecciona.

Hiperparámetros

Los hiperparámetros principales usados para entrenar distintos árboles de clasificación se describen en el cuadro 2.4. Se eligen los que consigan el mejor árbol óptimo de todas

las combinaciones de hiperparámetros.

Hiperparámetros	Descripción
Parámetro de complejidad (CP)	Ayuda a la comparación entre diferentes tamaños de árboles
Min split	Número mínimo de observaciones que deben existir en un nodo para que se intente una división
Max depth	Número máximo de nodos entre un nodo hoja y el nodo raíz

Cuadro 2.4: Hiperparámetros Árboles de Clasificación. (Extraído de Rstudio)

Predicción

Una expresión para clasificar la observación x en el m-ésimo hipercubo R_m de un árbol de clasificación es la expresión 2.3.5

$$\hat{f}(x) = \sum_{m=1}^{M} k(m) I(x \in R_m) \tag{2.3.5}$$

Donde $k(m)$ es la categoría mayoritaria en el m-ésimo hipercubo R_m y $I(x \in R_m)$ es la función indicadora. Si x pertenece al m-ésimo hipercubo R_m el valor predictivo de x será la clase mayoritaria en ese hipercubo porque la función indicadora hará cero las demás hipercubos.

Ventajas Árboles de Clasificación

1. Puede aplicarse a cualquier tipo de predictores, cuantitativas o cualitativas.
2. No tiene problemas para trabajar con datos perdidos.
3. Es un clasificador no paramétrico.

Desventajas Árboles de Clasificación

1. No es fácil elegir un árbol idóneo por el problema de alta varianza.
2. Tienden a estar sesgados a favor de predictoras cualitativas.
3. Es sensible al sobreajuste.

2. Bosque Aleatorio

En la sección 2.3.1 se menciono que Bagging (abreviatura de agregación bootstrap) puede reducir drásticamente la varianza en métodos inestables de aprendizaje automático, puesto que, Bagging promedia o elige la clase mayoritaria entre diferentes predicciones sobre muestras bootstrap, lo que lleva a una mejor predicción. Breiman introduce Bagging en 1996, el cual consta de tres pasos básicos [41]

- Seleccione de los datos de entrenamiento B muestras aleatorias con reemplazo (técnica bootstrap). El tamaño de cada muestra B es del tamaño de los datos de entrenamiento.
- Entrene un modelo en esas B muestras del paso anterior.

- Cuando haga predicciones dependiendo del problema a trabajar (regresión o clasificación) tome el promedio o la categoría mayoritaria en el paso anterior.

Como los árboles son inestables se puede usar Bagging para reducir su inestabilidad, donde el modelo que se ajusta en cada muestra bootstrap es un árbol de regresión o clasificación dependiendo del problema, para este proyecto sería de clasificación (ver sección 2.1). Aquí surge un inconveniente, los árboles ajustados resultan estar correlacionados, debido a que, tendrán muchas predictoras y puntos de división en común. El origen del Bosque Aleatorio se da por ese inconveniente en el año 2001 por Breiman [18], su nombre se dio en virtud de dos componentes que posee, muchos arboles ajustados en diferentes muestras Bootstrap (Bosque) y el elemento de aleatorización en el proceso de construcción de los árboles (Aleatorio). El elemento de aleatorización seda en cada nodo de cada árbol en el bosque aleatorio para reducir la correlación entre los árboles que resultan de las muestras bootstrap. Para ello, en cada nodo se extrae una muestra aleatoria de las predictoras de tamaño $m = \sqrt{p}$ [1] (p son todas las predictoras de la base de datos) que serán usadas para determinar con cuál de las predictoras y punto de división dividir los datos. El valor de m se mantiene constante durante la construcción del bosque aleatorio para contrarrestar el efecto de la correlación de los árboles construidos, [13] evitando en promedio que $\frac{p-m}{p}$ de las divisiones no consideren la predictora más significativa, y así, permitiendo que otras predictoras tengan más posibilidades de ser elegidas.

El Bosque Aleatorio hereda las ventajas y una de las desventajas de los árboles de clasificación discutidas en la sección 2.3.1, pues tiende a estar sesgado a favor de predictoras cualitativas, dado que, una predictora cualitativa puede tener una gran cantidad de divisiones potenciales y si el numero de categorías es lo suficientemente grande puede conducir al sobreajuste. Los árboles de clasificación se pueden visualizar pero el bosque aleatorio no, puesto que, tendrá igual cantidad de arboles como muestras bootstrap generadas. Una interesante ventaja del bosque aleatorio es que son una buena alternativa cuando se tienen más predictoras que número de observaciones [17].

Construcción del Bosque Aleatorio

La construcción del Bosque Aleatorio comienza seleccionando muestras aleatorias con reemplazo (muestras Bootstrap) a partir del conjunto de datos de entrenamiento del mismo tamaño del conjunto de datos de entrenamiento [42]. En promedio, cada muestra bootstrap selecciona alrededor de dos tercios de las observaciones para ajustar cada árbol del bosque aleatorio. La tercera parte restante que no se usa, se denomina observaciones fuera de la bolsa (de sus siglas en ingles OOB) [13]. En cada muestra Bootstrap se construye un árbol de clasificación dejándose crecer completamente sin poda. En cada nodo se elige un subconjunto aleatorio de predictoras de tamaño m ($m = \sqrt{p}$) para determina en esos mismos nodos con cuál de las predictoras y punto de división dividir los datos. La elección de la predictora de un subconjunto de predictoras permite reducir el tiempo de ejecución del bosque aleatorio y si se implementa en paralelo se puede obtener un entrenamiento relativamente rápido [42]. El cuadro 2.5 resume como se construye un bosque aleatorio. Considere N datos (filas) de entrenamiento y p predictoras (columnas).

[1] Si m=p equivale a usar bagging

Algoritmo. *Construyendo un Bosque Aleatorio*

1. Elija 1 muestra bootstrap de tamaño N de los datos de entrenamiento.
2. Construya un árbol de clasificación en la muestra del paso 1 y en cada nodo seleccione una predictora de las $m = \sqrt{p}$ predictoras elegidas aleatoriamente.
3. Repita el paso 1 al 2 en paralelo.

Cuadro 2.5: Construcción de un Bosque Aleatorio. (Adaptado de [13])

OOB (Out of Bag)

Hay una forma muy sencilla de estimar el error de prueba en el entrenamiento de un bosque aleatorio sin necesidad de realizar validación cruzada [13]. Esa forma es usando la tercera parte restante que no se usa para ajustar cada árbol del bosque aleatorio, denominada observaciones fuera de la bolsa (OOB). Las OBB se usan para calcular el error de clasificación (error OOB) en el entrenamiento con la expresión 2.3.6 que representa la proporción de observaciones que fueron incorrectamente clasificadas por el bosque aleatorio, donde $\hat{f}_{oob}(x)$ es la predicción de una observación x del bosque aleatorio (ver sección que sigue) en OOB.

$$Error_{OOB} = \frac{1}{N}\sum_{i=1}^{N} I(Y_i \neq \hat{f}_{oob}(x)) \tag{2.3.6}$$

Predicción

Una expresión para clasificar una observación x es la expresión 2.3.7

$$\hat{f}_B(x) = \text{categoría mayoritaria}\{\hat{f}_b(x)\}_1^B \tag{2.3.7}$$

Donde B es el numero de muestras Bootstrap o árboles de clasificación y $\hat{f}_b(x)$ es la predicción en el b-ésimo árbol del bosque aleatorio.

Hiperparámetros

Los hiperparámetros principales usados para entrenar distintos bosques aleatorios se describen en el cuadro 2.6. Se eligen los que consigan minimizar el error OBB de todas las combinaciones de hiperparámetros.

Hiperparámetros	Descripción
mtry	Número de variables seleccionadas aleatoriamente en cada división
Num trees	Número de árboles del bosque
Min node size	Número de observaciones mínimo en los nodos terminales de los árboles

Cuadro 2.6: Hiperparámetros Bosque Aleatorio. (Extraído de Rstudio)

Ventajas Bosque Aleatorio

1. Es resistente al sobreajuste porque reduce la inestabilidad de los árboles.

2. Al usar Bagging con árboles se reduce la correlación entre los árboles.

3. Es una buena alternativa cuando se tienen más predictoras que número de observaciones.

4. Es relativamente rápido.

5. No tiene problemas para trabajar con datos perdidos.

Desventajas Bosque Aleatorio

1. La implementación en R no funciona para predictoras cualitativas con más de 32 categóricas.

2. Tienden a estar sesgado a favor de predictoras cualitativas.

3. No es posible obtener una representación gráfica sencilla del método.

3. Stochastic Gradient Boosting

En 1988 Michael Kearns formula la pregunta ¿Puede un conjunto de clasificadores débiles crear un clasificador robusto? [44], más tarde, Robert Schapire en 1990 [45] responde afirmativamente a esa pregunta, dicha respuesta, llevó al desarrollo de boosting, y con ello, al desarrollo de varias metodologías o algoritmos que usan boosting, la primera conocida como AdaBoost por Yoav Freund y Robert Schapire en 1997, seguido de Gradient Boosting y Stochastic Gradient Boosting por Jerome Friedman en 1999 y 2000, y la más reciente XGBoost de sus siglas en ingles eXtreme Gradient Boosting por Tianqi Chen en 2016, entre otras. Este proyecto utiliza la metodología Stochastic Gradient boosting y se describe a continuación.

Recuerde que Bagging (ver sección 2.3.1) implica crear varias muestras bootstrap del conjunto de datos de entrenamiento para ajustar un modelo (que puede ser un árbol) en esas muestras en paralelo, y luego combina todos los modelos para crear un solo modelo predictivo. Boosting funcionan de manera similar excepto que los modelos, llamados weak learners, se ajustan secuencialmente utilizando la información de los modelos ajustados previamente [13]. Es decir, ajuste de manera secuencial cada weak learner para que aprenda de los errores del anterior y de esta manera impulse (como se refleja en el término "boosting") el rendimiento predictivo al formar un conjunto de modelos colectivamente poderoso [46]. Es particularmente común aplicar boosting con árboles limitados a pocas divisiones, si es solo una división los árboles se conocen como stump [46]. Boosting es otra alternativa poderosa de cómo combinar modelos para lograr un mayor rendimiento [17] y parece dominar a bagging, bosque aleatorio y árboles en la mayoría de los problemas, generalmente es un método que sobre sale entre los métodos de aprendizaje automático, y en consecuencia, se ha convertido en la opción preferida [20].

Como Stochastic Gradient Boosting aplica la idea de boosting al crear una secuencia de modelos donde cada uno intenta aportar una mejora al rendimiento del conjunto de entrenamiento logrado por sus predecesores, se diferencia de las demás metodologías que

usan boosting porque extrae diversas muestras al azar (sin reemplazo) de los datos de entrenamiento para ajustar los modelos secuencialmente a los residuos del modelo previamente creado en lugar de ajustar los modelos a los valores de la variable objetivo cuando crea un modelo posterior [47]. Esta metodología se presenta en la siguiente sección. En este proyecto los modelos que usa el Stochastic Gradient Boosting son árboles de regresión limitados a pocas divisiones, y en consecuencia, se heredan las ventajas y una de las desventajas de los árboles discutidas en la sección 2.3.1, pues tiende a estar sesgado a favor de predictoras cualitativas, dado que, una predictora cualitativa puede tener una gran cantidad de divisiones potenciales y si el numero de categorías es lo suficientemente grande puede conducir al sobreajuste. De esta manera, aunque el stochastic gradient boosting sea una metodología potente que reduce el tiempo de computo en cada muestra al azar y en muchos casos produce un modelo más preciso, se puede sobreajustar [20].

Construcción del Stochastic Gradient Boosting

La construcción del stochastic gradient boosting inicia con un modelo constante óptimo [20], que es un solo nodo terminal o un solo valor constante inicial determinado por la expresión 2.3.8.

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \gamma) \tag{2.3.8}$$

Donde $L(y_i, \gamma)$ representa la función de pérdida, que es una función que permite calcular los errores de predicción. Luego se extrae una fracción $0 < \eta \leq 1$ de las observaciones de entrenamiento sin reemplazo para formar una muestra de tamaño n<N. Sobre esa muestra calcule los residuales o seudo residuales (se pueden pensar como la probabilidad observada menos la probabilidad predicha) con la expresión 2.3.9, que se conoce como el gradiente negativo, compuesto por la derivada de la función de pérdida con respecto a la predicción $f(x_i)$, evaluada en la predicción mas reciente $f(x) = f_{m-1}(x)$.

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad i = 1, \ldots, n \tag{2.3.9}$$

Donde i es la fila (o dato) y m es el árbol que se esta construyendo. Después, ajuste un árbol de regresión a esos seudo residuales con sus respectivas predictoras y puntos de división, que dará como resultado a j-ésimos hipercubos R_{jm} con $j = 1, \ldots, J_m$. Seguidamente, calcule el valor de γ_{jm} por cada j-ésimo hipercubo que minimice la función de pérdida cuando se añade la predicción mas reciente $f_{m-1}(x_i) + \gamma$ con la expresión 2.3.10.

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \tag{2.3.10}$$

Por último haga nuevas predicciones basadas en las predicciones previas usando la expresion 2.3.11

$$f_m(x) = f_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \tag{2.3.11}$$

con v la tasa de aprendizaje que reduce el sobreajuste y escala la contribución de cada árbol por un factor entre $0 < v < 1$. El cuadro 2.7 resume como se construye un stochastic gradient boosting. Considere N datos (filas) de entrenamiento.

Algoritmo. *Construyendo Stochastic Gradient Boosting*

1. Dar un valor inicial como modelo constante.
2. Extraer una muestra de tamaño n<N al azar sin reemplazo de los datos de entrenamiento.
3. Calcule los residuales o seudo residuales.
4. Ajuste un árbol de regresión al paso anterior.
5. Hacer predicciones con el modelo del paso anterior.
6. Repita el paso del 2 al 5 m veces

Cuadro 2.7: Construcción de Stochastic Gradient Boosting. (Adaptado de [20])

Predicción

Una expresión para clasificar una observación x es la expresión 2.3.12

$$\hat{f}(x) = \sum_{m=1}^{M} f_m(x) \tag{2.3.12}$$

Donde $f_m(x)$ es la predicción en el m-ésimo árbol.

Hiperparámetros

Los hiperparámetros principales usados para entrenar distintos stochastic gradient boosting se describen en el cuadro 2.8. Se eligen los que consigan minimizar la función de pérdida de todas las combinaciones de hiperparámetros.

Hiperparámetros	Descripción
N trees	Número de arboles para ajustar
Interaction depth	Profundidad máxima de cada árbol
Shrinkage	Tasa de aprendizaje
N min obs in node	Número mínimo de observaciones en los nodos terminales de los árboles
Bag fraction	Fracción de observaciones seleccionadas al azar para proponer el siguiente árbol

Cuadro 2.8: Hiperparámetros Stochastic Gradient Boosting. (Extraído de Rstudio)

Ventajas Stochastic Gradient Boosting

1. Es un método poderoso cuando se pueden encontrar buenos weak learners.
2. Tiende a tener un desempeño sobresaliente entre los métodos de aprendizaje automático.

3. No tiene problemas para trabajar con datos perdidos.

Desventajas Stochastic Gradient Boosting

1. Puede sobreajustarse.
2. No es posible obtener una representación gráfica sencilla del método.

2.3.2. Basado en Probabilidad

1. Naïve Bayes

Naïve Bayes o Bayes Ingenuo es un método de clasificación basado en la probabilidad bayesiana. Es ingenuo porque asume predictores independientes. Es simple y sencillo de aplicar. Dado un vector de entradas $(X_1 = u_1, X_2 = u_2, \ldots, X_m = u_m)$ el objetivo es elegir la clase c_i que maximice la probabilidad posterior de la expresión 2.3.13. Es decir, [17] la probabilidad posterior de esa clase dadas las predictoras.

$$P(Y = c_i|X_1, X_2, \ldots, X_m) = \frac{P(X_1, X_2, \ldots, X_m|Y = c_i)P(Y = c_i)}{P(X_1, X_2, \ldots, X_m)} \tag{2.3.13}$$

Como el denominador es la probabilidad conjunta de todas las predictoras y no está influenciado por la clase elegida, entonces, maximizar la probabilidad de la clase posterior equivale a maximizar el numerador, obteniendo la expresión 2.3.14, donde u_j representa las observaciones por fila asociadas a cada predictora.

$$\begin{aligned} c &= \underset{c}{\operatorname{argmax}}\, P(X_1, X_2, \ldots, X_m|Y = c)P(Y = c) \\ &= \underset{c}{\operatorname{argmax}}\, P(X_1 = u_1|Y = c) \ldots P(X_m = u_m|Y = c)P(Y = c) \\ &= \underset{c}{\operatorname{argmax}}\, P(Y = c) \prod_{j=1}^{m} P(X_j = u_j|Y = c) \end{aligned} \tag{2.3.14}$$

Muchas predictoras en el producto pueden producir valores demasiados grandes, entonces, se sugiere usar logaritmos, obteniendo la expresión 2.3.15.

$$\begin{aligned} c &= \underset{c}{\operatorname{argmax}}\, log(P(Y = c) \prod_{j=1}^{m} P(X_j = u_j|Y = c)) \\ &= \underset{c}{\operatorname{argmax}}\, log(P(Y = c)) + \sum_{j=1}^{m} log(P(X_j = u_j|Y = c)) \end{aligned} \tag{2.3.15}$$

Si X_j es una predictora cualitativa $P(X_j = u_j|Y = c) = \frac{\text{Número de observaciones con} X_j = u_j \text{en la clase c}}{\text{Número de observaciones en la clase c}}$. Es decir, la frecuencia relativa. Si X_j es una predictora cuantitativa, hay dos alternativas

- Aplicar un método de discretización (Transformar datos cuantitativos a cualitativos).
- Asumir una distribución para cada predictora, en general, Gausiana con media y varianza estimada de los datos en la clase c o estimar la distribución usando un método como el kernel.

Corrección de Laplace

En muchas ocasiones los datos no contienen muestras de todas las combinaciones de las predictoras posibles. Es decir, supongamos que se desea clasificar una palabra .[es]tupendoçomo positiva o negativa. Esta palabra no esta en el datos, entonces, $P(\text{estupendo}|Y = c) = 0$. Esto hará que la probabilidad posterior sea igual a cero sin importar las otras probabilidades obtenidas ya que multiplicamos todas las probabilidades. Esto se conoce como el problema de probabilidad cero.

La corrección de Laplace o suavizado de Laplace [24] es una técnica que resuelve dicha problemática. Calcula esa probabilidad que hace cero las demás como se muestra en la expresión 2.3.16.

$$P(X_j = u_j|Y = c) = \frac{\text{Número de observaciones con} X_j = u_j \text{en la clase c} + \alpha}{h + \alpha * k} \quad (2.3.16)$$

Donde h es el número de observaciones en la clase c y k es el número de predictoras y α representa el parámetro de suavizado.

Hiperparámetros

Los hiperparámetros principales usados para entrenar distintos Naïve Bayes se describen en el cuadro 2.9. Se eligen los que consigan maximizar la probabilidad posterior (expresión 2.3.13) de todas las combinaciones de hiperparámetros.

Hiperparámetros	Descripción
Use kernel	Estimación de densidad
fL	Suavizado de Laplace
adjust	Ajustar el ancho de banda de la densidad del kernel (números más grandes significan una estimación de densidad más flexible)

Cuadro 2.9: Hiperparámetros Naïve Bayes. (Extraído de Rstudio)

Ventajas Naïve Bayes

1. Sencillo.
2. Rápido, no tiene problema para trabajar con 10,000 predictoras.

Desventajas Naïve Bayes

1. Las probabilidades cero afectan al clasificador.
2. No tiende a ser el método de aprendizaje automatizado con mejor desempeño.

2. Regresión Logística

La regresión logística [17] es un tipo de modelo lineal generalizado (GLM). Es un modelo que generaliza los conceptos y habilidades de los modelos lineales a casos cuando la variable respuesta sigue una distribución binomial. Este modelo obtiene como salida valores entre $[0, 1]$. Cada predictora se escala linealmente y este resultado es la entrada para una trasformación no lineal, la función logística 2.3.17. Asegurando que la salida se pueda

interpretar como una probabilidad, la probabilidad de que la entrada (de la combinación lineal de las predictoras) pertenezca a la clase o categoría 1 (expresión 2.3.17).

$$P(y_i = 1|\mathbf{X}) = \frac{e^{\beta^T \mathbf{X}}}{e^{\beta^T \mathbf{X}} + 1} \tag{2.3.17}$$

Este modelo [17] ya no requiere supuesto de normalidad para los residuos y tampoco necesitamos el supuesto homocedástico. Las predictoras no necesitan ser independientes, sin embargo, en la práctica el modelo enfrentará problemas si las predictoras tienen un alto grado de multicolinealidad. Cuando la variable respuesta Y tiene dos categorías se puede modelar usando la distribución Bernoulli $y_i \sim Ber(p)$, donde $p = P(y_i = 1|\mathbf{X})$ y p esta relacionado con las predictoras $\mathbf{X}$ a través de la función de enlace.

Función de Enlace

Una función de enlace es una función usada para conectar las salidas de una función a una función lineal. Por ejemplo, la regresión logística obtiene como resultados (valores de salida) entre $[0, 1]$, para relacionar estos valores de salida con las combinaciones lineales de las predictoras (valor de entrada de la función logística) se usa la función log-odds o función logit (expresión 2.3.18).

$$log\left(\frac{P(y_i = 1|\mathbf{X})}{1 - P(y_i = 1|\mathbf{X})}\right) = \beta^T \mathbf{X} \tag{2.3.18}$$

Por ejemplo, dada una función lineal como entrada para la función logística se obtiene un valor entre $[0, 1]$. Ese valor será la entrada para la función logit que produce una salida entre $(-\infty, \infty)$. La salida de la función logit en otras palabras proviene de la función lineal que fue entrada de la función logística. Con esto, si se desea interpretar los parámetros (coeficientes) de la regresión logística se debe hacer en términos de la función logit de la variable respuesta.

Parámetros

Como cada predictora se escala linealmente, cada predictora tendrá un coeficiente o parámetro β que debe ser estimado. La estimación de los β se hace mediante máxima verosimilitud dado que [13] tiene mejores propiedades estadísticas. La idea de máxima verosimilitud es

- Encontrar primero la función de densidad conjunta o la distribución de probabilidad de todas las observaciones. Esa función será la función de verosimilitud.
- Aplique el logaritmo a esa función del paso anterior. Esta función será la función logarítmica de la verosimilitud.
- La estimación del parámetro es el parámetro que maximice el paso anterior.

Como los datos para este proyecto tiene variable respuesta graduación Y (si) y N (no), dos categorías, entonces, la distribución de probabilidad para cada y_i (función de verosimilitud) es una distribución Bernoulli, es decir, $Y \sim Ber(p)$ (la expresión 2.3.19).

$$f(y_i; p) = p^{y_i}(1 - p)^{1-y_i}, i = 1, 2, \ldots, N \tag{2.3.19}$$

Sea [20] $p = P(y_i = 1|\mathbf{X}) = p(\mathbf{Y};\beta)$ para expresar la función logarítmica de la verosimilitud (log-likelihood function) como se muestra en la expresión 2.3.20

$$l(\beta;\mathbf{Y}) = logL(\beta;\mathbf{Y}) = \sum_{i=1}^{N} logf(y_i;p) = \sum_{i=1}^{N} \{y_i log(p(\mathbf{Y};\beta)) + (1-y_i)log(1-p(\mathbf{Y};\beta))\} \tag{2.3.20}$$

Entonces la estimación de los β's se realiza maximizando $l(\beta;\mathbf{Y})$ 2.3.20.

Ventajas Regresión Logística

1. Funciona bien para datos linealmente separables.

2. Fáciles de implementar.

Desventajas Regresión Logística

1. Las estimaciones de los parámetros son inestables cuando las clases están bien separadas.

2. Limita el tipo de asociaciones a explorar entre la variable respuesta y las predictoras.

2.3.3. Basado en Redes Neuronales Artificiales

1. TabNet

Las investigaciones han demostrado un éxito contundente de las redes neuronales profundas (Deep Learning Networks (DNN)) con [22] imágenes, [23] texto y [21] audio. Ejemplos como Amazon Rekognition, BERT, Alexa de Amazon y Siri de Apple demuestran su éxito, entre otras. Amazon Rekognition detecta objetos, escenas y rostros; extrae texto, reconoce a personas famosas e identifica contenido inapropiado en imágenes, también le permite realizar búsquedas y comparar rostros. BERT usada en el buscador de Google, procesa la manera en que nos expresamos. Alexa de Amazon y Siri de Apple asistentes virtuales que permiten la interacción humano-computadora. Sin embargo, no se ha obtenido tanto éxito con datos tabulares, así que, en agosto del 2019 Sercan O. Arik y Tomas Pfister [7] investigadores de Google proponen una nueva arquitectura de red neuronal profunda para datos tabulares llamada Tabnet, diseñada para heredar los beneficios de los métodos basados en árboles (la explicabilidad (importancia de las predictoras)) y los beneficios de los métodos basados en aprendizaje profundo (optimización basada en el descenso del gradiente). Esta nueva arquitectura se construye secuencialmente en varios pasos, entre 3 y 10, más pueden producir sobreajuste. Su construcción permite un aprendizaje más eficiente a diferencia de la arquitectura de redes neuronales profundas y otras integraciones de DNN con árboles de decisión (DT), ya que elige las predictoras más relevantes a través de la atención secuencial (transformador atento), impidiendo que en el aprendizaje se escojan variables irrelevantes. Además, tabnet se puede generalizar más allá de los datos de entrenamiento para conjuntos de datos donde la mayoría de las

predictoras son redundantes [7]. TabNet es fácil de implementar, pero no se han confirmado sus componentes teóricos, y en consecuencia, no se ha publicado en algún jornal. Su arquitectura se ilustra en la figura 2.6

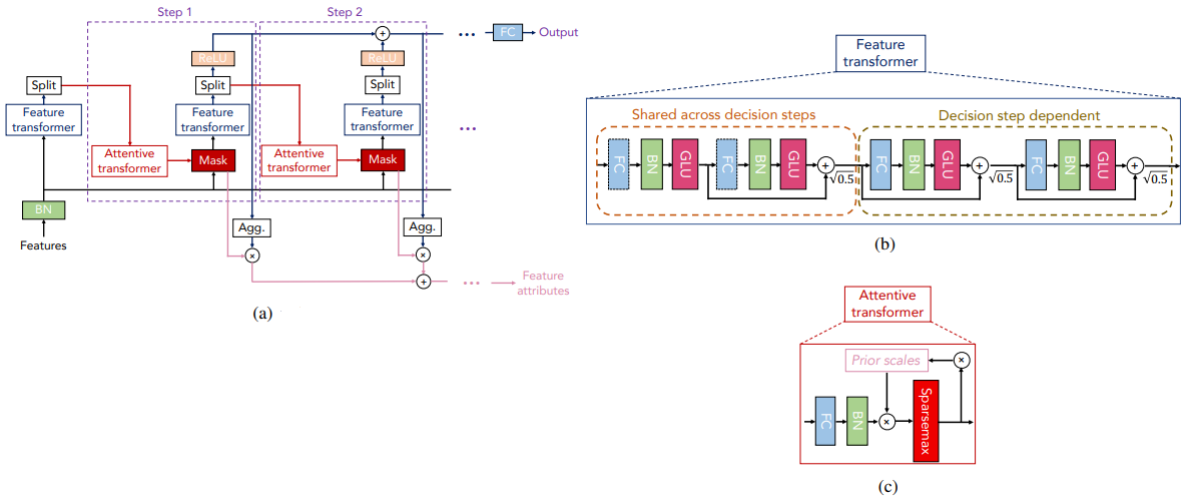

Figura 2.6: (a) Arquitectura Tabnet compuesta por un (b) transformador de predictoras y (c) un transformador atento. (Extraída de [7])

Funcionamiento

El funcionamiento de TabNet usando la figura 2.6 se puede describir de la siguiente manera.

1. Las predictoras se normalizan o se re-escalan usando una normalización por lote (BN) que consiste en seleccionar muestras disjuntas, calcular la media y varianza en cada muestra, y para cada observación que pertenezca a cada muestra se resta por la media de esa muestra y se divide entre la varianza de esa muestra, por último cada observación se transforma linealmente. Esto permite usar el beneficio promedio de baja varianza.

2. Se procesan las predictoras normalizadas con un transformador de predictoras (Feature Transformer).

3. Se dividen (split) para el paso de decisión en la salida y otro para la información del siguiente paso.

4. Entran al trasformador atento (Attentive Transformer) para generar un peso o coeficiente por predictora.

5. El peso entra a la mascara dispersa multiplicativa (Mask) para multiplicarse con la predictora a la que corresponde el peso.

6. Repita varias veces (step) desde 2 hasta 5.

Donde el transformador de predictoras (Feature Transformer) y el trasformador atento (Attentive Transformer) son bloques de DNN. Para un aprendizaje robusto y eficiente el transformador de predictoras (Feature Transformer) debe incluir dos capas, una capa compartida y una capa dependiente, ambas capas se componen de, una capa completamente conectada (FC), es decir, tiene conexiones completas con todas las activaciones en la capa anterior, como se ve en las redes neuronales normales. Una normalización por lote (BN) y una función de activación GLU (de sus siglas en ingles Gated Linear Unit) para que la red decida cuanta información dejar fluir, puesto que, GLU tiene valores de salida entre $[0, 1]$. El $\sqrt{0.5}$ es una normalización que ayuda a estabilizar el aprendizaje, al garantizar que la varianza de la red no cambie drásticamente. El trasformador atento (Attentive Transformer) se usa como una máscara de aprendizaje para la selección de predictoras más destacadas, para ello, debe incluir una capa completamente conectada (FC), una normalización por lote (BN), una multiplicación por un valor a priori que denota cuanto una predictora se ha usado anteriormente y una función de activación sparsemax para establecer los valores más pequeños de entrada como cero, dando resultado a una selección de predictoras más relevantes.

Hiperparámetros

Los hiperparámetros principales usados para entrenar distintos TabNet se describen en el cuadro 2.10. Se eligen los que consigan maximizar la sensibilidad (ver sección 2.4) de todas las combinaciones de hiperparámetros.

Hiperparámetros	Descripción
Max epochs	Número máximo de épocas para el entrenamiento
Batch size	Tamaño del lote a normalizar. Se recomiendan lotes grandes
Virtual batch size	Tamaño del lote a normalizar en el Feature Transformer y el Attentive Transformer
Patience	Número de épocas consecutivas sin mejora antes de realizar una parada anticipada

Cuadro 2.10: Hiperparámetros TabNet. (Extraído de [7])

Ventajas TabNet

1. Tiene un aprendizaje más eficiente a diferencia de otras DNN y otras integraciones de DNN con árboles de decisión (DT).

2. Se puede generalizar para conjuntos de datos donde la mayoría de las predictoras son redundantes.

3. En el entrenamiento se eligen las predictoras más relevantes.

Desventajas TabNet

1. Puede sobreajustarse.

2. Si hay que ajustar hiperparámetros esto puede tomar tiempo.

3. Aún se requiere mas evidencia para mostrar si verdaderamente tiene una desempeño competitivo en comparación a otros métodos de aprendizaje automatizado.

Se resume en el cuadro 2.11 las ventajas y desventajas de los métodos descritos anteriormente.

Metodos	Ventajas	Desventajas
Árboles de Clasificación	1. Puede aplicarse a cualquier tipo de predictores, cuantitativas o cualitativas. 2. No tiene problemas para trabajar con datos perdidos. 3. Es un clasificador no paramétrico, esto significa que no requiere suposiciones.	1. No es fácil elegir el árbol óptimo por el problema de alta varianza. 2. Tienden a estar sesgados a favor de predictoras cualitativas. 3. Es bastante sensible al sobreajuste.
Bosque Aleatorio	1. Es resistente al sobreajuste porque reduce la inestabilidad de los árboles. 2. Al usar Bagging con árboles se reduce la correlación entre los árboles. 3. Es una buena alternativa cuando se tienen más predictoras que número de observaciones. 4. Es relativamente rápido. 5. No tiene problemas para trabajar con datos perdidos.	1. La implementación en R no funciona para predictoras cualitativas con más de 32 categóricas. 2. Tienden a estar sesgados a favor de predictoras cualitativas. 3. No es posible obtener una representación gráfica sencilla del método.
Stochastic Gradient Boosting	1. Es un método poderoso cuando se pueden encontrar buenos weak learners. 2. Tiende a tener un desempeño sobresaliente entre los métodos de aprendizaje automático. 3. No tiene problemas para trabajar con datos perdidos.	1. Puede sobreajustarse. 2. No es posible obtener una representación gráfica sencilla del método.
Naïve Bayes	1. Sencillo. 2. Rápido, no tiene problema para trabajar con 10,000 predictoras.	1. Las probabilidades cero afectan al clasificador. 2. No tiende a ser el método de aprendizaje automatizado con mejor desempeño.
Regresión Logística	1. Funciona bien para datos linealmente separables. 2. Fáciles de implementar.	1. Las estimaciones de los parámetros son inestables cuando las clases están bien separadas. 2.Limita el tipo de asociaciones a explorar entre la variable respuesta y las predictoras.
TabNet	1. Tiene un aprendizaje más eficiente a diferencia de otras DNN y otras integraciones de DNN con árboles de decisión (DT). 2. Se puede generalizar para conjuntos de datos donde la mayoría de las predictoras son redundantes. 3. En el entrenamiento elige las variables más destacadas.	1. Puede sobreajustarse. 2. Si hay que ajustar hiperparámetros esto puede tomar tiempo. 3. Aún se requiere mas evidencia para mostrar si verdaderamente tiene una desempeño competitivo en comparación a otros métodos de aprendizaje automatizado.

Cuadro 2.11: Ventajas y Desventajas de los métodos.

2.4. Métricas de Evaluación

Las métricas de evaluación [6] son medidas que permiten calificar el rendimiento de un método de aprendizaje automático. Una forma de representar ese rendimiento es la matriz de confusión (cuadro 2.12). La matriz de confusión es una herramienta que permite visualizar de manera práctica el desempeño de un modelo, colocando en varias casillas los valores predichos por el modelo correctamente o incorrectamente para el caso positivo y el caso negativo en comparación con los valores reales. Para este proyecto se elige como caso positivo a los estudiantes que no se graduaron y como el caso negativo a los estudiantes que si graduaron. Con esta elección se tienen dos errores que un método de aprendizaje automático podría cometer.

- El método establece que un estudiante es capaz de graduarse, pero éste no se gradúa.

- El método establece que un estudiante no es capaz de graduarse, pero éste si se gradúa.

Que un método establezca que el estudiante no se gradúo cuando realmente si se gradúo no es trascendental, dado que, cuando se le brinde servicio necesario se verificará que cumplió con su grado académico. Mientras que si el método establece que el estudiante si se gradúo cuando realmente no lo hizo impedirá que al estudiante se le provea servicio necesario y así aumente las posibilidades de que el estudiante complete su grado. Dicho esto, es más importante minimizar el error de establecer que un estudiante es capaz de graduarse, pero éste no se gradúa. Para ello, se deben tener métricas de evaluación que permitan cuantificar ese error y métricas que permitan identificar correctamente a los estudiantes que no se graduaron. En consecuencia, se definen a continuación dos probabilidades, la sensibilidad y el área bajo la curva de características operativas del receptor (ROC). El método elegido será aquel que minimice esas probabilidades y que maximice la sensibilidad.

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP) Type II Error
Predicted Negative	False Negative (FN) Type I Error	True Negative (TN)

Cuadro 2.12: Matriz de Confusión para métodos de aprendizaje automatizado.

2.4.1. Sensibilidad

La Sensibilidad, también conocido como el Recall, es la proporción de casos positivos que fueron correctamente identificados por el modelo, se calcula con la expresión 2.4.1. A continuación se presenta un ejemplo.

$$\frac{TP}{TP + FN} \tag{2.4.1}$$

Supongamos que tenemos 200 animales, 90 están enfermos y 110 no lo están. De los 90 enfermos el modelo acertó 80 y en 10 se equivocó. Los 110 que no estaban enfermos el modelo acertó en 100 casos y en 10 se equivocó. La matriz de confusión resultante se refleja en el cuadro 2.13.

	Actual Positive	Actual Negative
Predicted Positive	80	10
Predicted Negative	10	100

Cuadro 2.13: Ejemplo matriz de confusión.

Calculando el Recall formulado en la expresión 2.4.1 sería $\frac{80}{80+10} = \frac{80}{90} = 0.888$. Una interpretación es que el modelo aproximadamente el 89 % de los animales enfermos los identifico correctamente.

2.4.2. Área Bajo la Curva

La curva ROC (Receiver Operating Characteristic Curve) [6] se utilizó por primera vez en la segunda guerra mundial por Estados Unidos para mejorar la tasa de detección de aviones japoneses a partir de las señales de su radar. El propósito de la curva ROC es analizar el poder predictivo en la detección de la mayor cantidad de verdaderos positivos y la menor cantidad de falsos positivos (ver figura 2.7). Un clasificador perfecto tendrá una curva desde el origen hasta el punto (0,1), que corresponde a una tasa de 100 por ciento de verdaderos positivos y una tasa de 0 por ciento de falsos positivos [17]. A la curva ROC se le puede calcular el área y está área se conoce como el AUC (Area Under the Curve) que es la proporción de casos positivos y negativos predichos correctamente por el modelo. Por ejemplo, para este proyecto una interpretación sería, el modelo predice correctamente un tanto por ciento a los estudiantes que no se graduaron y que si se graduaron. Si ese modelo esta sobre la linea roja $y = x$ (ver figura 2.7), indica que el modelo cuando predice esta adivinando, ya que estamos calculando el área bajo la línea $y = x$ que corresponde al 50 %; en este proyecto correspondería adivinar al predecir si un estudiante se gradúa o no.

Para construir la curva ROC se consideran posibles puntos de cohorte con el fin de facilitar la construcción y el cálculo del área bajo esa curva (AUC). Para cada punto de cohorte se calcula la sensibilidad con la expresión 2.4.1 y 1-especificidad con $1 - \frac{TN}{FP+TN}$ (ver matriz de confusión - cuadro 2.12) que representarán los puntos del eje y y x respectivamente. Por otro lado, para el área bajo esa curva ROC (AUC) se toman esas mismas coordenadas, se construyen sub-áreas con cada par de coordenadas y al final se suman. A continuación se presenta un ejemplo en el que se calcula la curva ROC y el AUC.

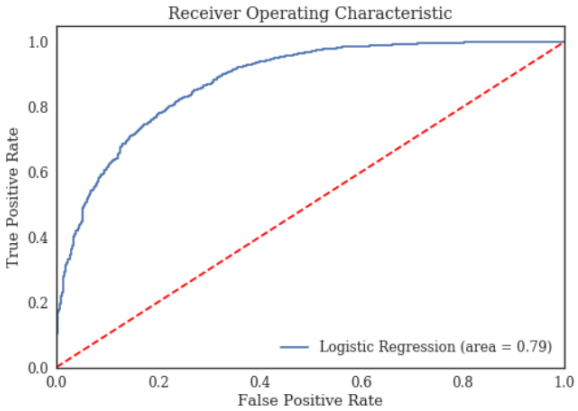

Figura 2.7: Curva ROC. (Extraída de [6])

Supongamos que después de aplicar una prueba que pretende identificar personas con demencia senil en China se obtienen los resultados del cuadro 2.14.

Puntaje	No Demencia	Demencia
0-5	0	2
6-10	0	1
11-15	3	4
16-20	9	5
21-25	16	3
26-30	18	1
Total	46	16

Cuadro 2.14: Datos del ejemplo para la curva ROC y el AUC.

Para construir la curva ROC se consideran los siguientes posibles puntos de cohorte ($\leq 5, \leq 10, \leq 15, \leq 20, \leq 25, \leq 30$). De manera especifica para el cohorte ≤ 20 se obtiene el cuadro 2.15 que permite encontrar 1-especificidad $= 1 - \frac{34}{46} = 0.26$ y la sensibilidad $= \frac{12}{16} = 0.75$. Siguiendo este proceso para cada punto de cohorte se obtiene el cuadro 2.16 y con ese cuadro 2.16 se obtiene la figura 2.8 que representa la curva ROC.

Cohorte	No Demencia	Demencia
$\leq$20	12	12
$>$20	34	4
Total	46	16

Cuadro 2.15: Matriz de confusión para el cohorte $\leq$20.

Cohorte	x = 1-especificidad	y = Sensibilidad
$\leq$5	0	0.125
$\leq$10	0	0.1875
$\leq$15	0.065	0.4375
$\leq$20	0.26	0.75
$\leq$25	0.60	0.93
$\leq$30	1	1

Cuadro 2.16: Coordenadas para graficar la curva ROC.

Para calcular el AUC se puede seguir el siguiente enfoque. Primero construya sub-áreas usando dos rectas y la integral definida para un intervalo en x. La primer recta será la recta que pasa por cada par de puntos y la segunda será la recta $y = x$. La integral permite encontrar el área entre esas dos rectas. Segundo sume todas las sub-áreas y añada el área de la recta $y = x$. De manera especifica para el primera sub-área a se encuentra la recta $y = 3.84x + 0.187$ que pasa por los puntos $(0, 0.1875)$ y $(0.065, 0.4375)$, al restarle a esa recta la recta $y = x$ y usar el intervalo $[0, 0.065]$ en x se obtiene la primera sub-área sombreada con verde en la figura 2.8. La expresión 2.4.2 determina el valor de la sub-área $a = 0.017$. Este proceso se repite para las demás sub-áreas y se obtiene $b = 0.083$, $c = 0.1422$ y $d = 0.0643$. Con lo anterior, el $AUC = a+b+c+d+0.5 = 0.017+0.083+0.1422+0.0643+0.5 = 0.806$ indicando que la prueba identifica correctamente el 80.6 % de los pacientes que tienen demencia senil y que no tienen demencia senil.

$$a = \int_0^{0.065} ((3.84x + 0.187) - x)\, dx = 0.017 \tag{2.4.2}$$

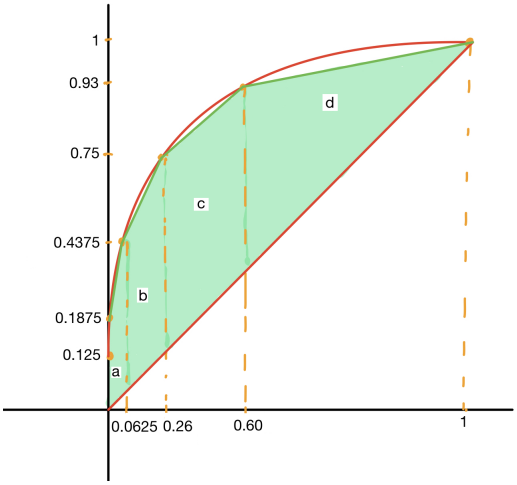

Figura 2.8: Curva ROC y Sub-áreas para calcular el AUC. (Elaboración propia)

2.4.3. Probabilidad de Graduación

Se definen las siguientes expresiones:

$$P1 = \frac{FN}{FN + TP} \qquad (2.4.3) \qquad\qquad P2 = \frac{FN}{FN + TP + FP + TN} \qquad (2.4.4)$$

P1 cuantifica la probabilidad de que el modelo predice que el estudiante si se gradúa dado que realmente el estudiante no se gradúo. P1 también se puede calcular como $1 - Recall$. De todas las predicciones hechas P2 cuantifica la probabilidad de que el modelo predice incorrectamente que el estudiante si se gradúa.

Capítulo 3

Resultados

3.1. Datos

Los datos para este proyecto fueron suministrados por la oficina de planificación, investigación y mejoramiento institucional (OPIMI). Estos datos representan estudiantes subgraduados de la Universidad de Puerto Rico Recinto Mayagüez que comienzan sus estudios en el periodo comprendido entre los años 1999 hasta 2010. La graduación considerada es al 150 % de la duración de su programa académico, por ejemplo, para programas de 4 años el tiempo de finalización al 150 % sería 6 años y para programas de 5 años sería 7.5 años pero se consideran 8 años. Para este proyecto que un estudiante no complete su G150, quiere decir que no se gradúo, sin embargo, tomar la graduación hasta el 150 % no indica que los estudiantes ya no se van a graduar es posible que algunos estudiantes se gradúen luego del 150 % de la duración de su programa académico.

Se obtuvo 24,432 estudiantes subgraduados, de los cuales 19,546 estudiantes son para entrenamiento (80 %) y 4,886 estudiantes (20 %) son para prueba. Se hicieron dos análisis dado que la información contenía valores faltantes. En el primer análisis para entrenamiento se usaron 14,418 estudiantes al eliminar los valores faltantes de los 19,546 estudiantes de entrenamiento. Para prueba 3,627 estudiantes al eliminar los valores faltantes de los 4,886 estudiantes de prueba. En el segundo análisis se usó MICE (ver sección 2.2) como método de imputación de datos para los 19,546 estudiantes de entrenamiento y para prueba se usaron los mismos 3,701 estudiantes sin valores faltantes. El software utilizado para el desarrollo de los modelos fue Jupyter Notebook que es una aplicación cliente-servidor que permite editar y ejecutar lenguaje de programación Python y Rstudio que es un Integrated Development Environment (IDE) para el lenguaje de programación R.

El 28 de diciembre del 2020 se obtuvo un permiso especial para este proyecto. Dado que estos datos involucran información controversial sobre seres humanos. Se solicitó por la Oficina del Comité para la Protección de los Seres Humanos en la Investigación (CPSHI) un certificado que permita el manejo de la información.

3.1.1. Variables

La Universidad dispone de una variedad de información sobre cada estudiante, no obstante, para propósitos de este proyecto se considera la siguiente información:

Variable	**Valor**	**Explicación**
Año	Cuantitativo	Año de admisión
Facultad	Cualitativo	Facultad de admisión
Programa académico	Cualitativo	Programa de admisión
Aptitud Verbal	Cuantitativo	Extraído del College Board
Aprovechamiento Matemáticas	Cuantitativo	Extraído del College Board
Aptitud Matemáticas	Cuantitativo	Extraído del College Board
Aprovechamiento Español	Cuantitativo	Extraído del College Board
Aprovechamiento Inglés	Cuantitativo	Extraído del College Board
Ingreso Familiar	Cualitativo	Ingreso Familiar
Educación Padre	Cualitativo	Educación del Padre
Educación Madre	Cualitativo	Educación de la Madre
Genero	Cualitativo	Genero (Masculino, Femenino)
Tipo de Escuela	Cualitativo	Tipo de Escuela (privada, pública, otra)
GPA Primer año	Cuantitativo	GPA del primer año en la universidad
Graduación	Cualitativo	Graduación (si, no)
Rel Estudiante GPA	Cuantitativo	Preparación del estudiante
Rel Escuela GPA	Cuantitativo	Calidad de la escuela

Se codificaron las siguientes predictoras:

- Ingreso Familiar en cinco categorías, I1-L12.5 (Menos de \$500 - \$12, 499), I2-B12.5A20 (\$12, 500 - \$19, 999), I3-B20A30 (\$20, 000 - \$29, 999), I4-B30A50 (\$30, 000 - \$49, 999), I5-O50 (Más de \$50, 000).
- Educación del padre y educación de la madre en seis categorías, del menor al mayor grado alcanzado. None (Ninguna), LHS (No completó escuela superior), HS (Completo Escuela Superior), Assoc or Less (Obtuvo grado asociado o menos al asistir a la universidad), College (Bachillerato), Grad (Maestría o Doctorado).

Las predictoras Apt Verbal, Aprov Matem, Apt Matem, Aprov Espanol, Aprov Ingles, miden el conocimiento del estudiante según la prueba College Board en un rango de puntaje que oscila entre 200 a 800 puntos.

Rel Estudiante GPA

El promedio de calificaciones (GPA, Grade Point Average) de escuelas superiores no es uniformemente medido porque las escuelas tienen estándares académicos distintos. Por tanto, usar solo el GPA de escuelas superiores se vuelve problemático [10], y esto limita su uso en nuestros modelos. Por ejemplo, un estudiante de escuela A con un GPA de escuela superior de 3.2, no necesariamente está mejor preparado para la universidad que un estudiante de escuela B con un GPA de escuela superior de 3, debido a que el GPA

de escuela superior es dependiente de la dificultad de los cursos en la escuela. El Rel Estudiante GPA pretende aliviar esta limitación del GPA de escuela superior.

El Rel Estudiante GPA se obtiene al dividir el GPA de escuela superior del estudiante al solicitar admisión, entre el promedio del GPA de escuela superior de todos los estudiantes admitidos a la universidad. De esta manera, la predictora explica qué tan preparados están los estudiantes para la universidad en relación a otros estudiantes admitidos. Por ejemplo, el estudiante Juan Lopez y Carlos Murillo provienen de la escuela American Military y Eugenio Maria de Hostos, respectivamente. Juan tiene un GPA de 3.2 y Carlos un GPA de 3. El promedio del GPA de todos los estudiantes admitidos a la universidad es de 3.1. Entonces, el Rel Estudiante GPA de Juan es $3.2/3.1 \approx 1.03$ y el de Carlos es $3/3.1 \approx 0.97$. Claramente el estudiante Juan Lopez de la escuela American Military esta mejor preparado para la universidad que Carlos Murillo de la escuela Eugenio Maria de Hostos dado que $1.03 > 0.97$. Estudiantes con el mismo Rel Estudiante GPA están igualmente preparados para la universidad.

Rel Escuela GPA

Rolke (2014 [10]) introduce la predictora Rel Escuela GPA que pretende medir la calidad de una escuela superior en relación a otras escuelas superiores. El Rel Escuela GPA se obtiene dividiendo el promedio del GPA de primer año de universidad de los estudiantes de una misma escuela superior entre el promedio del GPA de la escuela de esos mismos estudiantes [10]. Por ejemplo, los estudiantes que han terminado su primer año en la UPRM que provienen de la escuela Luis Muñoz Rivera tiene un promedio de GPA de primer año de 1 y el promedio del GPA de la escuela de esos mismos estudiantes es de 3.2, entonces la Rel Escuela GPA = $1/3.2$=0.31. Por otra parte, consideremos los estudiantes que han terminado su primer año en la UPRM provenientes de la escuela Cecilio Lebron Ramos con un promedio de GPA de primer año de 3 y el promedio GPA de la escuela de esos mismos estudiantes es de 3.5, entonces la Rel Escuela GPA = $3/3.5$=0.85. Claramente la escuela Cecilio Lebron es de mejor calidad que la escuela Luis Muñoz Rivera dado que $0.85 > 0.31$.

3.2. Análisis Exploratorio

El Recinto Universitario de Mayagüez desde hace muchos años se ha dado a conocer por su facultad de ingeniería y coloquialmente como el recinto de las "matemáticas fuertes". La facultad de ingeniería es la facultad que tiene la tasa más alta de graduación al 150 %, cercana del 68 %, seguida, por la facultad de artes y ciencias con un 58 %. En contraste, la facultad que tiene la tasa más baja es la facultad de ciencias agrícolas cercana del 39 %. Se esperaría que los programas que pertenecen a las facultades con una tasa de graduación alta, tengan una tasa de graduación sobresaliente por programas en comparación con los demás.

Quien lidera la lista de programas con una tasa de graduación alta es Biotecnología Industrial con un 73 %, seguido de ingeniería mecánica con un 71 %, pertenecientes a la

facultad de artes y ciencias, y la facultad de ingeniería respectivamente. Aunque Matemáticas está cerca del 31 % Economía Agraria es la que tiene la tasa más baja cercana del 17 % como se muestra en la figura 3.1. En específico, alrededor del 83 %, 81 % y 78 % de los estudiantes del programa de Economía Agraria, Agronegocios y Artes plásticas, respectivamente, no completan su G150. Esto muestra la necesidad de un servicio necesario para aumentar las posibilidades de que el estudiante complete su grado.

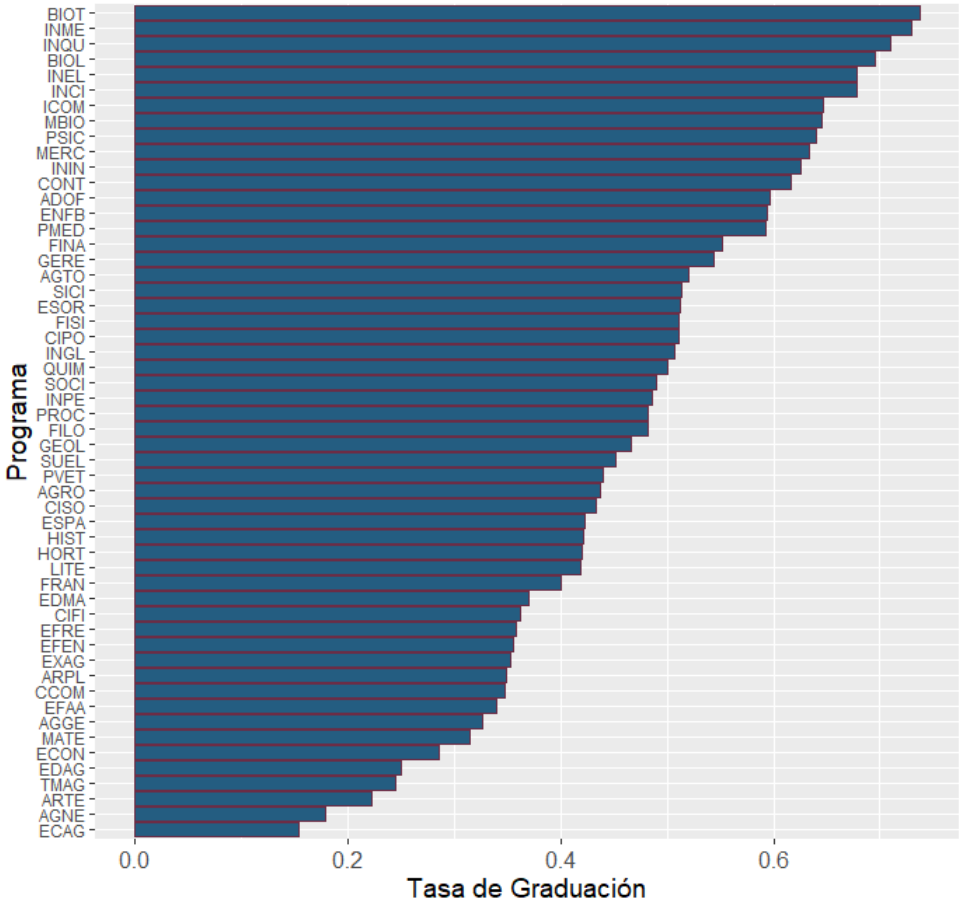

Figura 3.1: Tasa de Graduación exitosa por programa.

La tasa de graduación en la Universidad de Puerto Rico Recinto Mayagüez se mantuvo entre el 56 % y 61 % para años de ingreso entre 1999 y 2008, con un promedio del 57.56 %. Para el año 2009 se obtiene una disminución dramática en las tasas de graduación cercana del 7.56 % y un aumento sustancial del 15 % para el 2010 en comparación con el 2009, como se muestra en la figura 3.2. Posiblemente el causante del gran descenso del 2009 fue una huelga que duró aproximadamente 4 meses en ese mismo año, retrasando algunos estudiantes en su camino hacia la graduación exitosa.

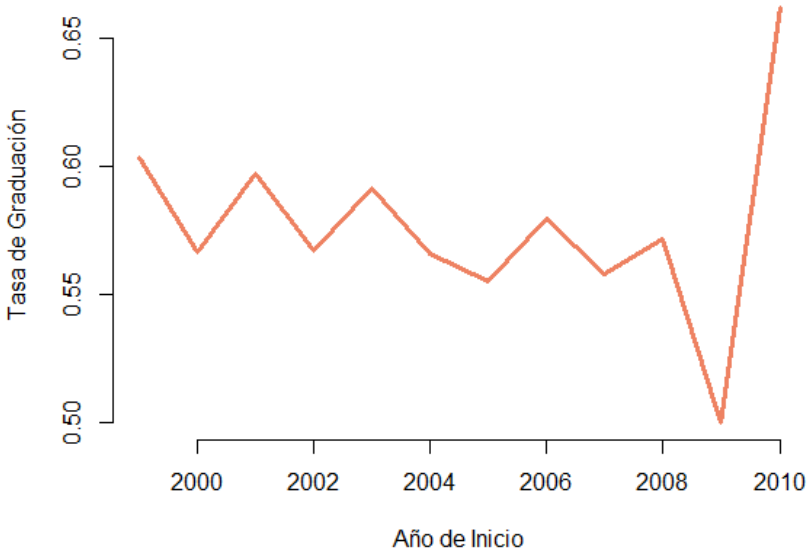

Figura 3.2: Tasa de Graduación por Año de Ingreso.

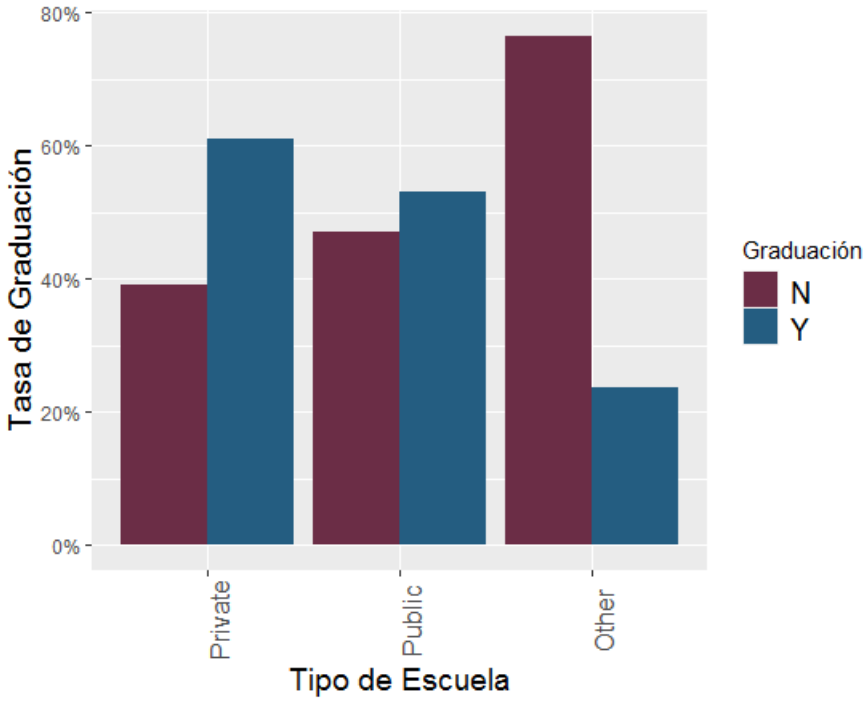

Figura 3.3: Tasa de Graduación por Tipo de Escuela.

El 56.21 % de los estudiantes provienen de escuelas públicas y el 43.7 % de escuelas privadas. En la figura 3.3 se detalla que los estudiantes de escuelas privadas tienen una

tasa de graduación más alta, cercana del 62 %, entre tanto, los de escuelas públicas y otras escuelas se sitúan un 8 % y 41 % por debajo de esa tasa de graduación, respectivamente, destacando a la categoría otras escuelas con la tasa más alta de no graduación cercana del 79 %. Esto indica que parece existir una asociación entre escuelas privadas y la graduación exitosa.

El ingreso familiar I5-O50 (Más de $50,000) en la figura 3.4 es el que tiene una tasa de graduación mas alta casi del 65 %, seguido por I4-B30A50 ($30,000 - $49,999), I3-B20A30 ($20,000 - $29,999) e I2-B12.5A20 ($12,500 - $19,999) con una tasa de graduación cercana del 61 %, 58 % y 52 %, respectivamente. Si se compara a I5-O50 con I1-L12.5 (Menos de $500 - $12,499), se obtiene que I1-L12.5 tiene una tasa de graduación un 14 % por debajo de I5-O50, además, en los ingresos familiares I1-L12.5 sobresale la tasa de no graduación. Note que en la en la figura 3.4 a medida que los ingresos familiares disminuyen las tasas de graduación disminuyen, indicando que parece existir una asociación entre los ingresos familiares altos y la graduación exitosa.

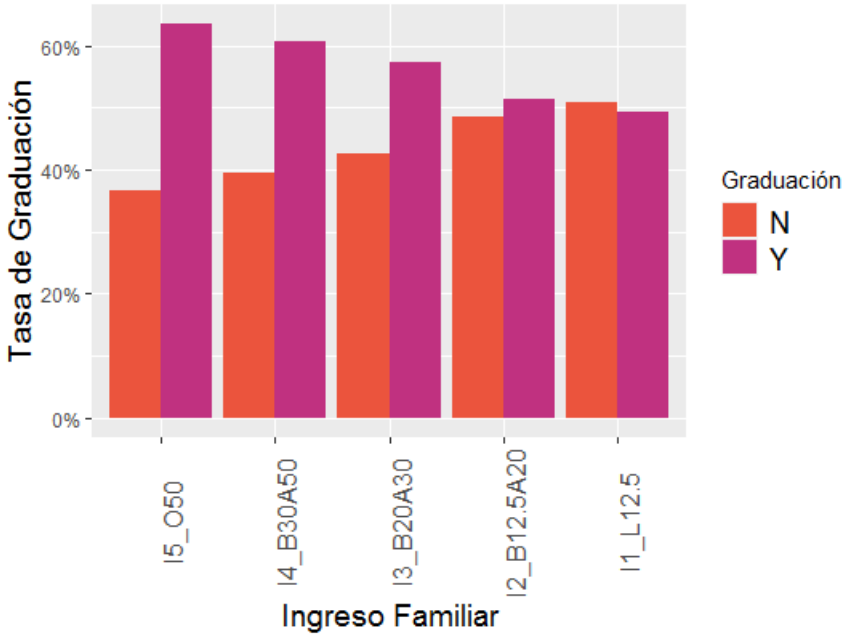

Figura 3.4: Tasa de Graduación por Ingreso Familiar.

En la figura 3.5 se puede observar que la categoría del nivel de educación de la madre Grad (Maestría o Doctorado) es la que posee una tasa de graduación más alta cercana del 62 % y muy próxima esta la categoría College (Bachillerato) con una tasa cercana del 61 %, mientras que, la tasa de graduación de la categoría Assoc or Less (Obtuvo grado asociado o menos al asistir a la universidad) esta por debajo un 9 % y un 8 % en comparación con Grad y Bachillerato, respectivamente. Las categorías HS (Completo Escuela Superior), LHS (No completó escuela superior) y None (Ninguna) tienen una tasa de graduación muy similar, cercanas del 50 %, 51 % y 50 %, respectivamente. Por otro lado, las tasas de graduación según el nivel de educación del padre tienen un comportamiento muy parecido a las tasas de graduación según el nivel de educación de la madre. Con la

figura 3.6 se puede notar que la categoría del nivel de educación del Padre Grad (Maestría o Doctorado) es la que posee una tasa de graduación más alta cercana del 63 % y muy próxima con 62 % esta la categoría College (Bachillerato), mientras que, Assoc or Less (Obtuvo grado asociado o menos al asistir a la universidad) esta por debajo un 11 % y un 10 % en comparación con Grad y College, respectivamente. Aparentemente existe un patrón en las tasas de graduación según el nivel de educación de los padres, debido a que a mayor nivel de educación de los padres mayor es la tasa de graduación.

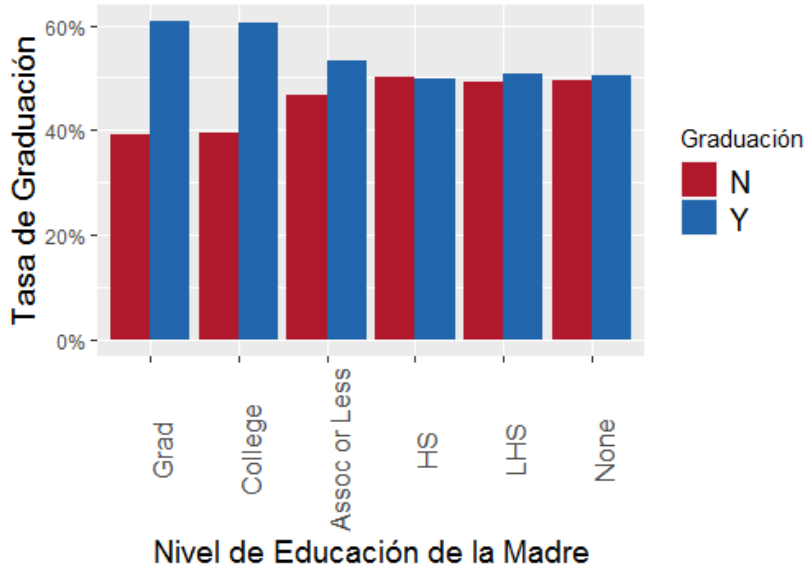

Figura 3.5: Tasa de Graduación por Nivel de Educación de la Madre.

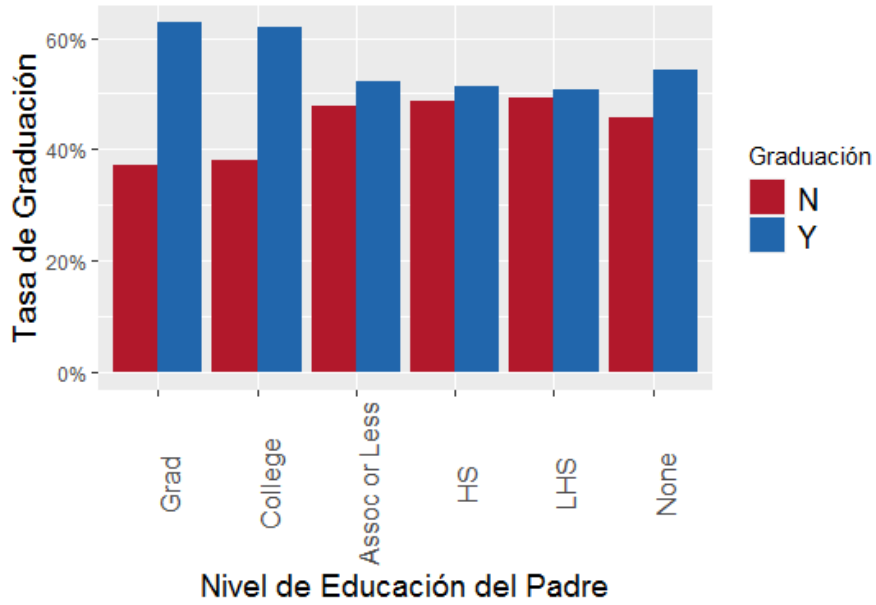

Figura 3.6: Tasa de Graduación por Nivel de Educación del Padre

Ahora bien, en el nivel de educación de los padres no se esperaría que la categoría None

(Ningún tipo de educación) tenga una tasa de graduación más alta o incluso igual que un nivel de educación HS (Escuela Superior) y Assoc or Less (Grado asociado), se esperaría que la tasa de graduación de None sea sustancialmente baja y más alta para bachillerato, maestría o doctorado, debido a que no hace sentido que padres con un nivel de educación más alto que None (ningún tipo de educación) posean tasas de graduación más bajas que padres sin ningún tipo de educación (None). Sin embargo, para la madre en la figura 3.5 None tiene la misma tasa de graduación que escuela superior (HS) pero la categoría "no completó escuela superior"(LHS) esta por encima que None y HS, cercana del 51 % para LHS y 50 % para None y HS. Para el padre en la figura 3.6 la categoría None (Ningún tipo de educación) tiene una tasa de graduación más alta que escuela superior (HS) o incluso que grado asociado (Assoc or Less), cerca del 54 % versus 51 % y 52 % respectivamente. Dicho lo anterior, es posible que los estudiantes al momento de suministrar la información para solicitar al RUM la categoría None para el nivel de estudios del padre o de la madre no la interpretaron bien, no la consideraron importante o no se acordaron del nivel de educación de los padres, y en consecuencia, colocaron como ningún tipo de educación (None), generando información poco certera con la realidad. Por tanto, se considera la categoría None como valores perdidos (de sus siglas en ingles NA) para el análisis de los valores faltantes y el rendimiento predictivo.

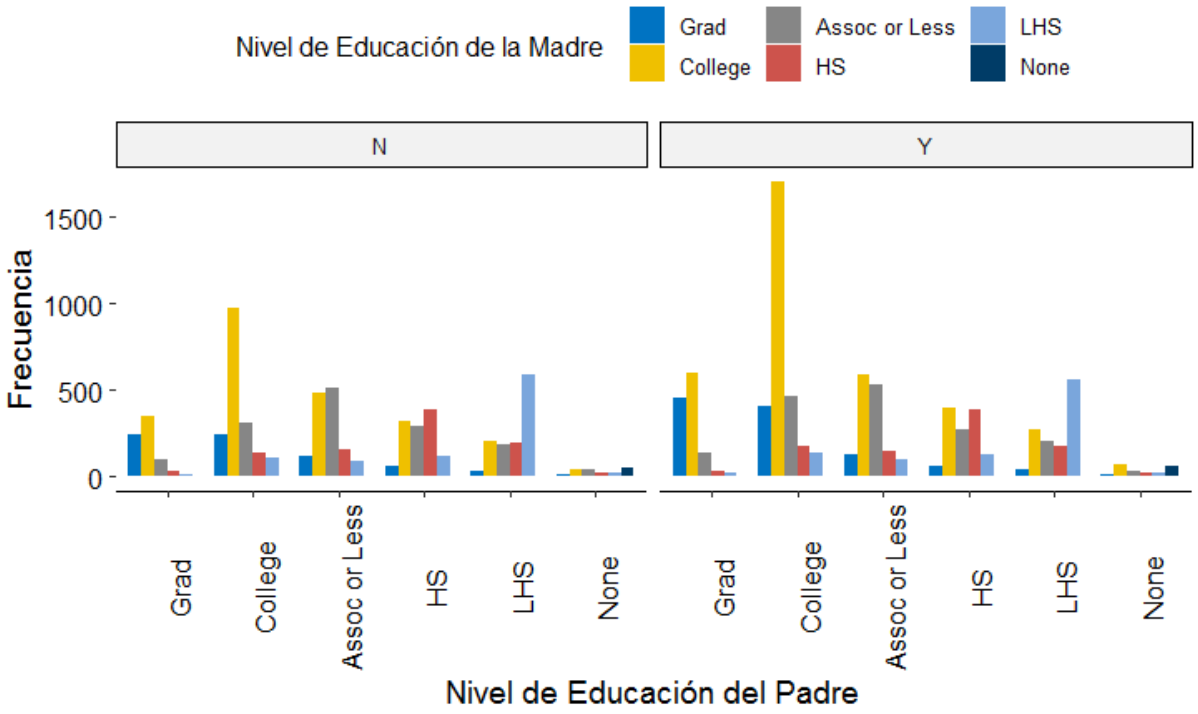

Figura 3.7: Efecto de la Educación del Padre con la Educación de la Madre.

Se puede determinar exploratorimente con la figura 3.7 que el efecto del nivel de educación del padre sobre la graduación exitosa de un estudiante depende del nivel de educación de la madre. Note que dentro del "si se gradúan"(Y) las distribuciones por cada nivel de educación del padre están cambiando (las alturas de los rectángulos cambian en cada nivel de educación), indicando que existe interacción con el nivel de educación

de la madre, en caso contrario, se esperaría distribuciones relativamente uniformes por cada nivel de educación. Si el nivel de educación del padre es escuela superior (HS) que un estudiante se gradúe depende del nivel de educación de la madre, observe que es más frecuente que un estudiante se gradúe cuando la madre tiene un bachillerato a un grado asociado o menos (Assoc or Less). Ahora si el padre tiene un nivel de educación grado asociado o menos (Assoc or Less) es más frecuente que un estudiante se gradúe cuando la madre tiene un bachillerato. Aunque el padre tenga un nivel de educación alto como maestría o doctorado el efecto sobre la graduación exitosa de un estudiante sigue dependiendo del nivel de educación de la madre. En contraste, dentro del "no se gradúan"(N) es más frecuente que los estudiantes no se gradúen cuando ambos padres tienen el mismo nivel de educación. En adición, la figura 3.7 también demuestra que los padres a menudo tienen nivel de educación similar, no necesariamente igual. Por ejemplo, si el padre posee un nivel de educación College (Bachillerato) es más frecuente que la madre también tenga un nivel de educación College (Bachillerato) o si el padre posee un nivel de educación Grad (Maestría o Doctorado) es más frecuente que la madre tenga un nivel de educación College (Bachillerato), señalando que es muy poco frecuente en los padres de estudiantes subgraduados tener un nivel de educación muy alto para el padre y un nivel de educación muy bajo para la madre o viceversa.

3.2.1. Análisis Valores Faltantes

En el análisis de los valores faltantes se usaron los datos de entrenamiento. Se encontró que la predictora con más valores faltantes era Educación Padre con 3934 observaciones que corresponde a 20.13 % de los estudiantes subgraduados. Seguido por Ingreso Familiar y Educación Madre con 3123 y 2596 correspondientes a 15.98 % y 13.28 % como se muestra en el cuadro 3.1 y 3.2. En total se tienen 9960 datos faltantes. Como al menos una predictora tiene entre el 5 % al 20 % de valores faltantes se requiere de métodos sofisticados para imputarlos. MICE es una buena alternativa (ver sección 2.2).

Gradacuación	Año	Facultad	Programa académico	Apt Verbal	Aprov Matemáticas
0	0	0	0	0	2
Apt Matemáticas	Aprov Español	Aprov Inglés	Ingreso Familiar	Educación Padre	Educación Madre
0	2	2	3123	3934	2596
Genero	Tipo de Escuela	GPA Primer año	Rel Estudiante GPA	Rel Escuela GPA	Total
0	88	213	0	0	9960

Cuadro 3.1: Conteo Valores Faltantes por Predictora.

Gradacuación	Año	Facultad	Programa académico	Apt Verbal	Aprov Matemáticas
0	0	0	0	0	0.01
Apt Matemáticas	Aprov Español	Aprov Inglés	Ingreso Familiar	Educación Padre	Educación Madre
0	0.01	0.01	15.98	20.13	13.28
Genero	Tipo de Escuela	GPA Primer año	Rel Estudiante GPA	Rel Escuela GPA	
0	0.45	1.09	0	0	

Cuadro 3.2: Porcentaje Valores Faltantes por Predictora.

La figura 3.8 representa todos los valores faltantes en los datos de entrenamiento. Con color rojo se exponen los valores que hacen falta y con color azul los que no. La

primera columna de la derecha representa el número de predictoras que faltan en el patrón. La primera columna izquierda representa el número de veces que ocurre el patrón de datos disponibles y datos faltantes. Por ejemplo, la primera línea indica que para 14,265 estudiantes todas variables están disponibles (representando el 73 % de la base de datos de entrenamiento) mientras que la segunda línea indica que para 1,412 estudiantes el nivel de educación del padre no está disponible.

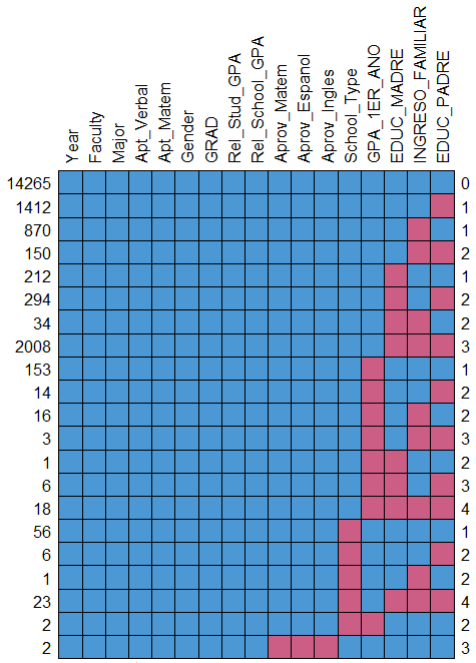

Figura 3.8: Valores Faltantes en los datos de entrenamiento.

Las predictoras Educación Madre, Ingreso Familiar y Educación Padre faltan simultáneamente en 2,008 observaciones correspondiente al 10.27 %, convirtiéndose en el número de observaciones más alto que hacen falta. Una posible explicación es que los padres de esos 2,008 estudiantes son muy cuidadosos con la información de sus ingresos familiares y por consiguiente no brindaron su nivel de educación, dado que, los ingresos se pueden aproximar a través de su nivel de educación. Con esto, se sabe que hay posibilidades de que los padres no brinden información de sus ingresos ni de su nivel académico, entonces, para no perder estos datos se pueden generar canales de comunicación que permitan concientizar a los padres del tratamiento que se le da a la información que ellos suministren a la universidad. En consecuencia, disminuir los valores perdidos.

3.3. Rendimiento Predictivo

Antes de analizar el rendimiento predictivo de los métodos de aprendizaje automático examinemos el funcionamiento del árbol de clasificación para entender como funcionarían dichos métodos. En la figura 3.9 se tiene un ejemplo adaptado de un árbol de clasificación. Ese árbol de clasificación no es el mejor modelo pero logra diferenciar las tasas de graduación usando la predictora GPA de primer año como la que clasifica si el estudiante

se gradúa o no con menos impureza. Si un estudiante tiene un GPA de primer año ≥ 3.01 ese estudiante tiene una probabilidad del 80 % de graduarse al 150 % de la duración de su programa académico. Por otra parte, si tiene un GPA de primer año > 1.79 y < 2.085 tiene una probabilidad del 31 % de graduarse al 150 % de la duración de su programa académico. Ahora si otro estudiante tiene un GPA de primer año entre $[2.085, 2.405)$, un año de admisión menor al 2003 y un puntaje de aprovechamiento matemático ≥ 649, entonces dicho estudiante tiene una probabilidad del 65 % de graduarse al 150 % de la duración de su programa académico.

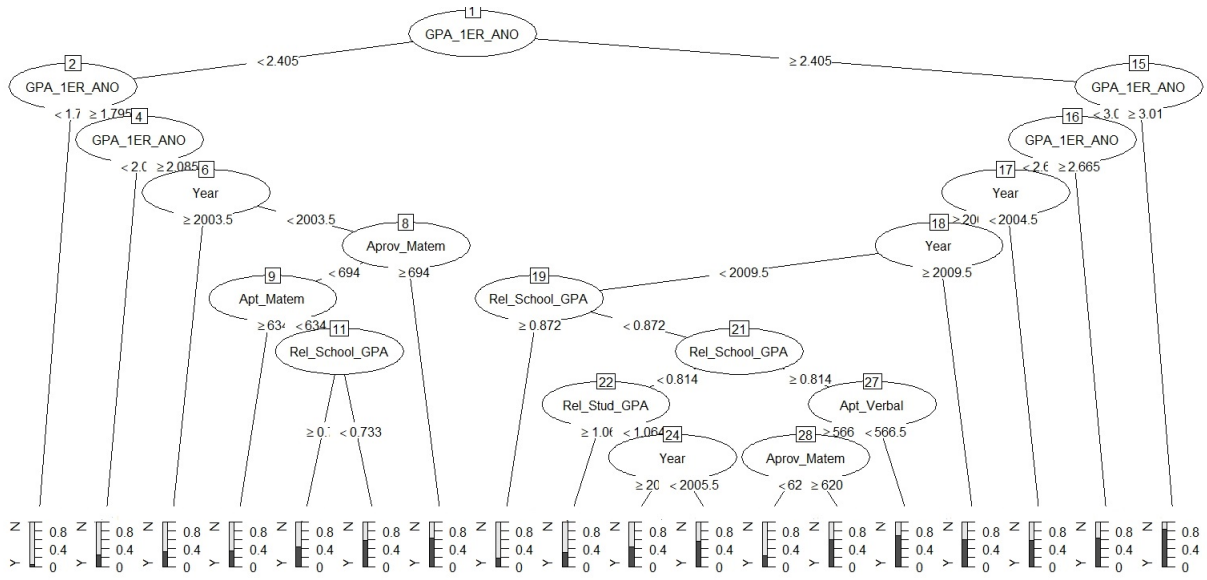

Figura 3.9: Ejemplo Árbol de clasificación.

En el rendimiento predictivo se tuvo en cuenta dos versiones de cada modelo. La primera versión (primer modelo) "un modelo para antes de entrar a la universidad" es para una detección temprana de los estudiantes en riesgo de graduarse y la segunda versión (segundo modelo) "un modelo luego del primer año en el RUM" permite después de un año verificar si los estudiantes siguen en riesgo. En la primera versión del modelo no se consideran las predictoras Rel Escuela GPA y GPA primer año, mientras que, en la segunda versión del modelo si se consideran. Para ambas versiones se hacen dos análisis, uno al eliminar los valores faltantes y el otro al imputar esos valores faltantes. Los modelos de aprendizaje automático a aplicar son Árboles de Clasificación, Bosque Aleatorio, Stochastic Gradient Boosting, Naïve Bayes, Regresión Logística y TabNet, los cuales son descritos en la sección 2.3.

Métrica	Árbol de Clasificación	Bosque Aleatorio	Stochastic Gradient Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.4896	0.4469	**0.7260**	0.5217	0.4500	**0.7017**
AUC	0.5953	0.6672	0.6868	0.6679	0.7023	0.6707
P(Modelo predice si se gradúa | Realmente no se gradúa)	0.5104	0.5531	**0.2740**	0.4783	0.5500	**0.2983**
P(Modelo predice si se gradúa pero realmente no se gradúa)	0.2205	0.2389	**0.1184**	0.2066	0.2376	**0.1265**

Cuadro 3.3: Modelos antes de entrar a la universidad.

Métrica	Árbol de Clasificación	Bosque Aleatorio	Stochastic Gradient Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.6166	0.6223	**0.7756**	0.6298	0.6675	**0.8046**
AUC	0.8155	0.8286	0.8373	0.7831	0.8392	0.8168
P(Modelo predice si se gradúa | Realmente no se gradúa)	0.3834	0.3777	**0.2244**	0.3702	0.3325	**0.1954**
P(Modelo predice si se gradúa pero realmente no se gradúa)	0.1656	0.1632	**0.0969**	0.1599	0.1436	**0.0829**

Cuadro 3.4: Modelos luego del primer año en el RUM.

Al eliminar los valores faltantes el método que mejor trabaja para el primer modelo es Stochastic Gradient Boosting en comparación con los demás métodos como se muestra en el cuadro 3.3. Tabnet es bastante competitivo pero Stochastic Gradient Boosting obtiene una sensibilidad sobresaliente, dado que, identifica correctamente el 72.6 % de los estudiantes que no se graduaron, además, son identificados correctamente el 68.68 % de los estudiantes que no se graduaron y que si se graduaron. Para el segundo modelo Stochastic Gradient Boosting y Tabnet trabajan mejor en comparación con los demás métodos como se muestra en el cuadro 3.4. Sin embargo, Tabnet sobresale, identifica correctamente el 80.46 % de los estudiantes que no se graduaron y el 81.68 % de los estudiantes que no se graduaron y que si se graduaron.

Métrica	Árbol de Clasificación	Bosque Aleatorio	Stochastic Gradient Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.5135	0.4944	**0.7591**	0.5168	0.4516	0.341
AUC	0.6154	0.6639	0.6936	0.6558	0.6896	0.5304
P(Modelo predice si se gradúa | Realmente no se gradúa)	0.4865	0.5056	**0.2409**	0.4832	0.5484	0.6589
P(Modelo predice si se gradúa pero realmente no se gradúa)	0.1996	0.2075	**0.0988**	0.1983	0.2250	0.2860

Cuadro 3.5: Modelos antes de entrar a la universidad - datos imputados.

Métrica	Árbol de Clasificación	Bosque Aleatorio	Stochastic Gradient Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.6485	0.6372	**0.7650**	0.6149	0.6702	**0.7462**
AUC	0.7611	0.8218	0.8307	0.7705	0.8312	0.8195
P(Modelo predice si se gradúa | Realmente no se gradúa)	0.3515	0.3621	**0.235**	0.3851	0.3298	**0.2537**
P(Modelo predice si se gradúa pero realmente no se gradúa)	0.1442	0.1486	**0.0964**	0.1580	0.1353	**0.1099**

Cuadro 3.6: Modelos luego del primer año en el RUM- datos imputados.

Al imputar los valores faltantes el método que mejor trabaja para el primer modelo es Stochastic Gradient Boosting en comparación con los demás métodos como se muestra en el cuadro 3.5. Stochastic Gradient Boosting obtiene una sensibilidad sobresaliente, identifica correctamente el 75.91 % de los estudiantes que no se graduaron. Además, son identificados correctamente el 69.36 % de los estudiantes que no se graduaron y que si se graduaron. Para el segundo modelo Stochastic Gradient Boosting y Tabnet trabajan mejor en comparación con los demás métodos como se muestra en el cuadro 3.6. Sin embargo, Stochastic Gradient Boosting sobresale. Identifica correctamente el 76.50 % de los estudiantes que no se graduaron y el 83.07 % de los estudiantes que no se graduaron y que si se graduaron.

Después de la imputación de los datos se puede notar que para el primer modelo (un modelo para antes de entrar a la universidad) Stochastic Gradient Boosting fue el único método que obtuvo una leve mejora de 3 centésimas porcentuales en la sensibilidad, pasando de 72.6 % a 75.91 %. Esta mejora hace que Stochastic Gradient Boosting sobresalga entre los métodos para el primer modelo en la comparativa con datos imputados y con datos no imputados. Stochastic Gradient Boosting con datos imputados identifica incorrectamente el 24.09 % de los estudiantes que no se graduaron. Dicho de otro modo, la

probabilidad de que Stochastic Gradient Boosting predice que el estudiante si se gradúa dado que realmente el estudiante no se gradúo es del 24.09 % y de todas las predicciones hechas la probabilidad de que Stochastic Gradient Boosting predice incorrectamente que el estudiante si se gradúa es del 9.88 %. Para el segundo modelo (un modelo luego del primer año) no fue relevante la imputación de los datos faltantes, dado que, no hubo una mejora significativa en ningún método. Sin embargo, Tabnet sin datos imputados sobresale entre los métodos para el segundo modelo en la comparativa con datos imputados y con datos no imputados, porque tiene una sensibilidad del 80.46 %. Tabnet sin datos imputados identifica incorrectamente el 19.54 % de los estudiantes que no se graduaron. Dicho de otro modo, la probabilidad de que Tabnet predice que el estudiante si se gradúa dado que realmente el estudiante no se gradúo es del 19.54 % y de todas las predicciones hechas la probabilidad de que Tabnet predice incorrectamente que el estudiante si se gradúa es del 8.29 %.

De esta manera, se selecciona Stochastic Gradient Boosting como el modelo antes de entrar a la universidad (con datos imputados) y Tabnet para el modelo luego del primer año en el RUM (sin datos imputados), puesto que, son los métodos que minimizan la probabilidad de predecir que el estudiante si se gradúa dado que realmente el estudiante no se gradúo.

3.3.1. Métodos Finalistas

De los métodos seleccionados finalmente se indican los hiperparámetros, el cálculo de las métricas de evaluación y las predictoras importantes. Se debe tener cuidado con el análisis de las predictoras importantes porque esta importancia depende del método que se usa, puesto que, las predictoras importantes son determinadas por la cantidad de veces que un método usa esa predictora a medida que se entrena el modelo.

Stochastic Gradient Boosting - Modelo Antes de Entrar a la Universidad

Hiperparámetros

Se usa un espacio de búsqueda para cada hiperparámetro. El número de árboles (N trees) se elige de $\{2000, 2500, 3000\}$. La profundidad máxima de cada árbol (Interaction depth) de $\{3, 5, 7\}$. La tasa de aprendizaje (Shrinkage) de $\{0.01, 0.1, 0.3\}$. El número mínimo de observaciones en los nodos terminales del árbol (N min obs in node) de $\{5, 10, 15\}$ y la fracción de observaciones seleccionadas al azar (Bag fraction) para proponer el siguiente árbol de $\{0.65, 0.8, 1\}$. Finalmente Stochastic Gradient Boosting elige N trees= 99, Interaction depth= 3, Shrinkage= 0.10, N min obs in node= 15 y Bag fraction= 1.

Métricas de Evaluación

Usando la matriz de confusión (cuadro 3.7) se calculan las métricas de evaluación

- Sensibilidad= $\frac{1153}{1153+366} = 0.759$
- P(Modelo predice si se gradúa | Realmente no se gradúa)= $\frac{366}{1153+366} = 0.2409$

- P(Modelo predice si se gradúa pero realmente no se gradúa)= $\frac{366}{1153+366+1099+1083}$ = 0.0988

	Actual N	Actual Y
Predicted N	1153	1099
Predicted Y	366	1083

Cuadro 3.7: Matriz de confusión Stochastic Gradient Boosting.

Predictoras Relevantes

La figura 3.10 permite visualizar que predictoras son importantes para Stochastic Gradient Boosting. La predictora facultad no fue importante para Stochastic Gradient Boosting, ya que esa predictora casi no fue usada. La predictora Programa académico (Major) es la mas importante, dado que, el 59.12 % de las veces Stochastic Gradient Boosting utiliza Major. La segunda predictora más importante es Aprovechamiento Matemáticas con un 15.47 % de veces utilizada y la tercera más importante es Genero con un 11 % de veces utilizada.

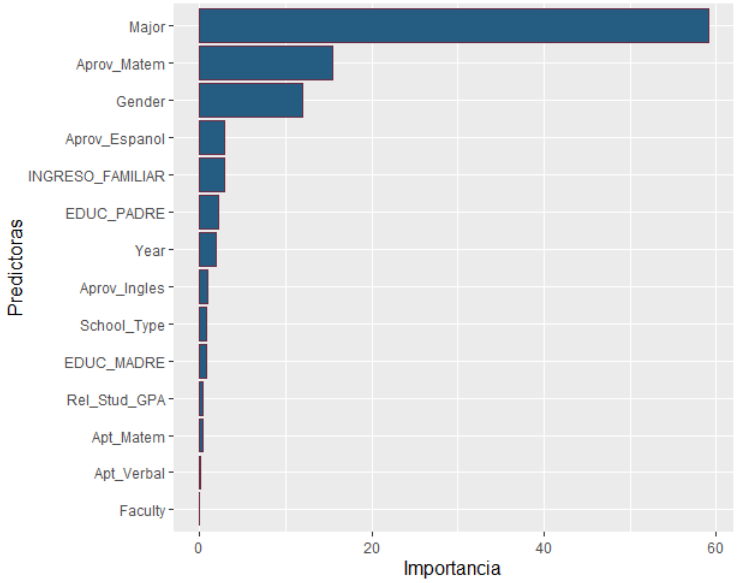

Figura 3.10: Predictoras Importantes para Stochastic Gradient Boosting.

TabNet - Modelo Luego del Primer Año en el RUM

Hiperparámetros

Se usa un espacio de búsqueda para cada hiperparámetro. El número máximo de épocas de entrenamiento (Max epochs) se elige de $\{20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30\}$. El tamaño de lote a normalizar (Batch size) de $\{1024, 1025, 1026, 1027, 1028, 1029, 1030\}$. El tamaño de lote para la normalización de lotes fantasma (Virtual batch size) de $\{128, 129, 130, 131, 132, 133, 134\}$ y el número de épocas consecutivas sin mejoría antes de realizar una

parada anticipada (Patience) de $\{20, 21, 22, 23, 24, 25, 26\}$. Finalmente TabNet elige Max epochs= 22, Batch size= 1029, Virtual batch size=131 y Patience= 22.

Métricas de Evaluación

Usando la matriz de confusión (cuadro 3.8) se calculan las métricas de evaluación

- Sensibilidad= $\frac{626}{626+152} = 0.8046$
- P(Modelo predice si se gradúa | Realmente no se gradúa)= $\frac{152}{626+152} = 0.1953$
- P(Modelo predice si se gradúa pero realmente no se gradúa)= $\frac{152}{626+364+152+691} = 0.0829$

	Actual N	Actual Y
Predicted N	626	364
Predicted Y	152	691

Cuadro 3.8: Matriz de confusión TabNet.

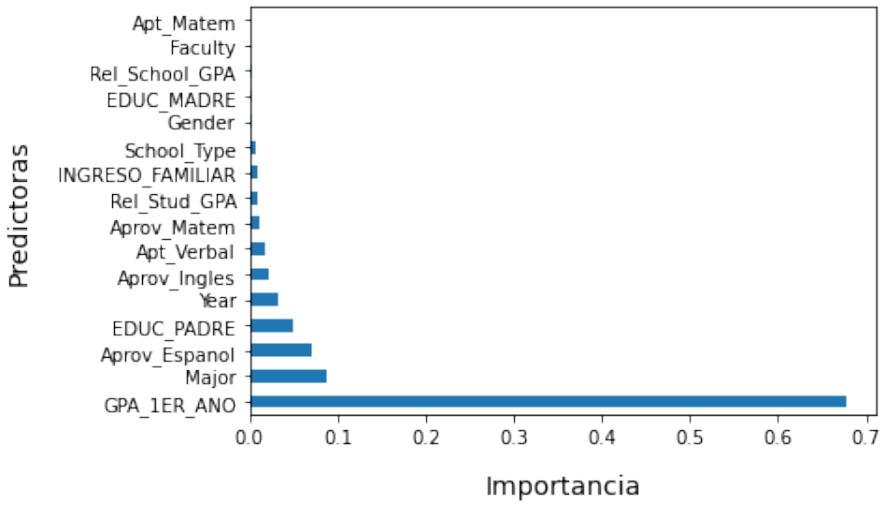

Figura 3.11: Predictoras Importantes para Tabnet.

Predictoras Relevantes

La figura 3.11 permite visualizar que predictoras son importantes para Tabnet. La predictora Aptitud Matemáticas es una predictora irrelevante porque es elegida cerca del 0.0004 % de las veces. Sin embargo, la predictora GPA Primer año es la más importante, dado que, el 67.69 % de las veces TabNet la utiliza. Seguida por Programa académico (Major) con un 8.79 % y Aprovechamiento Español con un 7 %.

Capítulo 4

Conclusión

Puerto Rico presenta una tasa de graduación universitaria alarmantemente baja, en promedio, cerca del 45 % de los estudiantes subgraduados obtienen su título al 150 % de la duración de su programa académico [43]. La baja tasa de graduación se hace aún más dramática si se mira en conjunto con la tasa de deserción escolar. Como se mencionó en el capitulo 1, el 40 % de los estudiantes no terminan la escuela superior y el 60 % restante que la terminan no todos van a la universidad. Pero supongamos que si de ese 60 % que ingresan a la universidad después de la escuela superior solo el 45 % se gradúa al 150 % de la duración de su programa académico. Aunque algunas universidades ofrecen talleres y consejería a estudiantes para facilitar que se gradúen, no existe una manera objetiva de predecir si una estudiante en particular completará su grado. Las bajas tasas de graduación universitaria son un problema que afecta a los estudiantes al tener menor probabilidad de acceder a planes de jubilación [16], tres veces más probable de vivir en pobreza [15], limita el acceso a la educación post-secundaria, afecta los cupos académicos y en consecuencia, afecta la universidad al perder ingresos por matrícula y cuotas [1], además, impacta el presupuesto federal y estatal, estudiantes no graduados pagan menos impuestos. Esta problemática exige a las instituciones de educación post-secundaria, desarrollar estrategias y modelos predictivos que permitan aumentar la cifra de estudiantes que completan sus grados exitosamente.

Por tales motivos, este proyecto implementó modelos predictivos utilizando métodos de aprendizaje automático con el objetivo de predecir si un estudiante subgraduado se gradúa al 150 % de la duración de su programa académico de la Universidad de Puerto Rico Recinto Mayagüez. Esos modelos, descritos en la sección 2.3 se aplicaron a los datos suministrados por OPIMI (la Oficina de planificación, investigación y mejoramiento institucional) que representan 24,432 estudiantes subgraduados admitidos al Recinto Universitario de Mayaguez (RUM) entre el 1999 hasta el 2010; el cual incluía las variables, Año de admisión, Facultad de admisión, Programa de admisión, Aptitud Verbal, Aprovechamiento Matemáticas, Aptitud Matemáticas, Aprovechamiento Español, Aprovechamiento Inglés, Ingreso Familiar, Educación Padre, Educación Madre, Genero, Tipo de Escuela, GPA Primer año, Rel Estudiante GPA, Rel Escuela GPA y Graduación.

Los métodos de aprendizaje automático implementados fueron 6, Árboles de Clasificación, Bosque Aleatorio, Stochastic Gradient Boosting, Naïve Bayes, Regresión Logística y

Tabnet, con estos, se hicieron dos análisis considerando dos versiones de cada modelo, un modelo para antes de entrar a la universidad y un modelo luego del primer año en el RUM. El primer modelo permite predecir si los estudiantes completarán su grado en el momento en el que fueron admitidos para una identificación temprana y el segundo modelo permite predecir luego del primer año si esos mismos estudiantes tienen probabilidades altas de no completar su grado. Un análisis se hizo al eliminar los valores faltantes y el otro al imputar esos valores faltantes usando MICE (imputación multivariante por ecuaciones encadenadas). La imputación de los datos mejoró las predicciones solo para Stochastic Gradient Boosting en el primer caso (un modelo para antes de entrar a la universidad) y en el segundo caso (un modelo luego del primer año en el RUM) no hubo una mejora significativa en ningún método de aprendizaje automático.

Cuando un método de aprendizaje automático predice incorrectamente que el estudiante si se gradúa hace que el método establezca que el estudiante si se graduó cuando realmente no lo hizo, impidiendo que al estudiante se le provea servicio necesario y así aumente las posibilidades de que el estudiante complete su grado. Por consiguiente, evaluar la probabilidad de que un método predice incorrectamente que el estudiante si se gradúa permite buscar el método que minimice el error de establecer que un estudiante es capaz de graduarse, pero éste no se gradúa. Al comparar los métodos de aprendizaje automático, Árboles de Clasificación, Bosque Aleatorio, Stochastic Gradient Boosting, Naïve Bayes, Regresión Logística y Tabnet entre los dos casos (un modelo para antes de entrar a la universidad y un modelo luego del primer año) se obtuvo que Stochastic Gradient Boosting tiene un mejor desempeño para el primer caso con datos imputados (ver sección 3.3, cuadro 3.5) porque identifica correctamente el 75.91 % de los estudiantes que no se graduaron. Además, la probabilidad de que Stochastic Gradient Boosting predice incorrectamente que el estudiante si se gradúa es del 9.88 %. Tabnet tiene un mejor desempeño para el segundo caso pero sin datos imputados (ver sección 3.3, cuadro 3.4) porque identifica correctamente el 80.46 % de los estudiantes que no se graduaron. Además, la probabilidad de que Tabnet predice incorrectamente que el estudiante si se gradúa es del 8.29 %.

En general, Stochastic Gradient Boosting y TabNet tienen un buen desempeño aunque TabNet tiene un tiempo de cómputo más costoso. El desempeño de los modelos no fue perfecto pero son útiles al predecir si un estudiante subgraduado se gradúa al 150 % de la duración de su programa académico de la Universidad de Puerto Rico Recinto Mayagüez, porque minimizan la probabilidad de predecir que el estudiante si se gradúa dado que realmente el estudiante no se graduó.

Adicionalmente, el análisis de los valores faltantes y el análisis de predictoras importantes para Stochastic Gradient Boosting y Tabnet sugiere a los funcionarios de admisiones y/o oficiales encargados de usar esta herramienta que deberían resaltar a los solicitantes la importancia de que el RUM cuente con los datos para Aprovechamiento Matemáticas, Aprovechamiento Español y GPA Primer año, puesto que, los análisis indican que son predictoras que tienen valores faltantes y son importantes para obtener un buen desempeño predictivo. Otras predictoras que se podrían tener en cuenta para obtener un buen desempeño predictivo son, beneficiario de beca federal (si o no), aprobó o no matemáti-

ca básica (precalculo I y II), aprobó o no cálculo diferencial y aprobó o no matemática financiera, debido a que, son importantes para estimar si un estudiante completará o no sus estudios universitarios [40] [35].

Cabe resaltar que este proyecto propone predecir la posibilidad de graduación para estudiantes ya admitidos al RUM, no es una herramienta para decidir si se admite un estudiante o no. No sería ético si se usa para admisiones, pues seria un acto de discriminación, como se vio en la sección 3.2 el ingreso familiar y el nivel de educación de los padres afecta las posibilidades de que un estudiante se gradúe, entonces, de usarse para admisiones se estaría sesgando a que personas con mejores beneficios sean admitidas.

4.1. Alcance y Limitaciones del Proyecto

Este proyecto brinda una herramienta que permitirá detectar estudiantes con bajas probabilidades de graduación, que se logre dependerá de la implementación efectiva de un programa de intervención utilizando esta herramienta. Ese programa podría contar con asistencia en manejo de tiempo y estrés, consejería académica, y orientación de programas en el RUM que ayuden al estudiante. Dicha detección e intervención disminuiría el impacto de las bajas tasas de graduación, y en consecuencia, brindaría mayores oportunidades al estudiante, permitiría aumentar los ingresos de la universidad [1] y medir la eficacia de la institución ante los organismos de acreditación y el gobierno [4]. Este proyecto no pretende diseñar el programa de intervención, sin embargo, se brinda un ejemplo de un programa de intervención utilizando la herramienta que brinda este estudio en la sección 4.2.

Este proyecto tiene limitaciones importantes

- El estudio no considera si un estudiante se traslada de un programa a otro y luego de ese traslado en ese nuevo programa se gradúe al 150 % de la duración de ese nuevo programa.
- Los métodos de aprendizaje automático requieren mantenimiento una vez las tasas de graduación se ajusten por la intervención. Esto se describe con más detalle en la sección 4.2.
- No tiene en cuenta si la modalidad (virtual o presencial) del estudiantes afecta o no las tasas de graduación.
- No tiene en cuenta la posibilidad de que otras variables permitan predecir si un estudiante se va a graduar. Por ejemplo, distancia/tiempo de viaje a la universidad por el estudiante, carga académica, asistencia a centros de tutoría y bibliotecas, beneficiario de beca federal (si o no), aprobó o no matemática básica (precalculo I y II), aprobó o no cálculo diferencial y aprobó o no matemática financiera.

4.2. Trabajo Futuro

El programa de intervención debe ser desarrollado por los oficiales del RUM y debe actuar según la predicción del modelo. Dicho programa de intervención debe ser bien

diseñado e implementado para lograr en conjunto con la herramienta que se brinda en este proyecto una detección efectiva de estudiantes con bajas probabilidades de graduación. Para evaluar la efectividad de la intervención se sugiere comparar la tasa de retención antes y después de la intervención y comparar el promedio académico del estudiante antes y después de la intervención. Es importante resaltar que se debe tratar de convencer a la administración de la UPR para que adopten la herramienta que se brinda en este proyecto de investigación, y en consecuencia, se puede adoptar para todos los recintos de la UPR, esta herramienta no es estrictamente para el RUM.

El desempeño de los métodos de aprendizaje automático cambiará con el tiempo, asumiendo que un programa de intervención aumente las posibilidades de que estudiantes se gradúen, las personas empezaran a graduarse, los datos cambiaran y los métodos ya no funcionarán. Supongamos que estudiantes con un GPA de primer año insuficiente son identificados con una probabilidad del 10 % de completar su grado, entonces, la intervención de los oficiales del RUM hará aumentar las probabilidades de que el estudiante se gradúe y con el tiempo de ajustarse las tasas de graduación por la intervención, se necesitará implementar nuevos modelos, y en consecuencia, la dinámica de los métodos debe cambiar. Por ejemplo, monitorear en tiempo real los cursos que toman los estudiantes [31]. Desarrollar estrategias similares a el University College de la Universidad de Maryland donde se usen los sistemas de gestión de aprendizaje como Moodle para medir el número de clic en materiales de clase, el tiempo activo en línea, la participación en discusiones [6], la frecuencia con la que los estudiantes usan los recursos del campus, como los centros de tutoría y las bibliotecas [31]. Tener a disposición estudiantes de semestres avanzados (de séptimo u octavo semestre para programas de 4 años y noveno o décimo semestre para programas de 5 años) para que recomienden un plan de estudio óptimo y conecten a los estudiantes con los servicios universitarios disponibles [27]. Con lo anterior, se proporcionaría información esencial que aún no se ha rastreado y así hacerle frente a las nuevas dinámicas que se puedan presentar a lo largo del tiempo.

Es importante resaltar que como los métodos de aprendizaje automático tienen que ser eficientes la intervención de los oficiales también lo debe ser. En consecuencia, se sugiere a los oficiales usar Stochastic Gradient Boosting para predecir si los estudiantes completarán su grado en el momento en el que fueron admitidos para una identificación temprana y luego del primer año aplicar Tabnet para predecir si esos mismos estudiantes tienen probabilidades altas de no completar su grado. Con esta información y su GPA de primer año se pueden programar reuniones con el estudiante y sus padres para brindar una evaluación precisa del desempeño actual del estudiante [27] y recomendaciones basadas en las estrategias que se sugieren en la sección 4.1 y en el segundo párrafo de la sección 4.2.

Describamos un ejemplo hipotético que refleje la implementación de la herramienta que brinda este proyecto con un programa de intervención. Hay una persona que fue admitido al programa de matemáticas puras con las siguientes características, Año de admisión=2024, Facultad=artes y ciencias, Programa académico= Matemáticas, Apt Verbal=245, Aprov Matemáticas=286, Apt Matemáticas=301, Aprov Español=250, Aprov Inglés=223, Ingreso Familiar=40,000, Educación Padre=Bachillerato, Educación Madre=

Bachillerato, Genero=Masculino, Tipo de Escuela=Privada, Rel Estudiante GPA=1.16. Los oficiales detectaron al aplicar Stochastic Gradient Boosting que el estudiante tiene una probabilidad de graduación del 21 %. Se envía un correo electrónico al estudiante para organizar una reunión con él y sus padres de manera compulsoria. El día de la reunión los oficiales le dejan saber al estudiante y a sus padres que sus sistemas predictivos detectaron de manera temprana una alta probabilidad de no completar su grado al 150 % del programa de matemáticas, se hacen recomendaciones basadas en las estrategias que se sugieren en la sección 4.1 y en el primer párrafo de la sección 4.2. Por ejemplo, asistir a los centros de tutoría para precalculo uno y dos, si llega a presentar estrés o siente que necesita una guía de como organizar su tiempo entonces tome una cita por psicología. Se les informa que se le dará seguimiento al estudiante y para ello el siguiente año académico debe haber otra reunión. Luego del primer año los oficiales aplican Tabnet con dos características más GPA Primer año=2.5 y Rel Escuela GPA=0.31. La probabilidad de graduación disminuye un 10 % pasando al 11 %. El día de la segunda reunión los oficiales le dejan saber al estudiante y a sus padres que la probabilidad de graduación disminuyó un 10 %. Se hacen nuevas recomendaciones, por ejemplo, reúnase con el estudiante "tal" de octavo semestre de su programa para que le recomiende un plan de estudio óptimo y lo conecte con los servicios universitarios disponibles [27], y si la conversación va bien se podría conducir al estudiante a inscribirse en otro programa [31]. A inicios del segundo año se evalúa el éxito de la primera intervención, se compara el promedio académico del estudiante de sus dos primeros semestres y se compara la tasa de retención para estudiantes que no se benefician de la intervención y para estudiantes que si se benefician. El éxito de la segunda intervención se evalúa a inicios del tercer año comparando el promedio académico del estudiante de sus dos años académicos consecutivos y se compara la tasa de retención para estudiantes que no se benefician de la intervención y para estudiantes que si se benefician.

En futuros estudios se podría implementar XGBoost un método de ensamble de aprendizaje automático que ha superado a otros métodos en diferentes competiciones de Kaggle para verificar su desempeño en la predicción de G150 de la Universidad de Puerto Rico Recinto Mayagüez. Se debería estudiar si la modalidad (virtual o presencial) del estudiantes afecta o no las tasas de graduación y la posibilidad de que un estudiante se traslade de un programa a otro y luego de ese traslado en ese nuevo programa se gradúe al 150 % de la duración de su programa académico. Además, los modelos desarrollados en éste proyecto se pueden entrenar añadiendo el recinto de la UPR al que pertenece el estudiante, es muy probable que el recinto sea un factor importante para predecir la tasa de graduación.

Bibliografía

[1] Barshay, J. y S. Aslanian. 2019. Colleges are using big data to track students in an effort to boost graduation rates, but it comes at a cost. *Enlace del Hechinger Report*.

[2] Oficina de Planeación, Investigación y Mejoramiento Instituciona (OPIMI) [En Línea]. Disponible *aquí*.

[3] Rodriguez, A. 2010. Retención de Estudiantes [En Línea]. Disponible: http://ponce.inter.edu/wp-content/uploads/documentos/retencion%20estudiantil/retencion%20%20estds%202010.pdf

[4] Karamouzis, S.T. y A. Vrettos. 2008. An artificial neural network for predicting student graduation outcomes. in Proceedings of the World Congress on Engineering and Computer Science, in San Francisco, USA.

[5] Bassi, J. S., E. G. Dada, A. A. Hamidu, y M. D. Elijah. 2019. Students Graduation on Time Prediction Model Using Artificial Neural Network. *IOSR-JCE*, Vol. 21, Issue 3. PP 28-35.

[6] Ploutz, E. C. 2018. Machine Learning Applications in Graduation Prediction at the University of Nevada, Las Vegas. Tesis B.S., Universidad de Nevada, Las Vegas, N.V., USA. http://dx.doi.org/10.34917/13568668

[7] Arık, S. O. y T. Pfister. 2020. TabNet: Attentive Interpretable Tabular Learning, *arXiv, [cs.LG]*, V. 4. https://arxiv.org/pdf/1908.07442.pdf

[8] Income for Recent Graduates the Highest in Over a Decade. 2016. *The Wall Street Journal*.

[9] Tasa de Graduación Universidades de Estados Unidos. 2020. *IPEDS*. https://cutt.ly/mzFhwTf

[10] Rolke, W. 2014. Identifying Students at Risk.

[11] U.S. Department of Education, National Center for Education Statistics. (2020). *The Condition of Education 2020* (NCES 2020-144). https://nces.ed.gov/ipeds/TrendGenerator/app/answer/7/20.

[12] U.S. Bureau of Labor Statistics. 2019. https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

[13] Tibshirani R., T. Hastie, D. Witten, y G. James. An Introduction to Statistical Learning with applications in R, First Edition, Springer New York Heidelberg Dordrecht London, 441 pp.

[14] Aprendizaje basado en árboles de decisión. https://es.wikipedia.org/wiki/Aprendizaje_basado_en_%C3%A1rboles_de_decisi%C3%B3n

[15] El costo creciente de no ir a la universidad. Pew Research Center. 2014. https://www.pewresearch.org/social-trends/2014/02/11/the-rising-cost-of-not-going-to-college/

[16] ¿Vale la pena una educación universitaria?. ProCon.org. 2020. https://college-education.procon.org/

[17] Miller J., R. Forte. Mastering Predictive Analytics with R, Second Edition, Packt Publishing Ltd, Livery Place, 35 Livery Street, Birmingham B3 2PB, UK, 448 pp.

[18] Breiman, L. 2001. Random Forests, *Kluwer Academic Publishers*, Machine Learning, 45:5–32.

[19] Fortmann-Roe, S. Understanding the Bias-Variance Tradeoff. 2012. https://scott.fortmann-roe.com/docs/BiasVariance.html

[20] Tibshirani R., T. Hastie, y F. Jerome. The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Second Edition, Springer New York, 745 pp.

[21] Amodei D., Anubhai R., Battenberg E., Case C., Casper, J., et al. 2015. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, *arXiv, [cs.LG]*, V. 1. https://arxiv.org/pdf/1512.02595.pdf

[22] He K., Zhang X., Ren S., y Sun J. 2015. Deep Residual Learning for Image Recognition, *arXiv, [cs.LG]*, V. 1. https://arxiv.org/pdf/1512.03385.pdf

[23] Lai S., Xu L., Liu K., y Zhao J. 2015. Recurrent Convolutional Neural Networks for Text Classification, *In AAAI.*

[24] Jayaswal V. 2020. Laplace smoothing in Naïve Bayes algorithm [En Línea]. Disponible: https://towardsdatascience.com/laplace-smoothing-in-na%C3%AFve-bayes-algorithm-9c237a8bdece

[25] Lorberfeld A. 2019. Machine Learning Algorithms In Layman's Terms, Part 1 [En Línea]. Disponible: https://towardsdatascience.com/machine-learning-algorithms-in-laymans-terms-part-1-d0368d769a7b

[26] Mera-Gaona M., Neumann Ú., Vargas-Canas R., López D. 2021. Evaluating the impact of multivariate imputation by MICE in feature selection, *Front. Neurorobot* [En Línea]. Disponible: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0254720

[27] Sweeney M., Rangwala H., Lester J., Johri A. 2016. Next-Term Student Performance Prediction: A Recommender Systems Approach, *arXiv, [cs.LG]*, V. 1. `https://arxiv.org/pdf/1604.01840.pdf`

[28] CAHSI Caribbean Circuit `https://ece.uprm.edu/cahsiecosystem/practices.html`

[29] Engineering PEARLS `https://www.uprm.edu/engineering/pearls/`

[30] R2DEEP `https://www.uprm.edu/engineering/r2deep/`

[31] Rathmell P. 2016. Colleges Raise Graduation Rates With New Statistical Tools. *Enlace del U.S. News*.

[32] Goodfellow I., Bengio Y., y Courville A. 2016. Deep Learning. *The MIT Press*.

[33] Bhande A. 2018. What is underfitting and overfitting in machine learning and how to deal with it [En Línea]. Disponible: *Enlace*

[34] López K. 2016. Pocos terminan sus estudios universitarios [En Línea]. Disponible: *Enlace*

[35] Choque Y., Acuña E. 2015. Modelo de clasificación y predicción en dos etapas: utilizando árboles de clasificación y el análisis de regresión multivariada. Tesis M.S., Universidad de Puerto Rico Recinto Mayagüez, PR.

[36] Alex. 2018. Meaning of Surrogate Split [En Línea]. Disponible: *Enlace*

[37] Timofeev, R. 2004. Classification and regression trees (cart). theory andapplications. Tesis Maestría, CASE - Center of Applied Statistics and Economics. Humboldt University, Berlin

[38] Breiman, L., Friedman, J., Olshen, R., y Stone, C. 1984. Classification and Regression Trees. Boca Raton London New York Washington, D.C. CHAPMAN & HALL, CRC.

[39] Trujillo A., y Macchiavelli R. 2020. Análisis estadístico de los requisitos matemáticos para estudiantes de nuevo ingreso en la UPRM. Tesis M.S., Universidad de Puerto Rico Recinto Mayagüez, PR.

[40] Lagman A., y Ambat S. 2015. Predictive Analytics of Student Graduation Using Logistic Regression and Decision Tree Algorithm. The Second International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2015). Islamic Azad University, UAE Branch, Dubai, United Arab Emirates

[41] Breiman, L. 1996. Bagging Predictors, *Kluwer Academic Publishers.*, Machine Learning, 24, 123-140.

[42] López L., y Lorenzo E. 2019. Modelo para el análisis de los factores asociados con el tipo de parto aplicando bosques aleatorios y regresión logística. Tesis M.S., Universidad de Puerto Rico Recinto Mayagüez, PR.

[43] Consejo de Educación de Puerto Rico. Compendio Estadístico sobre la Educación Superior de Puerto Rico Año académico 2020-2021. *Enlace*

[44] Kearns M. 1988. Thoughts on Hypothesis Boosting, Unpublished manuscript (Machine Learning class project) [En Línea]. Disponible: *Enlace*

[45] Schapire R. 1990. The Strength of Weak Learnability. *Boston, MA: Kluwer Academic Publishers*, Machine Learning, 5 (2), 197-227.

[46] Cichosz P. Data Mining Algorithms: Explained Using R, Wiley, Poland 718 pp.

[47] Friedman J. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, Volume 38, Issue 4, 367-378. `https://doi.org/10.1016/S0167-9473(01)00065-2`