ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Modelo predictivo para admisión en una institución de educación superior en las etapas de asignación y aceptación de cupo.

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

Javier David Díaz Romero

GUAYAQUIL - ECUADOR

Año: 2024

DECLARACIÓN EXPRESA

Yo, Javier David Díaz Romero acuerdo y reconozco que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al autor que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 09 de diciembre del 2024.

Firmado electrónicamente por:
JAVIER DAVID DIAZ ROMERO

Javier David Díaz Romero

EVALUADORES

Firmado electrónicamente por:
EDUARDO SEGUNDO
CRUZ RAMIREZ

Eduardo Segundo Cruz Ramírez

Tutor de proyecto

Firmado electrónicamente por:
ALLAN ROBERTO
AVENDANO SUDARIO

Allan Roberto Avendano Sudario

Revisor de proyecto

RESUMEN

El proceso de selección de estudiantes ha sido un factor crítico de éxito para las instituciones de educación superior. Históricamente, las universidades han buscado admitir a estudiantes calificados y comprometidos con sus programas académicos. Sin embargo, desde el 2023 con la implementación del reglamento del sistema nacional de nivelación y admisión en Ecuador, la presión por alcanzar metas de matrícula, igualdad de oportunidades, equidad, entre otros, ha llevado a que instituciones no tengan autonomía en el método de asignación de cupos y ha llevado a que se asignen cupos que podrían ser rechazados. Este fenómeno, ha generado desafíos en términos de calidad académica y sostenibilidad financiera.

Por otro lado, cuando los estudiantes recibían múltiples ofertas de admisión, su decisión final se veía influenciada por una variedad de factores, como la reputación institucional, la calidad de los programas académicos y la disponibilidad de recursos. En este contexto, las instituciones educativas han reconocido la importancia de comprender mejor el comportamiento de los estudiantes durante el proceso de selección.

Al analizar estos desafíos y oportunidades, se ha evidenciado la necesidad de desarrollar herramientas y modelos que permitan a las universidades optimizar sus procesos de admisión. A través de una mejor comprensión de los factores que influyen en las decisiones de los estudiantes, las instituciones pueden tomar decisiones más informadas y estratégicas, mejorando tanto la calidad de su alumnado como su posición competitiva en el mercado educativo.

Con la ayuda de modelos de aprendizaje automático, CatBoost, Random Forest y Multilayer perceptron fueron implementados comparando su rendimiento, relación con las variables del dataset y se logró implementar un modelo predictivo para la etapa de aceptación de cupos con una exactitud mayor al 90% y priorizando la necesidad de la universidad.

Palabras Clave: admisión universidad, modelo predictivo, aceptación cupo.

ABSTRACT

The student selection process has been a critical success factor for higher education institutions. Historically, universities have sought to admit qualified students who are committed to their academic programs. However, since 2023 with the implementation of the regulations of the national leveling and admission system in Ecuador, the pressure to achieve enrollment goals, equal opportunities, equity, among others, has led to institutions not having autonomy in the method of assigning places and has led to the assignment of places that could be rejected. This phenomenon has generated challenges in terms of academic quality and financial sustainability.

On the other hand, when students received multiple admission offers, their final decision was influenced by a variety of factors, such as institutional reputation, the quality of academic programs, and the availability of resources. In this context, educational institutions have recognized the importance of better understanding student behavior during the selection process.

By analyzing these challenges and opportunities, the need to develop tools and models that allow universities to optimize their admission processes has become evident. By better understanding the factors that influence student decisions, institutions can make more informed and strategic decisions, improving both the quality of their students and their competitive position in the educational market.

With the help of machine learning models, CatBoost, Random Forest and Multilayer perceptron were implemented, compared their performance and relationship with the dataset variables, and a predictive model was implemented for the stage of acceptance of places more than 90% accuracy and prioritizing the needs of the university.

Keywords: university admission, predictive model, quota acceptance.

ÍNDICE GENERAL

ABREVIATURAS

ESPOL	Escuela Superior Politécnica del Litoral
LOES	Ley Orgánica de Educación Superior
SENESCYT	Secretaría de Educación Superior, Ciencia, Tecnología e Innovación
IES	Institución de Educación Superior
KDD	Knowledge Discovery in Databases
SPSS	Statistical Package for the Social Sciences
ML	Machine Learning
CNN	Convolutional neural network
SVN	Support vector machine
ROC	Receiver operating characteristic curve
TP	True positive
TN	True negative
FP	False positive
FN	False negative
SHAP	SHapley Additive exPlanations
CSS	Cascading Style Sheets

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. Introducción

1.1. Descripción del problema

Desde los cambios implementados por la Constitución de la República del Ecuador y la Ley Orgánica de Educación Superior (LOES) reformada en 2010 y 2019, el acceso a una institución de educación superior está vinculada con la aplicación de una evaluación como mecanismo de ingreso a las Universidades, Escuelas Politécnicas e Institutos Técnicos, Tecnológicos y Conservatorios públicos del país junto a las notas obtenidas durante los años de colegio y la situación de vulnerabilidad que presente. [1]

La codificación del reglamento del sistema nacional de admisión y nivelación mediante el acuerdo Nro. SENESCYT-SENESCYT-2023-0003-AC, “tiene por objeto regular, coordinar y monitorear el proceso de acceso para las y los aspirantes a la educación superior en las Universidades y Escuelas Politécnicas públicas, la aplicación de dicho reglamento es obligatorio para las IES antes mencionadas” [2]. El producto principal es el proceso de admisión, el cual finaliza con la etapa de aceptación o rechazo de cupos asignados.

Una IES que enfrenta un alto índice de rechazo de cupos asignados y una tasa de deserción significativa entre los estudiantes matriculados se encuentra ante una serie de desafíos complejos. El rechazo de cupos asignados implica una pérdida de oportunidades para llenar los programas académicos y puede generar una disminución en la cantidad de estudiantes matriculados. Esto, a su vez, puede afectar los ingresos de la institución y comprometer la sostenibilidad de ciertos programas o carreras universitarias.

Además, tener una gran cantidad de cupos rechazados o sin asignar, puede afectar la reputación de la universidad, si se percibe como una institución con baja demanda o con dificultades para llenar sus programas. Por otro lado, la alta tasa de rechazo de cupos puede indicar problemas en el proceso de admisión o en la información proporcionada a los postulantes. Esto podría deberse a una mala comunicación de los requisitos académicos, a la falta de claridad sobre las

oportunidades de desarrollo profesional o a discrepancias entre las expectativas de los estudiantes y la realidad de la oferta académica.

Una vez realizada la asignación de cupos, los postulantes tienen la potestad de aceptar o rechazar dichos cupos que fueron asignados a una de las carreras que eligieron en la etapa de postulación, en las cuales tuvieron desde 1 hasta 5 opciones dependiendo de la IES. [2] Para esto, tienen un periodo de tiempo para realizar esta etapa ya que una vez lo tenga no puede modificarlo o invalidarlo, sin embargo, es aquí donde su decisión se ve afectada por razones como que el título secundario ya no es suficiente para acceder a un puesto de trabajo. Además, existen investigaciones donde se proponen diferentes factores que alteran esta elección desde la situación económica hasta como se relacionan con su entorno familiar.

Luego que termina la etapa de aceptación de cupos, las IES evalúan si es necesario abrir otra etapa desde la postulación o asignación de cupos según la cantidad de cupos remanentes que existan, esto se ampara en el hecho que según el artículo 42 del reglamento mencionado al inicio, las Universidades y Escuelas Politécnicas públicas podrán desarrollar estrategias para lograr una asignación de cupos eficiente, con base en un análisis técnico del comportamiento de la demanda vigente, con la finalidad de agotar la oferta de cupos disponibles además del artículo 43 el cual indica que las Universidades y Escuelas Politécnicas públicas deberán establecer mecanismos que permitan la eficiente colocación de oferta remanente en los casos que corresponda en función del orden de asignación, puntaje final de postulación y libertad de elección de carrera. [2]

Debido a las políticas antes mencionadas, existe un ciclo al final del proceso de admisión entre las etapas de postulación, asignación y aceptación de cupo para intentar minimizar la cantidad de cupos rechazados o no usados, el cual en el caso de la IES que se toman los datos para este proyecto, este ciclo genera que el proceso de admisión tome más tiempo de lo programado, entre 4 a 6 semanas más. Además, crece la inseguridad entre los postulantes que no se les asigna un cupo y deben de esperar a que el ciclo antes mencionado inicie nuevamente. Debido a la competencia de obtener a la mayor cantidad de estudiante, el tiempo

es un factor crucial, entre más rápido el proceso de admisión se ejecute, mejor gestión y confianza genera entre los postulantes.

Es así como surge la necesidad de diseñar un modelo predictivo que permita a las instituciones de educación superior anticipar el comportamiento de los postulantes en la etapa de aceptación de cupo ya que, así se podrían ajustar las estrategias de captación y retención de estudiantes, optimizar el proceso de admisión y asignación de cupos, y promover políticas que reduzcan las tasas de rechazo de cupo por razones ajenas al desempeño académico.

1.2. Justificación del problema

El acceso a la educación superior es considerado como un derecho debido a que se considera como un pilar esencial de los derechos humanos, para el desarrollo personal, social y cultural [3]. Por esta razón, en el país se busca facilitar el acceso a universidades e institutos superiores, promoviendo políticas y sistemas de admisión más inclusivos para asegurar que todos los jóvenes, independientemente de su origen o condición, tengan la oportunidad de continuar su formación académica.

Actualmente, existen diversas políticas públicas para el ingreso a la educación superior basados en la autonomía universitaria, de manera que, se garantiza un cierto nivel de control para el manejo de sus procesos. Esto resalta su papel crucial en la consolidación del Estado constitucional de derechos y justicia, especialmente en la formación de ciudadanos conscientes de sus derechos y responsabilidades, al fomentar el uso de la razón y la proyección de su presente y futuro, estas instituciones contribuyen significativamente al desarrollo personal y social, promoviendo la producción e intercambio de conocimientos desde una perspectiva intercultural [4], este proyecto comprobará el nivel de incertidumbre que tenga la etapa de asignación de cupos ya que esto en base a las políticas públicas debería de ser un proceso igualitario para todos los postulantes.

La etapa de aceptación de cupo debe alinearse con las políticas públicas de educación superior, basadas en la autonomía universitaria, lo que permite a las instituciones gestionar de manera controlada y flexible los procesos de admisión. Este enfoque contribuye a la consolidación del Estado constitucional de

derechos y justicia, fundamental para cultivar una ciudadanía comprometida y consciente., al fomentar el uso de la razón y la proyección de su presente y futuro, estas instituciones contribuyen significativamente al desarrollo personal y social, promoviendo la producción e intercambio de conocimientos desde una perspectiva intercultural.

La investigación es de interés para el Estado, las universidades, las diferentes autoridades comprometidas con la educación y los jóvenes, ya que permite comprobar la eficiencia y homogeneidad de la asignación de cupos, optimizar el uso de los recursos y asegurar que los jóvenes reciban un cupo que se ajuste a sus necesidades y oportunidades educativas.

La alta tasa de rechazo de cupos sugiere la existencia de problemas subyacentes en el proceso de admisión, como una comunicación inadecuada de los requisitos académicos o una discrepancia entre las expectativas de los estudiantes y la realidad de la oferta académica. Además, factores externos como la situación económica de los estudiantes o las oportunidades laborales pueden influir en su decisión de aceptar o rechazar un cupo.

El marco normativo que rige el proceso de admisión, con sus políticas de acción afirmativa, orden de asignación y puntaje de postulación compuesto, introduce una complejidad adicional. Si bien estas políticas buscan garantizar la equidad y la eficiencia en la asignación de cupos, también pueden generar incertidumbre y prolongar el proceso de admisión.

La existencia de un ciclo iterativo entre las etapas de postulación, asignación y aceptación de cupos, con el objetivo de minimizar los cupos rechazados, genera una serie de inconvenientes. Por un lado, este ciclo prolonga significativamente el proceso de admisión, generando incertidumbre y ansiedad entre los postulantes. Por otro lado, la competencia por los cupos se intensifica, lo que puede llevar a una mayor presión sobre los estudiantes y afectar su bienestar.

En este contexto, el desarrollo de un modelo predictivo que permita anticipar la probabilidad de que un estudiante acepte o rechace un cupo asignado se presenta como una solución prometedora. Al identificar los factores que influyen en la decisión de los postulantes, la institución podrá tomar decisiones más

informadas sobre la asignación de cupos, optimizar sus estrategias de comunicación y mejorar la experiencia de los postulantes.

Debido a la metodología que se plantea, se tendrá el respaldo de datos reales y recientes de una de las IES con mayor cantidad de estudiantes en Ecuador, incluyendo información sobre la demanda de carreras, las características de los estudiantes y el contexto socioeconómico, así como de fuentes secundarias válidas y disponibles. Por tanto, la investigación es factible, ya que se cuenta con el tiempo para desarrollar el modelo y se dispone de un soporte bibliográfico que permite un análisis exhaustivo de los datos, así como herramientas tecnológicas adecuadas.

1.3. Objetivos

1.3.1. Objetivo General

Implementar un modelo predictivo para la optimización del proceso de admisión de grado de una institución de educación superior, logrando una precisión mínima de 80% en la etapa de aceptación de cupos.

1.3.2. Objetivos Específicos

1. Analizar los factores claves que influyen en la decisión de los postulantes al momento de aceptar y rechazar un cupo.
2. Identificar las carreras con alta tasa de rechazo en la etapa de aceptación de cupos.
3. Comparar múltiples algoritmos predictivos mediante métricas de precisión para la estimación de la probabilidad de la aceptación de cupos en las diferentes carreras.

1.4. Solución propuesta

Se propone implementar algoritmos de ML para la predicción de datos en dos etapas para optimizar el proceso de admisión de una institución de educación superior (IES). En primer lugar, se usarán modelos que prediga la asignación de cupos a los postulantes, considerando factores como su perfil académico, socioeconómico y las características de los programas académicos ofertados.

En segundo lugar, se predecirá la aceptación del cupo asignado, considerando variables relacionadas con la decisión final del postulante.

Para realizar el proyecto se usó el lenguaje de programación Python en un archivo de cuaderno Jupyter usando principalmente scikit-learn como librería para ciencia de datos, las liberías pandas y numpy para el procesamiento de datos, matplotlib y seaborn para crear gráficos, finalmente shap para la interpretabilidad.

1.5. Dataset

Se cuenta con datos históricos de tres procesos de admisión de una IES desde el formato de ingreso a las universidades implementado por SENESCYT a inicios del 2023, Por lo tanto, se utilizarán datos de esos tres periodos académicos como base para el análisis y el desarrollo del proyecto.

Inicialmente existen más de 200 mil registros con 54 variables, de las cuales 2 son predictivas; se asigna cupo al postulante y acepta el cupo asignado. Estas variables y la descripción de cada una se muestran en la Tabla *1* Descripción de Dataset.

Tabla 1 Descripción de Dataset

#	VARIABLE	TIPO DE DATO	DESCRIPCION
1	PERIODO_ACADEMICO	NUMÉRICO	CODIGO PERIODO ACADEMICO
2	FECHA_NACIMIENTO	TEMPORAL	FECHA DE NACIMIENTO GTM -5
3	CODIGO_PAIS_RESIDE	NUMÉRICO	CODIGO DEL PAIS DE RESIDENCIA
4	PAIS_RESIDE	CATEGÓRICO	NOMBRE DEL PAIS DE RESIDENCIA
5	CODIGO_PROVINCIA_RESIDE	NUMÉRICO	CODIGO DE LA PROVINCIA DE RESIDENCIA
6	PROVINCIA_RESIDE	CATEGÓRICO	NOMBRE DE LA PROVINCIA DE RESIDENCIA
7	CODIGO_CANTON_RESIDE	NUMÉRICO	CODIGO DEL CANTON DE RESIDENCIA
8	CANTON_RESIDE	CATEGÓRICO	NOMBRE DEL CANTON DE RESIDENCIA
9	CODIGO_PARROQUIA_RESIDE	NUMÉRICO	CODIGO DE LA PARROQUIA DE RESIDENCIA
10	PARROQUIA_RESIDE	CATEGÓRICO	NOMBRE DE LA PARROQUIA DE RESIDENCIA

11	CODIGO_GENERO	NUMÉRICO	CODIGO DEL GENERO
12	GENERO	CATEGÓRICO	NOMBRE DEL GENERO
13	CODIGO_AUTOIDENTIFICACION	NUMÉRICO	CODIGO DE LA AUTOIDENTIFICACION
14	AUTOIDENTIFICACION	CATEGÓRICO	NOMBRE DE LA AUTOIDENTIFICACION
15	CODIGO_DISCAPACIDAD	NUMÉRICO	CODIGO DE DISCAPACIDAD
16	DISCAPACIDAD	CATEGÓRICO	TIENE O NO DISCAPACIDAD SEGÚN SENESCYT
17	CODIGO_CAMPO_CONOCIMIENTO	NUMÉRICO	CODIGO DEL CAMPO DE CONOCIMIENTO
18	CAMPO_CONOCIMIENTO	CATEGÓRICO	CAMPO DE CONOCIMIENTO AL CUAL PERTENECEN VARIAS CARRERAS
19	CODIGO_CARRERA	NUMÉRICO	CODIGO DE LA CARRERA ELEGIDA
20	CARRERA	CATEGÓRICO	NOMBRE DE LA CARRERA ELEGIDA
21	CODIGO_MODALIDAD	NUMÉRICO	CODIGO DE LA MODALIDAD DE LA CARRERA ELEGIDA
22	MODALIDAD	CATEGÓRICO	MODALIDAD DE LA CARRERA
23	CODIGO_JORNADA	NUMÉRICO	CODIGO DE LA JORNADA ELEGIDA
24	JORNADA	CATEGÓRICO	JORNADA DE LA CARRERA
25	PRIORIDAD_POSTULACION	NUMÉRICO	ORDEN DE LAS CARRERAS QUE LA PERSONA ELIGIÓ
26	INSTANCIA_ASIGNACION	NUMÉRICO	N CICLO DEL PROCESO DE ASIGNACIÓN/ACEPTACIÓN DE CUPO
27	NOTA_GRADO	NUMÉRICO	NOTA FINAL CON LA CUAL SE GRADUARON DEL COLEGIO
28	EVALUACION	NUMÉRICO	NOTA DE LA PRUEBA DE CONOCIMIENTOS QUE TOMA LA UNIVERSIDAD
29	ACCIONES_AFIRMATIVAS	NUMÉRICO	PUNTAJE EXTRA QUE LA DÁ EL GOBIERNO
30	CONDICION_SOCIOECONOMICA	NUMÉRICO	CONDICIÓN DE POBREZA
31	RURALIDAD	NUMÉRICO	ESTUDIOS EN IEP PÚBLICAS RURALES
32	TERRITORIALIDAD	NUMÉRICO	RESIDE EN PARROQUIA CON ALTO ÍNDICE DE POBREZA
33	BONO_JGL	NUMÉRICO	RECIBE BONO POR DISCAPACIDAD
34	VICTIMA_VIOLENCIA	NUMÉRICO	HA SIDO VÍCTIMA DE VIOLENCIA SEXUAL O DE GÉNERO
35	FEMICIDIO	NUMÉRICO	HIJAS E HIJOS DE MUJERES VÍCTIMAS DE FEMICIDIO
36	ENFERMEDADES_CATASTROFICAS	NUMÉRICO	ADOLECE DE ALGUNA ENFERMEDAD CATASTRÓFICA
37	CASA_ACOGIDA	NUMÉRICO	FUE INGRESADAS EN

			UNA UNIDAD DE ATENCIÓN DE ACOGIMIENTO INSTITUCIONAL
38	MIGRANTES_RETORNADOS	NUMÉRICO	MIGRANTE QUE HA RETORNADO AL PAIS
39	INTERNET_DOMICILIO	NUMÉRICO	TIENE INTERNET EN SU HOGAR
40	COMPUTADORA_DOMICILIO	NUMÉRICO	TIENE COMPUTADORA EN SU HOGAR
41	PPL	NUMÉRICO	ES UNA PERSONA PRIVADA DE LA LIBERTAD
42	TIENE_CUARTO_NIVEL	NUMÉRICO	TIENE TÍTULOS REGISTRADOS EN SENESCYT DE CUARTO NIVEL
43	TIENE_TERCER_NIVEL	NUMÉRICO	TIENE TÍTULOS REGISTRADOS EN SENESCYT DE TERCER NIVEL
44	TIPO_UNIDAD_EDUCATIVA	CATEGÓRICA	TIPO DE UNIDAD EDUCATIVA QUE SE GRADUÓ
45	VULNERABILIDAD_SOCIOECONOMICA	NUMÉRICO	ES DENOMINADO CON VULNERABILIDAD SOCIOECONOMICA
46	MERITO_ACADEMICO	NUMÉRICO	FUE ABANDERADO O ESCOLTA
47	BACHILLER_PUEBLOS_NACIONALIDADES	NUMÉRICO	BACHILLER PERTENECIENTE A ALGUN PUEBLO O NACIONALIDAD INDÍGENA
48	BACHILLER_PERIODO_ACADEMICO	NUMÉRICO	BACHILLER RECIÉN GRADUADO EN EL PERIODO ACADÉMICO PASADO
49	POBLACION_GENERAL	NUMÉRICO	POBLACIÓN SIN NINGUNA CARACTERÍSTICA ESPECIAL
50	PUNTAJE_POSTULACION	NUMÉRICO	PUNTAJE FINAL = 50% NOTA DE GRADO + 50% EVALUACION + ACCIONES_AFIRMATIVA
51	GRUPO_ASIGNACION	NUMÉRICO	EL CUPO SE ASIGNA EN BASE A LEYES, REGLAS DEL GOBIERNO, ESA ENTIDAD DECIDE A QUÉ GRUPO DE ASIGNACIÓN O PRIORIDAD PERTENECE CADA PERSONA
52	MODALIDAD_EVALUACION	CATEGÓRICO	PRESENCIAL O VIRTUAL
53	PREDICT_ASIGNACION_CUPO	NUMÉRICO	1 = CUPO ASIGNADO 0 = CUPO NO ASIGNADO
54	PREDICT_ACEPTACION_CUPO	NUMÉRICO	1 = CUPO ACEPTADO 0 = CUPO RECHAZADO

CAPÍTULO 2

2. Estado del Arte

2.1. Marco referencial

2.1.1. Reglamento del sistema nacional de nivelación y admisión en Ecuador

El acuerdo Nro. SENESCYT-SENESCYT-2023-0003-AC de 07 de julio de 2023 fue publicado en el Cuarto Suplemento del Registro Oficial Nro. 363 de 28 de julio de 2023 y, sus reformas emitidas mediante Acuerdo Nro. SENESCYT-SENESCYT-2023-0025-AC de 20 de septiembre de 2023, fueron publicadas en el Registro Oficial. [2]

Dicho acuerdo regula el proceso de admisión para las universidades y escuelas politécnicas públicas del Ecuador, el cual tendrá una aplicación obligatoria para las entidades antes mencionadas. Dicho se regirá por los “principios de méritos, igualdad de oportunidades, equidad, y libertad de elección de carrera”. [2]

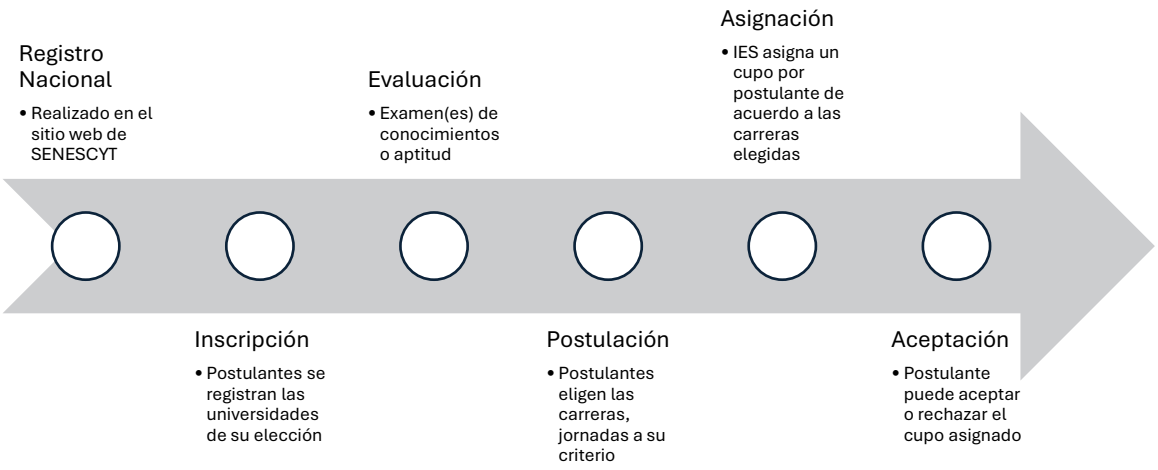

Figura 1 Etapas del proceso de admisión generalizado

Entre las diferentes etapas del proceso de admisión, se puede encontrar en la Sección VI – Evaluación para el proceso de acceso que las IES podrán elegir

la modalidad de aplicación de la evaluación, el tipo de evaluación y aceptar o no puntajes de evaluación correspondientes a otras IES. [2]

Luego en la sección VII – Postulación, se indica que las IES tienen la potestad de establecer la cantidad de etapas de postulación, puntaje mínimo de postulación y el porcentaje que llevará cada componente de dicho puntaje.

En el artículo 39. Asignación de cupos, para garantizar una distribución equitativa de los cupos, se ha establecido un orden de asignación que prioriza a ciertos grupos poblacionales, como aquellos en situación de vulnerabilidad socioeconómica, los estudiantes con alto rendimiento académico y los pertenecientes a pueblos y nacionalidades indígenas. Este orden busca promover la inclusión y la diversidad en las instituciones de educación superior. [2]

En caso de empate en el puntaje de postulación, las instituciones de educación superior establecerán mecanismos adicionales para resolver esta situación, siempre y cuando no se exijan requisitos de evaluación no contemplados en la normativa vigente. Asimismo, se promueve el desarrollo de estrategias para optimizar la asignación de cupos y agotar la oferta disponible.

La aceptación del cupo asignado es un acto voluntario y libre por parte del postulante. Una vez aceptado el cupo, este no podrá ser modificado ni anulado, y el estudiante deberá matricularse en el periodo correspondiente [2]. Para garantizar el acceso universal a la educación superior y el uso eficiente de los recursos estatales, se establece que un estudiante que acepte un cupo en un determinado periodo no podrá participar en el siguiente proceso de admisión, salvo excepciones debidamente justificadas [2].

2.1.2. Machine Learning en procesos de admisión a universidades

Podemos encontrar que el artículo “Prediction Probability of Getting an Admission into a University using Machine Learning”, establece un modelo de aprendizaje automático que tiene en cuenta límites como la puntuación GRE, la puntuación TOEFL, la clasificación universitaria, la declaración de propuesta y la carta de recomendación, el promedio de calificaciones de pregrado y la experiencia de estudio. Después de obtener todos los datos, predice la

probabilidad de admisión. En ocasiones de prueba poco conocidas, el modelo preparado tiene hallazgos fácticos sustanciales para la estimación (similar) de la probabilidad de confirmación y, en consecuencia, ofrece una impresión imparcial de la medición. [5]

Regresión Lineal, Random Forest y CatBoost son los algoritmos que decidieron usar y encontraron que CGPA es la variable con 80% de importancia, al evaluar los modelos, sin afinar parámetros, hay una precisión del 95% con el algoritmo CatBoost y es el resultado cuantitativo más alto de un modelo de expectativa de incentivo confirmatorio hasta ahora. [5]

En el 2021 se publicó “A Comparison of Classification Models in Predicting Graduate Admission Decision”, en este artículo se presentan comparaciones de varios modelos de clasificación basados en aprendizaje automático, incluidos Naïve Bayes, Regresión logística, Perceptrón multicapa y Modelos de árbol de decisiones, para predecir el resultado de la admisión de candidatos con un conjunto de parámetros conocidos utilizando un conjunto de datos de 400 registros de solicitantes. [6]

Al comparar las métricas de rendimiento de estos métodos, el estudio concluye que Naïve Bayes es el modelo más preciso para este tipo de conjunto de datos. Los modelos predictivos como los que se analizan en este artículo pueden ser una herramienta valiosa para los futuros estudiantes a la hora de seleccionar universidades en su proceso de solicitud. El estudio también propone un marco que incorpora la clasificación basada en aprendizaje automático en el proceso de decisión de admisión. La implementación de dichos métodos puede ayudar a los comités de admisión de posgrado a agilizar grandes grupos de solicitudes o a observar y comprender las tendencias en sus decisiones de admisión anteriores. [6]

A nivel internacional, se tiene el trabajo de “Modelo predictivo para la selección de postulantes destacados a una institución de educación superior” del autor Bugueño presentado en 2017 [7], que presenta el objetivo general de desarrollar un modelo predictivo que determina el perfil de los postulantes destacados para los estudiantes secundarios que postulan a la Facultad de Economía y Negocios de la Universidad de Chile, para ello, se utilizó una metodología basada en el Knowledge Discovery in Databases (KDD), junto a

cuatro enfoques de solución con diferente complejidad, utilizando así la selección de variables, clúster y combinación de modelos. Entre los resultados resalta la efectividad del enfoque Secuencial Combinado con un 80,4% de precisión y una identificación de hasta un 89,4% de los mejores postulantes, utilizando variables como notas escolares, género, edad, puntajes de lenguaje y NEM, aunque se cuestiona la importancia tradicional del puntaje de matemáticas en el proceso de selección, ya que, una vez alcanzado un nivel mínimo, otros factores como el puntaje de lenguaje cobran más relevancia. Los resultados obtenidos demuestran que es factible aumentar considerablemente el número de candidatos altamente calificados que acepten la postulación de su primera prioridad, mediante la optimización de los procesos de captación de la institución.

Además, se cuenta con el artículo de investigación titulado como “Factores asociados a la aceptación de un cupo en las universidades de los bachilleres de Manabí” de los autores Benavides, Maldonado & Buenaño publicado en el año 2018 [8] basado en el objetivo de identificar que tanto predicen las variables individuales, experienciales y las contextuales, la aceptación de un cupo en las carreras de las universidades públicas del Ecuador en términos de probabilidad de ocurrencia, a través del modelo de regresión logística determinado con la herramienta SPSS. El diseño metodológico es de tipo mixto, dada la parte cualitativa que permitió la identificación de las variables a partir de la teoría cognitiva social y la racional de decisiones, que son buenos predictores son prioridad asignada a la carrera por el estudiante mientras que la parte cuantitativa contempló los datos de 4648 estudiantes.

Los resultados mostraron que las áreas donde se otorgaron cupos corresponden a 26,9% en educación comercial y derecho, el 18% en salud, y el 14,6% en ingeniería, industria y construcción y en cuanto a conseguir un cupo se dio en un 50,4% fue su primera opción, el 20,7% en su segunda, y el 13,5% en su tercera prioridad, en total, el 79% aceptó el cupo asignado, mientras que el resto lo rechazó. Por tanto, se identificaron los factores que influyen en la decisión de aceptar el cupo asignado, como la prioridad de la carrera, la ubicación de la institución, el número de intentos de prueba, estado civil,

género, modalidad de estudio y edad también se revela cómo la inequidad socioeconómica limita el acceso a la educación superior para personas de bajos recursos.

2.2. Marco teórico

2.2.1. Aprendizaje Automático (Machine Learning)

El aprendizaje automático (ML) es una convención que consiste en utilizar algoritmos específicos para obtener datos, adquirir conocimiento a partir de ellos y hacer una predicción sobre algo de la naturaleza. En el aprendizaje automático, se utilizan algoritmos de regresión para anticipar valores de salida basándose en los atributos de entrada de un conjunto de datos. Seleccionar un buen algoritmo de ML será útil para predecir mejores resultados. También está influenciado por varios parámetros como el tamaño de los datos, la calidad de los resultados, el tiempo de entrenamiento, etc. [9]

2.2.2. Aprendizaje Automático supervisado

Random Forest

Los bosques aleatorios son una técnica de aprendizaje automático que consiste en construir un conjunto de árboles de decisión, cada uno entrenado en una muestra aleatoria de los datos y utilizando un subconjunto aleatorio de las características. Esta aleatorización reduce la varianza y mejora la capacidad de generalización del modelo. Al combinar las predicciones de múltiples árboles, los bosques aleatorios logran una mayor precisión y robustez que los árboles de decisión individuales. Además, proporcionan herramientas para evaluar la importancia de las variables y detectar problemas como el sobreajuste, lo que los convierte en una herramienta versátil para una amplia gama of aplicaciones. [10]

CatBoost

Boosting es una técnica de aprendizaje automático que combina múltiples modelos simples (aprendices débiles) para crear un modelo más poderoso. CatBoost es una técnica de aprendizaje automático que ha demostrado ser altamente efectiva en una amplia gama de aplicaciones. Su principio

fundamental radica en combinar de manera secuencial modelos predictivos simples (como árboles de decisión) para crear un modelo más complejo y preciso. Esta técnica ha logrado resultados sobresalientes en diversas áreas, desde la búsqueda en línea hasta la predicción meteorológica, gracias a su capacidad para capturar patrones complejos en datos heterogéneos y ruidosos. En esencia, el aumento de gradiente optimiza un modelo de manera iterativa, ajustándolo gradualmente para minimizar el error de predicción.

Multilayer perceptron

Una red neuronal artificial es un sistema de aprendizaje automático que imita el funcionamiento del cerebro humano. Está compuesto por unidades interconectadas llamadas neuronas artificiales, organizadas en capas. La información se transmite a través de estas capas, donde cada neurona procesa la entrada y genera una salida que afecta a las siguientes neuronas.

Un perceptrón multicapa (MLP) es un tipo específico de AAN con múltiples niveles de neuronas. Estas neuronas emplean funciones de activación no lineales, lo que les permite reconocer patrones complejos en los datos. Los MLP son herramientas poderosas en el aprendizaje automático debido a su capacidad de identificar relaciones no lineales en los datos, siendo útiles para tareas como clasificación, predicción y reconocimiento de patrones.

2.2.3. Métricas de evaluación

Precision

La precisión se refiere a las opiniones que el sistema ha identificado correctamente como parte de una clase, coincidiendo con el criterio del experto, se calcula dividiendo los verdaderos positivos entre la suma de verdaderos positivos y falsos positivos [11], por tanto, muestra el porcentaje de muestras correctamente clasificadas en relación con el total de muestras asignadas a esa clase, en caso de tener una precisión baja corresponde a la presencia de un elevado número de falsos positivos [12].

Recall

La cobertura mide el porcentaje de opiniones del conjunto de evaluación que el sistema ha clasificado correctamente, por ello, se calcula dividiendo los verdaderos positivos entre la suma de verdaderos positivos y falsos negativos, si el sistema no identifica correctamente ninguna opinión como falsa, su cobertura es 0 y no supera la evaluación [11]. Se calcula utilizando la siguiente fórmula, donde [13]:

VP= Verdaderos Positivos

FN = Falsos Negativos

$$REC = \frac{VP}{FN + VP}$$

F1-score

Es una métrica que combina precisión y recall en un único valor, siendo particularmente útil en situaciones donde existe un desequilibrio entre clases en los datos. Se calcula como la media armónica de las dos medidas indicadas, proporcionando un equilibrio entre ambos indicadores, utilizando la siguiente fórmula:

$$F1 = \frac{2 * Precisión * Recall}{Precision + Recall}$$

Matriz de confusión

Se considera como la forma más sencilla de evaluar el rendimiento, es una tabla con dos dimensiones: "Real" y "Predicho", cada una de las cuales contiene "Verdaderos Positivos (TP)", "Verdaderos Negativos (TN)", "Falsos Positivos (FP)" y "Falsos Negativos (FN)", considerando los siguientes puntos:

- TP: Cuando la clase real y la predicha son 1.
- TN: Cuando la clase real y la predicha son 0.
- FP: Cuando la clase real es 0 y la predicha es 1.
- FN: Cuando la clase real es 1 y la predicha es 0 [14].

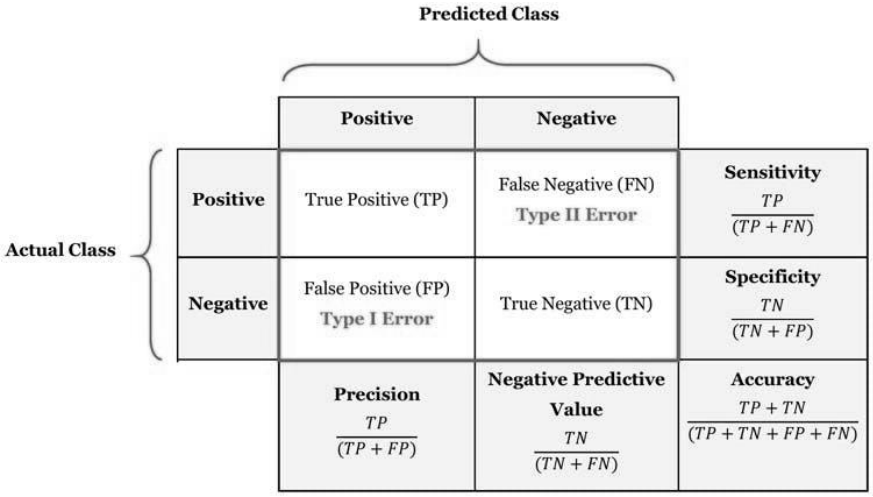

Figura 2: Matriz de confusión *[14]*

La Figura *2*: Matriz de confusión representa visualmente la relación entre la clase real y la clase predicha para un conjunto de datos determinado. La matriz se divide en cuatro cuadrantes: verdadero positivo (TP), verdadero negativo (TN), falso positivo (FP) y falso negativo 1 (FN). Los verdaderos positivos se producen cuando el modelo predice correctamente una clase positiva, mientras que los verdaderos negativos representan predicciones correctas para la clase negativa. Los falsos positivos se producen cuando el modelo predice incorrectamente una clase positiva para una instancia negativa real, y los falsos negativos se producen cuando el modelo predice incorrectamente una clase negativa para una instancia positiva real. [14]

La matriz de confusión incluye métricas de rendimiento calculadas a partir de estos cuadrantes, como la sensibilidad (sensitivity) (tasa de verdaderos positivos), la especificidad (specificity) (tasa de verdaderos negativos), la precisión (valor predictivo positivo), el valor predictivo negativo 2 y la exactitud (acurracy). Estas métricas brindan información sobre la capacidad del modelo para clasificar correctamente las instancias, su tendencia a cometer errores de falsos positivos o falsos negativos y su desempeño general en términos de clasificar correctamente las instancias positivas y negativas. [14]

Curva ROC

El análisis con base en curvas en inglés Receiver Operating Characteristic (ROC) constituye un método estadístico para determinar la exactitud diagnóstica de test que utilizan escalas continuas [15]. También es concebida

para evaluar el rendimiento óptimo de una prueba diagnóstica en relación con diferentes valores de una variable predictora, donde el punto de corte ideal es el que maximiza el área bajo la curva, lo que asegura la mejor sensibilidad y especificidad, reflejando así el mejor desempeño de la prueba, por ejemplo, se tiene la siguiente figura que demuestra la razón de verosimilitud positiva la cual es Sensibilidad/1 – Especificidad [16]

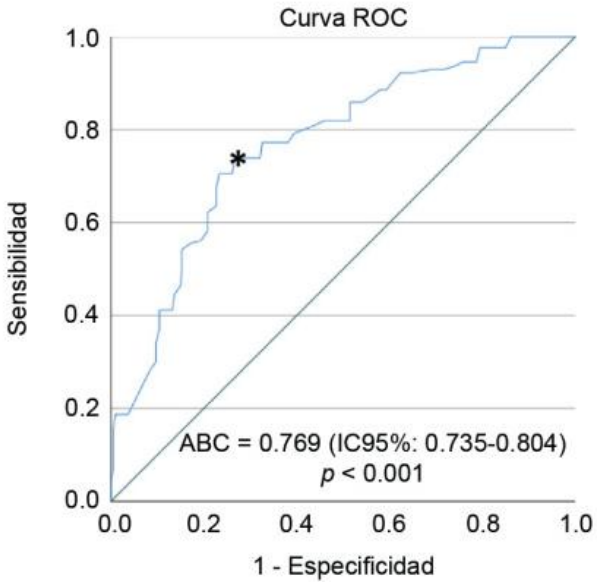

Figura 3: Curva ROC *[16]*

Por ejemplo, en la Figura *3*: Curva ROC , en el eje de las abscisas (horizontal) se representa 1 - especificidad, que indica la proporción de falsos positivos, mientras que en el eje de las ordenadas (vertical) se representa la sensibilidad, que corresponde a la proporción de verdaderos positivos. La línea diagonal que cruza la gráfica representa el rendimiento de un clasificador aleatorio, donde la sensibilidad es igual a 1 - especificidad para todos los umbrales.

El área bajo la curva ROC (AUC) es un valor numérico que resume el rendimiento global del modelo. Un valor de AUC cercano a 1 indica un modelo con un excelente poder de discriminación, mientras que un valor cercano a 0.5 indica un modelo con un poder de discriminación similar al azar. En la Figura *3*: Curva ROC , el valor de AUC es de 0.769, lo que sugiere un buen rendimiento del modelo. Además, el intervalo de confianza del 95% para el AUC se encuentra entre 0.735 y 0.804, lo que indica que este resultado es estadísticamente significativo ($p < 0.001$).

2.2.4. Afinamiento de Hiperparámetros

El diseño de modelos predictivos es un proceso complejo que requiere seleccionar el algoritmo adecuado y ajustar los hiperparámetros, los cuales influyen en el rendimiento del modelo durante el entrenamiento. Sin embargo, con frecuencia se omite este ajuste y se utilizan los valores predeterminados establecidos por las bibliotecas de aprendizaje automático, ya sea por falta de conocimiento sobre su impacto o no se dispone de los mecanismos adecuados de configuración [17].

Por otro lado, se consideran como configuraciones externas que no se estiman directamente a partir de los datos, sino que deben establecerse antes de entrenar el algoritmo, por ello, son variables que deben ser seleccionadas por el usuario o determinadas de forma heurística, para adaptarse a un problema específico, ya que no se pueden conocer de antemano. Se suelen probar diferentes combinaciones mediante técnicas como búsqueda en grilla y se evalúan con una función de pérdida para identificar la mejor opción [18].

2.2.5. Grid Search Cross-Validation

Es una potente técnica para ajustar los hiperparámetros de los modelos de aprendizaje automático. Permite buscar sistemáticamente a través de un conjunto predefinido de valores de hiperparámetros para encontrar la combinación que resulta en el mejor rendimiento del modelo. [19]

En el aprendizaje automático, los hiperparámetros son configuraciones del modelo. A diferencia de los parámetros del modelo, que se aprenden durante el entrenamiento, los hiperparámetros se establecen antes del entrenamiento y guían el proceso de aprendizaje. Elegir los hiperparámetros correctos es crucial para lograr un rendimiento óptimo del modelo. [19]

Seleccionar manualmente los mejores hiperparámetros puede ser desafiante y llevar mucho tiempo. GridSearchCV automatiza este proceso realizando una búsqueda exhaustiva en una cuadrícula de hiperparámetros especificada. Evalúa sistemáticamente el rendimiento del modelo para cada combinación de valores de hiperparámetros, ayudando a identificar el conjunto óptimo. [19]

Entre los beneficios de usar GridSearchCV está que esta técnica explora sistemáticamente las combinaciones de hiperparámetros, ahorrando tiempo en comparación con el ajuste manual, además identificar los mejores hiperparámetros a menudo conduce a un mejor rendimiento del modelo en datos no vistos. La validación cruzada asegura que el rendimiento del modelo se generalice a diferentes subconjuntos de los datos, reduciendo el sobreajuste y automatiza el proceso de ajuste de hiperparámetros, haciéndolo accesible y reproducible. [19]

2.2.6. SHAP values

Los valores SHAP, originados en la teoría de juegos cooperativos por Shapley en 1953, explican la contribución de cada jugador, se adaptaron en relación con el aprendizaje automático para calcular la contribución de una variable a una predicción, aunque el proceso es computacionalmente costoso, sin embargo, para optimizarlo, desarrollaron un algoritmo aproximado, implementando el concepto de TreeSHAP, un método rápido y exacto para modelos basados en árboles, que permite explicar eficientemente millones de predicciones al aprovechar la estructura de los árboles [20].

2.2.7. Lenguaje de programación

Python

Es un lenguaje de programación de alto nivel, orientado a objetos e interpretado con semántica dinámica. Sus estructuras de datos integradas de alto nivel, combinadas con tipado y vinculación dinámicos, lo hacen muy atractivo para el desarrollo rápido de aplicaciones, así como para su uso como lenguaje de scripting o de unión para conectar componentes existentes. [21]

La sintaxis simple y fácil de aprender de Python enfatiza la legibilidad y, por lo tanto, reduce el costo de mantenimiento del programa. Python admite módulos y paquetes, lo que fomenta el modularidad del programa y la reutilización del código. El intérprete de Python y la amplia biblioteca estándar están disponibles en forma de código fuente o binario sin cargo para todas las plataformas principales y se pueden distribuir libremente. [21]

2.2.8. Aplicación web

Jupyter Notebook

Es una aplicación web interactiva para crear y compartir documentos computacionales. El proyecto se denominó primero IPython y luego se renombró Jupyter en 2014. Es un producto de código abierto y los usuarios pueden usar todas las funciones disponibles de forma gratuita. Admite más de 40 lenguajes, incluidos Python, R y Scala. [22]

Un notebook es un archivo mutable guardado en formato ipynb. Jupyter Notebook tiene un panel de control de notebooks para ayudar a los usuarios a administrar diferentes notebooks. Los kernels también son parte de Jupyter notebooks. Los kernels son procesos que ejecutan código interactivo en un lenguaje de programación particular, devuelven el resultado al usuario, también responden a las solicitudes de introspección y de finalización de tabulación. [22]

2.2.9. Librerías de Python

Scikit-learn

También conocido como sklearn, es una biblioteca de modelado de datos y aprendizaje automático de código abierto para Python. Cuenta con varios algoritmos de clasificación, regresión y agrupamiento, incluidos máquinas de vectores de soporte, bosques aleatorios, potenciación de gradiente, k-means y DBSCAN, y está diseñado para interoperar con las bibliotecas de Python, NumPy y SciPy. [23]

Scikit-learn se lanzó por primera vez en 2010 y, desde entonces, ha ganado un lugar destacado en el ecosistema de aprendizaje automático de Python. Implementa numerosos algoritmos de aprendizaje automático y modelado de datos, y proporciona API de Python consistentes. Admite una interfaz de modelo estandarizada y concisa en todos los modelos. Por ejemplo, Scikit-learn utiliza un modelo de flujo de trabajo de ajuste/predicción simple para sus algoritmos de clasificación. [23]

Pandas

Es una biblioteca de software de código abierto creada sobre Python específicamente para la manipulación y el análisis de datos, y ofrece estructuras y operaciones de datos para un análisis y una manipulación de datos potentes, flexibles y fáciles de usar. Pandas fortalece a Python al brindarle al popular lenguaje de programación la capacidad de trabajar con datos similares a hojas de cálculo, lo que permite cargar, alinear, manipular y fusionar rápidamente, además de otras funciones clave. Pandas es apreciado por brindar un rendimiento altamente optimizado cuando el código fuente del back-end está escrito en C o Python. [24]

Pandas, que algunos han denominado un «cambio de paradigma» para el análisis de datos con Python, se encuentra entre las herramientas más populares y utilizadas para el denominado «tratamiento de datos» o «munging». Esto describe un conjunto de conceptos y una metodología que se utilizan para sacar datos de formas inutilizables o erróneas y llevarlos a los niveles de estructura y calidad necesarios para el procesamiento analítico moderno. Pandas se destaca por su facilidad para trabajar con formatos de datos estructurados, como tablas, matrices y datos de series temporales. También funciona bien con otras bibliotecas científicas de Python. [24]

Numpy

Es una biblioteca de código abierto, potente y optimizada para el lenguaje de programación Python, que agrega soporte para matrices multidimensionales grandes (también llamadas matrices o tensores). NumPy también viene equipado con una colección de funciones matemáticas de alto nivel para trabajar en conjunto con estas matrices. Estas incluyen álgebra lineal básica, simulación aleatoria, transformadas de Fourier, operaciones trigonométricas y operaciones estadísticas. [25]

Como biblioteca principal para computación científica, NumPy es la base para bibliotecas como Pandas, Scikit-learn y SciPy. Se usa ampliamente para realizar operaciones matemáticas optimizadas en matrices grandes. [25]

Matplotlib

Es una biblioteca multiplataforma de visualización de datos y representación gráfica (histogramas, diagramas de dispersión, gráficos de barras, etc.) para Python y su extensión numérica NumPy. Como tal, ofrece una alternativa viable de código abierto a MATLAB. Los desarrolladores también pueden utilizar las API de matplotlib (interfaces de programación de aplicaciones) para incorporar gráficos en aplicaciones GUI. [26]

Seaborn

Es una biblioteca para crear gráficos estadísticos en Python. Se basa en matplotlib, se integra estrechamente con las estructuras de datos de pandas y ayuda a explorar y comprender sus datos. [27]

Sus funciones de representación gráfica operan en marcos de datos y matrices que contienen conjuntos de datos completos y realizan internamente el mapeo semántico y la agregación estadística necesarios para producir gráficos informativos. Su API declarativa orientada a conjuntos de datos le permite centrarse en lo que significan los diferentes elementos de sus gráficos, en lugar de en los detalles de cómo dibujarlos. [27]

Shap

SHAP es un marco de trabajo que se utiliza para interpretar los resultados de los modelos de aprendizaje automático. La idea clave detrás de los valores SHAP se basa en la teoría de juegos cooperativos y el concepto de valores Shapley. [28]

A diferencia de otros métodos, SHAP nos brinda una comprensión detallada de cómo cada característica contribuye a las predicciones. Esto no solo garantiza la imparcialidad, sino que también facilita la comprensión para todos. [28]

SHAP es útil porque nos muestra la importancia de cada característica para hacer predicciones. Al proporcionar valores Shapley, nos ayuda a comprender modelos complejos y cómo las características de entrada afectan las predicciones. [28]

CAPÍTULO 3

3. Metodología

3.1. Pipeline de la solución propuesta

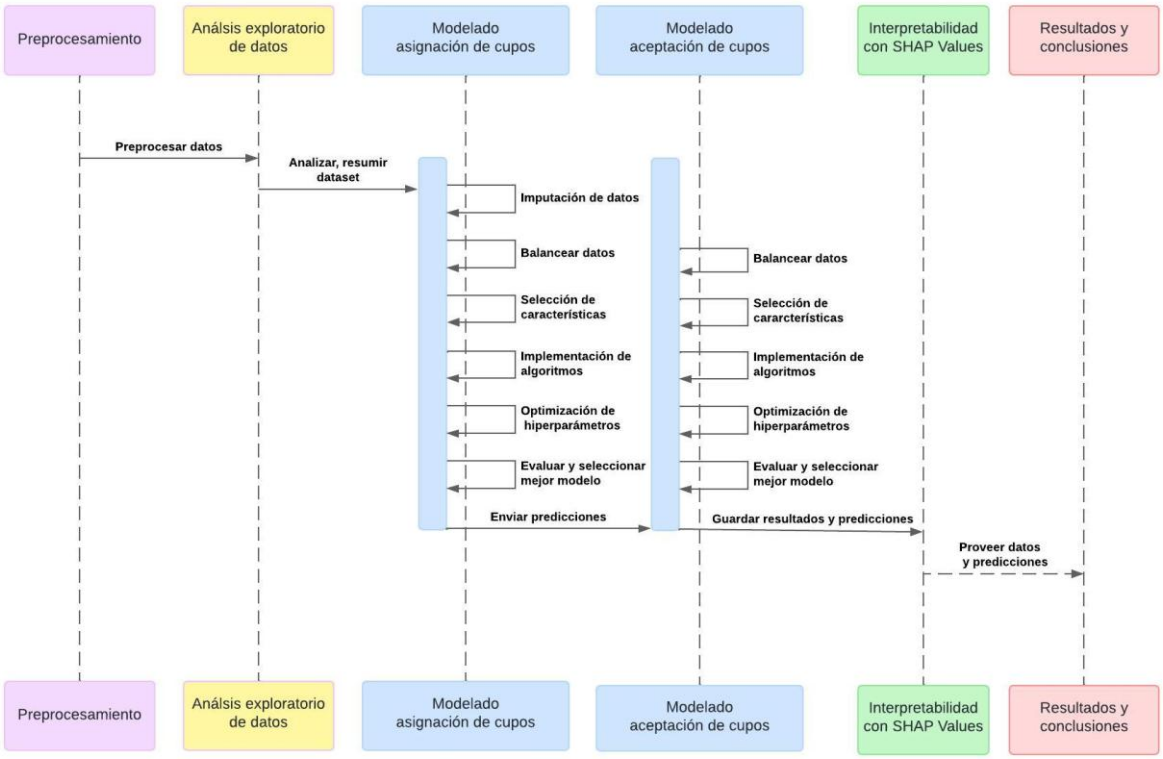

Figura 4 Pipeline de la solución propuesta

El pipeline de la Figura *4* comienza con el preprocesamiento de los datos, donde se limpian y transforman para adecuarlos al modelo. Luego, se realiza un análisis exploratorio de los datos para comprender su estructura y características. Posteriormente, se procede al modelado de asignación de cupos, donde se entrena un modelo para predecir la asignación de cupos. Este proceso incluye la imputación de datos faltantes, el balanceo de las clases si es necesario, la selección de las características más relevantes y la implementación y optimización de algoritmos de machine learning.

A continuación, se lleva a cabo el modelado de aceptación de cupos, donde se entrena otro modelo para predecir si un cupo será aceptado o no. Este proceso sigue una metodología similar al anterior.

Una vez obtenidos los modelos, se realiza una interpretabilidad con SHAP values para entender cómo cada característica influye en las predicciones. Finalmente, se presentan los resultados y conclusiones del proceso de modelado.

3.2. Preprocesamiento

Se importó el conjunto de datos y se realizó una exploración exhaustiva de sus características, identificando variables numéricas, categóricas y valores faltantes. A continuación, se corrigieron el tipo de datos de las columnas, eliminaron duplicados, se aplicó técnicas de imputación para tratar los datos faltantes y se eliminaron los outliers que puedan sesgar los resultados.

Luego de realizar esta exploración, se puede considerar que el dataset no está distribuido normalmente sin embargo este es un comportamiento correcto en este caso debido a que la cantidad de postulantes que inician el proceso versus los que se les asigna y acepta cupo.

3.3. Análisis exploratorio de datos

3.3.1. Asignación de cupos

Figura 5 Distribución de cupos asignados

Como se expone en la Figura *5* Distribución de cupos asignadosFigura *5*, se encontró que la cantidad de postulantes a las cuales se les asigna un cupo es del 8%, lo cual evidencia una marcada desigualdad en la asignación de cupos, con una gran mayoría de postulantes quedando fuera del proceso. Esta situación plantea interrogantes sobre la equidad del proceso de selección y la

necesidad de implementar medidas para mejorar las oportunidades de acceso a los programas educativos.

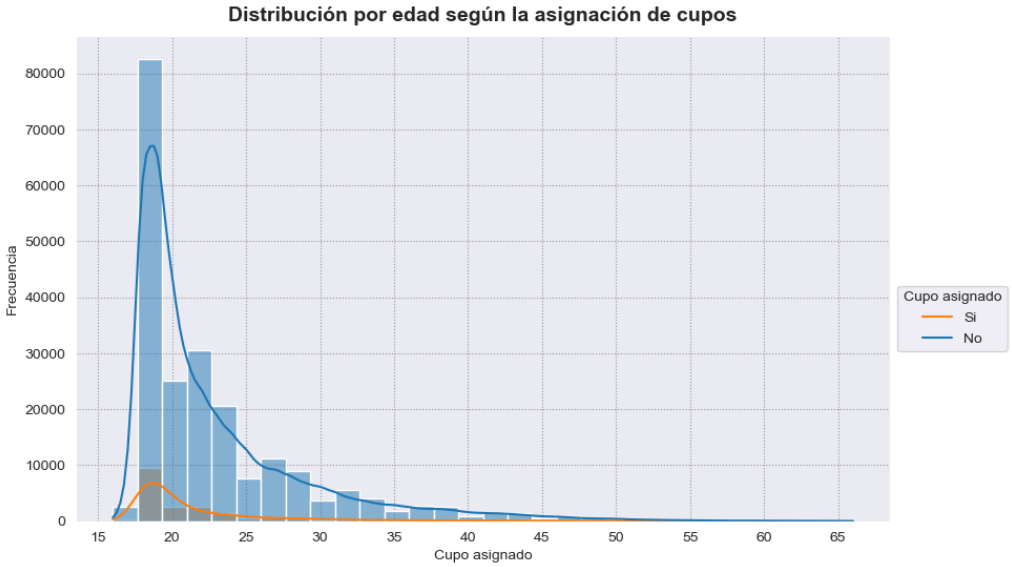

Figura 6 Distribución por edad según la asignación de cupos

Se observa en la Figura 6 Distribución por edad según la asignación de cupos, que la mayoría de los postulantes se encuentran en un rango de edad comprendido entre 15 y 25 años, lo cual es coherente con el perfil típico de los estudiantes que ingresan a la educación superior. Se aprecia una clara tendencia; la frecuencia de asignación de cupos es significativamente mayor en los grupos de edad más jóvenes, especialmente entre los 18 y 22 años. A medida que aumenta la edad, la probabilidad de obtener un cupo disminuye de manera considerable.

Esto sugiere que la edad es una variable relevante que considerar en el estudio de los factores que influyen en la asignación de cupos. Además, ponen de manifiesto la necesidad de explorar políticas y estrategias que promuevan el acceso a la educación superior para los grupos de edad mayores, garantizando así una mayor equidad y oportunidades para todos los postulantes.

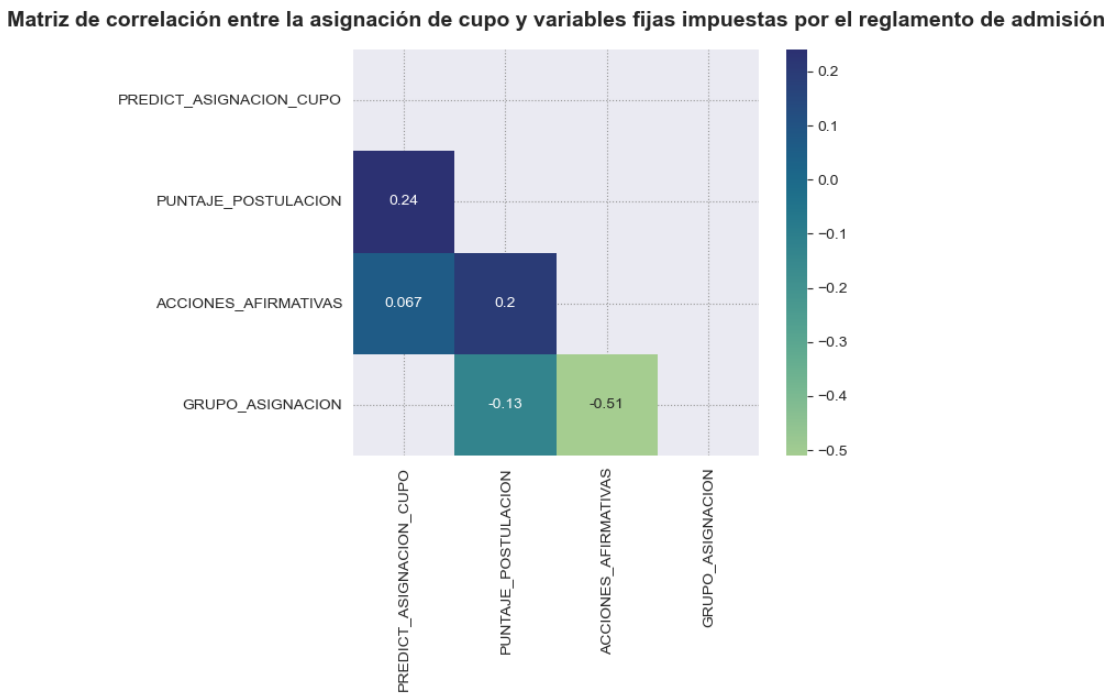

Figura 7 Matriz de correlación entre la asignación de cupo y variables fijas

En la Figura 7 Matriz de correlación entre la asignación de cupo y variables fijas, se puede realizar un análisis de correlación en la cual se puede concluir que el puntaje de postulación es un factor determinante debido a que el análisis confirma que el puntaje de postulación es un factor clave en la decisión de asignar un cupo. Los estudiantes con mejores calificaciones académicas tienen una ventaja significativa en el proceso de selección.

Las acciones afirmativas tienen un impacto limitado: Aunque las políticas de acción afirmativa buscan promover la equidad, su impacto en la asignación de cupos parece ser limitado según esta matriz de correlación.

Existen desigualdades: La correlación negativa entre el grupo de asignación y la asignación del cupo evidencia la existencia de desigualdades en el acceso a la educación superior.

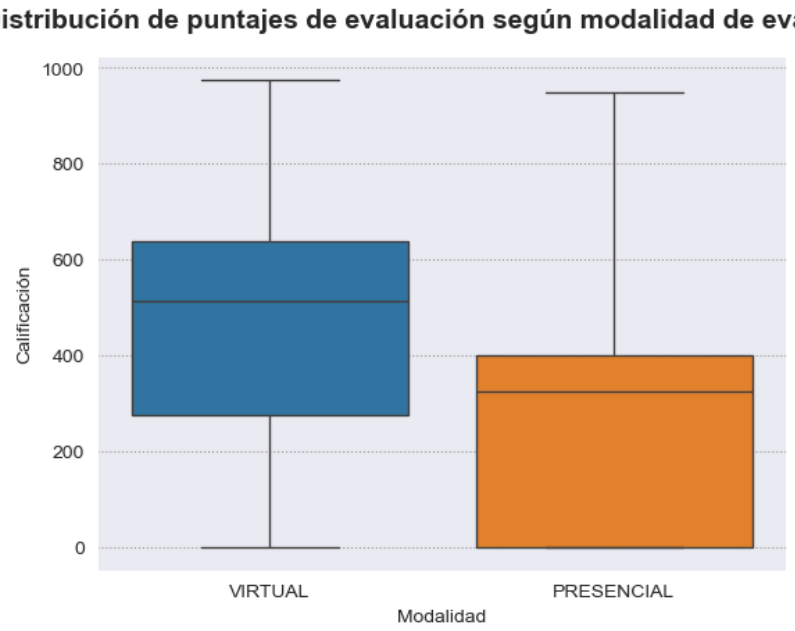

Figura 8 Distribución de puntajes de evaluación según la modalidad de evaluación

A partir de la Figura *8* Distribución de puntajes de evaluación según la modalidad de evaluación, podemos inferir que la modalidad de evaluación influye en el cálculo del puntaje final de postulación ya que La caja en cada modalidad representa el rango intercuartílico, que abarca el 50% central de los datos. El IQR de la modalidad virtual es ligeramente más amplio que el de la modalidad presencial, lo que sugiere una mayor dispersión en los puntajes de los estudiantes que realizaron las evaluaciones de manera virtual.

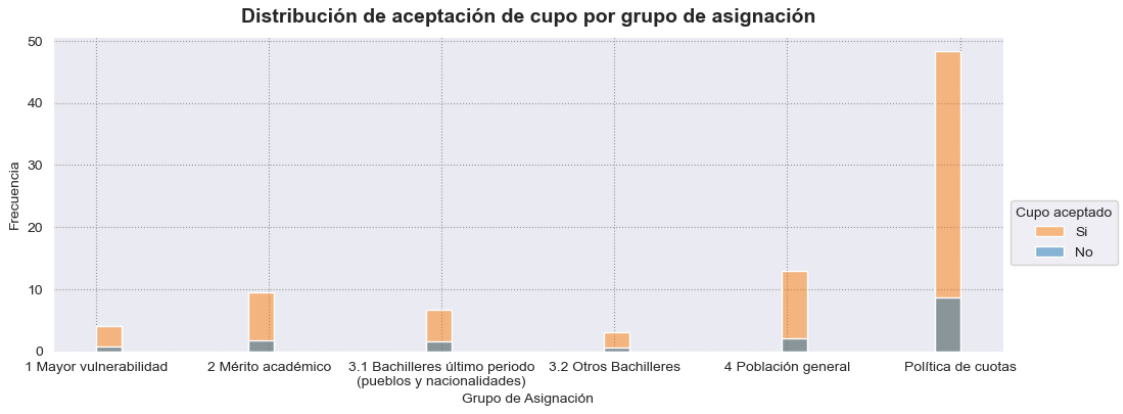

Figura 9 Distribución de aceptación de cupo por grupo de asignación

En la Figura *9* Distribución de aceptación de cupo por grupo de asignación, se observa una clara heterogeneidad en la tasa de aceptación de cupos entre los diferentes grupos de asignación. Algunos grupos presentan una tasa de aceptación significativamente mayor que otros. Los grupos con las tasas de aceptación más altas parecen ser aquellos asociados a políticas de acción afirmativa o a criterios específicos de selección, como el "2. Mérito académico" y "5. Política de cuotas". Esto sugiere que estas políticas podrían estar teniendo un impacto positivo en el acceso a la educación superior para ciertos grupos de estudiantes.

Por otro lado, los grupos con las tasas de aceptación más bajas podrían estar experimentando barreras adicionales en el proceso de admisión. Es importante en un trabajo a futuro las razones de estas diferencias para identificar posibles desigualdades y diseñar estrategias para mejorar el acceso equitativo a la educación.

3.3.2. Aceptación de cupos

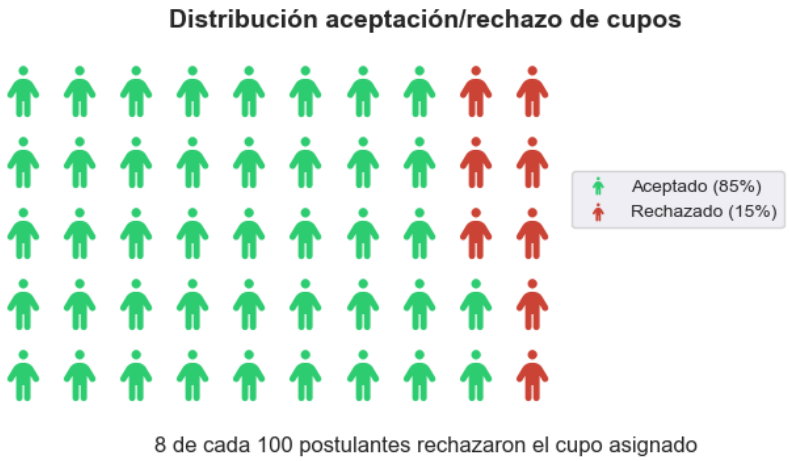

Figura 10 Distribución aceptación/rechazo de cupos

La Figura *10* Distribución aceptación/rechazo de cupos ha presentado evidencia una tasa de rechazo de cupos que, aunque minoritaria, merece atención debido a que el rechazo de un cupo puede tener implicaciones tanto para el postulante como para la institución. Para el postulante, puede significar postergar sus estudios o buscar alternativas académicas en otras instituciones. Para la institución, estos rechazos se van acumulando poco a poco y esto es una de las causas por lo cual se realizan múltiples asignaciones de cupos.

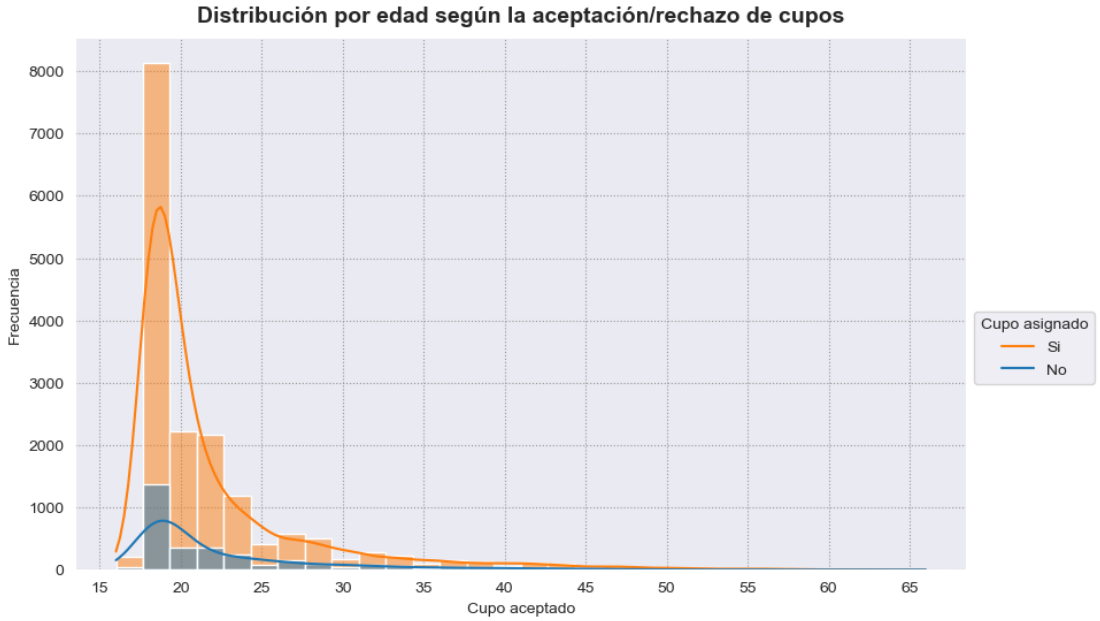

Figura 11 Distribución por edad según la aceptación/rechazo de cupos

En la Figura *11* Distribución por edad según la aceptación/rechazo de cupos, existe una concentración en edades jóvenes ya que a mayoría de los postulantes se encuentran en un rango de edad comprendido entre 15 y 25 años. Esto sugiere que la oferta educativa está principalmente dirigida a jóvenes que recién culminan sus estudios secundarios.

Se observa una tendencia clara: los postulantes más jóvenes (entre 15 y 25 años) presentan una mayor tasa de aceptación de los cupos asignados. La curva de la línea que representa a los que aceptaron el cupo alcanza su pico máximo en este rango de edad. A medida que aumenta la edad, se aprecia un incremento gradual en la proporción de postulantes que rechazan el cupo. Esto indica que los postulantes mayores tienen una menor probabilidad de aceptar la oferta.

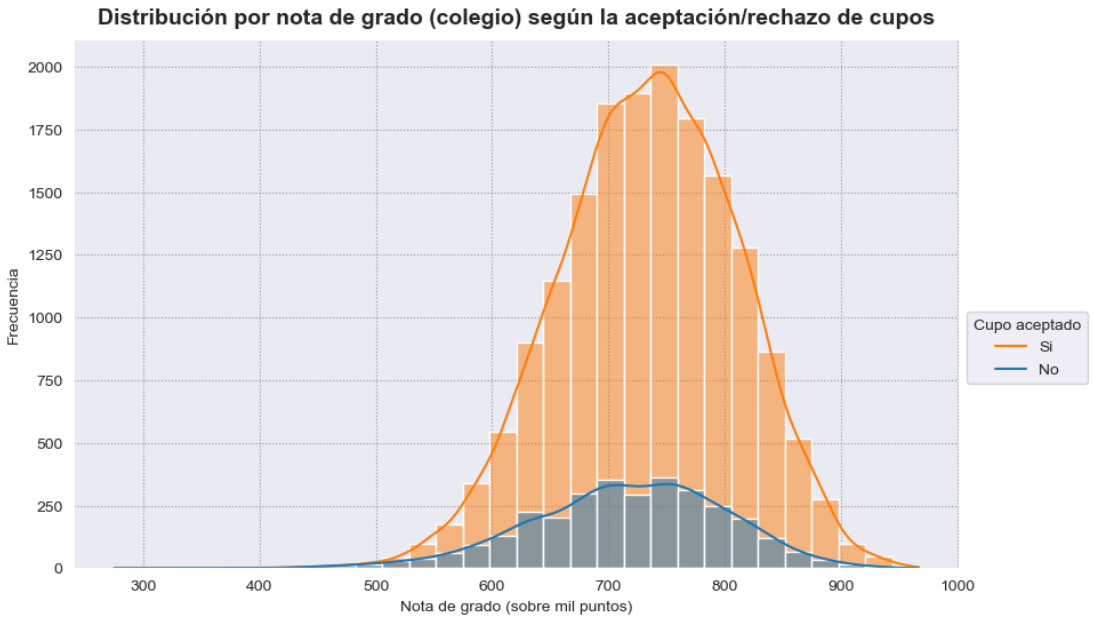

Figura 12 Distribución por nota de grado según la aceptación de cupos

Tanto para los que aceptaron como para los que rechazaron el cupo, se aprecia en la Figura *12* una mayor concentración de estudiantes en el rango de notas entre 600 y 800 puntos. Esto sugiere que la mayoría de los postulantes a la universidad cuentan con un buen rendimiento académico colegial.

Se visualiza que los estudiantes que obtuvieron notas superiores a 600 puntos presentan una mayor probabilidad de aceptar el cupo asignado. La curva de la línea que representa a los que aceptaron el cupo alcanza su pico máximo en este rango de notas. A medida que disminuye la nota, aumenta la proporción de estudiantes que deciden rechazar el cupo. Esto indica que los estudiantes con notas más bajas tienen una menor probabilidad de aceptar la oferta universitaria.

3.4. Modelado asignación de cupos

De acuerdo con la variable predictora “PREDICT_ASIGNACION_CUPO”, existían 217.219 mil postulantes sin cupo asignado y 20.114 mil con cupo asignado, por lo cual se decidió balancear los datos usando resampling, se realizaron pruebas con undersampling y oversampling.

Luego se aplicó Target Encoding solo en el conjunto de entrenamiento a las columnas categóricas y se dividió el dataset en partes de entrenamiento (70%) y test (30%). Posteriormente se implementaron los algoritmos usando siempre una misma semilla inicial para para garantizar la reproducibilidad de los resultados.

3.4.1. Random Forest

Con la función GridSearchCV, se exploraron diferentes combinaciones de parámetros, como se detalla en la Tabla *2*. Los hiperparámetros ajustados incluyen el número de árboles en el bosque (n_estimators), la profundidad máxima de cada árbol (max_depth), el número mínimo de muestras requeridas para dividir un nodo interno (min_samples_split) y una hoja (min_samples_leaf), y si se utiliza o no muestreo con reemplazo (bootstrap) para construir cada árbol.

Tabla 2 Grid de hiperparámetros para Random Forest - Asignación de cupos

Parámetro	**Valores usados**
n_estimators	100, 200
max_depth	10, 20
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 5, 10
bootstrap	True, False

Tras evaluar 20 combinaciones diferentes de hiperparámetros, utilizando validación cruzada de 5 pliegues, se determinó que la configuración óptima para el modelo Random Forest, detallada en la Tabla *3*, es aquella que emplea 100 estimadores, una profundidad máxima de 10, un mínimo de 2 muestras por división y 1 muestra por hoja, con muestreo con reemplazo.

Tabla 3 Mejores hiperparámetros para Random Forest - Asignación de cupos

Parámetro	**Mejor valor**
n_estimators	100
max_depth	10
min_samples_split	2

min_samples_leaf	1
bootstrap	True

3.4.2. CatBoost

Con la cuadrícula de hiperparámetros establecidas en la Tabla *4*, se procedió a instanciar el modelo y a realizar una búsqueda exhaustiva de la mejor combinación de hiperparámetros. Los parámetros usados fueron el número de estimadores bases (iterations) y profundidad (depth) que controlan la complejidad del modelo y la velocidad de aprendizaje, respectivamente. Para medir el error de los resultados se establecieron dos posibles funciones de pérdida (loss_function) y para calcular la diferencia entre la salida prevista y la real de una red neuronal (custom_loss) se usaron cuatro métricas.

Tabla 4 Grid de hiperparámetros para CatBoost - Asignación de cupos

Parámetro	**Valores usados**
iterations	100.0, 200, 300
depth	4, 5, 6
loss_function	Logloss, CrossEntropy
custom_loss	AUC, Accuracy, F1, Recall

Mediante la aplicación de GridSearchCV, se exploraron diversas combinaciones de hiperparámetros antes especificados. Los resultados de este proceso se resumen en la Tabla *5*. La configuración óptima encontrada implica 100 iteraciones del algoritmo, una profundidad máxima de los árboles de decisión de 4 niveles, la utilización de la función de pérdida Logloss y la optimización de la métrica AUC.

Tabla 5 Mejores hiperparámetros para CatBoost - Asignación de cupos

Parámetro	**Valores usados**
iterations	100.0
depth	4
loss_function	Logloss
custom_loss	AUC

3.4.3. MLP

Se creó un diccionario que especifica los diferentes valores que se probarán para cada hiperparámetro de la red neuronal, esto se evidencia en la Tabla *6*. Los hiperparámetros usados son configuraciones que controlan el comportamiento de la red, como la arquitectura (número de capas ocultas y neuronas por capa), la función de activación, la tasa de aprendizaje (learning_rate) y el número máximo de iteraciones del algoritmo de optimización (max_iter).

Tabla 6 Grid de hiperparámetros para MLP - Asignación de cupos

Parámetro	**Valores usados**
hidden_layer_sizes	(50,), (100,), (50, 50)
activation	identity, logistic, tanh, relu
learning_rate	constant, invscaling, adaptive
max_iter	100, 200

Gracias a un exhaustivo proceso de búsqueda de hiperparámetros utilizando GridSearchCV, se identificó la configuración óptima para el modelo MLP. Se evaluaron 72 combinaciones diferentes de hiperparámetros, resultando en un total de 360 ajustes. Como se detalla en la Tabla *7*, la mejor configuración incluye una capa oculta con 50 neuronas, función de activación identidad, tasa de aprendizaje constante y un máximo de 100 iteraciones.

Tabla 7 Mejores hiperparámetros para MLP - Asignación de cupos

Parámetro	**Mejor valor**
hidden_layer_sizes	(50,)
activation	identity
learning_rate	constant
max_iter	100

3.4.4. Evaluación

Luego que se encontraron los mejores parámetros para todos los modelos, se ajustaron a los datos de entrenamiento utilizando la búsqueda en cuadrícula y se realizaron predicciones sobre el conjunto de prueba con el modelo entrenado con los mejores hiperparámetros.

Para evaluar el desempeño de los modelos, se emplearon diversas métricas de evaluación apropiadas, como precision, acurracy, recall, F1-score y curva ROC.

3.5. Modelado aceptación de cupos

De acuerdo con la variable predictora “PREDICT_ACEPTACION_CUPO”, existían 16.999 mil postulantes que aceptaron cupo y 3.115 mil con cupo rechazado, por lo cual se decidió balancear los datos usando resampling, se realizaron pruebas con undersampling y oversampling.

Luego se aplicó Target Encoding solo en el conjunto de entrenamiento a las columnas categóricas y se dividió el dataset en partes de entrenamiento (70%) y test (30%). Posteriormente se implementaron los algoritmos usando siempre una misma semilla inicial para para garantizar la reproducibilidad de los resultados.

3.5.1. CatBoost

Para optimizar el rendimiento del modelo CatBoost, se utilizó GridSearchCV para evaluar una amplia gama de hiperparámetros. La Tabla *8* presenta las diferentes combinaciones exploradas, incluyendo un rango de valores para el número de iteraciones, la profundidad de los árboles, la función de pérdida, la regularización de las hojas, el número de iteraciones de estimación de hojas, la tasa de aprendizaje, el nivel de registro y una semilla aleatoria.

Tabla 8 Grid de hiperparámetros para CatBoost - Aceptación de cupos

Parámetro	**Valores usados**
iterations	100.0, 200, 300
depth	4, 5, 6
loss_function	Logloss, CrossEntropy
custom_loss	AUC, Accuracy, F1, Recall

Tras evaluar múltiples combinaciones de valores para el número de iteraciones, la profundidad máxima de los árboles, la función de pérdida y la métrica de evaluación, se concluyó que la configuración óptima, presentada en la Tabla *9*, correspondía a 300 iteraciones, una profundidad máxima de 6, la función de pérdida Logloss y la optimización del AUC. Esta configuración permitió obtener el mejor equilibrio entre el ajuste del modelo a los datos de entrenamiento y su capacidad de generalización a nuevos datos.

Tabla 9 Mejores hiperparámetros para CatBoost - Aceptación de cupos

Parámetro	**Mejor valor**
iterations	300
depth	6
loss_function	Logloss
custom_loss	AUC

3.5.2. Random Forest

Como se muestra en la Tabla 10, se exploraron diversas combinaciones de hiperparámetros para el modelo de Random Forest, incluyendo el número de árboles en el bosque (n_estimators), la profundidad máxima de cada árbol (max_depth), el número mínimo de muestras requeridas para dividir un nodo interno y una hoja. (min_samples_split y min_samples_leaf)

Tabla 10 Grid de hiperparámetros para Random Forest - Aceptación de cupos

Parámetro	**Valores usados**
n_estimators	100, 200
max_depth	10, 20
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 5, 10

Tras evaluar 48 combinaciones de hiperparámetros a través de 5 pliegues de validación cruzada, se determinó que la configuración óptima para el modelo de Random Forest, detallada en la Tabla 11, corresponde a 200 estimadores, una profundidad máxima de 20, un mínimo de 2 muestras para dividir un nodo y 1 muestra por hoja.

Tabla 11 Mejores hiperparámetros para Random Forest - Aceptación de cupos

Parámetro	**Mejor valor**
n_estimators	200
max_depth	20
min_samples_split	2
min_samples_leaf	1

3.5.3. MLP

Los hiperparámetros de una red neuronal, que controlan aspectos como su arquitectura y el proceso de aprendizaje, fueron explorados de manera exhaustiva. La Tabla *12* presenta el rango de valores considerados para cada hiperparámetro, incluyendo el número de capas ocultas (hidden_layer_sizes), la función de activación (activation), la tasa de aprendizaje (learning_rate) y el número máximo de iteraciones (max_iter).

Tabla 12 Grid de hiperparámetros para MLP - Aceptación de cupos

Parámetro	**Valores usados**
hidden_layer_sizes	(50,), (100,), (50, 50)
activation	‘identity’, ‘logistic’, ‘tanh’, ‘relu’
learning_rate	‘constant’, ‘invscaling’, ‘adaptive’
max_iter	100, 200

El ajuste de hiperparámetros para la red neuronal evaluó 72 combinaciones diferentes a través de 5 pliegues de validación cruzada, lo que resultó en un total de 360 ajustes. La configuración óptima, detallada en la Tabla *13*, indica que la mejor performance se obtuvo con una capa oculta de 50 neuronas, función de activación ReLU, tasa de aprendizaje adaptativa y un máximo de 100 iteraciones.

Tabla 13 Mejores hiperparámetros para MLP - Aceptación de cupos

Parámetro	**Mejor valor**
hidden_layer_sizes	(50,)
activation	relu
learning_rate	*adaptive*
max_iter	100

3.5.4. Evaluación

Luego que se encontraron los mejores parámetros para todos los modelos, se ajustaron a los datos de entrenamiento utilizando la búsqueda en cuadrícula y se realizaron predicciones sobre el conjunto de prueba con el modelo entrenado con los mejores hiperparámetros.

Para evaluar el desempeño de los modelos, se emplearon diversas métricas de evaluación apropiadas, como precision, acurracy, recall, F1-score y curva ROC.

3.6. Interpretabilidad con SHAP Values

3.6.1. Modelo de asignación cupos

Al aplicar la técnica de SHAP, se descompuso el modelo de predicción para determinar la importancia relativa de cada característica en la asignación de cupos. Los resultados, visualizados en la Figura *13*, indican que las variables 'Grupo de asignación' y 'Puntaje de postulación' tienen el mayor efecto marginal sobre la salida del modelo. Estos hallazgos corroboran la importancia de estas variables, tal como se establece en el reglamento de admisión y nivelación.

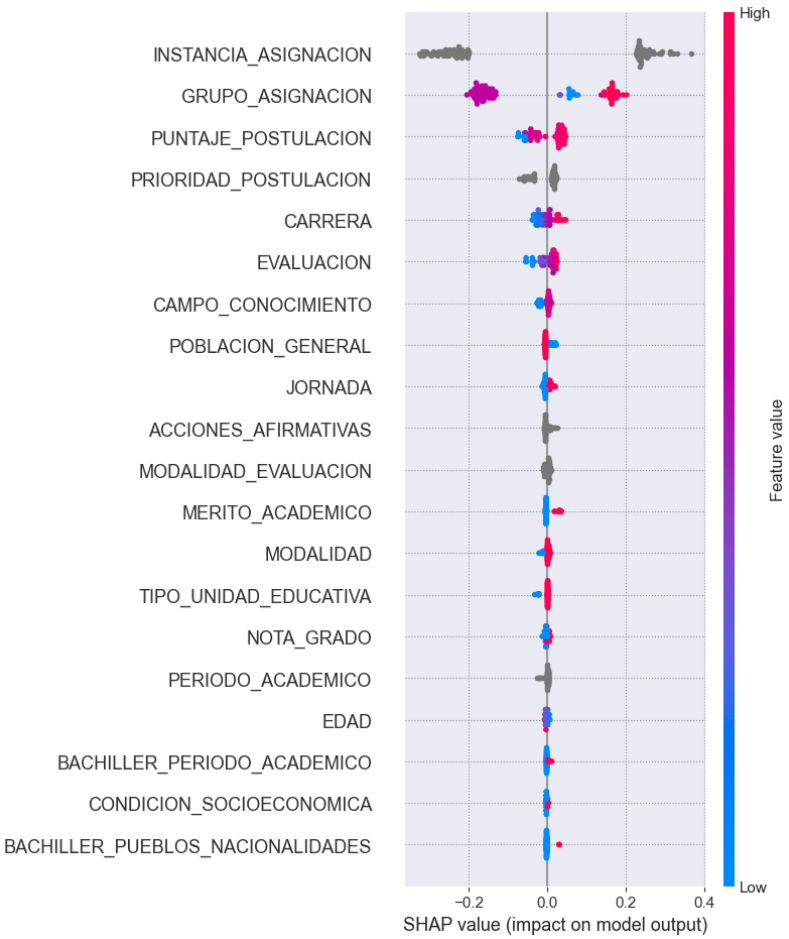

Figura 13 Importancia de variables usando SHAP values en la etapa de asignación de cupos

3.6.2. Modelo aceptación cupos

El impacto de las variables para este modelo de predicción se midió con la importancia de estas y lo podemos ver gráficamente en la Figura 14 Importancia de variables usando SHAP values en la etapa de aceptación de cupos, de las cuales se eligieron 5 como las más representativas y el resto no tiene impacto en el resultado; nota de grado, puntaje de postulación, carrera, cantón de residencia y edad. Estas características tienen una fuerte influencia positiva en la predicción de aceptación.

Los estudiantes con notas y puntajes más altos tienen una mayor probabilidad de ser admitidos, la carrera elegida también parece influir en la decisión de admisión, aunque su impacto es menos claro debido a la dispersión de los puntos. Podría haber carreras más competitivas o con cupos más limitados.

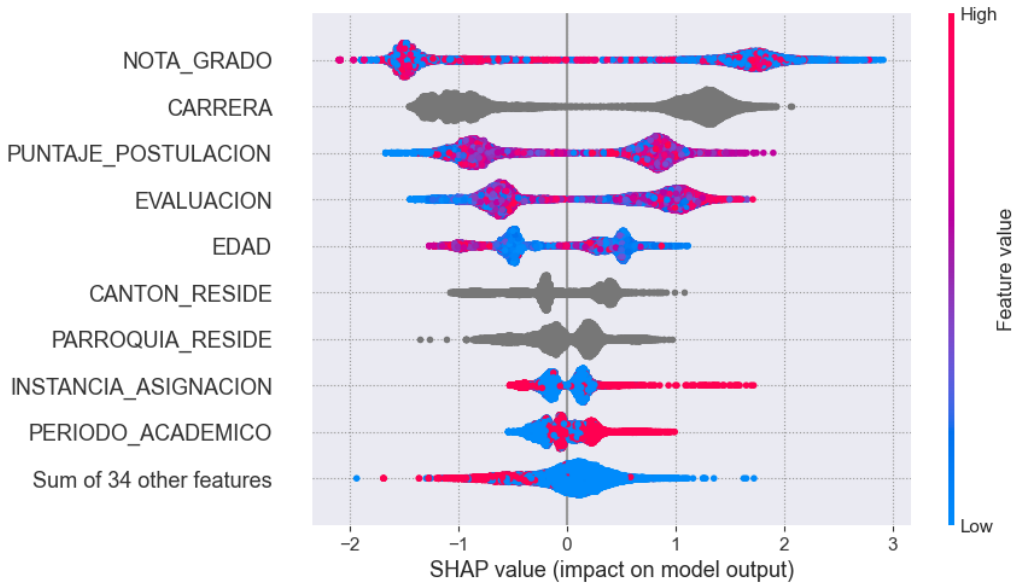

Figura 14 Importancia de variables usando SHAP values en la etapa de aceptación de cupos

CAPÍTULO 4

4. Análisis de resultados

4.1. Asignación de cupos

Los modelos de Random Forest, CatBoost y MLP, evaluados bajo diferentes configuraciones de hiperparámetros y técnicas de remuestreo, alcanzaron un desempeño “perfecto” en la predicción de la asignación de cupos. Como se evidencia en la Tabla *14*, todas las métricas evaluadas; precisión, recall, F1-score y el área bajo la curva ROC, obtuvieron un valor de 1.0, indicando una clasificación correcta del 100% de las instancias.

Tabla 14 Resultado de métricas de evaluación para los modelos de asignación de cupos

Métrica	Random Forest	CatBoost	MLP
precision	1.0	1.0	1.0
recall	1.0	1.0	1.0
F1	1.0	1.0	1.0
acurracy	1.0	1.0	1.0
curva ROC	1.0	1.0	1.0

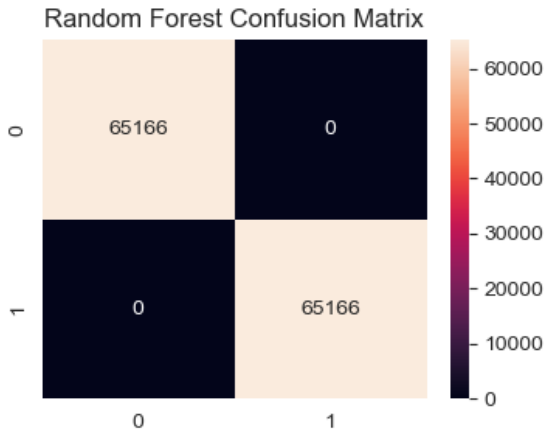

Figura 15 Matriz de confusión usando Random Forest en la predicción de asignación de cupo

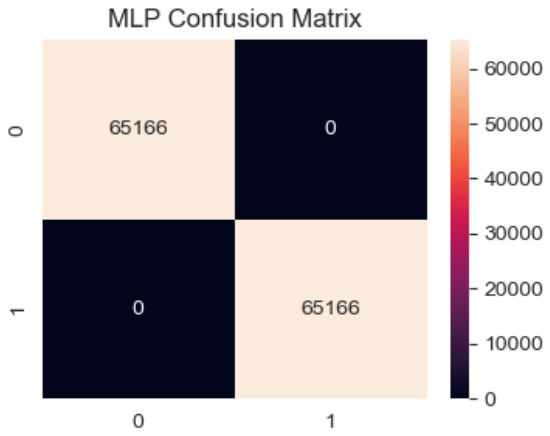

Figura 16 Matriz de confusión usando MLP en la predicción de asignación de cupo

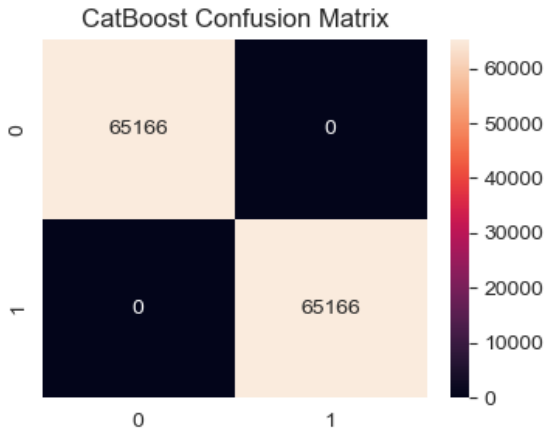

Figura 17 Matriz de confusión usando CatBoost en la predicción de asignación de cupo

Como se observa en las matrices de confusión de las figuras 15, 16 y 17 para los modelos de asignación de cupo, todas son iguales debido a que en las métricas de evaluación todas obtuvieron un puntaje de 1.0, al igual que si graficamos la curva ROC con los 3 modelos, podremos ver como lo muestra la Figura *18* que se sobreponen en los mismos puntos debido a su precisión.

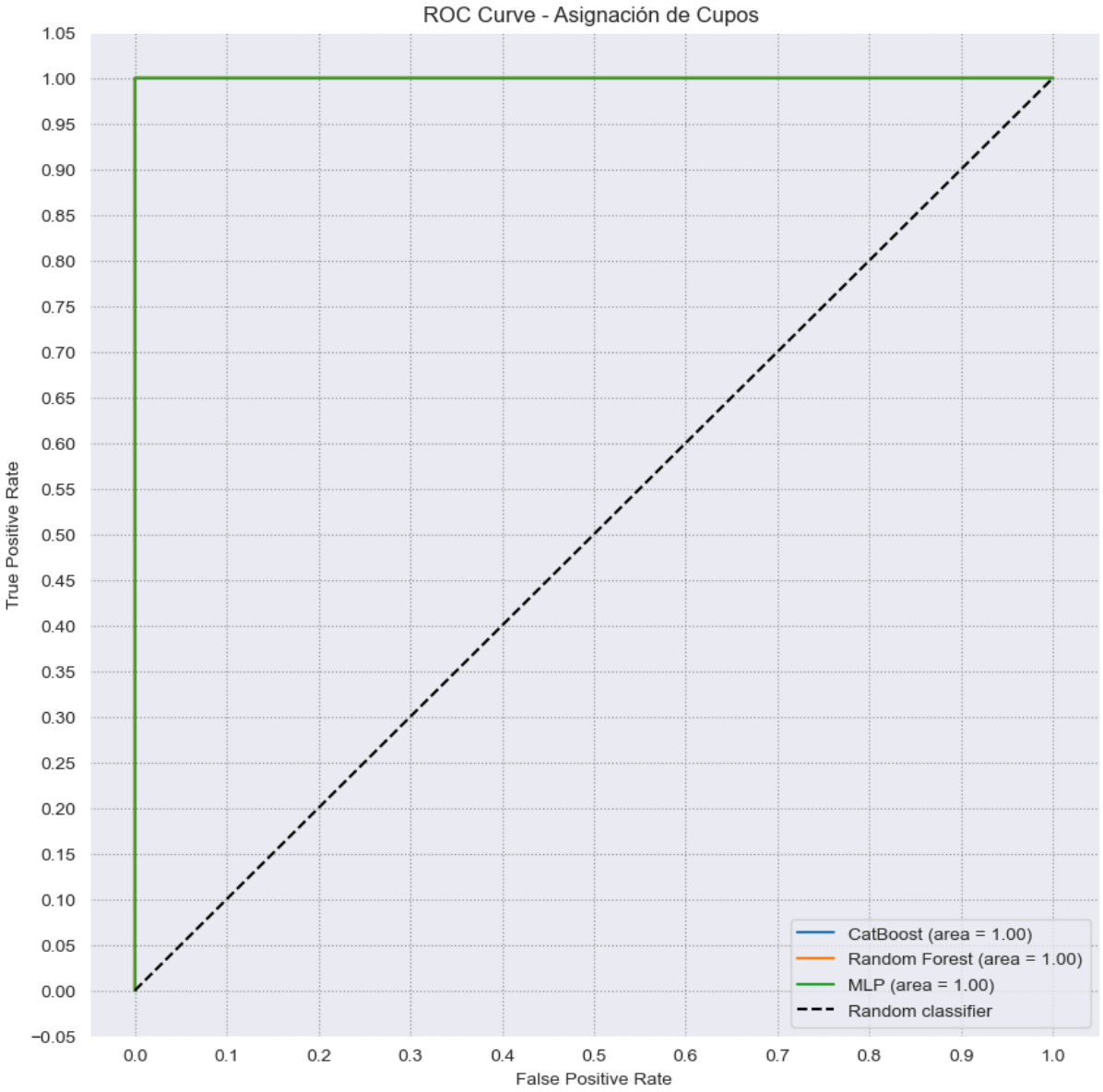

Figura 18 Curva ROC de los modelos implementados para la asignación de cupos

4.2. Aceptación de cupos

Los modelos de Random Forest, CatBoost y MLP fueron evaluados en la tarea de predicción de la aceptación de cupos. Se puede evidenciar en la Tabla *15* los resultados obtenidos, los cuales muestran que tanto Random Forest como CatBoost alcanzaron altas tasas de precisión, recall, F1-score y accuracy, con valores superiores a 0.95. Por su parte, el modelo MLP obtuvo un desempeño notablemente inferior en todas las métricas evaluadas. El área bajo la curva ROC también fue superior para Random Forest y CatBoost en comparación con MLP.

Tabla 15 Resultado de métricas de evaluación para los modelos de aceptación de cupos

Métrica	Random Forest	CatBoost	MLP
precision	0.9579	0.9574	0.6539
recall	0.8512	0.9534	0.6568
F1	0.9014	0.9537	0.6543
acurracy	0.9073	0.9539	0.6557
curva ROC	0.9824	0.9844	0.6801

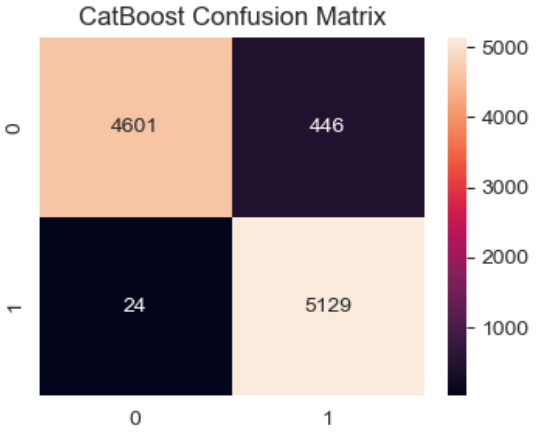

Figura 19 Matriz de confusión usando CatBoost en la predicción de aceptación de cupo

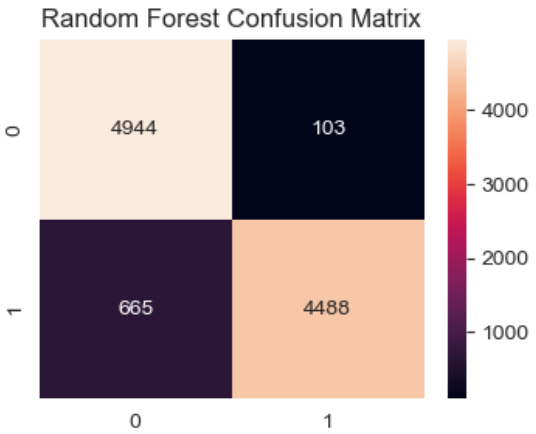

Figura 20 Matriz de confusión usando Random Forest en la predicción de aceptación de cupo

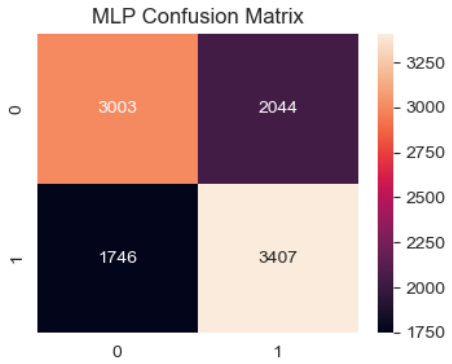

Figura 21 Matriz de confusión usando MLP en la predicción de aceptación de cupo

Como se muestra en la figura 19 y 20, tanto CatBoost como Random Forest muestran un rendimiento significativamente superior al modelo MLP. Esto se evidencia en el número de predicciones correctas y en la menor cantidad de falsos positivos y falsos negativos. Ambos modelos logran clasificar correctamente una gran proporción de las instancias, lo que sugiere una alta precisión en sus predicciones.

Por otro lado, el modelo MLP presenta un mayor número de falsos positivos y falsos negativos en comparación con los otros dos modelos como se puede observar en la Figura *21*. Esto indica que el MLP tiende a confundir las clases con mayor frecuencia, lo que puede ser un problema en aplicaciones donde el costo de los falsos positivos o falsos negativos es elevado.

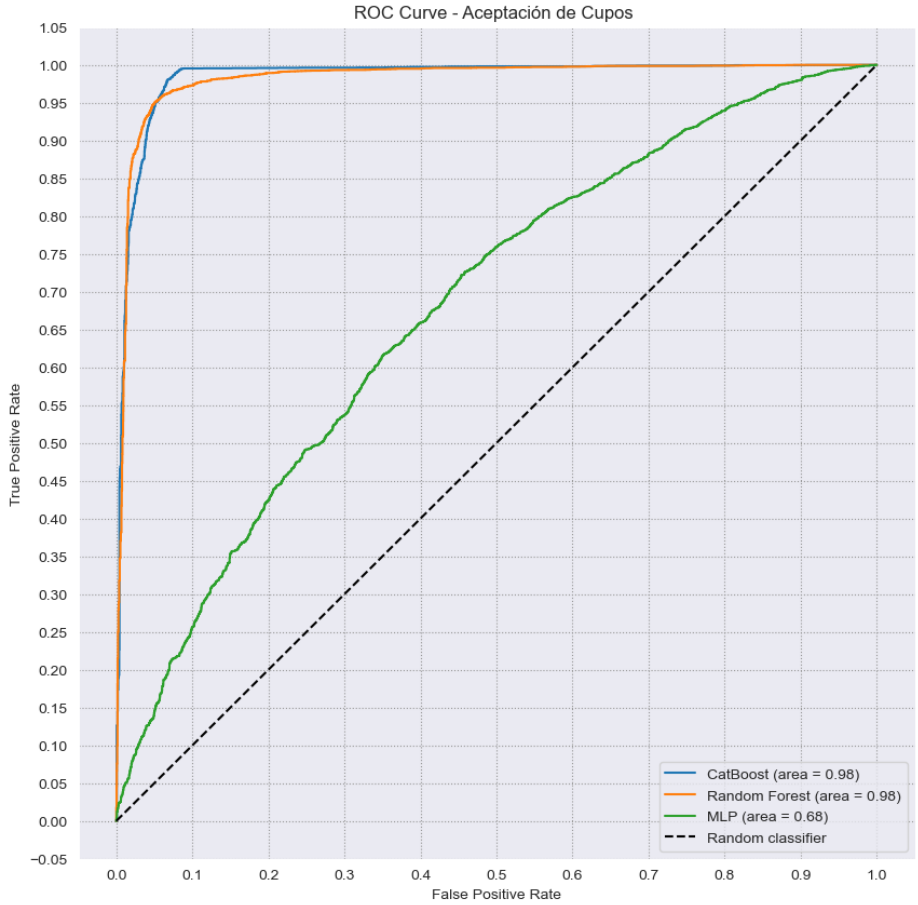

Figura 22 Curva ROC para los modelos implementados en la aceptación de cupos

La Figura *22* presenta las curvas ROC (Receiver Operating Characteristic) de los modelos de clasificación implementados para la aceptación de cupos. Estas curvas permiten evaluar el desempeño de cada modelo en términos de su capacidad para distinguir entre las clases positivas (aceptados) y negativas (rechazados). El área bajo la curva (AUC) proporciona una medida cuantitativa del rendimiento, siendo valores cercanos a 1 indicativos de un mejor desempeño, como se observa en esta figura y corroborando lo que se mostró en las matrices de confusión, MLP tiene un menor rendimiento en comparación con Random Forest y CatBoost.

Recall por clase

De los modelos CatBoost y Random Forest, para poder elegir el de mejor rendimiento es necesario obtener sus métricas por clase.

Tabla 16 Resultado de Recall por clase para la aceptación de cupos

	CatBoost	**Random Forest**
Clase	**Recall**	
0 (Rechaza cupo)	0.91	0.98
1 (Acepta cupo)	1.00	0.87

Como se expone en la Tabla *16*, Random Forest tiene un recall ligeramente superior a CatBoost (0.98 vs. 0.91). Esto significa que es ligeramente mejor en identificar correctamente los casos en los que se rechaza un cupo. CatBoost tiene un recall perfecto de 1.0. con lo cual identifica correctamente todos los casos en los que se acepta un cupo. Random Forest, por su parte, tiene un recall de 0.87, lo que indica que no identifica correctamente todos los casos positivos de esta clase.

Resultados específicos

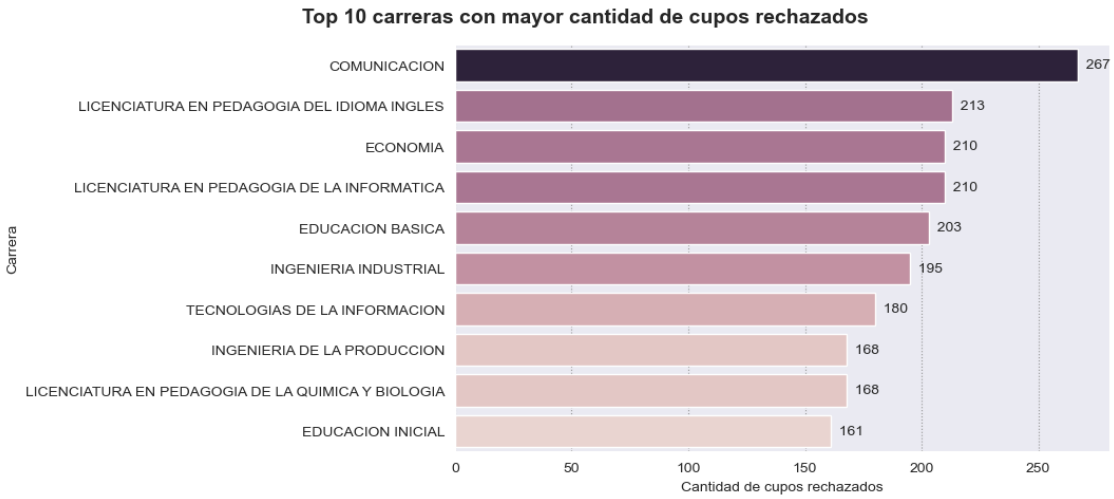

Figura 23 Top 10 carreras con mayor cantidad de cupos rechazados

Uno de los posibles resultados específicos que la IES puede obtener con este proyecto es el de la Figura *23* Top 10 carreras con mayor cantidad de cupos rechazados, la alta tasa de rechazo en estas carreras podría deberse a una

saturación del mercado laboral, a que los estudiantes podrían no tener información suficiente sobre las salidas laborales, planes de estudio u oportunidades de desarrollo profesional de estas carreras.

Una vez identificadas las carreras que se predice tendrían cantidades elevadas de rechazo, ante de iniciar el proceso de admisión, se podría realizar estudios más detallados para identificar las causas específicas del rechazo en cada carrera. Esto puede incluir encuestas a estudiantes, egresados y empleadores. Además, revisar y actualizar los planes de estudio para que sean más atractivos y respondan a las demandas del mercado laboral, así como también fortalecer los servicios de orientación vocacional para ayudar a los estudiantes a tomar decisiones informadas sobre su futuro académico.

CONCLUSIONES Y RECOMENDACIONES

5. Conclusiones

- Al combinar la interpretación que proporcionaron los valores SHAP con los algoritmos de ML usados, se obtuvo que, para el modelo de aceptación de cupos, la nota de grado, carrera asignada, puntaje de postulación, nota de evaluación realizada por la IES y edad del postulante son las características más influyentes en la decisión de aceptar o rechazar un cupo.

- Las carreras que se predice tienen la mayor cantidad de cupos rechazados son comunicación, licenciatura en pedagogía del idioma inglés, informática, química y biología, economía, educación básica e inicial, ingeniería industrial, tecnologías de la información y ingeniería de la producción. Estas carreras representan el 35% de cupos rechazados.

- Según el análisis comparativo del modelado de asignación de cupo, se descarta el modelo MLP por su bajo nivel de precisión (62%). CatBoost tuvo un 95% de precisión, analizado el recall por clase, es excelente en identificar todos los casos positivos cuando se acepta un cupo (1), sin embargo, podría estar perdiendo algunos casos positivos de cupos rechazados (0.91). Random Forest con un 92% de precisión es muy bueno en identificar casos positivos de rechazo de cupos (0.98), pero podría estar perdiendo algunos casos positivos de cupos aceptados (0.87), por lo cual este último es elegido el mejor modelo.
- El reglamento del SNNA (Sistema Nacional de Nivelación y Admisión) establece un proceso fijo de asignación de cupos para todas las IES, lo cual significa que los resultados se obtienen siguiendo reglas preestablecidas y no hay un patrón de la cual se puede establecer un modelo que aprenda el comportamiento de asignar cupos, esto es, usando los datos de la IES de este proyecto.

6. Recomendaciones

- Es necesario aumentar la granularidad de la predicción al considerar el proceso de admisión como una secuencia de eventos interrelacionados. Aumentar en varios “niveles” la predicción realizada debido a que el proceso de admisión el cual se contempla en este proyecto es la primera etapa de la vida educativa de un estudiante a nivel superior, se debe de considerar luego de las personas que aceptan un cupo, si se matriculan al curso de nivelación o preuniversitario, si aprueban dicho curso y logran matricularse en primer semestre/ciclo.

- Durante el proceso de admisión en IES estudiada, se empezaron a recopilar comentarios tanto de los estudiantes que rechazan como de los que aceptan un cupo. Los datos textuales proporcionados por los postulantes al rechazar o aceptar un cupo representan una valiosa fuente de información cualitativa. A futuro, se podría aplicar el procesamiento del lenguaje natural para extraer patrones y temas recurrentes, lo que permitiría comprender mejor las razones detrás de las decisiones de los postulantes y optimizar el proceso de admisión.

- Para futuras investigaciones, sería interesante explorar la incorporación de nuevas variables, como datos de redes sociales, para mejorar la precisión de las predicciones. Además, se podrían desarrollar modelos más complejos, como redes neuronales profundas, para capturar relaciones no lineales entre las variables.

REFERENCIAS

[1] A. Burneo y D. Yunga, «Acceso de los Jóvenes a la Educación Universitaria en el Ecuador: Reformas, Políticas y Progreso,» *Sísifo — Revista de Educación,* vol. 8, nº 2, pp. 70-85, 2020.

[2] SENESCYT, *ACUERDO Nro. SENESCYT-SENESCYT-2023-0003-AC,* 2023.

[3] J. Terán, «El acceso a la educación superior como derecho humano,» *Revista Jurídica Crítica Y Derecho,* vol. 1, nº 1, pp. 1-12, 2020.

[4] G. Terán Sevilla, «Guía de Jurisprudencia Constitucional,» 2023. [En línea]. Available: https://repositorio.dpe.gob.ec/bitstream/39000/3552/1/DEPE-DPE-001-2024.pdf.

[5] A.Sivasangari, V.Shivani y Y.Bindhu, «Prediction Probability of Getting an Admission into a University using Machine Learning,» *Fifth International Conference on Computing Methodologies and Communication,* 2021.

[6] A. Iman y X. Tian, «A Comparison of Classification Models in Predicting Graduate Admission Decision,» *Journal of Higher Education Theory and Practice,* vol. 21, nº 7, 2021.

[7] F. Bugueño, «Modelo predictivo para la selección de postulantes destacados a una institución de educación superior,» 2017. [En línea]. Available: https://repositorio.uchile.cl/handle/2250/147565.

[8] K. Benavides, B. Maldonado y J. Buenaño, «Factores asociados a la aceptación de un cupo en las universidades de los bachilleres de Manabí,» *Revista Internacional Administración & Finanzas,* pp. 63-76, 2018.

[9] S. Ahmad, M. Asgher, A. Malik, Z. Mushtaq, K. Khan y Z. Abbas, «Data Mining Methods and Obstacles: A Comprehensive Analysis,» *Journal of Computing & Biomedical Informatic,* vol. 6, nº 1, 2023.

[10] J. Becerra y E. Villarreal, « Data Mining para modelo predictivo de ventas y servicios de mantenimiento en un concesionario automotriz ligero,» 2021.

[11] J. Moré, «Evaluación de la calidad de los sistemas de reconocimientos de sentimientos,» 2019. [En línea]. Available:

https://openaccess.uoc.edu/bitstream/10609/148645/3/Modulo3_EvaluacionDeLaCalidadDeLosSistemasDeReconocimientoDeSentimientos.pdf.

[12] L. Polanía, «Evaluación de modelos de Machine Learning para Sistemas de Detección de Intrusos en Redes IoT,» 2021.

[13] A. Zapeta, G. Galindo, H. Juan y M. Martínez, «Métricas de rendimiento para evaluar el aprendizaje automático en la clasificación de imágenes petroleras utilizando redes neuronales convolucionales,» *Ciencia Latina Revista Científica Multidisciplinar,* vol. 6, nº 5, 2022.

[14] M. Bilal, G. Ali, M. Wassem, M. Anwar, M. Arshad y R. Abdul, «Auto-Prep: Efficient and Automated Data Preprocessing Pipeline,» 2022.

[15] J. Martínez y P. Pérez, «ROC Curve,» *Semergen,* vol. 49, nº 1, 2023.

[16] I. Roy, C. Paredes, J. Moreno, R. Rivas y A. Flores, «ROC curves: general characteristics and their usefulness in clinical practice,» *Rev Med Inst Mex Seguro Soc,* vol. 61, 2023.

[17] E. Sánchez, Y. Hernández, J. Ortiz, A. Martínez y H. Estrada, «Configuración de hiperparámetros mediante algoritmos de optimización: Aplicación en la predicción de enfermedades cardiovasculares,» *Research in Computing Science,* vol. 152, nº 8, 2023.

[18] G. Giménez, «Análisis y aprovechamiento de bases de datos agronómicas recurriendo al proceso “Knowledge discovery in databases” (KDD) y algoritmos de “data mining”(DM). Una Aplicación al pronóstico de producción de frutas de pepita en los Valles de Río Negro y Neuquén,» 2020.

[19] M. A. Azeemi, A. A. S. Alarood y M. I. Uddin, «Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models,» *PeerJ Computer Science,* 2022.

[20] S. Matthews y B. Hartman, «mSHAP: SHAP Values for Two-Part Models,» *Risks,* vol. 10, nº 1, 2022.

[21] G. Van Rossum, «Python 3 Reference Manual. Scotts Valley,» 2009. [En línea]. Available: https://www.python.org/doc/essays/blurb/.

[22] «Project Jupyter,» [En línea]. Available: https://jupyter.org/. [Último acceso: 22 11 2024].

[23] D. Cournapeau, «Scikit-learn: machine learning in Python,» [En línea]. Available: https://scikit-learn.org/. [Último acceso: 22 11 2024].

[24] «Pandas - Python Data Analysis Library,» [En línea]. Available: https://pandas.pydata.org/. [Último acceso: 22 11 2024].

[25] «NumPy: The fundamental package for scientific computing with Python,» [En línea]. Available: https://numpy.org/. [Último acceso: 22 11 2024].

[26] «Matplotlib,» [En línea]. Available: https://matplotlib.org/stable/. [Último acceso: 22 11 2024].

[27] «Seaborn: statistical data visualization,» [En línea]. Available: https://seaborn.pydata.org/. [Último acceso: 22 11 2024].

[28] «SHAP,» [En línea]. Available: https://shap.readthedocs.io/en/stable/. [Último acceso: 22 11 2024].