ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"DETECCIÓN TEMPRANA DE DESERCIÓN ACADÉMICA Y SU APLICACIÓN EN EL SISTEMA DE CONSEJERÍAS ACADÉMICAS DE LA ESPOL"

TRABAJO DE TITULACIÓN

Previa a la obtención del Título de:
MAGISTER EN CIENCIAS DE LA COMPUTACIÓN

Presentado por:
Estefanía Vanessa Heredia Jiménez

GUAYAQUIL - ECUADOR

AÑO: 2020

AGRADECIMIENTOS

Agradezco a Dios, familia y amigos por darme las fuerzas para superar los obstáculos y dificultades a lo largo de este camino.

A Centro de Tecnologías de la Información(CTI) y colegas por darme las facilidades para realizar la maestría.

Al proyecto LALA apoyado por la Comisión Europea (concesión no. 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP) y sus colaboradores por permitirme trabajar junto con ellos.

A mi tutor Ph.D. Gonzalo Mendez y colegas Ph.D. Margarita Ortiz, Msc. Alberto Jimenez por la enseñanza, paciencia, dedicación y apoyo en el presente trabajo.

DEDICATORIA

Dedico este trabajo primero a Dios por permitirme llegar hasta este momento importante en mi formación profesional. A mis padres y hermanos, por ser el pilar fundamental en mi vida. A la ESPOL y todo CTI por permitirme la oportunidad de trabajar con ellos y dejarme grandes enseñanzas. A mis amigos por su apoyo incondicional y motivación para esta culminación.

TRIBUNAL DE SUSTENTACIÓN

PhD. Gonzalo Méndez
Tutor

PhD. Vanessa Cedeño
Miembo del Tribunal

PhD. Javier Tibau
Miembo del Tribunal

DECLARACIÓN EXPRESA

"La responsabilidad por los hechos, ideas y doctrinas expuestas en este informe me corresponde exclusivamente; y, el patrimonio intelectual de la misma, a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL).

Vanessa Heredia Jiménez

RESUMEN

La meta de los estudiantes universitarios es finalizar su carrera y obtener un título que les permita mejorar en el ámbito laboral. Pero, existen estudiantes que por diversos factores abandonan sus estudios. Por ello, las Instituciones de Educación Superior buscan ayudar a sus estudiantes aplicando diferentes métodos, que permitan reducir el porcentaje de abandono o su contraparte aumentar el nivel de retención. Sin embargo, los esfuerzos no son suficientes para reconocer de forma rápida aquellos estudiantes que están próximos abandonar su carrera. Esta tesis propone un modelo que predice la probabilidad de deserción/retención académica de un estudiante, de cualquier semestre finalizado, en cualquier carrera de pregrado. Además, a diferencia de otros modelos disponibles, el modelo presentado permite identificar a los estudiantes en riesgo incluso cuando solo han finalizado un semestre académico. Por otro lado, este modelo es usado como parte de una herramienta web de orientación académica de la ESPOL, que muestra el nivel de deserción/retención académica de un estudiante y una breve explicación sobre las características que influyen en el resultado del modelo predictivo. Con esto, los consejeros académicos pueden identificar y tomar mejores decisiones para orientar a los estudiantes y finalmente minimizar la deserción. Como caso de estudio, utilizamos los datos de los estudiantes de todas las carreras de pregrado de la ESPOL desde 2000 hasta 2020.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ABREVIATURAS

TIC - Tecnología de información y comunicación

EDM - Education Data Mining

LA - Learning Analytics

LMS - Learning Management System

ML - Machine Learning

RFC - Ramdom Forest Classification

CAPÍTULO 1

1. Introducción

El nivel de retención académica –o, su contraparte, la tasa de deserción académica– es uno de los indicadores de éxito (o fracaso) más importantes en la misión de enseñanza de una institución de educación superior (IES). Una baja tasa de deserción (o una tasa alta de retención) representa una ganancia no solo para una universidad, sino también para la sociedad que invierte recursos en la formación de los estudiantes del sistema de educación pública.

Por estas y otras razones, las IESs a menudo invierten recursos de varios tipos en la identificación de las causas por las que un estudiante abandona su carrera y en la implementación de medidas preventivas que reduzcan las probabilidades de que esto ocurra. Esta tesis describe uno de estos esfuerzos realizados en la ESPOL y su sistema de consejerías académicas. Más específicamente, esta tesis se enfoca en la identificación automática de

estudiantes en riesgo de desertar. Para esto, esta tesis utiliza técnicas de *Machine Learning* y *Learning Analytics* para la generación de un modelo de predicción que permita a los consejeros académicos de la ESPOL identificar tempranamente a estudiantes en riesgo de desertar. Todo esto es posible mediante el uso de datos históricos de rendimiento de los estudiantes de la ESPOL y de el uso de la historia académica particular de cada estudiante.

La investigación reportada en este documento se enmarca dentro del proyecto LALA[1] (Learning Analytics for Latin America). Este proyecto busca mejorar la calidad y eficiencia de la Educación Superior en América Latina, desarrollando capacidades para crear, adaptar, implementar y adoptar herramientas de Learning Analytics que mejoren la toma de decisiones académicas [1]. La implementación en la ESPOL de una herramienta de detección temprana de deserción fue uno de los objetivos del proyecto LALA.

Por lo expuesto arriba, algunos de los esfuerzos reportados en esta tesis fueron realizados en colaboración con investigadores del Centro de Tecnologías de Información (CTI) de la ESPOL y con un equipo de la Universidad Carlos III de Madrid. Esto incluye, principalmente, las secciones 3.1.1 y 4.1. Asimismo, producto de estas colaboraciones, algunas partes de esta tesis han sido publicadas como *workshop papers* en las conferencias ECTEL 2020 ([2]) y LALA 2020 ([3,4])[2]. Finalmente, para reflejar la naturaleza colaborativa inherente a la investigación científica, parte del texto de esta tesis utiliza el pronombre “nosotros” y verbos conjugados en la tercera persona del plural.

El resto de este capítulo discute los antecedentes y cifras que motivan esta investigación. Posteriormente, presenta las preguntas de investigación que se abordan en esta tesis, el alcance de la misma y el contexto específico en que ésta se sitúa.

[1] https://www.lalaproject.org

[2] Las publicaciones de esta conferencia aún no han sido indexadas

1.1. Antecedentes y justificación

La tasa de deserción académica promedio en los países que conforman la Organización para la Cooperación y el Desarrollo Económico, (OECD, por sus siglas en inglés), bordea el 24 % [5]. En el caso específico de América Latina, el Banco Mundial reporta una cifra similar—22 % de deserción académica [6]. Según el Diario El Telégrafo, en Ecuador los índices alcanzan el 40 % y la deserción ocurre predominantemente durante los primeros años de formación académica [7, 8].

En la ESPOL, el 43 % de los estudiantes registrados en el periodo [2000 - 2019] lograron egresar o graduarse de alguna carrera de pregrado. Aunque no se dispone de una cifra oficial, este valor sugiere que la tasa de deserción en la ESPOL podría ser relativamente alta en comparación con el porcentaje de estudiantes que logran graduarse. Esto resulta preocupante, pues detrás de estas cifras no solo existe inversión de recursos del Estado y la universidad, sino también la voluntad de muchos estudiantes, y de sus familias, de lograr un título universitario. Este hecho afirma la importancia de identificar a los estudiantes en riesgo de abandonar sus estudios y los factores por los que el riesgo de deserción podría incrementarse.

A menudo, las universidades implementan varios tipos de estrategias para reducir las probabilidades de que sus estudiantes deserten. Esto típicamente incluye programas de orientación [9], incentivos económicos o becas [10], uso de TICs en sistemas de tutorías [11], y acciones de seguimiento académico. No obstante, la evidencia sugiere que estas estrategias son insuficientes para reducir significativamente las tasas de deserción académica [9], puesto que difícilmente permiten identificar **de manera sistemática y temprana** a los estudiantes en riesgo de desertar. Asimismo, a menudo,

estas estrategias requieren de una gran cantidad de recursos (humanos, financieros y logísticos), lo cual las hace costosas, poco prácticas, y no escalables.

Trabajo previo en el área de Learning Analytics sugiere que los datos histórico sobre rendimiento académico de estudiantes representan una oportunidad para identificar los factores que podrían derivan en deserción académica. Los trabajos previos en este sentido utilizan algoritmos de predicción con el fin de tomar medidas preventivas para reducir las probabilidades de deserción de los estudiantes en riesgo [12, 13]. Sin embargo, la vasta mayoría de estos trabajos tienen limitaciones de distinta naturaleza: a) caracterizan la deserción académica solo a través de estadísticas descriptivas [14], b) consideran únicamente datos generados durante los primeros años de estudio [15], c) construyen modelos predictivos que funcionan solo para una carrera [16] o para un curso [17] específicos, o d) generan resultados demasiado ajustados o sesgados [18]—lo cual los hace poco generalizables.

Por otro lado, las salidas que producen este tipo de modelos predictivos usualmente se comunican a las partes interesadas (profesores y estudiantes) mediante visualizaciones (por ejemplo, [19, 20]). El uso de representaciones visuales permite al usuario final obtener una visión *rápida* de la situación de un estudiante sin tener que preocuparse por el funcionamiento interno de los algoritmos utilizados. Estas visualizaciones se enfocan, por ejemplo, en el desempeño del estudiante en un curso específico [21] o en su carrera [19]. Otros estudios que también recurren al uso de visualización para mostrar los resultados de modelos predictivos, en cambio omiten qué características influyen en su modelo [22], lo cual limita su aplicación en otros escenarios.

1.2. Preguntas de investigación y alcance

El Sistema de Consejerías Académicas de la ESPOL tiene por objetivo acompañar a los estudiantes a lo largo de su carrera. Este acompañamiento es realizado por profesores, quienes asesoran a los estudiantes en varios aspectos de su vida académica (por ejemplo, sus opciones de registro). Las sesiones de consejerías se apoyan con una herramienta web[3] que pone a disposición de los consejeros el historial académico de los estudiantes, su desempeño en las materias que cursan actualmente, y otros datos relevantes. Toda esta información es presentada mediante estadísticas y visualizaciones que ilustran el progreso y desempeño académico de los estudiantes.

Antes de la realización de esta tesis, la herramienta web del sistema de consejerías de la ESPOL no disponía de un mecanismo para estimar la probabilidad de deserción de los estudiantes. Esta tesis contribuyó con funcionalidad al respecto abordando las siguientes preguntas de investigación:

- **Q1:** ¿Qué tipo de características de la historia académica de los estudiantes de la ESPOL tienen un rol importante en la predicción de deserción académica?
- **Q2:** ¿Cómo podrían comunicarse los resultados de un modelo predictivo de deserción académica a los consejeros del sistema de consejerías académicas de la ESPOL?
- **Q3:** ¿Cuál es el beneficio percibido por los consejeros de la ESPOL al ser expuestos a predicciones de deserción académica y a explicaciones sobre éstas?

Para responder estas preguntas, esta tesis emplea técnicas de minería

[3]Disponible en www.consejerias.espol.edu.ec

de datos y algoritmos de aprendizaje automático para implementar modelos que predicen la probabilidad de deserción académica de los estudiantes. Dichos modelos utilizan información demográfica y datos históricos de rendimiento para predecir la probabilidad de deserción de los estudiantes de **cualquier carrera** de la ESPOL y **en cualquier punto** de sus estudios. Los modelo predictivos descritos en esta tesis se enfocan únicamente en las carreras de pregrado; es decir, no son aplicables a programas de maestría, ni doctorales de la ESPOL.

Adicionalmente, esta tesis recurre a técnicas de visualización de datos para **integrar la salida de los modelos predictivos construidos en la herramienta web del sistema de consejerías académicas de la ESPOL**. Esto incluye, además, una explicación sobre la contribución relativa de cada una de las características de entrada utilizadas. Aunque estas visualizaciones son insuficientes para abrir completamente las "cajas negras" de los algoritmos que se utilizan en esta tesis, ciertamente son efectivas para "traducir" los resultados del modelo a una representación que los consejeros puedan comprender sin mayores dificultades.

1.3. Contexto y Marco Teórico

La contribución de esta tesis se sitúa en la intersección de dos grandes áreas (Figura 1.1): Educational Data Mining (EMD) y Learning Analytics (LA). La Sociedad Internacional de Minería de Datos Educativos[4] define A la EDM como una disciplina que se ocupa de desarrollar métodos para explorar datos provenientes de entornos educativos y que los utiliza para comprender mejor a los estudiantes y los entornos en los que estos aprenden [23, 24].

[4] http://educationaldatamining.org

Por otro lado, la Sociedad para la Investigación de la Analítica del Aprendizaje[5] define la LA como la medición, recopilación, análisis y notificación de datos sobre los alumnos y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que éste se produce [25, 26]. Estas dos definiciones enfatizan la obtención de datos, métodos de análisis y presentación de resultados en el campo educativo [27, 28], para comprender a los estudiantes y mejorar las condiciones en que su aprendizaje tiene lugar.

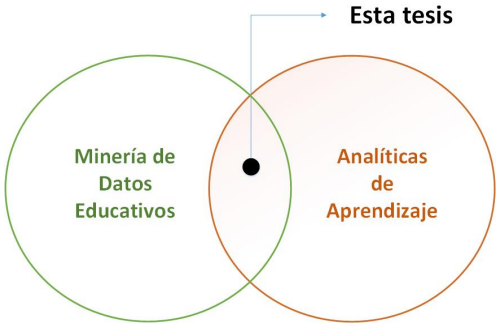

Figura 1.1: Contexto y alcance de la investigación abordada en esta tesis

Una de las aplicaciones más exploradas de estas dos áreas es la predicción académica [29]. Esto se debe, en parte, al gran volumen de información que usualmente se registra sobre los estudiantes en los contextos educativos. Siguiendo esta línea de investigación, esta tesis produce modelos de predicción de deserción académica. El objetivo final de estos modelos es tomar medidas preventivas que minimicen los riesgos de abandono de estudios. Esta tesis también utiliza técnicas básicas de visualización de datos para presentar los resultados de los modelos predictivos. Sin embargo, es importante resaltar que la contribución principal de esta tesis se encuentra en la intersección de la EDM y LA.

[5] https://www.solaresearch.org

1.4. Organización del documento

El resto del presente documento se encuentra estructurado de la siguiente manera:

- El capítulo 2 presenta una revisión de trabajos existentes relacionados a modelos de predicción de deserción académica, y la manera en que se muestran las salidas de dichos modelos.
- El capítulo 3 describe el origen de los datos académicos utilizados en esta tesis, las tareas de procesamiento que se realizaron, el diseño del modelo predictivo, y el diseño de las variables de predicción utilizadas.
- El capítulo 4 expone los resultados experimentales generados a partir de la aplicación de distintas versiones del modelo predictivo. Este capítulo también presenta comparaciones del rendimiento de estos modelos.
- Finalmente, el capítulo 5, presenta una discusión del trabajo realizado en esta tesis, las conclusiones que se derivan del mismo, y sugiere posibles direcciones para trabajos futuros en esta área.

CAPÍTULO 2

2. Marco Teórico

Como se indicó en el Capítulo 1, la investigación reportada en esta tesis se encuentra en la intersección de dos grandes disciplinas (EDM y LA), cada una de las cuales hace uso extensivo de algoritmos de predicción y de visualización de datos. En este contexto, este capítulo primero detalla algunas definiciones fundamentales que se utilizarán a lo largo de esta tesis. Esto es seguido de la presentación de varios trabajos relevantes a esta investigación. Esto incluye estudios orientados a la construcción de modelos de predicción de deserción académica, y trabajos que adoptan un enfoque visual para presentar resultados a usuarios de herramientas que utilizan datos educativos.

2.1. Definiciones

Los conceptos de deserción y retención académica están profundamente relacionados [30]. Estos términos pueden aplicarse a diferentes escenarios y, por tanto, significar algo distinto dependiendo del contexto en que se los utilice. A continuación se explican algunas connotaciones comúnmente asociadas a estos términos en la literatura académica.

La deserción académica puede ser entendida de acuerdo a las siguientes definiciones [30]:

- Deserción total: abandono definitivo de la institución.
- Deserción discriminada por causas: según la causa de la decisión, puede ser social, psicológica, o económica.
- Deserción por facultad: el estudiante cambió de facultad.
- Deserción por carrera: el estudiante cambió de carrera (dentro de una misma facultad o a otra distinta).
- Deserción en el primer semestre de la carrera: ocurre a menudo por inadecuada adaptación al ambiente universitario.
- Deserción acumulada: es el conjunto de deserciones que tienen lugar en una institución educativa.

Por otro lado, la retención académica, se refiere al acto de mantenerse inscrito en una universidad. También existen varias acepciones relacionadas a la retención académica [30]:

- Retención para graduación: Caracterizada por tres distintos comportamientos: a) graduación dentro de los plazos normales establecidos en

la institución; b) graduación de la facultad donde el estudiante se registró originalmente; y c) graduación de la carrera en la cual el estudiante se registró originalmente.

- Retención para la finalización: Se refiere a la finalización de un curso o periodo académico, usualmente durante el primer y segundo año de estudio.

- Retención para el logro de objetivos: Se refiere a la retención cuando el objetivo del estudiante no es la graduación.

Para los objetivos que persigue esta tesis y el contexto de la ESPOL, la deserción académica se define al evento de que **un estudiante abandone sus estudios por un periodo mayor a cinco años consecutivos desde su último registro**. Esto está en línea con las directrices y regulaciones internas de la ESPOL[1]. Como contraparte, la retención académica queda definida como el hecho de que un estudiante se mantenga inscrito en su carrera hasta obtener su título universitario.

2.2. Predicción de Deserción Académica

En la literatura académica existen diferentes modelos y características de entrada para la predicción de deserción. Algunos de estos trabajos se enfocan en el análisis de un solo tipo de características (por ejemplo, solo la información de rendimiento académico), mientras que otros mezclan diferentes grupos de características (por ejemplo, datos demográficos + indicadores socio económicos).

En el grupo de los trabajos que usan un solo tipo de características, Ting

[1] http://reglamentos.espol.edu.ec

y Man [31] utilizaron características psicosociales con un modelo estadístico aplicado a 690 estudiantes. Otro estudio de Viloria y Parody predice deserción analizando características académicas de desempeño, como la nota final del primer parcial y el porcentaje de inasistencias acumuladas de 171 alumnos [32]. Un estudio de Kappe y Van der Flier [33] combinó características de inteligencia, personalidad y motivación de 137 estudiantes.

En la categoría de trabajos que combinan características de distintos tipos, Aulck y sus colegas utilizaron información demográfica, información de ingreso preuniversitario, y el historial universitario de 69.116 estudiantes registrados entre 1998 y 2006. Este trabajo generó un total de más de 700 características, con una precisión promedio del 66 % [34]. Junto a otros investigadores, Ameri utilizó características demográficas, historia familiar, información financiera, escuela secundaria, año de matrícula universitaria y créditos semestrales de 11.121 estudiantes [35]. La precisión de este modelo osciló entre el 71 % y 82 %. Otros autores exploraron el uso de información sobre el comportamiento del estudiante en la primera parte de la carrera de pregrado de 498 estudiantes registrados entre 2001 y 2004, prescindiendo de atributos socio-demográficos clásicos [16]. Este trabajo derivó en una precisión de entre el 74 % y 78 %. Otro modelo predictivo [15], utilizó el rendimiento académico y datos recopilados de un Sistema de Gestión de Aprendizaje (LMS, por sus siglas en inglés) entre el 2016 y 2017, obteniendo una precisión del 79 % al 92 %.

Los modelo mencionados arriba abordan el problema de predicción de deserción académica con precisiones comparativamente buenas. Sin embargo, los estudios mencionados también evidencian limitaciones de distintos tipos. Por ejemplo, utilizan muestras pequeñas, se enfocan exclusivamente en los primeros años académicos, o lidian con cursos o carreras específicas

de pregrado. Estas limitaciones hacen que los modelos mencionados, y otros que existen en la literatura, no puedan ser aplicados a nivel global dentro de una misma institución educativa (como la ESPOL). Por otro lado, combinar pocas o demasiadas características usualmente produce resultados sesgados o predicciones menos precisas [36].

Las limitantes discutidas arriban resaltan la necesidad de modelos predictivos que puedan ser aplicados a diferentes carreras, con características idóneas que permitan predicciones precisas, y que puedan ser entendidos por los consejeros a fin de tomar acciones que permitan mitigar riesgos de deserción. Esta tesis contribuye en esta sentido, desarrollando modelos de predicción de deserción académica que permiten detectar, de forma temprana, a estudiantes en riesgo.

2.3. Visualización de Predicción de Deserción Académica

La literatura incluye una amplia diversidad en torno a la visualización de la salida de modelos de predicción de deserción académica. Por ejemplo, Gkontzis y otros investigadores muestran paneles de visualización en Moodle para monitorear a los estudiantes que están comprometidos con el contenido del curso y facilitar al tutor señalar a los estudiantes que posiblemente lo hayan abandonado. (Figura 2.1). Los paneles son vistos por los tutores en diferentes instancias del tiempo. Algunos son mostrados diariamente, semanalmente y otros tardan más dos horas en generarse debido a que tiene que procesar una gran cantidad de elementos que componen el material educativo de los cursos [21].

Dempere utiliza una visualización donde se enlistan los estudiantes que

Figura 2.1: Ejemplo de visualizaciones proporcionada por el estudio de Gkontzis

han sido identificados por su sistema aprendizaje automático con riesgo de abandono. El asesor puede ver el porcentaje de deserción (Figura 2.2). Sin embargo, la visualización no muestra más información sobre la predicción, como por ejemplo el nivel de confianza. Tampoco muestra las características que influyeron en los resultados de la predicción [22].

Lucio y otros investigadores utilizan 20 paneles de visualización. Algunos de ellos se refieren a las tasas de graduación y retención, pero se encuentran a nivel de carrera y ubicación del campus de estudio. Aunque el estudio menciona una visualización sobre graduación y retención, no muestra imágenes al respecto. Por otro lado, este trabajo no utiliza modelos de predicción, lo

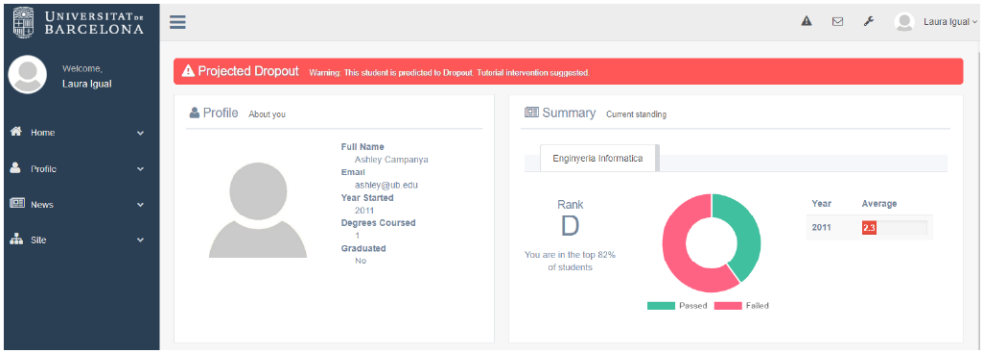

Figura 2.2: Ejemplo de visualización proporcionada por el estudio de Dempere, donde se muestra como advertencia la probabilidad de abandono de un estudiante

que impide identificar de manera inmediata a los estudiantes en riesgo de desertar [19].

Gutiérrez y otros colegas presentan la herramienta llamada LADA [20], la cual muestra información del estudiante a través de diferentes visualizaciones que permiten a los consejeros realizar análisis comparativos (Figura 2.3). Una de estas visualizaciones muestra la probabilidad de éxito, que podría indicar riesgo académico cuando el valor estimado es bajo. Este trabajo sugiere que futuros esfuerzos en esta área deberían presentar, además del valor de la predicción, un indicador de calidad de la misma. Esto serviría para incrementar la transparencia y confianza de los consejeros en la predicción presentada por la herramienta de visualización.

Las visualizaciones discutidas arriba están diseñadas para mostrar probabilidades de deserción o retención académica. Estos estudios, sin embargo, están limitados en varios frentes. Por ejemplo, se enfocan en mostrar únicamente el resultado de las predicciones, pero no las razones detrás de la predicción (es decir, cómo las características de entrada influyen en la salida). Tampoco muestran el nivel de confianza de la predicción, lo que podría introducir incertidumbre en los consejeros y hasta desconfianza respecto al

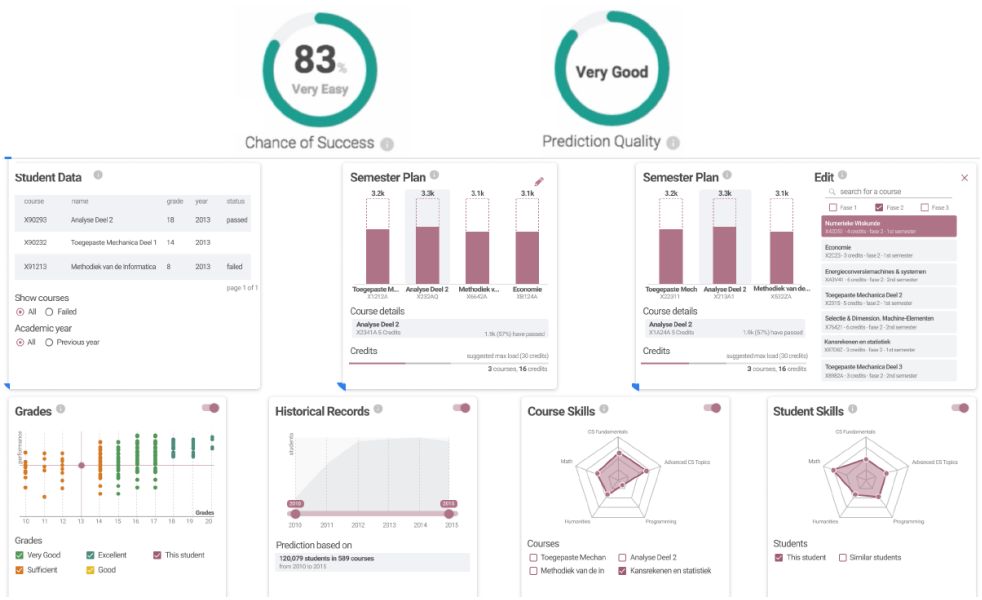

Figura 2.3: Descripción general de LADA: un panel de analítica de aprendizaje utilizado para asesoramiento académico

resultado de la predicción. Otra limitación que existe en la literatura actual es que muy pocos trabajos presentan evaluaciones de uso o indicadores de aceptación de las visualizaciones que proponen. Estas limitaciones sugieren que las visualizaciones de deserción o retención académica deberían: a) poder ser utilizadas para estudiantes de cualquier carrera; b) mostrar de forma eficaz no sólo el resultado de la predicción, sino también qué características influyeron en la salida de la misma; c) mostrar el nivel de confianza de la predicción; d) ser evaluadas con respecto al uso o aceptación por parte de los consejeros.

Esta tesis combina técnicas de EDM y LA con visualizaciones básicas para contribuir a superar las limitaciones listadas arriba. Los resultados de esta tesis han sido incorporados en la herramienta web del sistema de consejerías académicas de la ESPOL. El modelo y las visualizaciones aquí descritas han sido utilizadas ya durante 3 periodos académicos ordinarios (desde

el segundo término académico de 2019). Los siguientes capítulos detallan la construcción de los modelos predictivos que se han utilizado en el sistema de consejerías, los distintos procesos de validación que se realizaron, y algunas cifras en cuanto al uso de estas funcionalidades por parte de los consejeros de la ESPOL.

CAPÍTULO 3

3. Construcción del modelo predictivo

Esta tesis aborda el problema de deserción académica como un problema de predicción, a través de la construcción de un modelo predictivo. Con el fin de que la salida del modelo sea la probabilidad de un estudiante de desertar o no en un futuro cercano, y a su vez sirva al consejero para la toma de decisiones. A continuación, el capítulo describe la metodología usada, los datos proporcionados para la construcción del modelo, las tareas de preprocesamientos llevadas a cabo, el diseño de las características de entrada al modelo, y finalmente describe el modelo predictivo resultante.

3.1. Metodología

El diseño de características y la construcción del modelo predictivo generado utiliza datos socio-demográficos y el historial académico de los estudiantes proporcionados por la Gerencia de Tecnologías y Sistemas de la Información de la ESPOL[1]. Se siguió una metodología iterativa, el cual fue implementado a lo largo de 3 periodos académicos distintos entre el año 2019 y 2020. Se generaron un total de 3 iteraciones debido a los procesos de acreditación de la carreras de pregrado que han tenido lugar en dichos periodos académicos. Por tanto, cada iteración tiene ciertas mejoras y ajustes en relación a la iteración previa. Otra razón por la cual hay 3 iteraciones es porque la iteración *n+1* fue informada de la validación de la iteración *n*. Las validaciones llevadas a cabo son con respecto al modelo y al despliegue del modelo dentro de la herramienta web del sistema de Consejerías Académicas. Esencialmente, para cada iteración se utilizaron datos distintos, el número de características de entrada al modelo es variado y ocurren cambios tanto a nivel del modelo predictivo como a nivel de inclusión de la salida del modelo en la herramienta web del Sistema de Consejerías Académicas. La Figura 3.1 muestra el esquema de la metodología usada que consistió en las siguientes iteraciones:

3.1.1. Iteración 1

La información usada para esta iteración representa los datos sociodemográficos e historial académico de 29.983 estudiantes matriculados a partir del primer término académico del año 2000 hasta el primer término académico del año 2019, distribuidos en un total de 66 carreras de pregrado (50

[1] https://www.serviciosti.espol.edu.ec/

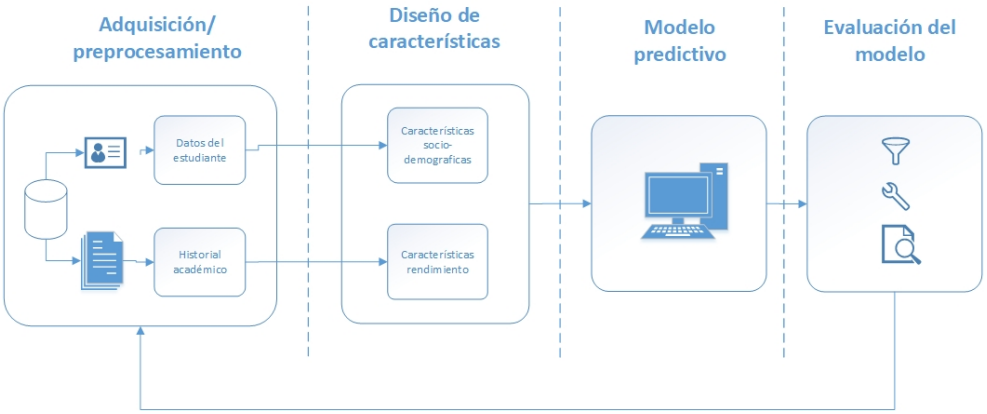

Figura 3.1: Metodología iterativa de la construcción del modelo predictivo

ingenierías, 10 licenciaturas, 6 tecnológicas). Una vez realizadas las tareas de pre-procesamiento, se procedió con el diseño de las características para ser usadas como entradas al modelo predictivo. Como caso de uso escogimos de manera aleatoria una de las carreras que poseía el mayor número de estudiantes, 753 estudiantes, con el fin de obtener una buena muestra con respecto a toda la población. El detalle del diseño de características y la construcción del modelo son discutido en las subsecciones 3.3 y 3.4 de este capitulo, respectivamente. Luego de la construcción del modelo, se realizó un proceso de validación de características para establecer cuáles eran las características con mayor influencia en el modelo y descartar la menos influyentes. Debido a este proceso de validación, las características socio-demográficas fueron descartadas y se usó las características de rendimiento académico del estudiante. Después de la evaluación del modelo para el caso de uso, se aplicó el modelo para todas las demás carreras de la ESPOL. El modelo usó el mismo número de características de entrada para todas las carreras. Luego, la salida del modelo fue incorporada en la herramienta web del sistema de Consejerías Académicas para recibir una retroalimentación de los consejeros y realizar mejoras en el diseño.

3.1.2. Iteración 2

Esta iteración usa datos de un semestre adicional, un total de 26.763 estudiantes desde el primer término académico del año 2000 hasta el segundo término académico del año 2019, distribuidos en un total de 28 carreras de pregrado (22 ingenierías, 5 licenciaturas, 1 tecnológica). El numero de estudiante y carreras es diferente de la iteración anterior, porque en esta iteración se aplicaron procesos de unificación de carreras y de verificación de la cantidad de estudiantes por carrera que se puede predecir. En base a la retroalimentación recibida en la primera iteración respecto a las visualizaciones, el modelo de esta iteración estima la probabilidad de retención como contra parte de la deserción. El motivo del cambio radica en la percepción de las visualizaciones que pueden generar un impacto negativo en los consejeros, y podrían desalentar a los estudiantes. Por tanto, para esta iteración el modelo predice la probabilidad de retención de los estudiantes, como contraparte de la deserción. La carrera usada como caso de uso, las tareas de pre-procesamiento y las características de entrada usadas fueron las mismas de la primera iteración. Se realizó un segundo proceso de validación de características distinto, y encontramos que algunas características perdían poder de predicción en relación a la primera iteración. Como resultado se tiene un modelo predictivo que usa un número diferente de características entrada cada vez que predice; es decir, en cada predicción se realiza la evaluación de las características y se descartan las menos influyentes. Esta iteración incorpora el cálculo del porcentaje de importancia de las características usadas. Al igual que la iteración anterior, después de la evaluación y validación del modelo para el caso de uso, se aplicó el modelo a las demás carreras de la ESPOL. Luego la salida del modelo fue incluida en la herramienta web de Consejerías para recibir una nueva retroalimentación de los

consejeros.

3.1.3. Iteración 3

En esta iteración se agregó otro semestre adicional, un total de 28.932 estudiantes desde el primer término académico del año 2000 hasta el primer término académico del año 2020, distribuidos en un total de 46 carreras de pregrado (30 ingenierías, 10 licenciaturas, 6 tecnológicas). La carrera usada como caso de uso, las tareas de pre-procesamiento y las características de entrada usadas fueron las mismas de la segunda iteración. Sin embargo, se agregó una nueva característica de entrada al modelo predictivo, siendo el factor económico. Para la primera y segunda iteración existieron retrasos en el proceso de entrega de esta información o era inconsistente, por tal motivo no fue incluido. El factor económico pasó por un proceso de conversión porque el estado financiero de los estudiante tiene rangos y nomenclaturas distintas en diferentes periodos de tiempo. En ESPOL el factor económico para los estudiantes matriculados antes del año 2000 se denomina Factor P, para los matriculados desde el 2000 al 2014 se denomina GMS, y para los matriculados desde el 2015 hasta la actualidad se denomina ISE, ver tabla 3.1. Para homogeneizar este información antes de proveerlo como característica de entrada al modelo predictivo, se convirtió a los diferentes rangos de estados financieros a una misma escala de 0 a 1 según la escala actual.

Año	Tipo estado financiero	Rango
<2000	Factor P	[0 - 40]
2000-2014	GMS	[-0,6 - 10]
>2014	ISE	[0 - 1]

Tabla 3.1: Tipos de estados financieros existentes en la ESPOL y sus respectivos rangos

Al igual que la iteración anterior, los procesos de validación se mantuvie-

ron, por lo que el modelo predice solo con las características de entrada más importantes y calcula su porcentaje de importancia. Después de los procesos de validación y evaluación para el caso de uso, se aplicó el modelo a las demás carreras de la ESPOL, y la salida del modelo es incorporada en el sistema de Consejerías.

Las siguientes secciones reportan la estrategia que se utilizó para la generación del modelo, que es común al modelo generado en cada una de las iteraciones. Las excepciones específicas a cada iteración, el proceso de validación de cada una y las evaluaciones dentro del Sistema de Consejerías Académicas con consejeros reales se describen en el Capítulo 4.

3.2. Datos y pre-procesamiento

Los datos utilizados para cada iteración representan los datos socio-demográficos y el historial académico de los estudiantes. Tienen como mínimo cubrir cinco años de estudios completos; es decir, cinco años transcurridos desde el momento en que un estudiante comienzan su primer semestre académico hasta la obtención de un título profesional.

3.2.1. Datos socio-demográficas

Contiene información de los estudiantes como el estado del estudiante (*Graduado, Egresado, Perdió carrera, A prueba, Inactivo*). Aquellos estudiantes que no tenían un estado asignado son los que se encuentran cursando algún semestre; por lo tanto, se los considera como estudiantes *Activos*. Otros datos que contiene son: sexo (*Femenino/Masculino*), estado civil (*Casado, Unión libre, Soltero, Viudo/a*), estado laboral (Trabaja *Si/No*), lugar de residencia, financiamiento del colegio proveniente (*Particular,Liceo, Academia,*

Nacional, Fiscal, Instituto, Fisco-misional, Municipal), y factor económico. Estos datos fueron considerados ya que en el estudio [37], mencionan que son factores influyentes en la explicación de los riesgos de deserción y probabilidades condicionales de graduación. Además, no solo debe considerarse el merito académico, sino también las características socio-demográficas para diferenciar los sectores de la población menos favorecidos con los que disponen de mayores recursos económicos y superior nivel cultural en el núcleo familiar. La Figura 3.2 muestra los datos que fueron usados para este tipo de características.

estadocivil character var	trabaja characte	paisnacimiento character varying	provincianac character var	ciudadnacimiento character varying	colegio character varying (100)	tipocolegio character varyin	paisreside character v	provincia character	ciudadreside character varyi	gms double	sexo character vary	idcarreraactual smallint	codigocarr character v	nombrecarrera character varyin	facultadcar character va	creditostotalcarrera double precision	estadox characti	promed double	id [Pk] bigint
CASADO	SI	ECUADOR	LOS RIOS	BABAHOYO	PARTICULAR MIXTO MARI...	PARTICULAR	ECUADOR	GUAYAS	GUAYAQUIL	0.28	FEMENINO	165	LILSIN2	Licenciatura e...	Facultad ...	240		[null]	1
SOLTERO		ECUADOR	GUAYAS	GUAYAQUIL	ACADEMIA NAVAL ALMIRA...	PARTICULAR	ECUADOR	GUAYAS	GUAYAQUIL	0.32	MASCULINO	80	LIDWA	Licenciatura e...	Escuela d...	240	G	[null]	2
CASADO	SI	ECUADOR	GUAYAS	GUAYAQUIL	CATOLICO DANIEL COMBO...	PARTICULAR	ECUADOR	GUAYAS	GUAYAQUIL	0.18	MASCULINO	83	LIINFN2	Licenciatura e...	Facultad ...	207		[null]	3
CASADO	SI	ECUADOR	GUAYAS	GUAYAQUIL	ACADEMIA NAVAL ALMIRA...	PARTICULAR	ECUADOR	GUAYAS	GUAYAQUIL	0.25	MASCULINO	83	LIINFN2	Licenciatura e...	Facultad ...	207	G	[null]	4
CASADO		ECUADOR	GUAYAS	GUAYAQUIL	GUILLERMO ROHDE AROSE...	PARTICULAR	ECUADOR	GUAYAS	GUAYAQUIL	0.18	MASCULINO	83	LIINFN2	Licenciatura e...	Facultad ...	207	G	[null]	5
SOLTERO		ECUADOR	GUAYAS	NARANJITO	NARANJITO	NACIONAL	ECUADOR	GUAYAS	GUAYAQUIL	0.18	FEMENINO	83	LIINFN2	Licenciatura e...	Facultad ...	207	G	[null]	6
SOLTERO		ECUADOR	GUAYAS	GUAYAQUIL	NACIONAL RITA LECUMBE...	NACIONAL	ECUADOR	GUAYAS	GUAYAQUIL	0.12	FEMENINO	165	LILSIN2	Licenciatura e...	Facultad ...	240		[null]	7
SOLTERO		ECUADOR	GUAYAS	GUAYAQUIL	AMARILIS FUENTE ALCIVA...	FISCAL	ECUADOR	GUAYAS	GUAYAQUIL	0.12	FEMENINO	83	LIINFN2	Licenciatura e...	Facultad ...	207	G	[null]	8
CASADO	SI	ECUADOR	GUAYAS	GUAYAQUIL	28 DE MAYO GUAYAQUIL	FISCAL	ECUADOR	GUAYAS	GUAYAQUIL	0.08	FEMENINO	83	LIINFN2	Licenciatura e...	Facultad ...	207	G	[null]	9
SOLTERO		ECUADOR	GUAYAS	GUAYAQUIL	FISCAL DOCTOR CESAR B...	FISCAL	ECUADOR	GUAYAS	GUAYAQUIL	0.12	MASCULINO	80	LIDWA	Licenciatura e...	Escuela d...	240		[null]	10

Figura 3.2: Presenta los datos socio-demográficos mencionados anteriormente que fueron procesados y categorizados, la identidad de los estudiantes se ha ocultado para proteger su privacidad

3.2.2. Historial académico

Contiene la información sobre todas las carreras en las que el estudiante se registró, cada registro tiene: el código de las materia, estado de las materia (*Aprobada, Reprobada, Perdió por falta, Sancionado*), número de veces que un estudiante ha tomado la materia, año y término en el que tomó la materia, promedio final de la materia, créditos de la materia, el tipo de crédito de la materia(*Tomada, Convalidada, Acreditada*), identificador que permite diferenciar si una materia fue tomada o caso contrario el estudiantes dio un examen de suficiencia o la materia provino de otra malla académica, y el número total de créditos de la carrera. Debido a que se oculta la identidad del estudiante, el conjunto de datos del historial académico contiene una registro

con todas sus columnas en vació, para de esta manera hacer correspondencia al estudiante con su historial académico. Por ejemplo, el estudiante 1 tiene por historial académico todas las filas antes de llegar a una fila en vació, el estudiante 2 tiene por historial académico las siguientes filas antes de llegar a una fila en vació, el estudiante 3 tiene por historial académico las siguientes filas antes de llegar a una fila en vació y así sucesivamente. La Figura 3.3 muestra los datos que fueron usados para realizar el diseño de características de rendimiento del estudiante.

	codigomateriatc character varying(	nombremateriatomada character varying (200)	estadomateria character varyi	veztomada integer	anio charac	termino character	codigocarrera character varying	nota1 numeric	nota2 numeric	mejorami numeric (i	promedio numeric (5,:	nota numeric (i	creditosm double pr	tipoformacion character varyir	tipocredito character va	id [PK] bigint	codigocarreraactual text	tomada boolean
40	ICHE03541	EMPRENDIMIENTO E INNO...	AP	1	2011	2S	LILSIN2	[null]	[null]	[null]	[null]	6.80	4	[null]	T	40	LILSIN2	true
41	ICM01040	ESTADÍSTICA (IIT95)	AP	1	2012	1S	LILSIN2	[null]	[null]	[null]	[null]	7.40	4	[null]	T	41	LILSIN2	true
42	FIEC06445	SIMULACIÓN DE NEGOCIOS	AP	1	2012	1S	LILSIN2	[null]	[null]	[null]	[null]	7.40	4	[null]	T	42	LILSIN2	true
43	PRTC003244	METODOLOGÍA DE DESARR...	AP	1	2012	2S	LILSIN2	[null]	[null]	[null]	[null]	8.20	6	[null]	T	43	LILSIN2	true
44	CELEX00083	INGLÉS INTERMEDIO A	AP	2	2013	1S	LILSIN2	[null]	[null]	[null]	[null]	8.80	6	[null]	T	44	LILSIN2	true
45	FIEC04838	PLAN Y CONTROL DE PRO...	AP	1	2013	1S	LILSIN2	[null]	[null]	[null]	[null]	6.55	4	[null]	T	45	LILSIN2	true
46	FIEC05892	SISTEMAS DE GESTIÓN DE ...	AP	1	2013	1S	LILSIN2	[null]	[null]	[null]	[null]	8.80	4	[null]	T	46	LILSIN2	true
47	FICT03293	LEGISLACION PROFESIONAL	AP	1	2013	2S	LILSIN2	[null]	[null]	[null]	[null]	7.75	3	[null]	T	47	LILSIN2	true
48	CELEX00091	INGLÉS INTERMEDIO B	AP	1	2013	2S	LILSIN2	[null]	[null]	[null]	[null]	7.75	6	[null]	T	48	LILSIN2	true
49	CELEX00109	INGLÉS AVANZADO A	AP	1	2013	3S	LILSIN2	[null]	[null]	[null]	[null]	0.00	4	[null]	C	49	LILSIN2	false
50	CELEX00117	INGLÉS AVANZADO B	AP	1	2014	1S	LILSIN2	[null]	[null]	[null]	[null]	0.00	4	[null]	C	50	LILSIN2	false
51	PRTC001008	MATEMATICAS II 95/2	RP	1	2001	2S	LILSIN2	[null]	[null]	[null]	[null]	5.60	4	[null]	A	51	LILSIN2	true
52	PRTC000968	FUNDAMENTOS DE PROGR...	RP	1	2001	2S	LILSIN2	[null]	[null]	[null]	[null]	4.20	6	[null]	T	52	LILSIN2	true
53	PRTC001248	CONTABILIDAD II (1998-3)	RP	1	2002	1S	LILSIN2	[null]	[null]	[null]	[null]	5.45	4	[null]	T	53	LILSIN2	true
54	PRTC001248	CONTABILIDAD II (1998-3)	RP	2	2002	2S	LILSIN2	[null]	[null]	[null]	[null]	5.85	4	[null]	T	54	LILSIN2	true
55	PRTC001032	ORGANIZACIÓN DE COMP...	RP	1	2003	1S	LILSIN2	[null]	[null]	[null]	[null]	5.35	4	[null]	T	55	LILSIN2	true
56	PRTC001289	SIST.CLIENTE SERVIDOR	RP	1	2003	2S	LILSIN2	[null]	[null]	[null]	[null]	2.25	4	[null]	A	56	LILSIN2	true
57	PRTC001149	SISTEMAS OPERATIVOS (9...	RP	1	2004	2S	LILSIN2	[null]	[null]	[null]	[null]	1.90	6	[null]	A	57	LILSIN2	true
58	PRTC001115	DISEÑO DE SISTEMAS(95/2)	RP	1	2004	2S	LILSIN2	[null]	[null]	[null]	[null]	4.80	6	[null]	A	58	LILSIN2	true

Figura 3.3: Presenta los datos del historial académico que fueron procesados para la creación de características de rendimiento

Los datos disponible no fueron ideales y necesitaban ser tratados antes de poder ser utilizados en tareas de Machine Learning. Esto se debe a los procesos de acreditación de la carreras que han implicado cambios de mallas académicas, convalidaciones de materias, trabajo manual para identificar las materias de las mallas antiguas y mallas nuevas. Otras tareas de preprocesamiento que consistieron en eliminar espacios en blanco y caracteres especiales de todos los textos; descartar estudiantes sin historial académico; determinar la carrera más reciente cursada por un estudiante.

3.3. Diseño de características

Debido al origen de la información y por ser un trabajo exploratorio, se tomaron decisiones sobre los datos para el diseño de las características usadas como entrada para el modelo predictivo. Estas decisiones fueron las siguientes:

- Selección de la última carrera de pregrado cursada por los estudiantes. Al seleccionar la última carrera cursada, se tiene información actualizada respecto a los procesos de acreditación de las carreras.

- Correspondencia del historial con los datos demográficos debido al anonimato, ya que el historial académico contiene toda la vida estudiantil independiente de las carreras en las que se registró el estudiante.

- Como se mencionó en la iteración 3, la información sobre el factor económico pasó por un proceso de homogeneización, porque el estado financiero de los estudiante variaba de rangos en diferentes periodos de tiempo. Por lo que se convirtió a los diferentes rangos a una escala de 0 a 1 según la escala actual usada por ESPOL.

- Construcción de un coeficiente denominado coeficiente de ponderación (***Coef***) que decrece de manera lineal dependiendo del número de veces que un estudiante toma una misma materia, ver tabla 3.2. Con el fin de que la nota promedio con que contribuye una materia refleje la historia de reprobación del estudiante. El cálculo del promedio se detalla en la subsección 3.3.1.2, ecuación 3.7. Esta decisión de diseño fue motivada por una aplicación práctica de evaluación de promedio de la Universidad de Carlos III de Madrid [38] y adaptado del sistema euro-

peo de créditos y calificaciones en las titulaciones universitarias [39].

No. de veces tomadas de una materia	***Coef***
1	1
2	0,9
3	0,8
4	0,7
5	0,6
>= 6	0,55

Tabla 3.2: Coeficiente de ponderación según el número de veces que un estudiante toma una materia

- Las materias con estados *Perdió por falta* y *Sancionado*, fueron consideradas como materias *Reprobadas*, ya que son materias que el estudiantes las pierde y debe volver a tomarlas, similar al estado de Reprobada.

- La duración de los términos académicos de la ESPOL son variados respecto al tiempo (Figura 3.4) y se dividen en períodos ordinarios (Primer y Segundo Término) y período extraordinarios (Tercer Término). Por lo tanto, el modelo predictivo de esta tesis utiliza los períodos ordinarios por tener la misma duración, y ser términos académicos donde la mayor cantidad de estudiantes se registra, por lo que se puede tener mayor información para predecir.

- Enumeración secuencial de los semestres registrados en el historial académico, para reconocer los semestres de manera ordenada y consecutiva, independiente del año de matriculación de un estudiante.. Por ejemplo, un estudiante puede ingresar a su primer semestre en el primer o segundo término académico en cierto año, después realizar su segundo semestre en el primer o segundo término en un siguiente año

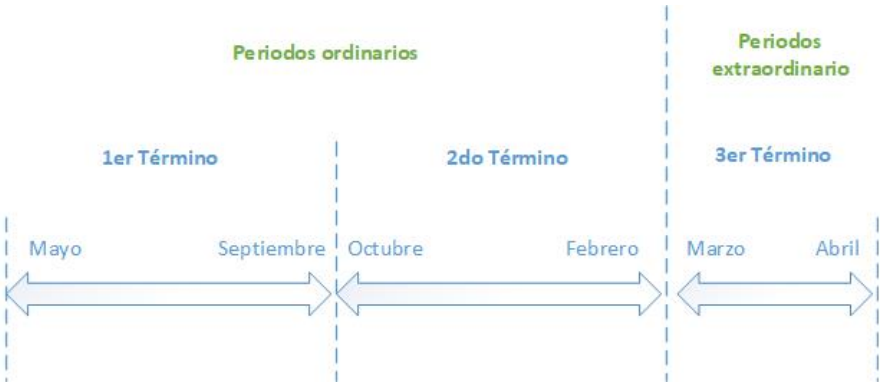

Figura 3.4: Periodos - Términos académicos

o en el mismo año y así sucesivamente.

- Conteo del total de los años y semestres cursados por los estudiantes. Es importante tener conocimiento del tiempo que le lleva a un estudiante finalizar su malla. Para esta tesis y demás variables relacionadas con el tiempo, los años son representados por números enteros, mientras que el semestre toma el valor de 0.5. Por ejemplo, si un estudiante tiene valor de 2.5 significa que le llevó 2 años y un semestre estudiar, total 5 semestres.

- Asignación a priori de etiquetado de estudiantes, esta asignación sirve para distinguir los grupos de estudiantes con estado *Graduados o Egresados*, de aquellos que han desertado de la carrera. La asignación se basa en tres criterios principales: el estado del estudiante, un criterio temporal que considera la cantidad de tiempo que el estudiante lleva inactivo en la ESPOL, y un criterio porcentual que indica el avance de la malla, dando como resultado lo siguiente:

  - Los estudiantes con estado *Graduados o Egresados* se los considera como un solo grupo, *Graduados/Egresados*.

- Cálculo del *Porcentaje de créditos hechos*($\%credHechos$), implica la división del sumatorio de los créditos de las materias con estado *Aprobadas* con o sin nota final, con créditos diferente de cero, para el Número total de créditos de la carrera, representada por la siguiente fórmula:

$$\%credHechos = \frac{\sum CredMatAP}{\#totalCredCarrera} \tag{3.1}$$

  Cuando el estudiante tiene este porcentaje con valor mayor o igual al 90 % se lo etiqueta dentro del grupo de los *Graduados/Egresados*. Este porcentaje se basa en otro estudio [40], que indica que los estudiantes con este nivel avanzado de carrera desertan por motivos ajenos a su desempeño académico.

- Cálculo del *Tiempo desde su último registro* ($\delta AniosSemUltReg$), representa los años-semestres que el estudiante lleva sin estudiar desde su último registro, si el resultado es cero significa que el estudiante esta actualmente cursando la carrera. Si el resultado es mayor o igual a cinco años se lo considera como estudiante *Desertor*, porque no se ha registrado en ninguna materia por cinco años o más según la definición de deserción del Capitulo 2, subsección 2.1.

3.3.1. Características de entrada del modelo

A través de proceso de prueba y error se generaron varias características de entrada para el modelo. Las cuales variaban en cada iteración debido a los procesos de validación de las mismas. El capítulo 4 detalla las características creadas y usadas en cada iteración de manera específica. Sin embargo, esta

sección menciona las principales características que fueron comunes a las 3 iteraciones trabajadas.

3.3.1.1. Características socio-demográficas

La sección 3.2.1 describe los datos socio-demográficos usados, ver tabla 3.3. Además, un proceso de categorización fue llevado a cabo de la siguiente manera: el estado civil según el tipo de convivencia (Solos, Acompañados), porque se vio la necesidad de distinguir si el estudiante tenía cargas familiares o no; cuidad de residencia (Guayaquil, Otras ciudades) porque la gran mayoría de los estudiantes de la ESPOL provienen de Guayaquil, mientras que alrededor del 21 % provienen de otras ciudades; financiamiento del colegio proveniente (Pagados, Semi-gratuitos/Gratuitos) ya que al provenir de diferentes tipos de colegios con distintos niveles académicos, la base de conocimiento de los estudiantes en su primer semestre académico puede variar y afectar en el rendimiento académico, además el 63 % proviene de colegios pagados. La tabla 3.4 muestra un resumen de esta categorización. Un total de 6 características fueron utilizadas.

3.3.1.2. Características de rendimiento académico

La sección 3.2.2 describe la información del historial académico capturada. Esta subsección detalla las características que nacieron a partir de estos datos. Un total de 11 características fueron creadas y agrupadas de la siguiente manera: numéricas o acumulativas (4) que capturan información a través de acumulados o contadores, las porcentuales (3) se refieren a proporciones, las promediadas (3), y las temporales (1) que tienen que ver con el tiempo. Estas características fueron creadas con el fin de capturar y reflejar la historia académica y para analizar el rendimiento de los estudiantes.

Dato	Valor
Sexo	Femenino
	Masculino
Trabaja	Si
	No
Factor socio-económico	Valor numérico
Estado del estudiante	Sin estado
	Graduado
	Egresado
	Perdió carrera
	A prueba
	Inactivos

Tabla 3.3: Datos socio-demográficos usados

Dato	Valor	Categoría
Estado civil	Casado	Conviven solos
	Unión libre	
	Soltero	Conviven acompañados
	Viudo/a	
Ciudad de residencia	Guayaquil	Vive en Guayaquil
	Otras ciudades	Vive fuera de Guayaquil
Financiamiento del colegio proveniente	Particular	Colegios pagados
	Liceo	
	Academia	
	Nacional	Colegios semi-gratuitos y gratuitos
	Fiscal	
	Instituto	
	Fisco-misional	
	Municipal	

Tabla 3.4: Categorización de datos socio-demográficos

□ ***Numéricas o acumulativas***

- *Número de materias total* (#$MatTotal$): es el número total de materias que el estudiante tiene en su historial académico; es decir, materias con estado *Convalidadas*, *Acreditadas*, y *Tomadas*.

- *Número de materias tomadas* ($\#MatTomadas$): es el número de materias **tomadas por el estudiante de manera presencial**, con la siguiente condición: si no posee materias tomadas pero si tiene *Número de materias total* diferente de cero se considera este último como *Número de materias tomadas*. Note que este cálculo, a diferencia del anterior, excluye a las materias con estado *Convalidadas y Acreditadas* ya que los estudiantes dan un examen de suficiencia o realizan procesos de acreditación de materias debido a sus cambios de carrera.
- *Número de materias de segunda matrícula* ($\#2daMat$): es el número de materias que un estudiante ha repetido por segunda vez. Se consideran las materias con promedio final diferente de cero, estado *Reprobada*, y con coeficiente de ponderación *Coef* igual a 0,9, note que el coeficiente de una materia de segunda matrícula de acuerdo a la tabla 3.2 es siempre igual a 0,9.
- *Número de materias de tercera matrícula* ($\#3eraMat$): es el número de materias que un estudiante ha repetido por tercera vez. Se consideran las materias con promedio final diferente de cero, estado *Reprobada*, y con coeficiente de ponderación *Coef* igual a 0,8, note que el coeficiente de una materia de tercera matrícula de acuerdo a la tabla 3.2 es siempre igual a 0,8.

□ ***Porcentuales***

- *Porcentaje de materias aprobadas* ($\%AP$): es la división del número de materias con estado *Aprobadas* con o sin promedio final para el Número de materias total, representada por la siguiente

formula.

$$\%AP = \frac{\#MatAP}{\#MatTotal} \tag{3.2}$$

- *Porcentaje de materias reprobadas* ($\%RP$): es la división del número de materias con estado *Reprobadas* con o sin promedio final para el Número de materias total, representada por la siguiente formula.

$$\%RP = \frac{\#MatRP}{\#MatTotal} \tag{3.3}$$

- *Porcentaje de materias anuladas* ($\%AN$): es la división del número de materias *Anuladas* con o sin promedio final para el Número de materias total, representada por la siguiente formula.

$$\%AN = \frac{\#MatAN}{\#MatTotal} \tag{3.4}$$

□ ***Promediadas***

- *Promedio de todas las materias Aprobadas y Reprobadas*: es la división del sumatorio de todas las materias con estado *Aprobadas* y *Reprobadas* por un estudiante particular con promedio final diferente de cero, para el Número total de materias con estado *Aprobadas* y *Reprobadas*, representada por la siguiente formula.

$$\bar{X}_i APRPTotal = \frac{\sum_{j=0}^{n} GPA_{ij}}{\#MatTotalARPR} \tag{3.5}$$

- *Promedio de las materias tomadas Aprobadas y Reprobadas*: es la división del sumatorio de **las materias tomadas de manera presencial** con estado con estado *Aprobadas* y *Reprobadas* por un estudiante particular con promedio final diferente de cero, para

el Número de materias tomadas *Aprobadas* y *Reprobadas*, representada por la siguiente formula.

$$\bar{X}_i APRPTomada = \frac{\sum_{j=0}^{n} GPA_{ij}}{\#MatTomadasARPR} \tag{3.6}$$

Note que a diferencia de la ecuación 3.5, este valor excluye a las materias que el estudiante Convalidó o Acreditó a través de un examen de suficiencia o porque vinieron de otras mallas académicas.

- *Promedio ponderado* ($\bar{X}_i Pond$): es la división de la multiplicación de todas las materias con estado *Aprobada* con promedio final y créditos diferente de cero, por el valor asignado del Coeficiente de ponderación *Coef*, para el sumatorio de los créditos de las materias con estado *Aprobada*, representada por la siguiente fórmula:

$$\bar{X}_i Pond = \frac{\coprod_{j=0}^{n} (GPA_{ij})(Coef_{ij})(Credij)}{\sum_{j=0}^{n} Credij} \tag{3.7}$$

donde:

  - GPA_{ij} es el promedio final obtenida por el estudiante i en la materia j.
  - $Coef_{ij}$ es el valor asignado de *Coef* de la materia j.
  - $Credij$ es el crédito asignado a la materia j.

□ ***Temporales***

- *Tiempo entre semestres* ($\delta AnioSemEsp$): representa el espacio de tiempo en años-semestres que le lleva al estudiante continuar un siguiente semestre.

3.4. Modelo predictivo

La deserción académica por ser considerada para esta investigación como un problema de predicción, el resultado del modelo es un número que indica la probabilidad de un estudiante de desertar o no en un futuro cercano. Por lo tanto, usamos el algoritmo de clasificación de bosque aleatorio (RFC, por sus siglas en inglés), porque trabaja con una serie de particiones, realizando predicciones futuras en base a estas particiones [41]. Adicionalmente, el RFC es conocido por su baja tendencia al sobre-entrenamiento y su alta precisión [42].

El proceso inicia con la ejecución de la toma de decisiones explicadas en la sección 3.3, donde se selecciona el historial académico de la última carrera del estudiante, se homogeneiza la información, se enumeran de forma ascendente los semestres, se obtiene el tiempo de estudio y el tiempo que pasa inactivo el estudiante. Luego, con la asignación a priori del etiquetado de los estudiantes, entre el grupo de *Graduados/Egresado* y el grupo de *Desertores*, basado en los criterios del estado del estudiante, el tiempo que lleva inactivo, y el porcentaje de avance de la malla. Los estudiantes etiquetados como *Graduados/Egresado*, tienen un valor de deserción igual a cero. Los estudiantes etiquetados como *Desertores*, tienen un valor de deserción igual a uno.

La estrategia acumulativa que se explica a continuación, se hace con el fin de capturar toda la historia previa de los estudiantes, para que el modelo pueda trabajar con cada conjunto, en lugar de trabajar con instancias individuales de semestres. Por tanto, luego de la asignación a priori, el proceso continua con la selección de los datos socio-demográficos e historial académico del primer semestre de los estudiantes, y calculando las características

de entrada al modelo. Una vez calculadas, ingresan al algoritmo de predicción. Luego se selecciona los datos del primer y segundo semestre de los estudiantes, y se calcula las características de entrada para ingresarlas al algoritmo. Luego se selecciona los datos del primer, segundo y tercer semestre para calcular las características de entrada e ingresarlas al algoritmo. Y así sucesivamente hasta el hasta el quinto semestre, que representa dos años y medio de estudios, ya que en En Ecuador, el promedio de deserción de los estudiantes se da entre el segundo y tercer año de estudios [43]. Luego del quinto semestre se selecciona todos los semestres en conjunto, para calcular las características de entrada e ingresarlas al algoritmo.

En cada iteración de semestres se seleccionan las características sociodemográficas y se calculan las características de rendimiento académico. Posterior a la generación de características, se provee al algoritmo RFC los datos etiquetado para separar el conjunto de datos de entrenamiento y pruebas del algoritmo. El modelo es entrenado con el grupo de estudiantes cuya deserción se definió previamente con 1 y 0. El modelo se prueba con el conjunto de estudiantes que no tenían definida una deserción y se predice la probabilidad. Después de finalizado el entrenamiento y las pruebas de modelo, se tienen los resultados de probabilidad de deserción académica. De los cuales, los valores de alta probabilidad significan que el estudiante tiene una alta probabilidad de abandonar la carrera, mientras que los valores bajos indican una alta probabilidad de terminar la carrera.

Una vez generado el modelo, se procede con la validación de rendimiento. Culminado el proceso de validación, el modelo es aplicado a todas las carreras de la ESPOL. La salida del modelo, es incorporada en la herramienta web del Sistema de Consejerías. La figura 3.5 muestra los pasos que se llevaron a cabo para la construcción del modelo.

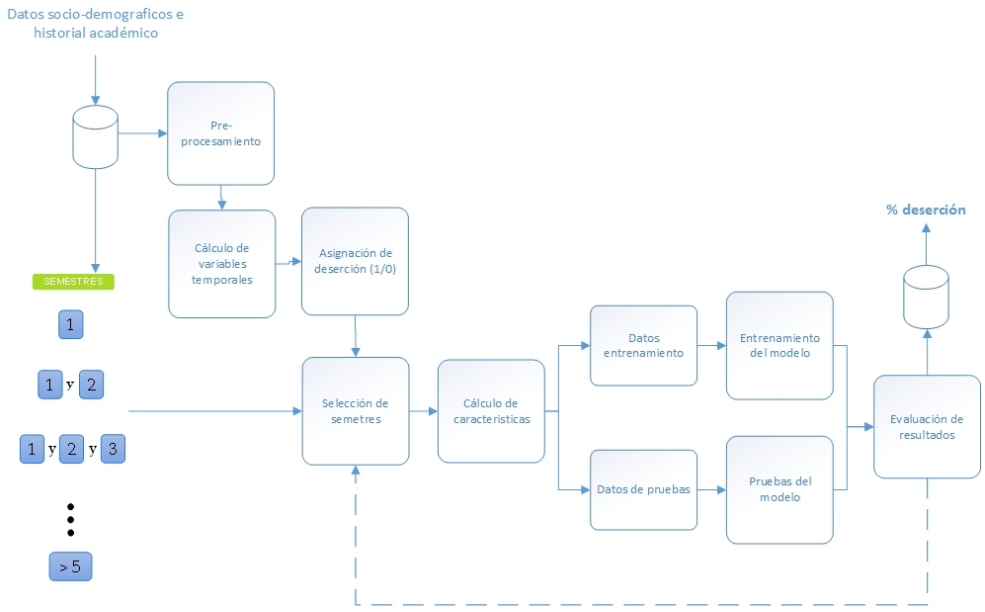

Figura 3.5: Modelo predictivo

Este capítulo explicó las decisiones llevadas a cabo, el diseño de las características y la generación del modelo predictivo de deserción académica utilizando los datos de la ESPOL, que fueron divididos en tres distintas iteraciones. La estrategia de generación del modelo es mayormente común a cada una de las iteraciones de esta investigación; sin embargo, hay ciertas excepciones que se explicarán más en detalle en el Capítulo 4, que además aborda los procesos de validación en cuestión de rendimiento de los modelos generados, y en relación a los estudios con usuarios que se hicieron para el Sistema de Consejerías.

CAPÍTULO 4

4. Evaluación del modelo y análisis de resultados.

Como se indicó en el Capítulo 3, usamos una metodología iterativa que derivó en 3 iteraciones debido a los procesos de validación de rendimiento del modelo, la integración de la salida del modelo en el Sistema de Consejerías y su uso por parte de los consejeros. Aunque el modelo generado es mayormente común para cada iteración, este capítulo detalla las excepciones específicas de cada iteración durante, las características creadas, las novedades encontradas y las razones para el uso o descarte de las características. También detalla los procesos de validación, y resultados obtenidos tanto del rendimiento del modelo como de las visualizaciones por parte de los consejeros.

4.1. Iteración 1

Como caso de uso trabajamos con la carrera (Ingeniería Industrial) seleccionada bajo los criterios indicados en la subsección 3.1.1. Se capturaron los datos sociodemográficos y el historial académico de 753 estudiantes. El proceso inició con la toma de decisiones, principalmente con el etiquetado a priori de los estudiantes entre el grupo de *Graduados/Egresado* y el grupo de *Desertores* explicado en la sección 3.3. Los estudiantes etiquetados como *Graduados/Egresado*, fueron asignados un valor de deserción igual a cero. Los estudiantes etiquetados como *Desertores*, fueron asignados un valor de deserción igual a uno.

4.1.1. Evaluación 1

Dada la naturaleza exploratoria de este trabajo, se crearon varias características de entrada para la generación de la primera versión del modelo predictivo. Primero se diseñaron características tipo promediadas, siendo el promedio de las materias *Aprobadas*, promedio de las materias *Reprobadas*, y promedio de las materias *Anuladas*, para sus valores ser ingresados al modelo junto con las variables socio-demográficas de la sub-sección 3.3.1.1. El factor socio-económico fue descartado ya que el proveedor de los datos había intentado homologar las distintas escalas que el factor socio-económico ha tomado a lo largo del tiempo, pero el proceso tenía errores. Por ejemplo, estudiantes con alto factor económico tenía un valor muy bajo en las escalas subsiguientes, eso hace que la información no sea confiable desde el punto de vista estadístico, y por tanto no ser usado en un modelo predictivo.

Se aplicó un matriz de correlación para analizar cómo afectan las diferentes características dentro del modelo propuesto. En consecuencia, las

características socio-demográficos fueron descartadas debido a que tenían un coeficiente de correlación menor a 0.1. En cuanto a las características de rendimiento, se utilizaron aquellas con un valor mayor a 0.5.

Caracteristicas	Sexo	Trabaja	Estado civil	Ciudad de residencia	Financimiento del colegio	Promedio materias AP	Promedio materias RP	Promedio materias AN
Sexo	1	0,056	0,099	-0,016	0,056	-0,144	0,083	0,003
Trabaja	0,056	1	0,037	-0,026	0,062	0,082	0,014	0,003
Estado civil	0,099	0,037	1	0,007	0,062	0,226	-0,065	0,034
Ciudad de residencia	-0,016	-0,026	0,007	1	-0,037	0,005	-0,095	0,032
Financimiento del colegio	0,056	0,062	0,062	-0,037	1	0,005	0,003	0,011
Promedio materias AP	-0,144	0,082	0,226	0,005	0,005	1	-0,921	0,362
Promedio materias RP	0,083	0,014	-0,065	-0,095	0,003	-0,921	1	0,321
Promedio materias AN	0,003	0,003	0,034	0,065	0,011	0,362	0,321	1

Figura 4.1: Matriz de correlación de las características. Los valores que se encuentran en la diagonal de la matriz contienen los valores máximos de correlación en donde r=1, ya que aquí cada característica se está comparando con sí mismo. En la medida que lo valores se acercan a cero la correlación es menor y en la medida que se aproximan a 1 y -1, la correlación aumenta. Los valores positivos indican que las relaciones son proporcionales de manera positiva y si los valores son negativos significa que la relación es inversamente proporcional, o sea que si el conjunto de valores de las características aumentan, en una proporcionalidad inversa (correlación negativa) los valores de las otras características tenderán a reducirse; si se trata de una correlación positiva, al aumentar los valores de unas características tendería a aumentar en las otras características.

Inspeccionando el historial académico de todos los estudiantes, se encontró que la cantidad de materias anuladas representaban el 10 % con respecto al todas las materias de todos los estudiantes. Por tanto, el promedio de materias Anuladas fue descartado. También se encontró que los resultados del modelo eran inconsistentes con los valores que las características

tomaban. Un ejemplo de esta situación se presentó con estudiantes donde su promedio de materias aprobadas era igual a 8 y el promedio de materias reprobadas igual a 6, pero el modelo daba como resultado una alta probabilidad de deserción, alrededor de 80 %. Por lo tanto, para esta versión del modelo queríamos diseñar una nueva característica que capturara la información de rendimiento de los estudiante todas las veces que tomaban una materia independiente si la aprobaban o reprobaban. Basado en este conocimiento nace la característica que combina el *promedio de las materias aprobadas y reprobadas* obtenidas del estudiante que están registradas en su historial académico.

Por otra parte, se decidió adoptar el cálculo del promedio ponderado de la Universidad de Carlos III, con el fin de reflejar el historial de reprobación de los estudiantes [38]. También se decidió diseñar variables porcentuales que permitan identificar el avance del estudiante con respecto a su malla.

4.1.2. Evaluación 2

Después de las pruebas y evaluación anterior, se decidió ingresar solo las características promediadas y porcentuales al modelo. Las características socio-demográficas volvieron a tener un valor bajo de correlación y fueron descartadas. Se resolvieron los casos que presentaban inconsistencia respecto a la evaluación anterior. Sin embargo, surgieron otros escenarios de falsos positivos, donde estudiantes con promedio de materias aprobadas y reprobadas era superior a 7, pero el modelo reportaba una alta probabilidad de deserción, alrededor de 80 %. Una inspección manual y detallada del historial académico de estos estudiantes reveló que tenían más materias con estado *Convalidadas* o *Acreditadas* que materias con estado *Tomadas* que podían ser *Aprobadas* o *Reprobadas*, lo que generaba conflictos al momento

de predecir. Por tanto, se diseñaron características numéricas que permitan diferenciar el número de materias total que tienen los estudiantes en el historial académico entre *Convalidadas*, *Acreditadas*, y *Tomadas*, respecto al número de materias con estado *Tomadas*.

También se encontraron casos donde los estudiantes tenían un promedio de aprobación y reprobación de 6, con porcentaje de materias reprobadas 70 %, pero con probabilidad de deserción baja de 40 %. Inspeccionando el historial académico, develó que este tipo de estudiantes tenían muchas materias reprobadas, que eran tomadas en los semestres siguientes y las aprobaban. Este nuevo conocimiento implicó crear más características numéricas que permitan reflejar el historial de reprobación de los estudiantes, y conocer la cantidad de veces que toman la mismas materias por segunda o tercera ocasión.

4.1.2.1. Indicadores a nivel de carrera de graduados y desertores

Con el fin de validar los valores que toman las características de rendimiento de tipo promedio con el resultado obtenido del modelo, se crearon indicadores a nivel de carrera de tipo promedio, para comprobar si la probabilidad predicha es consistente. Esta consistencia verifica si un estudiante es etiquetado como posible *Desertor* pero sus características de tipo promedio son bastante cercanas al indicador de la carrera que está asociada al grupo de los Graduados/Egresados. Si este es el caso se identifica esto como una inconsistencia porque el modelo predice una cosa pero las características del estudiante sugieren otra. Para el cálculo de estos indicadores se considera solo a los estudiantes que fueron previamente etiquetados.

Se evalúa el modelo con un grupo de estudiantes seleccionados de manera aleatoria. La información de estos estudiantes es enviada al modelo y

los resultados del modelo son inspeccionados de manera manual, teniendo en cuenta las características de estos estudiantes con los indicadores a nivel de carrera que se calculan previamente. A continuación, se detallan los indicadores de rendimiento a nivel de carrera:

- *Promedio de todas las materias Aprobadas y Reprobadas de los Graduados de la carrera*: es la división del sumatorio de todas la materias con estado *Aprobadas* y *Reprobadas* de los estudiantes etiquetados como *Graduado/Egresado* con promedio final diferente de cero, para el Número de estudiantes etiquetado como *Graduado/Egresado* de la carrera, representada por la siguiente formula.

$$\bar{X}_c APRPGradTotal = \frac{\sum_{i=0}^{n} \bar{X} APRPTotal_{cj}}{\#Grad} \tag{4.1}$$

- *Promedio de las materias tomadas Aprobadas y Reprobadas de los Graduados de la carrera*: es la división del sumatorio de **las materias tomadas de manera presencial con estado *Aprobadas* y *Reprobadas* de los estudiantes etiquetados como *Graduado/Egresado*** con promedio final diferente de cero, para el Número de estudiantes etiquetado como *Graduado/Egresado* de la carrera, representada por la siguiente formula.

$$\bar{X}_c APRPGradTomadas = \frac{\sum_{j=0}^{n} \bar{X} APRPTomadas_{cj}}{\#Grad} \tag{4.2}$$

  Note que a diferencia de la ecuación 4.1, este cálculo excluye a las materias que el estudiante Convalidó o Acreditó.

- *Promedio de todas las materias Aprobadas y Reprobadas de los Desertores de la carrera*: es la división del sumatorio de todas la materias

con estado *Aprobadas* y *Reprobadas* de los estudiantes etiquetados como *Desertores* con promedio final diferente de cero, para el Número de estudiantes etiquetado como *Desertores* de la carrera, representada por la siguiente formula.

$$\bar{X}_c APRPDesertTotal = \frac{\sum_{i=0}^{n} \bar{X} APRPTotal_{cj}}{\#Desert} \tag{4.3}$$

- *Promedio de las materias tomadas Aprobadas y Reprobadas de los Desertores de la carrera*: es la división del sumatorio de **las materias tomadas de manera presencial con estado *Aprobadas* y *Reprobadas* de los estudiantes etiquetados como *Desertores*** con promedio final diferente de cero, para el Número de estudiantes etiquetado como *Desertores* de la carrera, representada por la siguiente formula.

$$\bar{X}_c APRPDesertTomadas = \frac{\sum_{j=0}^{n} \bar{X} APRPTomadas_{cj}}{\#Desert} \tag{4.4}$$

  Note que a diferencia de la ecuación 4.3, este calculo excluye a las materias que el estudiante Convalidó o Acreditó.

4.1.3. Evaluación 3

En esta evaluación, ingresaron al modelo las características promediadas, porcentuales y numéricas, resolviendo los problemas encontrados en la anterior evaluación. El modelo de esta iteración usó las características que se listan en la tabla 4.1.

Debido a que en este punto no se encontraron más casos de falso positivos, se procedió con la evaluación de los parámetros usados para el algoritmo de RFC. Para cada evaluación de parámetros del algoritmo se utiliza-

Dato	Tipo	Valor
Características de rendimiento	Númericas	Número de materias total *#MatTotal*
		Número de materias tomadas *#MatTomadas*
		Número de materias de segunda matrícula *#2daMat*
		Número de materias de tercera matrícula *#3eraMat*
	Porcentuales	Porcentaje de materias aprobadas *%AP*
		Porcentaje de materias reprobadas *%RP*
		Porcentaje de materias anuladas *%AN*
	Promediadas	Promedio de todas las materias Aprobadas y Reprobadas $\bar{X}_i APRPTotal$
		Promedio de las materias tomadas Aprobadas y Reprobadas $\bar{X}_i APRPTomada$
		Promedio ponderado $\bar{X}_i Pond$
	Temporales	Tiempo entre semestres $\delta AnioSemEsp$

Tabla 4.1: La tabla contiene la características de entrada usadas en el modelo que fueron resultado de la validación en la iteración 1. Debido a las validaciones, se descartaron las características socio-demográficas.

ron valores de parámetros diferentes. Como método principal de evaluación usamos el método de validación cruzada, K-fold Cross-Validation, la técnica probablemente más común para la evaluación de modelos en la práctica de ML. Además sirve para estimar qué tan bien se generalizan a conjuntos de datos independientes [44]. Después de varias configuraciones de parámetros probadas, se escogió la configuración de parámetros que produjeron los mejores resultados en el procedimiento de validación cruzada, para usarlos

en el conjunto de entrenamiento completo para el ajuste del modelo. Como resultado final el algoritmo de RFC usó los siguientes parámetros: 300 árboles, cada uno con una profundidad de 6 niveles, con un número mínimo de muestras necesarias para dividir igual a 6, y con un número mínimo de muestras necesarias para estar en un nodo de hoja igual a 7.

Después se evaluó el rendimiento del modelo propuesto utilizando como métricas de evaluación los siguientes criterios [45]:

- *Área bajo la curva*: AUC, por sus siglas en ingles, es la proporción de alumnos que abandonaron la carrera.
- *Accuracy*: es la proporción de alumnos que abandonaron la carrera.
- *Recall*: es la proporción de alumnos que abandonaron la carrera, pronosticada correctamente por el clasificador en todos los estudiantes que abandonaron la carrera.
- *Precision*: es la proporción de alumnos que abandonaron la carrera, pronosticada correctamente por el clasificador.
- *F1-score*: es la media armónica de la precisión y el recall.

La tabla 4.2 presenta los resultados de los criterios de evaluación por semestre del modelo predictivo en base a las características finales descritas anteriormente para el caso de uso.

4.1.4. Aplicación del modelo a todas las carreras de la ESPOL

Debido a los resultados alcanzados en el caso de uso durante la fase exploratoria, con porcentaje de precisión alrededor del 80 %, se decidió aplicar el modelo con los datos de las demás carreras de la ESPOL. La información usada representa el historial académico de 29.983 estudiantes matriculados

Semestre	*AUC*	*Accuracy*	*F1*	*Recall*	*Precision*
1	0,88	81 %	0,66	0,68	0,65
2	0,77	80 %	0,44	0,55	0,55
3	0,78	76 %	0,43	0,52	0,47
4	0,79	77 %	0,42	0,54	0,35
5	0,74	71 %	0,32	0,39	0,28
>5	0,95	93 %	0,82	0,78	0,86

Tabla 4.2: Criterios de evaluación por semestre para el caso de uso en la iteración 1

a partir del primer término académico del año 2000 hasta el primer término académico del año 2019, distribuidos en un total de 66 carreras de pregrado(50 ingenierías, 10 licenciaturas, 6 tecnológicas).

El modelo predijo la deserción académica carrera por carrera siguiendo la estrategia acumulativa explicada en la sección 3.4. En la que se calculan los valores de las características de entrada al modelo para cada grupo de semestres y se obtiene la salida del modelo. El modelo trabajó con los parámetros de RFC indicados anteriormente. También se evaluó el rendimiento del algoritmo con las mismas métricas de evaluación usadas para el caso de uso. La precisión del modelo fue alrededor del 70 %. La Tabla 4.3 muestra un resumen de los cinco criterios de evaluación del modelo para las 10 carreras con el mayor número de estudiantes matriculados.

4.1.5. Visualización

Con el propósito de asistir a los consejeros para tomar mejores decisiones en sus reuniones con los estudiantes, se implementó una visualización con información acerca del abandono académico.

La visualización construida fue incorporada dentro de las ventanas de estadísticas del Sistema de Consejerías de la ESPOL. Estuvo disponible durante la segunda semana del segundo término académico del 2019; es

Código carrera	Carrera	*AUC*	*Accuracy*	*F1*	*Recall*	*Precision*
CI002	Ingeniería Química	0,88	81	0,67	0,68	0,74
CI005	Ingeniería Civil	0,79	75	0,48	0,52	0,65
CI007	Mecánica	0,4	70	0,42	0,54	0,62
CI009	Logística y Transporte	0,87	78	0,70	0,58	0,63
CI013	Computación	0,77	79	0,72	0,56	0,67
CI017	Telecomu-nicaciones	0,77	80	0.64	0,55	0,72
LITUR	Licenciatura en Turismo	0,79	81	0,75	0,57	0,74
INALL	Ingeniería en Alimentos	0,74	70	0,48	0,53	0,69
INACP	Ingeniería en Auditoría y Contaduría Pública Autorizada	0,84	82	0,76	0,65	0,66
LI002	Economía	0,84	75	0,70	0,62	0,68

Tabla 4.3: Criterios de evaluación para las carreras con mayor número de estudiantes matriculados durante el primer término académico del año 2000 hasta el primer término académico del año 2019 para la iteración 1

decir para 2393 sesiones de consejerías. Sin embargo, la visualización fue mostrado solo a los consejeros cuyos estudiantes tenían un porcentaje de deserción superior o igual al 50 %. La visualización consistía en mostrar un mensaje que indicaba si el estudiante se encontraba en riesgo de deserción. Se decidió solo mostrar este texto, mas no el valor del porcentaje, porque aun se tenía que trabajar en la forma de presentación para que sea algo mas amigable con el usuario. El mensaje de riesgo de abandono también mostraba un link de "Ver más", que permitía a los consejeros conocer las características que influyeron en el modelo (Figura 4.2).

Figura 4.2: Panel de visualización incorporado en el Sistema de Consejerías durante la iteración 1

Para evaluar la eficacia y utilidad de la visualización mostrada, se incluyó en el Sistema de Consejerías un mecanismo para capturar: el número de veces que esta parte de la interfaz fue mostrada a los consejeros, y el número de visitantes que ingresaron a la visualización para ver más detalles. Utilizamos Google Analytics para este propósito.

Los resultados indicaron que la visualización fue mostrada un 0,04 % de un total de 678 consejerías, y de ellos solo el 70 % hicieron clic en el link "Ver más". Debido al bajo porcentaje alcanzado de visitas a la visualización de deserción, se decidió reubicarlo a la ventana principal del Sistema de Consejerías.

Por otro lado, al presentar los resultados obtenidos de la evaluación de la visualización a las autoridades académicas de la ESPOL en una reunión presencial, recomendaron cambiar el enfoque de un panel de deserción a un panel de retención académica. La razón radica en la percepción de los paneles que pueden generar un impacto negativo en los consejeros. La literatura confirma esta intuición [46], indicando que los paneles de control deben fomentar el desempeño del estudiante en lugar de desanimarlo, e incluso considerar cómo el desempeño deficiente puede reflejarse mejor en los estudiantes para motivarlos. Por tal motivo, se decidió realizar mejoras a

la visualización, así como refinar el modelo para obtener en su lugar el nivel de retención académica y resultados más precisos.

4.2. Iteración 2

Para cumplir con recomendación recibida por las autoridades de la ESPOL en la iteración anterior, definimos la retención académica como la contra parte de la deserción, ver sección 2.1. Por tanto, a partir de esta iteración los resultados que se muestran se basan en la retención académica de los estudiantes, para generar un impacto positivo en los consejeros.

Para que el modelo funcione basado en retención académica, el etiquetado a priori de estudiantes mantuvo los mismos grupo (*Graduados/Egresado*, y *Desertores*) basados en los criterios del estado del estudiante, el tiempo que lleva inactivo, y el porcentaje de avance de la malla. Sin embargo, el etiquetado fue de la siguiente manera: los estudiantes etiquetados como *Graduados/Egresado*, fueron asignados un valor de retención igual a uno; mientras que los estudiantes etiquetados como *Desertores*, fueron asignados un valor de retención igual a caro.

4.2.1. Evaluación

Para tener una mejor selección de las características de entrada al modelo, se usó métodos de selección de características [47] para identificar las características más útiles del conjunto de datos. Las características innecesarias disminuyen la velocidad de entrenamiento, disminuyen la interpretabilidad del modelo y, lo que es más importante, disminuyen el rendimiento de generalización en el conjunto de prueba. Los métodos de selección de características utilizados fueron:

- Valores faltantes: retorna las características con una fracción de valores perdidos por encima de un umbral especificado, dichas características no ingresan al modelo. Para este trabajo el umbral fue de 60 %. Este porcentaje fue escogido en base a otro estudio [47], que indica que este valor puede servir como inicialización de pruebas del método.

- Características colineales: retorna las características colineales con una correlación mayor que un coeficiente de correlación especificado. Dicho coeficiente es el valor del coeficiente de correlación de Pearson. Para cada par de características con un coeficiente de correlación mayor al umbral, solo se identifica uno del par para su eliminación. El coeficiente usado para este trabajo fue de 0.98 basado en el estudio [47].

- Variables de valor único: se eliminan aquellas características que tengan un valor único. Una característica con un valor único no puede ser útil para el ML porque esta características tiene una variación de cero.

- Peso de importancia de las características: este método asignan una puntuación a las características de entrada de un modelo predictivo en función de su utilidad para predecir una variable objetivo, e indica la importancia relativa de cada característica al realizar una predicción. En base a esta importancia se eliminan aquellas características con importancia cero o casi cero que no contribuyan a la importancia acumulativa de una característica específica.

Para esta iteración se usó la misma carrera de caso de uso y las características indicadas en la tabla 4.1 de la iteración 1, donde el modelo predecía con la misma cantidad de características para cada semestre. En cambio, en la iteración 2 para cada predicción las características pasan previamente por los métodos de selección, teniendo como resultado solo las caracterís-

ticas más influyentes o importantes. Dichas características son ingresadas al modelo. Por ejemplo, para predecir el semestre 1, después de aplicados los métodos de selección, puede ocurrir que todas las características de la tabla 4.1 son importantes. Pero para predecir el semestres 2, después de aplicados los métodos, puede darse el caso que todas las características numéricas, promediadas y porcentuales sean importante, descartando las temporales. Para predecir el semestre 3, después de aplicados los métodos, puede suceder que solo ciertas características numéricas, ciertas promediadas, ciertas porcentuales y ciertas temporales sean importantes, y las demás sean descartadas. La tabla 4.4 presenta los resultados de los criterios de evaluación por semestre del modelo predictivo en base a las características finales descritas anteriormente para el caso de uso.

Semestre	***AUC***	***Accuracy***	***F1***	***Recall***	***Precision***
1	0,76	81 %	0,87	0,88	0,86
2	0,73	80 %	0,85	0,89	0,82
3	0,81	85 %	0,89	0,91	0,88
4	0,82	84 %	0,88	0,90	0,87
5	0,90	85 %	0,89	0,91	0,88
>5	0,90	91 %	0,94	0,92	0,95

Tabla 4.4: Criterios de evaluación por semestre para el caso de uso en la iteración 2

4.2.2. Aplicación del modelo a todas las carreras de la ESPOL

Debido a los resultados alcanzados para el caso de uso donde el modelo obtuvo un porcentaje de precisión alrededor del 84 %, se decidió aplicar el modelo para todos las carreras impartidas por ESPOL. Los datos usados representan el historial académico de 30.280 estudiantes matriculados desde el primer término académico del año 2000 hasta el segundo término académico del año 2019, distribuidos en 66 carreras.

Debido a la antigüedad de los datos seleccionados y por los resultados y la experiencia de la iteración anterior, se descubrió que las carreras presentaban variaciones en sus denominaciones a lo largo del tiempo, pero mantenían el núcleo de enseñanza. Por ejemplo una carrera denominada Acuicultura y otra carrera llamada Ingeniería en Acuicultura, son la misma carrera con nombres distintos. Por lo que se ejecutó un paso adicional de pre-procesamiento a los códigos de denominación de las carreras. Se llevó a cabo un proceso de unificación de códigos de las carreras para evitar la pérdida de información, debido a los cambios efectuados en las mallas por actualizaciones en el Reglamento de Régimen Académico a lo largo de los años. El proceso consistió en identificar las carreras que hayan cambiado su código y asociarlas a un código de carrera padre. Por ejemplo, de los datos seleccionados existían estudiantes con los siguientes códigos de carreras hijas LIPRO(Licenciatura en Diseño y Producción Audiovisual), LIDPRN2(Licenciatura en Diseño y Producción Audiovisual), LI006(Producción para Medios de Comunicación); sin embargo, todas pertenecen a la carrera de Licenciatura en Diseño y Producción Audiovisual con código de carrera padre LIDPP00, por lo tanto fueron asociadas.

Con el propósito de diferenciar el total de estudiantes Graduado / Egresados, el total de Desertores, y el total de estudiantes que el modelo debe predecir por carrera, se creó un proceso de verificación de cantidad de estudiantes. Este paso previo permite verificar si existe la cantidad suficiente de datos para predecir, lo cual reduce el tiempo de procesamiento de todo el modelo ya que se evitan cálculos innecesarios. Un ejemplo de este caso es la carrera Tecnología en Plástico, que posee un total de 135 estudiantes, de los cuales 121 están Graduados, 14 son Desertores y no existen estudiantes a los que se les deba predecir la probabilidad de retención, por tanto esta

carrera es descartada del entrenamiento del modelo.

Como resultado del pre-procesamiento de los códigos de las carreras y la verificación de la cantidad de estudiantes por carrera, se obtuvo un total de de 28 carreras de pregrado (22 ingenierías, 5 licenciaturas, 1 tecnológica) que representan un total de 26763 estudiantes usados por el modelo predictivo.

El modelo predijo la retención académica carrera por carrera. En cada grupo de semestres se calcula las características de rendimiento académico y, en consecuencia, se aplican los métodos de selección de características para que el modelo prediga solo con las características más importantes. Para luego definir el conjunto de datos para el entrenamiento y las pruebas. Las características ingresan al modelo para calcular el porcentaje de retención académica. También se evaluó el rendimiento del algoritmo con las mismas métricas de evaluación usadas para el caso de uso. La precisión del modelo fue alrededor del 84 %. La Tabla 4.5 se muestra un resumen de los cinco criterios de evaluación del modelo para 10 carreras seleccionadas de manera aleatoria.

Para un mejor interpretación de los resultados obtenidos por el modelo en base a los valores que tomaban las características, se implementó un módulo de interpretabilidad post-hoc. Para ello, se utilizó SHAP [48], que es una capa de explicación para los modelos ML. Se basa en la teoría de juegos para calcular la contribución de cada característica (llamada valor de Shapley de la característica) [49] a la respuesta de un modelo de caja negra. Por medio del uso de esta librería se calculó los valores de shap de las características, y en base a éste el peso o porcentaje de influencia de las características en la predicción.

Adicional, se implementó el cálculo de percentiles, con el objetivo de ubi-

Código carrera	Carrera	*AUC*	*Accuracy*	*F1*	*Recall*	*Precision*
CI002	Ingeniería Química	0,77	87	0,92	0,91	0,92
CI003	Minas	0,87	88	0,90	0,88	0,91
CI004	Petróleos	0,84	87	0,91	0,91	0,91
CI007	Mecánica	0,84	86	0,89	0,89	0,89
CI009	Logistica y Transporte	0.83	83	0,85	0.82	0,89
CI013	Computación	0,80	80	0,83	0,81	0,85
LI002	Economía	0,86	87	0,90	0,88	0,92
LI005	Diseño Gráfico	0,89	88	0,88	0,83	0,93
LI007	Administración de Empresas	0,84	83	0,83	0,77	0,90
LI009	Turismo	0,91	90	0,91	0,87	0,95

Tabla 4.5: Criterios de evaluación de 10 carreras seleccionadas de manera aleatoria, con estudiantes matriculados durante el primer término académico del año 2000 hasta el segundo término académico del año 2019 para la iteración 2

car a los estudiantes en un nivel de retención según una escala discreta construida en base a los percentiles, ver tabla 4.6, de acuerdo a su probabilidad de retención resultante del modelo. Esta escala es la que se muestra en la visualización incorporada en el Sistema de Consejerías de la ESPOL.

Percentil	Porcentaje retención	Nivel de retención
P1	[0-20)	Muy bajo
P2	[20-40)	Bajo
P3	[40-60)	Medio
P4	[60-80)	Alto
P5	[80-100)	Muy alto

Tabla 4.6: Percentiles usados en la iteración 2 para indicar el nivel de retención académica según la probabilidad obtenida del modelo

4.2.3. Visualización

De acuerdo a la retroalimentación y los resultados de la visualización en la iteración 1, el panel de visualización presenta el nivel retención académica de los estudiantes. Fue reubicado e incorporado en la ventana principal de la herramienta web del Sistema de Consejerías (Figura 4.3). Estuvo disponible para todos los consejeros durante las consejerías del primer término académico del 2020.

El panel fue rediseñado de la siguiente manera: muestra una escala de color para indicar el nivel de retención ordenados de izquierda a derecha como: verde oscuro (muy alto), verde claro (alto), amarillo (moderado), naranja (bajo) y rojo (muy bajo). Cuando los consejeros quieren más información, pueden hacer clic en el enlace "Ver más", ubicado en la parte inferior de la escala del nivel de retención (Figura 4.4). También en la parte inferior muestra el porcentaje de precisión de la predicción como un indicador de nivel de confianza.

Al dar clic en el enlace "Ver más"se muestra una ventana emergente (Figura 4.5), que contiene información detallada de las características de entrada más importantes que usó el modelo para predecir la probabilidad de retención. Esta información se muestra como un gráfico de barras, ordenado de mayor a menor dependiendo el porcentaje de influencia de las características. Al hacer clic en cada barra del gráfico, se muestra una explicación en la parte inferior derecha, que describe el valor que tomó la característica y de ser caso el indicador a nivel de carrera. La parte superior de la ventana emergente muestra una pregunta que permite a los consejeros responder si consideran o no consistente la información mostrada.

Se procedió con la evaluación de estas visualizaciones para capturar: el número de visitas a la sección que muestra el nivel de retención, número de

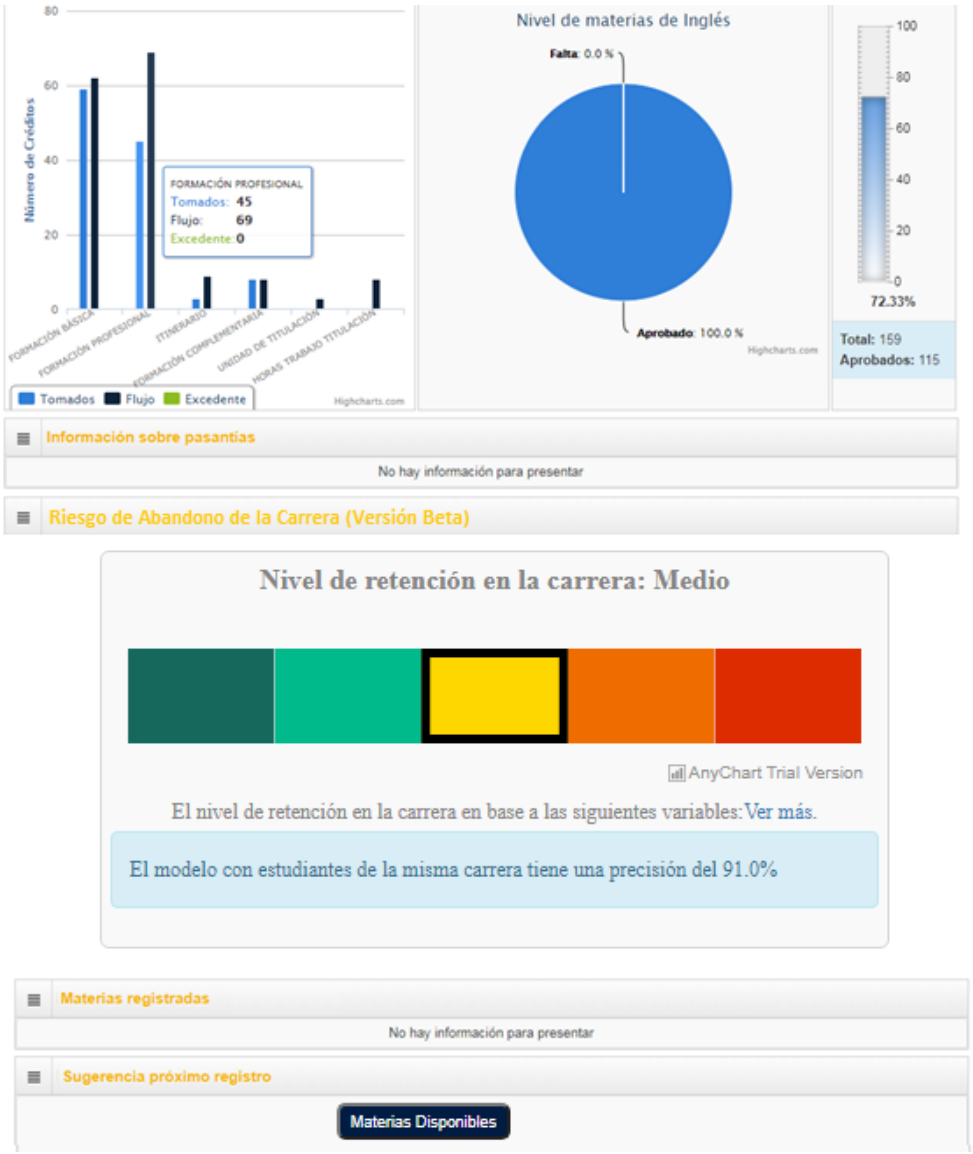

Figura 4.3: Panel de visualización incorporado en el Sistema de Consejerías durante la iteración 2

visitas a la ventana emergente que muestra las características, y número de veces que los usuarios querían ver la explicación de las características. La información cuantitativa también incluyó el número de profesores que estuvieron de acuerdo o en desacuerdo con la precisión de la predicción y los datos mostrados.

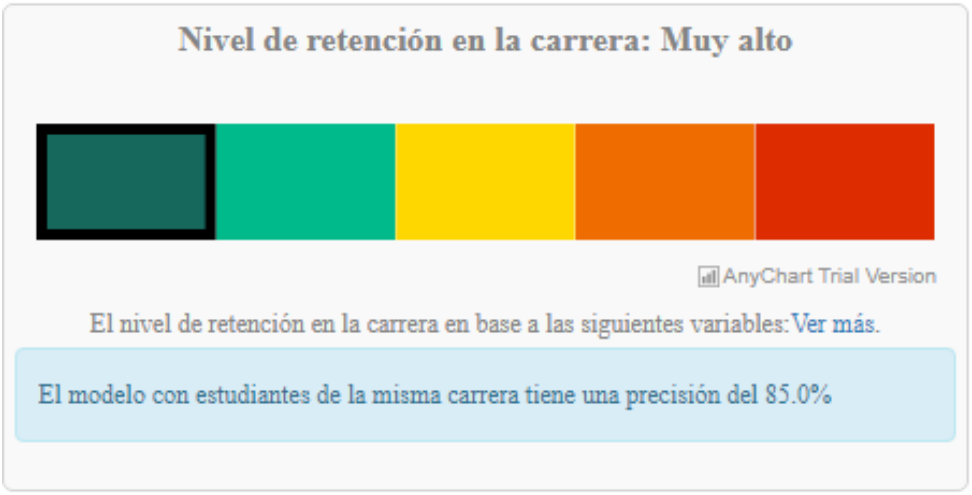

Figura 4.4: Panel de visualización enfocado en retención académica en iteración 2

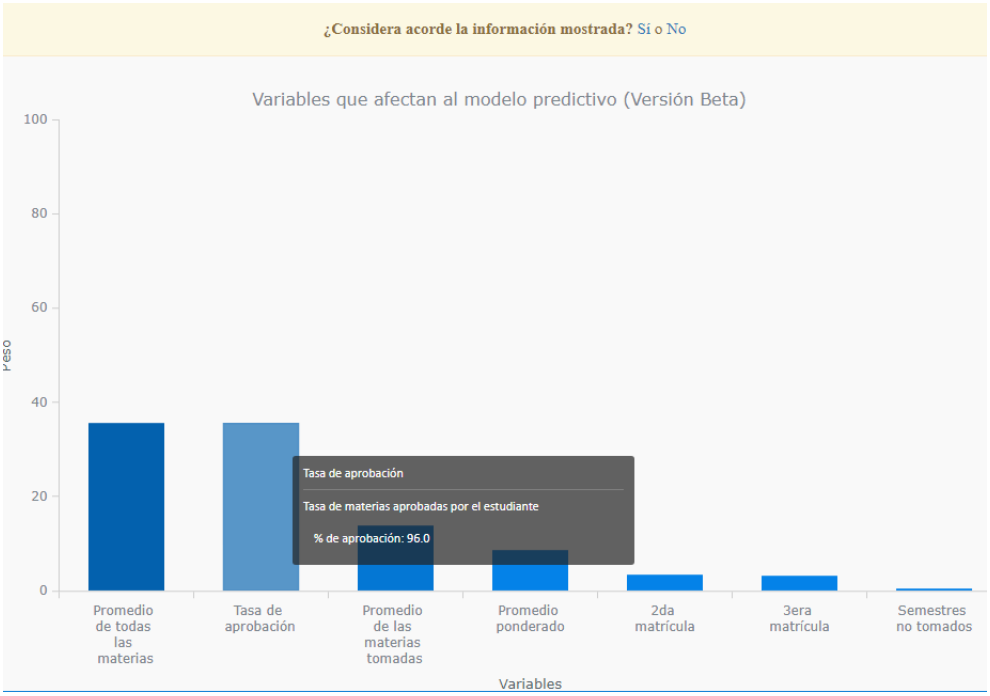

Figura 4.5: Ventana emergente que muestra información de las características ingresadas al modelo en la iteración 2

Los resultados indicaron que el panel de retención fue visto por el 76 % de las 3821 sesiones de consejerías. 27 de 297 consejeros hicieron clic 44

veces en el enlace "Ver más"para ver qué características influyeron en la predicción. De este número, 12 consejeros hicieron clic 131 veces en el gráfico de barras para ver la explicación de las características. En cuanto a la consistencia de la información, en 9 sesiones de consejerías (2 consejeros) estuvieron de acuerdo, mientras que en 3 sesiones de consejerías (1 consejero) discrepó con la información mostrada. Las razones para estar en desacuerdo se centran en cómo se ubicaron los estudiantes en los respectivos niveles de retención.

4.3. Iteración 3

En esta iteración se mantiene el concepto de retención académica definida previamente para el modelo. Se usó la misma carrera de caso de uso y las características indicadas en la tabla 4.1 de las iteraciones anteriores. Sin embargo, para esta iteración surgió un nueva característica de entrada, siendo el factor económico de los estudiantes. Esta característica fue incorporada porque la literatura resalta la importancia del estado financiero de los estudiantes, ya que este factor es uno de los motivos detrás de la decisión de los estudiantes de detener sus estudios [37]. En las anteriores iteraciones no fue incorporado por retrasos en la entrega de esta información o por la inconsistencia en la homologación del factor.

La información del factor económico entregada por el proveedor tuvo que preprocesar toda la información relacionada al factor económico porque la forma en la que éste se registraba en ESPOL había cambiado a lo largo del tiempo según el año de ingreso de los estudiantes, ver tabla 3.1. Por lo tanto, toda la información de estado financiero se la homogeneizó antes de proveerla como característica de entrada al modelo. Convirtiendo esta

información a una escala entre 0 a 1, que es la escala actual que utiliza la ESPOL.

4.3.1. Evaluación

El factor económico, al igual que las demás características, fue evaluada con los métodos de selección descritos en la iteración 2. Demostrando que es una característica influyente para el modelo, porque para cada predicción de grupos de semestres esta característica nunca fue descartada.

Por el contrario, la característica de tipo promedio denominada *Promedio de las materias tomadas Aprobadas y Reprobadas* fue descartada. Porque el valor que tomaba esta característica al pasar por los métodos de selección de valores faltantes o valores únicos provocaba que se descarte. Otra causa del descarte de esta característica se debe a los procesos de acreditación de las carreras que se produjo antes de la generación de esta iteración. Estos proceso de acreditación provocaron cambios en el estado de las materias de *Tomadas* a *Convalidadas*.

Por otro lado el cálculo de los percentiles también fue actualizado ya que existían estudiantes donde el porcentaje de retención eran igual para los percentiles P1 y P2, o para los percentiles P4 y P5, debido a la pequeña diferencia en la división de los grupos de percentiles. Por lo que, se decidió modificar los limites de inclusión y exclusión. Para realizar estos cambios, recibimos la ayuda de un experto en Estadística que nos proporcionó los detalles de inclusión y exclusión dependiendo de los casos, como se puede ver en la tabla 4.7.

La tabla 4.8 presenta los resultados de los criterios de evaluación por semestre del modelo predictivo en base a las características finales descritas para el caso de uso.

Percentil	Caso normal	Si P1==0	Si P5==1	Nivel de retención
P1	[0-20)	[0-20]	[0-20]	Muy bajo
P2	[20-40)	(20-40]	(20-40]	Bajo
P3	[40-60)	(40-60]	(40-60]	Medio
P4	[60-80)	(60-80]	(60-80)	Alto
P5	[80-100]	(80-100]	[80-100]	Muy alto

Tabla 4.7: Percentiles usados en la iteración 3 para indicar el nivel de retención académica según la probabilidad obtenida del modelo

Semestre	*AUC*	*Accuracy*	*F1*	*Recall*	*Precision*
1	0,75	93 %	0,96	0,98	0,94
2	0,70	93 %	0,96	0,99	0,94
3	0,74	90 %	0,94	0,96	0,93
4	0,73	92 %	0,95	0,98	0,93
5	0,76	91 %	0,95	0,96	0,94
>5	0,87	94 %	0,97	0,96	0,98

Tabla 4.8: Criterios de evaluación por semestre para el caso de uso en la iteración 3

4.3.2. Aplicación del modelo a todas las carreras de la ESPOL

Debido a los resultados alcanzados en el caso de uso donde el modelo obtuvo un porcentaje de precisión alrededor del 94 %, se decidió aplicar el modelo para todos las carreras impartidas por ESPOL. Los datos usados representan el historial académico de 30.576 estudiantes matriculados a partir del primer término académico del año 2000 hasta el primer término académico del año 2020, distribuidos en 66 carreras.

Como resultado del pre-procesamiento de los códigos de las carreras y la verificación de la cantidad de estudiantes por carrera, se obtuvo un total de de 46 carreras de pregrado (30 ingenierías, 10 licenciaturas, 6 tecnológica) que representan un total de 28932 estudiantes usados por el modelo predictivo.

El modelo predice la retención académica carrera por carrera. Para cada grupo de semestres se calcula las características de rendimiento académico

y, en consecuencia, se aplican los métodos de selección de características para que el modelo prediga solo con las características más importantes. Para luego definir el conjunto de datos para el entrenamiento y las pruebas. Las características ingresan al modelo para calcular el porcentaje de retención académica, el porcentaje de influencia de las características, y los percentiles.

Al realizar una inspección manual de los resultados del modelo con respecto a los valores que tomaron las característica y los indicadores a nivel de carrera, se encontraron casos atípicos. Por ejemplo las características de tipo promedio eran iguales a cero, y como salida del modelo una alta probabilidad de retención. Los datos de estos estudiantes (259) no fueron utilizados durante el entrenamiento del modelo. También se encontró casos donde las características de tipo promedio eran superiores a 8 y con porcentaje de materias Aprobadas mayor a 95 % pero etiquetados como estudiantes *Desertores* (1508), estos también fueron descartados del entrenamiento del modelo. Otro caso encontrado donde descartamos estudiantes, fue de aquellos(53) donde las características de tipo promedio eran mayor a los indicadores de tipo promedio de la carrera, y el modelo los etiquetaba como *Desertores*. Estos problemas no surgieron en las iteraciones anteriores porque para ese entonces no todas las carreras terminaban su proceso de actualización y acreditación.

Una vez resueltos los casos atípicos se ejecutó nuevamente la predicción carrera por carrera. La Tabla 4.9 se muestra un resumen de los cinco criterios de evaluación del modelo para las mismas 10 carreras de pregrado que fueron seleccionadas en la iteración 2.

Los resultados de las métricas de evaluación del modelo mejoraron con respecto a los resultado de las iteraciones anteriores. Podemos ver que las

Código carrera	Carrera	*AUC*	*Accuracy*	*F1*	*Recall*	*Precision*
CI002	Ingeniería Química	0,58	93	0,97	0,99	0,94
CI003	Minas	0,86	90	0,93	0,94	0,92
CI004	Petróleos	0,80	91	0,94	0,96	0,94
CI007	Mecánica	0,80	89	0,94	0,95	0,93
CI009	Logistica y Transporte	0.73	87	0,92	0.94	0.91
CI013	Computación	0,74	89	0,94	0,96	0,92
LI002	Economía	0,80	91	0,94	0,94	0,95
LI005	Diseño Gráfico	0,75	94	0,96	0,98	0,95
LI007	Administración de Empresas	0,63	88	0.94	0,97	0,91
LI009	Turismo	0,58	89	0,94	0,95	0,93

Tabla 4.9: Criterios de evaluación de las 10 carreras seleccionadas en la iteración 2, con estudiantes matriculados durante el primer término académico del año 2000 hasta el primer término académico del año 2020 para la iteración 3

características de rendimiento en conjunto con el factor económico, al ser evaluados por los métodos de selección de las características influye al momento de predecir. A continuación, se muestran los resultados obtenidos tras analizar la importancia de las características en cada uno de los modelos para las distintas carreras y semestres. Se describe la frecuencia de uso y los pesos que han obtenido las características en los diferentes modelos.

La tabla 4.10 indica que las características Promedio de todas las materias Aprobadas y Reprobadas, Porcentaje de materias aprobadas, Factor económico y Promedio ponderado se utilizan en todos los modelos de las carreras en todos los semestres. Se puede observar que la variable: Porcentaje de materias reprobadas solo se utiliza en el 1,1 % de todos los modelos, debido a su alta relación con la variable Porcentaje de materias aprobadas. Con respecto a los pesos promedio de las variables, las cuatro variables uti-

Características	Frecuencia de uso	Peso de importancia promedio
Promedio de todas las materias Aprobadas y Reprobadas	100 %	0.28
Porcentaje de materias aprobadas	100 %	0.27
Factor económico	100 %	0.16
Promedio ponderado	100 %	0.16
Número de materias de segunda matrícula	87,91 %	0.07
Tiempo entre semestres	80,22 %	0.02
Número de materias de tercera matrícula	75,82 %	0.05
Porcentaje de materias reprobadas	1,1 %	0.02
Porcentaje de materias anuladas	39,56 %	0.19

Tabla 4.10: Resultados generales de la importancia de las características

lizadas en todos los modelos tienen los pesos más altos.

La Figura 4.6 muestra los pesos medios de los modelos para todas las carreras segmentadas por semestre. En los estudiantes de primer año las cuatro características con mayor peso son las mismas cuatro que fueron las mas usadas por el modelo. A medida que el alumno avanza en su carrera, se incluyen otras características como importantes en el modelo. Por ejemplo, cuando el estudiante está en el sexto semestres o más, la características Porcentaje de materias reprobadas tiene un peso importante en el modelo mientras que la variable Factor económico tiene un peso de importancia baja.

La Figura 4.7 muestra el comportamiento de las características para 10 carreras seleccionadas de manera aleatoria (8 carreras de Ingeniería y 2 Licenciaturas). Se puede observar que existe un patrón similar entre las características que tienen más peso entre todos los modelos. Sin embargo, se

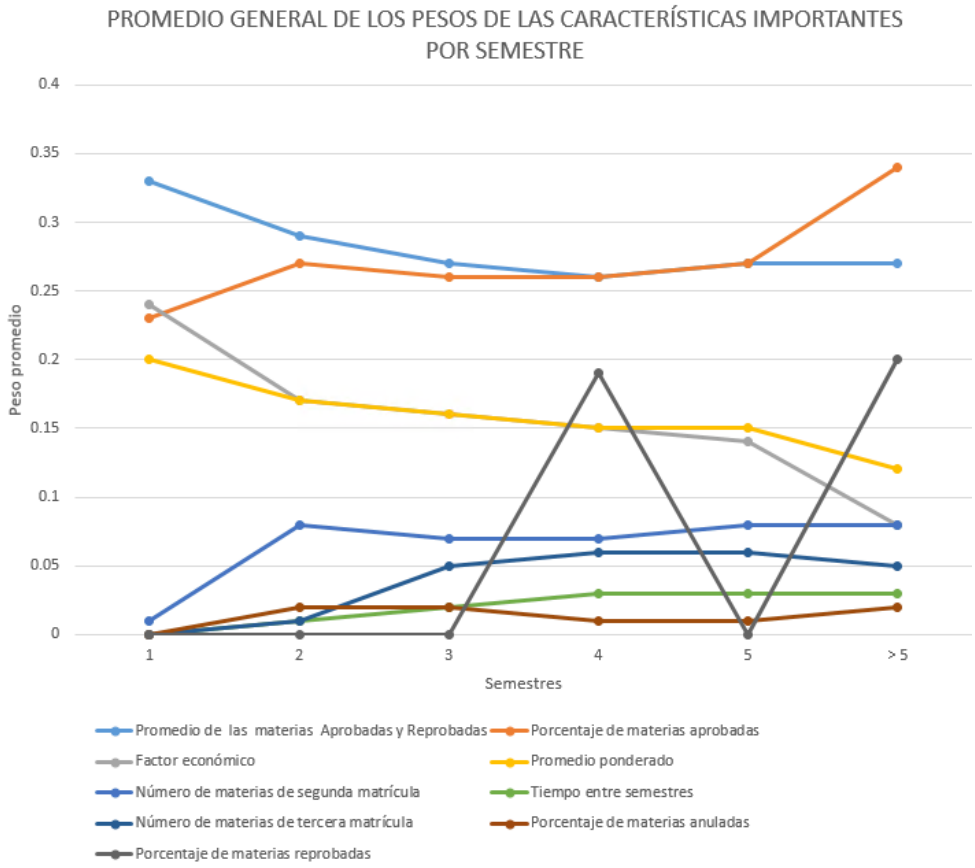

Figura 4.6: Promedio general del peso de las características de entra al modelo por semestres

puede ver que la característica Porcentaje de materias anuladas no fue importante para la carrera de ingeniería Mecánica (CI007), a diferencia de las demás carreras de Ingeniería y Licenciatura

4.3.3. Visualización

Al igual que la visualización en la iteración 2, el panel de visualización presenta el nivel de retención académica de los estudiantes. Estuvo disponible para todos los consejeros durante las consejerías del segundo término académico del 2020. El panel fue rediseñado cambiando el orden de la escala

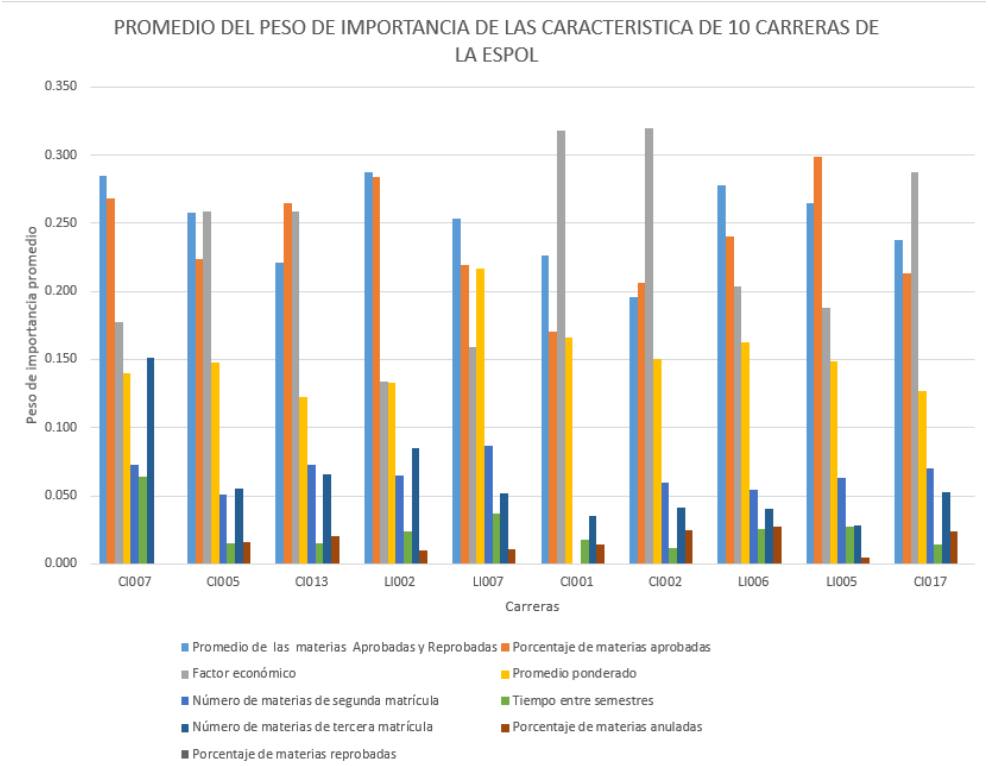

Figura 4.7: Promedio general del peso de importancia las características de entra al modelo para 10 carreras de la ESPOL

de colores que indica el nivel de retención, ordenado de izquierda a derecha como: rojo (muy bajo), naranja (bajo), amarillo (moderado), verde claro (alto) y verde oscuro (muy alto) (Figura 4.8).

Una vez que los consejeros hacen clic en el enlace, aparece una ventana emergente (Figura 4.9), mostrando la información detallada de las características de entrada más influyentes en el modelo. La información sigue mostrándose como un gráfico de barras, ordenado de mayor a menor dependiendo el porcentaje de influencia de las características. Al hacer clic en cada barra del gráfico, se muestra una explicación en la parte inferior derecha, que describe el valor que tomó la característica y de ser caso el indicador a nivel de carrera. La parte inferior de la ventana emergente muestra un texto indicando que

Figura 4.8: Panel de visualización incorporado en el Sistema de Consejerías durante la iteración 3

cuáles son las características socio-demográficas que no son consideradas en la predicción.

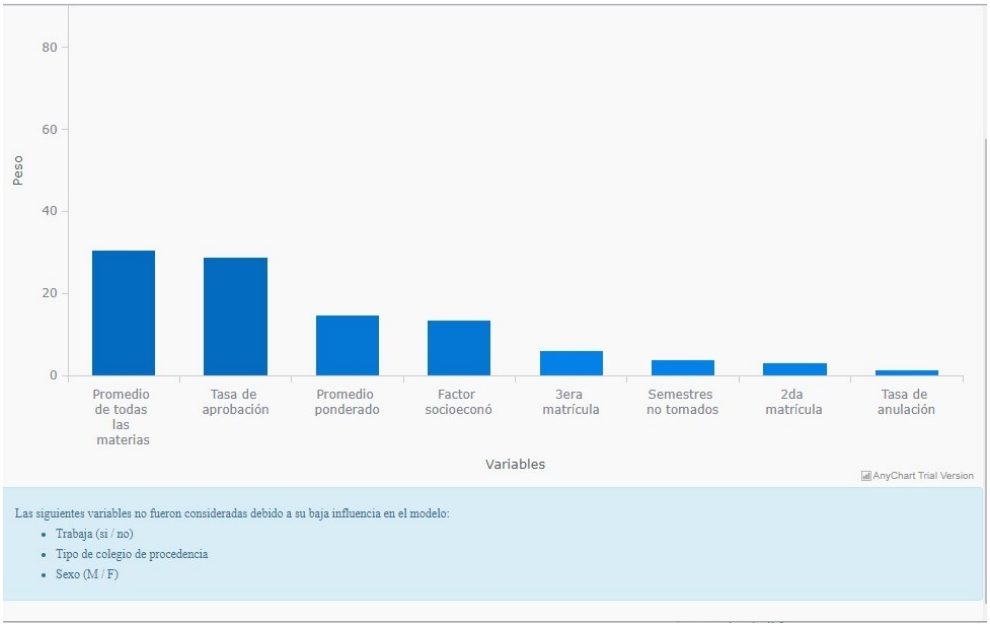

Figura 4.9: Ventana emergente que muestra información de las características de ingresadas al modelo en la iteración 3

Este capítulo al igual que el anterior describió la generación del modelo predictivo usando los datos académicos de la ESPOL, que derivó en tres distintas iteraciones. La explicación de la estrategia para la generación del modelo es común a cada una de las iteraciones de este trabajo. Aunque el modelo comienza prediciendo deserción académica y finaliza prediciendo retención, los conceptos son complementarios. Por lo tanto, el modelo puede generar como resultado tanto la probabilidad de deserción como su contra parte la probabilidad de retención. Se explicó también las diferencias que existen en cada iteración como la toma de decisiones, el proceso de diseño de características, las características de cada iteración y su justificación de uso o descarte. Se describieron los casos atípicos y su resolución, los procesos de validación y evaluación para cada iteración respecto al rendimiento del modelo predictivo. También, se describieron los resultados de la evaluación de las visualizaciones al incorporar la salida del modelo en el Sistema de Consejerías de la ESPOL.

CAPÍTULO 5

5. Discusión, trabajo futuro y conclusiones

Este capítulo presenta un resumen de los hallazgos de esta investigación e interpreta los resultados y las implicaciones generales. También resume las limitaciones de este trabajo y cómo fueron abordadas. Detalla el trabajo futuro que se pretende realizar. Y finalmente, describe las conclusiones y respuestas a la preguntas de investigación planteadas.

5.1. Discusión

Los datos disponibles sobre la información académica de los estudiantes permitió crear un modelo que permita predecir si un estudiante tiende o no

a desertar de una carrera. Los capítulos anteriores presentaron y discutieron varios hallazgos en torno al diseño de características, la implementación del modelo y la incorporación de los resultados del modelo en la herramienta web del Sistema de Consejerías de la ESPOL. Esta tesis describió tres iteraciones distintas pero con un modelo mayormente común para cada una.

Durante la iteración 1, el historial académico y las evaluaciones realizadas respecto al diseño de características, sirvieron como base para generar características de rendimiento académico distintas a las que se usan comúnmente en la mayoría de los estudios sobre deserción académica. Por ejemplo, en este trabajo se promediaron y combinaron las notas finales de las materias, independientemente si eran Aprobadas o Reprobadas, para generar una característica de rendimiento de tipo promedio. A diferencia de otros estudios, donde las notas a veces son valores cualitativos y difícilmente pueden ser promediadas entre ellas.

El resultado alcanzado en cuanto a precisión del modelo en la iteración 1 de manera general para todas las carreras—70 %, no es del todo “bajo”. Aunque indicó una señal de que el modelo debía ser mejorado. Este porcentaje de precisión puede deberse a que se predijeron 66 carreras de manera individual. En consecuencia, podían existir carreras que no tenían suficientes datos de entrenamiento para predecir; por lo tanto, obtenían una baja precisión de predicción. A diferencia de la iteración 2, donde recién se pudo notar que las carreras podían ser asociadas a carreras padres. Esto provocó tener menos carreras que predecir, pero con más datos de entrenamiento por carrera debido a la asociación. Al realizar la asociación de carreras en la iteración 2, la precisión del modelo mejoró con respecto a la iteración 1, alcanzando una precisión de alrededor del 84 % de manera general para todas las carreras.

Por otro lado, para la iteración 1 y 2 las características socio demográficas fueron descartadas según los resultado obtenidos por la matriz de correlación, que mostraba una baja correlación de estas características. Mientras que, para todas las iteraciones se usaron características rendimiento académico para predecir tanto la deserción de los estudiantes, como su contraparte la retención. En cuanto a la característica socio-demográfica del factor económico del estudiante siempre se la quiso usar, porque literatura académica demuestra que es un factor influyente en la deserción. Sin embargo, por los percances descritos en los capítulos anteriores no se pudo usar esta información sino hasta la iteración 3. Una vez que fue proveída la información del factor económico, inmediatamente fue incorporado al modelo predictivo. Al ingresar esta característica el porcentaje de precisión mejoró, alcanzando un promedio del 90 % precisión de manera general para todas las carreras. Al igual que otros estudios, este trabajo demostró que al añadir el estado financiero de los estudiantes, esta información es influyente en la predicción de la deserción.

Como se pudo observar en la iteración 3, las características de entrada al modelo no interactúan de la misma manera para todos los semestres ni para todas las carreras, debido a los métodos de selección de las características. Por ejemplo, habían características que eran descartadas para unas carreras al momento de predecir; mientras que para otras carreras era utilizadas, tales como el Porcentaje de materias Reprobadas, Porcentaje de materias Anuladas. Sin embargo, durante la predicción existieron características que fueron común para todas las carreras en todos los semestres, siendo: el Promedio de todas las materias Aprobadas y Reprobadas, el Porcentaje de materias aprobadas, el Factor económico y el Promedio ponderado. Estas características fueron las más usadas y al mismo tiempo obtuvieron los pesos más altos

de importancia.

Con respecto a los resultados de la evaluación de los paneles de visualización de deserción y retención académica, la iteración 1 no es comparable con la iteración 2. Esto se debe a que la visualización de la iteración 1 presenta algunas desventajas. Esta visualización se encontraba en la herramienta web del Sistema de Consejerías en un lugar al que los consejeros no acceden con mucha frecuencia. A diferencia de visualizaciones de las otras iteraciones, donde estos paneles fueron ubicados en la pagina principal de la herramienta. Durante la iteración 1 la visualización estuvo disponible por poco tiempo (1 semana) y solo para los consejeros que tenían estudiantes con porcentaje de deserción mayor o igual al 50 %. A diferencia de las otras iteraciones, donde las visualizaciones estuvieron disponibles durante todo el periodo de consejerías y para todos los consejeros.

En cuanto a la evaluación de la visualización de la iteración 2, esta tuvo bastante acogida por parte de los consejeros. Se encontró que interactuaron con la visualización incluso para conocer más información respecto a las razones detrás de los resultados de la predicción. Se puede suponer que el haber incluido el nivel de precisión de la predicción en la iteración 2 y 3, ayudó a los consejeros a confiar en la información mostrada. Sobretodo en la iteración 3, donde los niveles de precisión son altos. Por otro lado, incluir un pregunta cerrada para saber si los consejeros consideraban consistente la información presentada, ayuda a reconocer si tanto el modelo predictivo como la visualización deben ser mejorados. Por otro lado, los resultados de las visualizaciones contribuyen positivamente a una visión limitada que tiene la evaluación de la utilidad de los paneles.

Los resultados alcanzados en este trabajo extienden investigación previa en este espacio y supera las limitaciones de otros trabajos en varios sentidos.

Utiliza información más allá de solo los primeros semestres, ya que analiza al estudiante a lo largo de toda su carrera de pregrado. Aunque otros estudios utilizan diferente tipos de características, esta investigación utiliza datos mayormente académicos con una característica económica del estudiante. Mientras otros estudios se enfocan en una sola carrera o curso, este trabajo presenta un modelo que fue probado en diferentes carreras de pregrado.

5.1.1. Limitaciones

Las limitaciones surgen mayormente por los datos recopilados. Por ejemplo, cuando un estudiante se retira de sus estudios por un período largo y su carreras de pregrado cambia el código asignado. Por lo que, al retomar sus estudios, es modificado en la base de datos, asignándole el nuevo código de carrera, pero con todo su historial académico como un único semestre. La porción de esta población estudiantil con más de 10 asignaturas cursadas en un solo semestre fue un promedio de 6 % para cada iteración generada. Esta limitación se abordó eliminando a estos estudiantes del entrenamiento modelo, porque generaban alteraciones en los resultados de predicción.

El promedio de nota final obtenida por los estudiantes, usada por esta tesis, no es un indicadores confiables de rendimiento, ya que no necesariamente captura si un estudiante es inteligente o no. Solo captura si el estudiantes obtuvo una buena o mala nota, que no es lo mismo. Existen otros factores externos que pueden moldear la confianza académica de los estudiantes y, posteriormente, afectar sus resultados y éxito académico [50]. Por eso, los resultados del modelo no se muestran a los estudiantes, sino a los consejeros, ya que en sus consejerías pueden considerar otros factores como problemas familiares, de salud, sentimentales, entre otros, que el modelo de predicción no esta considerando.

Otra limitación encontrada se generó al realizar las validaciones de resultados en cada iteración, puesto que se seleccionaban estudiantes al azar para comparar los valores que tomaban las características con el resultado obtenido por el modelo. Lo que provocó invertir más tiempo en las pruebas y en solucionar los escenarios atípicos que surgían en ese momento. Por ejemplo, recién en la tercera iteración se encontró el escenario donde los estudiantes etiquetados como Desertores, tenían características de rendimiento de tipo promedio superior a 8, provocaban alteraciones en los resultados de la predicción, porque significa una contradicción para el modelo. Estos estudiantes fueron descartados del entrenamiento del modelo.

Otro problema presentado en cuanto a las evaluaciones del modelo y sobretodo de las visualizaciones en cada iteración, es que estos no pasaron por un periodo de pruebas pre-test riguroso con usuarios finales para todas las carreras antes de ser usado en las semanas de consejerías que la ESPOL tiene de manera reglamentaria. Lo que se aconseja es realizar pruebas pre y post test para evitar posibles errores o los casos especiales que surjan durante el proceso de consejerías.

5.2. Trabajo futuro

Como trabajo futuro se tiene como objetivo mejorar las técnicas en el modelo de predicción propuesto y utilizar otros algoritmos de ML para compara la eficiencia del modelo. Además, se plantea utilizar este modelo en carreras de posgrado y en conjuntos de datos académicos de otros tipos de universidades donde las tasas de deserción tienden a ser mucho más altas que las de la ESPOL.

Los cambios realizados para la iteración 3 sobretodo lo que tiene que

ver con la visualización aún no han sido evaluados, por lo que se pretende realizar las evaluaciones para una siguiente iteración. También se plantea mejorar las visualizaciones por medio del uso de técnicas de visualización con modelos explicativos.

Por otra parte, los resultados de las visualizaciones mostraron que los consejeros académicos sentían seguridad en la información mostrada, ya que la visualización incorporaba un indicador de nivel de confianza de la predicción. Sin embargo, encontramos que se requiere más transparencia en el cálculo de los percentiles para aumentar la confianza. En este sentido, se plantea aplicar métodos de ML para ubicar a los estudiantes con mayor precisión en su respectivo nivel. Por lo tanto, se espera que una nueva iteración mida estos cambios.

Como último punto, se debe estudiar el efecto psicológico de mostrar a los consejeros un panel de deserción frente a mostrar un panel de retención.

5.3. Conclusiones

La predicción de la deserción es un requisito previo esencial para realizar intervenciones de estudiantes en riesgo. Para abordar este problema, y ayudar a la ESPOL a reducir sus niveles de deserción académica, esta investigación propuso abordar tres preguntas de investigación (enumeradas en la sección 1.2 del capítulo 1. A continuación, se resumen las respuestas a cada una de las preguntas planteadas.

- **Q1: ¿Qué tipo de características de la historia académica de los estudiantes de la ESPOL tienen un rol importante en la predicción de deserción académica?**

  En esta investigación se realizaron varios análisis de las características

de entrada, proponiendo un modelo predictivo que en su mayor parte usó características de rendimiento académico. El modelo analiza las características de rendimiento semestre a semestre con el objetivo de capturar toda la historia previa de los estudiantes, para que el modelo pueda trabajar con cada conjunto de semestres, en lugar de instancias individuales, ver las secciones 3.3 y 3.3.1.2.

De todas las características de rendimiento, las que influyeron mayormente en cada iteración de este trabajo, fueron las de tipo promedio y porcentuales, específicamente *Promedio de todas las materias Aprobadas y Reprobadas* y el *Porcentaje de materias aprobadas*, puesto que su frecuencia de uso era de alrededor del 100 %, ver capítulo 4. En cuanto a la característica socio-demográfica del factor económico también alcanzó el 100 % de frecuencia de uso. De hecho, incorporar esta característica mejoró la precisión de la predicción. Lo que demuestra que el factor económico es una característica influyente en la decisión de un estudiante de continuar o no con sus estudios.

- **Q2: ¿Cómo podrían comunicarse los resultados de un modelo predictivo de deserción académica a los consejeros del sistema de consejerías académicas de la ESPOL?**

  Normalmente, cuando se usan modelos de predicción y algoritmos de ML en general, es difícil revelar al usuario final por qué el modelo predice lo que predice. En este caso, para el problema de deserción, es necesario entender las razones detrás de las predicciones por lo que se utilizan paneles de visualización como herramienta de soporte.

  En este trabajo los paneles de visualización implementados fueron incorporados en la herramienta web del Sistema de Consejerías de la

ESPOL, como un instrumento de ayuda para los consejeros. Se usó una escala de colores para representar de manera fácil el nivel de deserción/retención de los estudiantes. Se incorporó indicadores de nivel de precisión de la predicción del modelo para generar confianza en los consejeros. También, muestra un panel que permite ver en detalle las características usadas por el modelo predictivo y presentarlas como las razones por las que se obtuvo del nivel de deserción/retención presentado. Estos paneles de visualizaciones permiten a los consejeros ver información relevante de manera simplificada y entendible. Además, les ofrece la posibilidad de tomar decisiones más acertadas en base a la información proporcionada, ver las subsecciones de visualización de cada iteración del capítulo 4 para mayor información.

- **Q3: ¿Cuál es el beneficio percibido por los consejeros de la ESPOL al ser expuestos a predicciones de deserción académica y a explicaciones sobre éstas?**

  Este investigación presentó un panel de visualización de deserción académica, pero también presentó un panel de retención académica como contra parte de la deserción.

  Basados en las evaluaciones y las respuestas a la pregunta que se incorporó en el panel de visualización, los resultados mostraron que los consejeros consideraron consistente la información presentada sobre la predicción y las características mostradas como justificación de la predicción. Los consejeros encontraron estas visualizaciones de gran ayuda al momento de sugerir materias a los estudiantes, porque podían ver el nivel de retención y la carga académica de un estudiante en una sola presentación. Permitiéndoles indagar más en la situación del

estudiante y recomendar materias de acuerdo a su estado académico.

Nuestro modelo fue probado con datos reales de ESPOL de las diferentes carreras de pregrado. Los resultados experimentales mostraron que el modelo tiene buenos resultados comparado con las propuestas existentes en términos de características de entrada y precisión. En consecuencia, queda demostrado que el modelo propuesto en esta tesis permite predecir con una alta precisión. Al mismo tiempo, puede predecir la intención de deserción lo suficientemente temprano, es capaz de predecir incluso para estudiantes que terminan su primer semestre. Hasta donde se conoce, esta es la primera propuesta de un modelo predictivo capaz de predecir deserción incluso para estudiantes que tienen solo un semestre académico. El hecho de que se mencionen específicamente las características que se usaron, puede ayudar a otros estudios a replicar este modelo o adaptarlo a sus necesidades. Por otro lado, estos resultados preliminares pueden servir como pautas para diseñar y evaluar mejor un panel de visualización de abandono o retención académica. Finalmente, el modelo predictivo desarrollado en esta tesis está incorporado en la herramienta web del Sistemas de Consejerías de la ESPOL y ha sido usado por los consejeros en tres términos académicos ordinarios, desde el segundo término del 2019 hasta la actualidad.

BIBLIOGRAFÍA

[1] Jorge Maldonado-Mahauad, Isabel Hilliger, Tinne De Laet, Martijn Millecamp, Katrien Verbert, Xavier Ochoa, and Mar Pérez-Sanagustín. The lala project: Building capacity to use learning analytics to improve higher education in latin america. In *companion proceedings of the 8th international learning analytics & knowledge conference*, pages 630–637, 2018.

[2] Vanessa Heredia-Jimenez, Alberto Jimenez, Margarita Ortiz-Rojas, Jon Imaz Marın, Pedro Manuel Moreno-Marcos, Pedro J Munoz-Merino, and Carlos Delgado Kloos. An early warning dropout model in higher education degree programs: A case study in ecuador.

[3] Vanessa Heredia-Jimenez, Jhony Yaguana, Alberto Jimenez, and Margarita Ortiz-Rojas. Lessons learned of the design and evaluation of an academic dropout and retention dashboard: A case study in ecuador.

[4] Vanessa Heredia-Jimenez, Irving Valeriano, Danny Torres, Alberto Jimenez, and Margarita Ortiz-Rojas. Understanding users' needs: evaluating alearning analytics dashboard.

[5] OECD. *Education at a Glance 2019*. 2019.

[6] María Marta Ferreyra, Ciro Avitabile, Javier Botero Álvarez, Francisco Haimovich Paz, and Sergio Urzúa. At a crossroads : Higher education in latin america and the caribbean. 2017.

[7] Diario El telégrafo. La deserción universitaria bordea el 40 %. Technical report, Diario El telégrafo, 2016.

[8] Diario El telégrafo. La deserción universitaria tiene cuatro causas. Technical report, Diario El telégrafo, 2016.

[9] Johan J Vossensteyn, Andrea Kottmann, Benjamin WA Jongbloed, Franciscus Kaiser, Leon Cremonini, Bjorn Stensaker, Elisabeth Hovdhaugen, and Sabine Wollscheid. Dropout and completion in higher education in europe: Main report. 2015.

[10] Rong Chen. Financial aid and student dropout in higher education: A heterogeneous research approach. In *Higher education*, pages 209–239. Springer, 2008.

[11] Viviana Catalano, Damián Eduardo Murcia Pérez, Fabio Escudero, Paula Daiana Gerlo, and Antonio Pantoja. Influence of a tutorial system based on the use of ict in the decrease of dropout and academic failureof first-year students of the veterinary career of the juan agustín maza university. In *I Jornadas de Inclusión de Tecnologías Digitales en la Educación Veterinaria (La Plata, 2018)*, 2018.

[12] S Suganya and V Narayani. Analysis of students dropout forecasting using data mining. In *3rd Internaational Conference on Lastest Trends in Engineering, Science, Humanities and Management*, 2017.

[13] Marian Barbu, Ramon Vilanova, José Lopez Vicario, Maria João Pereira, Paulo Alves, Michal Podpora, Miguel Ángel Prada, Antonio Morán, Aldo

Torreburno, Simona Marin, et al. Data mining tool for academic data exploitation: literature review and first architecture proposal. *Projecto SPEET-Student Profile for Enhancing Engineering Tutoring*, 2017.

[14] Joel P Murphy and Shirley A Murphy. Get ready, get in, get through: Factors that influence latino college student success. *Journal of Latinos and Education*, 17(1):3–17, 2018.

[15] David Baneres, M Elena Rodriguez-Gonzalez, and Montse Serra. An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies*, 2019.

[16] Fernando Jimenez, Alessia Paoletti, Gracia Sanchez, and Guido Sciavicco. Predicting the risk of academic dropout with temporal multi-objective optimization. *IEEE Transactions on Learning Technologies*, 2019.

[17] Fisnik Dalipi, Ali Shariq Imran, and Zenun Kastrati. Mooc dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1007–1014. IEEE, 2018.

[18] Amelec Viloria and Omar Bonerge Pineda Lezama. Mixture structural equation models for classifying university student dropout in latin america. *Procedia Computer Science*, 160:629–634, 2019.

[19] R Lucio, M Campbell, M Detres, and H Johnson. Using dashboards to engage faculty in improving academic programs and courses. In *INTED2018: 12th International Technology, Education, and Development*

Conference—Conference proceedings, pages 1844–1852. IATED Academy Valencia, Spain, 2018.

[20] Francisco Gutiérrez, Karsten Seipp, Xavier Ochoa, Katherine Chiluiza, Tinne De Laet, and Katrien Verbert. Lada: A learning analytics dashboard for academic advising. *Computers in Human Behavior*, page 105826, 2018.

[21] Andreas F Gkontzis, CV Karachristos, F Lazarinis, EC Stavropoulos, VS Verykios, G Ubachs, and L Konings. A holistic view on academic wide data through learning analytics dashboards. In *Conference Proceedings: The Online, Open and Flexible Higher Education Conference*, pages 25–27, 2017.

[22] German Dempere Guillermo. Predictive data driven dashboard as an academic guidance support platform for mentors. 2018.

[23] Ashish Dutt, Maizatul Akmar Ismail, and Tutut Herawan. A systematic review on educational data mining. *Ieee Access*, 5:15991–16005, 2017.

[24] Mohammad Khalil and Martin Ebner. Learning analytics: principles and constraints. In *EdMedia+ Innovate Learning*, pages 1789–1799. Association for the Advancement of Computing in Education (AACE), 2015.

[25] Charles Lang, George Siemens, Alyssa Wise, and Dragan Gasevic. *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research, 2017.

[26] Dragan Gašević, Shane Dawson, and George Siemens. Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71, 2015.

[27] RSJD Baker, Erik Duval, John Stamper, David Wiley, and S Buckingham Shum. Panel: educational data mining meets learning analytics.

In *Proceedings Of International Conference On Learning Analytics & Knowledge*, volume 2, 2012.

[28] George Siemens and Ryan SJ d Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012.

[29] Ryan Shaun Baker and Paul Salvador Inventado. Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer, 2014.

[30] Luz Elba Torres Guevara. *Retención estudiantil en la educación superior: revisión de la literatura y elementos de un modelo para el contexto colombiano*. Editorial Pontificia Universidad Javeriana, 2012.

[31] Siu-Man Raymond Ting and R Man. Predicting academic success of first-year engineering students from standardized test scores and psychosocial variables. *International Journal of Engineering Education*, 17(1):75–80, 2001.

[32] Amelec Jesus Viloria Silva and Alexander Parody. Methodology for obtaining a predictive model academic performance of students from first partial note and percentage of absence. 2016.

[33] Rutger Kappe and Henk van der Flier. Predicting academic success in higher education: what's more important than being smart? *European Journal of Psychology of Education*, 27(4):605–619, 2012.

[34] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*, 2016.

[35] Sattar Ameri, Mahtab J Fard, Ratna B Chinnam, and Chandan K Reddy. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 903–912. ACM, 2016.

[36] Farshid Marbouti, Heidi A Diefes-Dux, and Krishna Madhavan. Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103:1–15, 2016.

[37] Elvis Roberto Sinchi Nacipucha and Gliceria Petrona Gómez Ceballos. Access and desertion in universities. Financing alternatives. *ALTERIDAD. Revista de Educación*, 13(2):274–287, 12 2018.

[38] UNIVERSIDAD CARLOS 111 DE MADRID. ResoluciÓn del rector relativa al procedimiento de determinaciÓn de la nota media del expediente de los alumnos de la universidad carlos iii de madrid.

[39] Boletín Oficial del Estado. Real decreto 1125/2003, de 5 de septiembre, por el que se establece el sistema europeo de créditos y el sistema de calificaciones en las titulaciones universitarias de carácter oficial y validez en todo el territorio nacional. *Boletín Oficial del Estado*, 18:34355–34356, 2003.

[40] Pizanán Alvarracín and Jairo Daniel. Identificación del perfil de egreso correspondiente a la licenciatura de la carrera de laboratorio clínico e histotecnológico de la universidad central del ecuador periodo 2017-2022. 2016.

[41] Ismael Ahrazem Dfuf. *Análisis de Sensibilidad Mediante Random Forest*. PhD thesis, Universidad Politécnica de Madrid, 2018.

[42] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[43] T Cevallos. Cuadernos del contrato social por la educación. *Cuaderno*, 10:34–46, 2014.

[44] Ram C Sharma, Keitarou Hara, and Hidetake Hirayama. A machine learning and cross-validation approach for the discrimination of vegetation physiognomic types using satellite based multispectral and multitemporal data. *Scientifica*, 2017, 2017.

[45] Jing Chen, Jun Feng, Xia Sun, Nannan Wu, Zhengzheng Yang, and Sushing Chen. Mooc dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Mathematical Problems in Engineering*, 2019, 2019.

[46] Lynne D Roberts, Joel A Howell, and Kristen Seaman. Give me a customizable dashboard: Personalized learning analytics dashboards in higher education. *Technology, Knowledge and Learning*, 22(3):317–333, 2017.

[47] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.

[48] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017.

[49] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[50] Susan Bickerstaff, Melissa Barragan, and Zawadi Rucks-Ahidiana. Experiences of earned success: Community college students' shifts in co-

llege confidence. *International Journal of Teaching and Learning in Higher Education*, 29(3):501–510, 2017.