Desarrollo de un modelo inteligente de análisis de datos que prediga los posibles casos de deserción de estudiantes en los primeros niveles del Instituto Superior Tecnológico Cotopaxi

Ing. Rosero Valdiviezo, José Luis

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

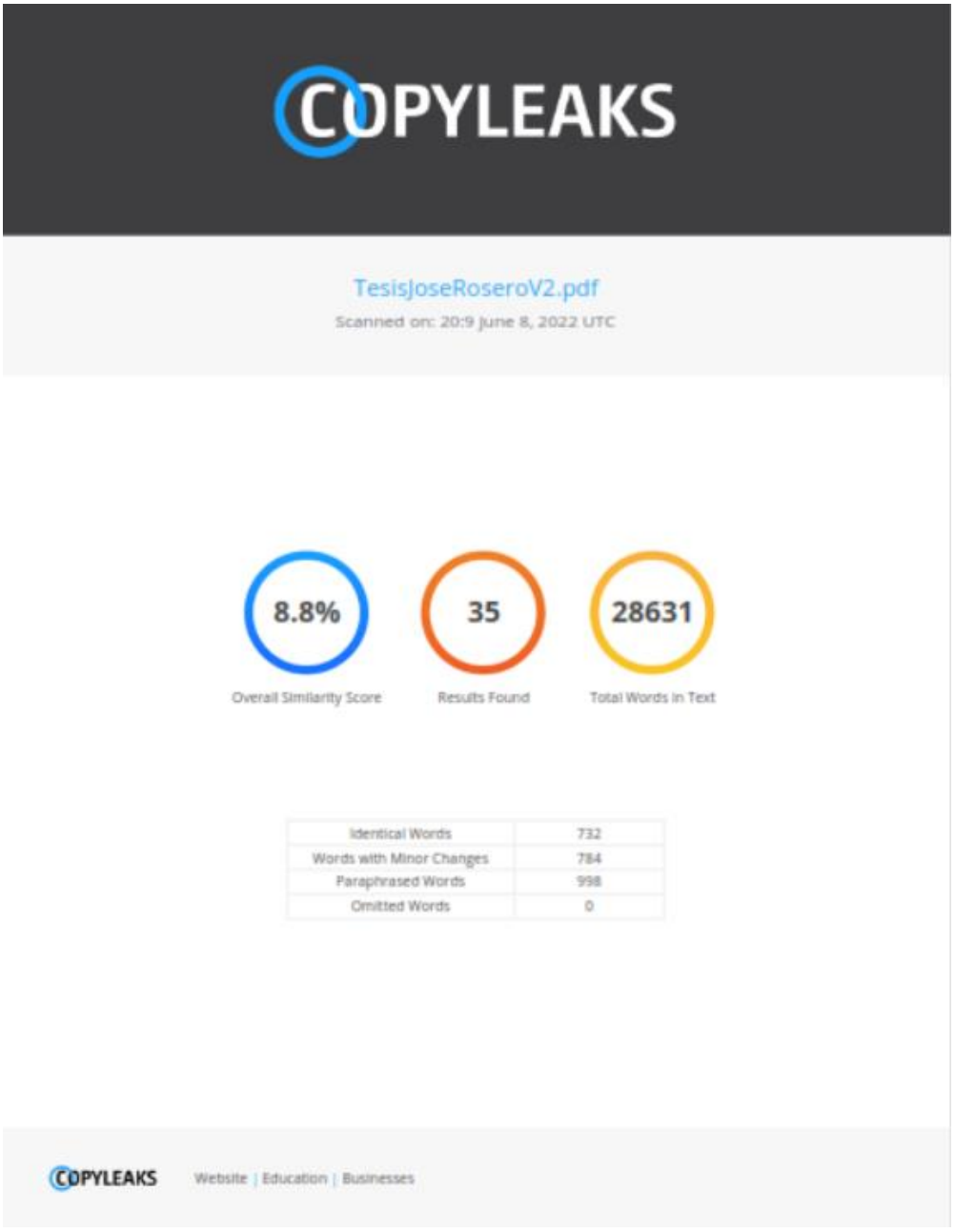
Maestría en Ingeniería en Software

Trabajo de titulación, previo a la obtención del título de Magister en Ingeniería en Software

Ing. Gallardo Corrales, Diego Eduardo

2 de diciembre de 2022

Latacunga – Ecuador



Firmado electrónicamente por:
DIEGO EDUARDO GALLARDO CORRALES

. . . . . .

Ing. Gallardo Corrales, Diego Eduardo

Director

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Certificación

Certifico que el trabajo de titulación: **"Desarrollo de un modelo inteligente de análisis de datos que prediga los posibles casos de deserción de estudiantes en los primeros niveles del Instituto Superior Tecnológico Cotopaxi"** fue realizado por el señor **Rosero Valdiviezo, José Luis**; el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Latacunga, 2 de diciembre de 2022

Firmado electrónicamente por:
DIEGO EDUARDO GALLARDO CORRALES

. . . . . .

Ing. Gallardo Corrales, Diego Eduardo

Director

C.C.: 0503201717

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Responsabilidad de Autoría

Yo **Rosero Valdiviezo, José Luis**, con cédula de ciudadanía N° 0502337157, declaro que el contenido, ideas y criterios del trabajo de titulación***: "*Desarrollo de un modelo inteligente de análisis de datos que prediga los posibles casos de deserción de estudiantes en los primeros niveles del Instituto Superior Tecnológico Cotopaxi"** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Latacunga, 2 de diciembre de 2022

JOSE LUIS ROSERO VALDIVIEZO

Firmado digitalmente porJOSE LUIS ROSERO VALDIVIEZO

.....

Rosero Valdiviezo, José Luis

C.C.: 0502337157

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Autorización de Publicación

Yo **Rosero Valdiviezo, José Luis**, con cédula de ciudadanía N° 0502337157 autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: ***"*Desarrollo de un modelo inteligente de análisis de datos que prediga los posibles casos de deserción de estudiantes en los primeros niveles del Instituto Superior Tecnológico Cotopaxi"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Latacunga, 2 de diciembre de 2022

JOSE LUIS ROSERO VALDIVIEZO

Firmado digitalmente porJOSE LUIS ROSERO VALDIVIEZO

. . . . . .

Rosero Valdiviezo, José Luis

C.C.: 0502337157

Dedicatoria

El presente trabajo lo dedico a mis Padres Oswaldo y Chelita quienes con su amor incondicional siempre me apoyan y alientan a cumplir mis metas.

A mi esposa, y a mis bellas hijas Sol y Paz quienes con sus abrazos y sonrisas me llenan la vida de optimismo y convicción para culminar el presente trabajo.

A mis queridos hermanos y hermanas quienes siempre están a mi lado apoyándome en todo.

José L. Rosero V.

Agradecimiento

A Dios y a la Virgen de la Merced por sus abundantes bendiciones.

A la Universidad de las Fuerzas Armadas ESPE por ser mi alma mater que me abrió sus puertas para poder educarme en sus aulas y llegar a ser un profesional.

A mi Tutor de Tesis por su apoyo y buena predisposición para sacar adelante el presente trabajo de titulación.

Al Instituto Superior Tecnológico Cotopaxi que me brindó la oportunidad de educar a sus estudiantes y el de poder desarrollar el presente trabajo.

José L. Rosero V.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Resumen

El abandono educativo es un fenómeno que se ha venido produciendo con más frecuencia en las Instituciones de educación superior, y los Institutos técnicos y tecnológicos no son la excepción. El abandono temprano provoca un impacto negativo emocional en los estudiantes que ven frustrados sus sueños de obtener un título de tercer nivel, produciéndose un incremento de este indicador de abandono en las instituciones de educación superior en el Ecuador. En la actualidad la mayoría de las instituciones de educación superior cuentan con un sistema informático académico para el registro de su gestión Institucional. La información del alumnado almacenada en las bases de datos del Instituto constituyó la base fundamental para el desarrollo del presente trabajo basado en el análisis y selección de la información de los estudiantes que ingresaron a los primeros niveles del Instituto en un DataSet, evaluarlo mediante la aplicación de los modelos supervisados de Machine Learning como son árboles de decisión, bosques aleatorios, y redes neuronales. Cada modelo de clasificación fue evaluado mediante sus respectivas métricas de precisión y el modelo ganador fue desplegado en la nube mediante la plataforma como servicio Heroku, cumpliendo de esta manera todas las fases de la metodología CRISP-DM. Se concluye que la aplicación de técnicas de Machine Learning en el ámbito educativo contribuye significativamente a identificar estudiantes con riesgo de deserción en el Instituto Superior Tecnológico Cotopaxi, esta identificación permitirá desarrollar planes de acción que ayuden a mitigar dichos riesgos, disminuyendo el porcentaje de deserción estudiantil.

Palabras Claves: Machine Learning, Árboles de Decisión, Bosques Aleatorios, Redes Neuronales, Abandono Escolar, Deserción Escolar.

Abstract

Educational dropout is a phenomenon that has been occurring more frequently in higher education institutions, and technical and technological institutes are no exception. Early dropout causes a negative emotional impact on students who see their dreams of obtaining a third level degree frustrated, producing an increase in this dropout indicator in higher education institutions in Ecuador. At present, most of the higher education institutions have a scholastic computer system for the registration of their institutional management. The student information stored in the Institute's databases constituted the fundamental basis for the development of the present work based on the analysis and selection of the information of the students who entered the first levels of the Institute in a DataSet, evaluating it through the application of supervised Machine Learning models such as decision trees, random forests, and neural networks. Each classification model was evaluated by means of its respective accuracy metrics and the winning model was deployed in the cloud through the Heroku platform as a service, thus fulfilling all the phases of the CRISP-DM methodology. It is concluded that the application of Machine Learning techniques in the educational field contributes significantly to identify students at risk of dropout in the Instituto Superior Tecnológico Cotopaxi, this identification will allow the development of action plans to help mitigate these risks, reducing the percentage of student dropout.

Keywords: Machine Learning, Decision Tree, Random Forest, Neural Network, School Dropout.

Capítulo I

Introducción

La deserción en la Educación Superior en el Ecuador ha tenido varios estudios, los cuales se enfocan principalmente en la deserción en las Universidades, sin embargo, los estudios de este fenómeno en los Institutos Superiores Técnicos y Tecnológicos son escasos.

La deserción escolar es un fenómeno que se produce en las Instituciones de educación superior en el Ecuador, y los Institutos no son la excepción. Debido a ello el Consejo de Aseguramiento de la Calidad CACES, detalla los elementos fundamentales que los Institutos deben acatar en referencia a este indicador en el proceso de acreditación y sobre todo de acompañamiento y permanencia de los estudiantes.

En el “MODELO DE EVALUACIÓN INSTITUCIONAL PARA LOS INSTITUTOS SUPERIORES TÉCNICOS Y TECNOLÓGICOS EN PROCESO DE ACREDITACIÓN 2020”, el indicador 6.1.1 Acompañamiento a estudiantes, su estándar indica “*El instituto diseña y aplica un proceso de acompañamiento a los estudiantes desde su ingreso (admisión y/o nivelación), que contribuye a su motivación para lograr un adecuado rendimiento académico, desarrollo integral y su permanencia en la institución hasta la culminación de sus estudios*.” (CACES, 2020)

Mitigar la deserción escolar en la educación superior es una actividad prioritaria para los órganos rectores de la educación superior, en donde los institutos deben formular planes de acción y encontrar los mecanismos adecuados para disminuir los índices de deserción. En este contexto el presente trabajo se apalancará del Machine Learning en el campo educativo para predecir posibles casos de deserción escolar entre otras buenas prácticas.

Antecedentes

La deserción estudiantil cuenta con sus primeros estudios por los años 1970 y 1980 (Spady, 1971). La preocupación por el estudio de esta temática incrementó a partir de la década de 1990, al experimentarse un incremento en las matrículas, en consecuencia, el índice de la deserción incrementó.

En los años 80 del siglo XX la sociedad explicaba la deserción universitaria debido a un bajo promedio en el bachillerato, el estado civil y la necesidad de los estudiantes de combinar estudios con la actividad laboral. (Ramirez, 2016)

De acuerdo con la literatura la deserción o la permanencia estudiantil está condicionada por una fuerte conjunción entre el estudiante y la institución. (Ramirez, 2016)

La ineficacia del sistema conlleva una pérdida de estudios y a una deserción muy alta de la población Institucional (…), situación que desemboca en que de cada 100 estudiantes que ingresan a la primaria únicamente tres culminan con un título universitario. (Pacheco, 2015)

Con la premisa que las Universidades reciben por parte del Estado un presupuesto para el desarrollo de sus actividades, la deserción es alta.

En la LOES se reconoce a los Institutos Técnicos y Tecnológicos como Instituciones de Educación Superior capaces de otorgar títulos de tercer nivel, convirtiéndose su oferta académica muy apetecida por los jóvenes en el proceso de obtener su título de tercer nivel, sin embargo al no contar con la asignación de un presupuesto para el cumplimiento de sus funciones, y especialmente para generar programas de retención y ayuda social a estudiantes con diferentes clases de problemas en la continuación de sus estudios.

A pesar de que los Institutos no cuenten con la asignación de un presupuesto para de alguna manera erradicar este fenómeno, las acciones para evitar el incremento de este

fenómeno de deserción estudiantil son mínimos, lo que hace que los indicadores de deserción crezcan cada periodo académico.

Justificación e importancia

En la actualidad el departamento de Bienestar Institucional del Instituto Superior Tecnológico Cotopaxi, revisa el índice de deserción producida en un periodo académico mediante el indicador de la asistencia emitido por el sistema informático SIGA, que los tutores de curso informan al terminar el periodo académico a la Coordinación indicada, lo cual no les permite a los responsables de la Unidad de Bienestar Institucional tomar acciones oportunas para lanzar proyectos de acompañamiento para con el estudiante, con el objetivo de brindar el asesoramiento y apoyo que éste requiere para continuar con sus estudios. Se hace imperante la necesidad de implementar un modelo inteligente de análisis de datos que se enfoque en determinar los posibles estudiantes desertores.

Ante esta realidad, se decide realizar el presente proyecto, cuyos fines inmediatos se componen en desarrollar la investigación, diseño e implementación de un modelo inteligente de análisis de datos con la técnica más apropiada de Machine Learning, y la metodología acorde a este tipo de proyectos, serializar el modelo, crear una Api y lanzarlo a la web mediante una plataforma como servicio.

Los beneficios del modelo inteligente para el Instituto Superior Tecnológico Cotopaxi y a su vez para la Unidad de Bienestar Institucional son que al contar con la API web se puede:

1. Predecir y alertar de los posibles casos de deserción de los estudiantes de primer nivel.
2. Realizar un seguimiento adecuado a los casos referidos por la API.
3. Realizar oportunas intervenciones en los casos identificados.
4. Tomar decisiones adecuadas para los casos alertados por la API.

5. Contar con las estadísticas y métricas por cada semestre que permita ir evaluando los resultados obtenidos a lo largo del tiempo.
6. Incrementar la tasa de retención de estudiantes.
7. Disminuir la tasa de deserción estudiantil.
8. Contribuir con los indicadores relacionados (retención y eficiencia terminal) en el Modelo de Evaluación de Institutos Superiores Técnicos y Tecnológicos emitido por el CACES.

De la revisión bibliográfica realizada en torno al tema planteado, existe escasa información de estudios realizados entorno a la deserción educativa en la educación superior a nivel nacional, por lo que el presente trabajo investigativo pretende ser un aporte significativo, materializándolo mediante el desarrollo del modelo inteligente de análisis de datos, para el Instituto Superior Tecnológico Cotopaxi y sea a su vez un referente para las demás instituciones de educación superior.

Alcance

Estudiantes del Instituto Superior Tecnológico Cotopaxi en el año 2022

Planteamiento del problema

En la actualidad se ha incrementado la demanda al acceso a la Educación Superior Pública, siendo esta mayor a la ofertada por parte de dichas Instituciones. Por citar un ejemplo para el primer periodo académico del 2020 según la SENESCYT, las Instituciones de educación superior ofertaron alrededor de 113,072 cupos siendo un 31% más que el primer periodo del año pasado. El 19% de la oferta corresponde a carreras técnicas y tecnológicas de tercer nivel, y el 81% a carreras de tercer nivel de grado que ofertan Universidades y Escuelas Politécnicas. [Fuente: El Universo 3 de marzo de 2020]. Mientras que alrededor de 284,239

estudiantes realizaron la prueba Ser bachiller en enero de 2020, y estarían aptos para participar en el proceso de admisión a la educación superior para el primer periodo 2020. Con estos datos podemos notar que aproximadamente existe un déficit del 60%, y tan sólo un 40% tiene acceso a la Educación Superior.

El Instituto Superior Tecnológico Cotopaxi es un Instituto público ubicado en la zona centro del país, y cuya demanda a sus carreras técnicas y tecnológicas al iniciar cada periodo académico se mantiene e inclusive se incrementa debido a sus fortalezas en infraestructura física, su cuerpo docente de cuarto nivel, entre otras fortalezas, seducen a sus postulantes para decidirse a ingresar a la Institución.

Formulación del problema

El presente trabajo, plantea el ¿Cómo predecir los posibles casos de deserción de estudiantes en los primeros niveles del Instituto Superior Tecnológico Cotopaxi?

Objetivos

Objetivo general

Desarrollar un modelo inteligente de análisis de datos por medio de un algoritmo de Machine Learning que permita la identificación de patrones de comportamiento o causalidad en los datos de forma temprana, sobre los posibles casos de deserción estudiantil y aplicarlo en un caso de estudio.

Objetivos específicos

- Determinar las variables de causalidad que formarán parte del set de datos.

- Determinar los requerimientos que debe cumplir el set de datos y el modelo inteligente de análisis de datos.
- Validar el modelo inteligente de análisis de datos con set de datos de diferentes periodos académicos.

Hipótesis

H0: Si se desarrolla un modelo inteligente de análisis de datos entonces se genera una predicción de los alumnos con riesgo de deserción en etapas tempranas al iniciar sus estudios en los primeros niveles en el Instituto Superior Tecnológico Cotopaxi.

Variables de la investigación

Variable independiente

Desarrollar un modelo inteligente de análisis de datos.

Conceptualización de la variable independiente.

Un modelo inteligente de análisis de datos es una herramienta desarrollada mediante técnicas de Machine Learning y la aplicación de la metodología CRISP-DM, con la finalidad de detectar la deserción temprana de estudiantes mediante el análisis de sus datos en DataSet.

Variable dependiente

Predecir los posibles casos de deserción en etapas tempranas de los estudiantes de los primeros niveles en el Instituto Superior Tecnológico Cotopaxi.

Conceptualización de la variable dependiente.

Los algoritmos de predicción a eventos futuros, son una herramienta útil en la toma efectiva de decisiones oportunas en beneficio de una colectividad, de un negocio, de una empresa, entre otros, para reducir el impacto e incertidumbre de posibles eventos en un mediano plazo.

Predecir los posibles casos de deserción en etapas tempranas de los estudiantes de los primeros niveles en el Instituto Superior Tecnológico Cotopaxi, contribuirá al Instituto Superior Tecnológico Cotopaxi y directamente a la Coordinación de Bienestar Institucional, a identificar los posibles casos de deserción en la comunidad académica que cursan su primer año de estudios, para crear medidas y acciones preventivas de acompañamiento a los estudiantes y lograr que estos continúen sus estudios.

Metodología de Investigación

Tipo de investigación

El tipo de investigación utilizada en el presente trabajo de investigación es la investigación aplicada ya que se centra en solucionar un problema concreto.

Niveles de Investigación

Investigación Exploratoria.

La investigación exploratoria ayuda a identificar el problema de deserción que tiene el Instituto con el objetivo de recabar información inicial para el desarrollo del presente trabajo.

Investigación Descriptiva.

La investigación descriptiva realiza un análisis de la población Institucional que ingresa a estudiar en los primeros niveles con el objetivo de ir seleccionando las variables demográficas en la construcción del modelo predictivo.

Métodos y técnicas de la investigación

Los métodos teóricos se utilizarán para el cumplimiento de las tareas, se utilizaron los siguientes métodos teóricos de investigación:

Método histórico lógico.

Determinar los antecedentes históricos de los modelos inteligentes de análisis de datos.

Método analítico sintético.

Para desarrollar la situación problemática, marco teórico, además se utiliza para procesar la información obtenida en la investigación de campo.

Método hipotético deductivo.

Presente en la hipótesis propuesta para realizar el proceso de investigación.

Método de modelación.

Para determinar la solución en el enfoque determinado del área de la educación.

Capítulo II

Fundamentación teórica y referencial

Introducción

El aprendizaje automático es parte de la inteligencia artificial, que en base a datos anteriores ayuda a las computadoras a aprender y a tomar decisiones inteligentes. El aprendizaje automático consiste en transformar información en una base estructurada de conocimiento para posteriormente ser utilizada en distintas áreas. En el área de la educación, los docentes pueden distribuir mejor su tiempo fuera de las aulas mediante la adopción del aprendizaje automático. Por ejemplo, los docentes pueden hacer uso de varias herramientas virtuales creadas justamente para establecer una comunicación de forma remota desde cualquier lugar para sus estudiantes. Por medio de estas herramientas se mejora la experiencia de enseñanza con los estudiantes y ellos progresivamente aumentan su rendimiento. El aprendizaje automático fomenta el aprendizaje personalizado en el contexto de la difusión de la educación, además permite a los docentes comprender mejor cómo progresan sus estudiantes en el aprendizaje. Facilitando a los docentes a crear un plan de estudios de aprendizaje diferente, en pos de cubrir todas las expectativas de los alumnos. Cuando se emplea en el contexto de la educación, la IA puede fomentar la moderación de la inteligencia. Es a través de esta plataforma que se hace posible el análisis de datos por parte de tutores y moderadores humanos.

Un ejemplo concreto de un programa que maneja inteligencia artificial (IA) y que actualmente ha sido implementado en la mayoría de los Institutos Públicos del país, y el Instituto Superior Tecnológico Cotopaxi no es la excepción, es el programa ALEKS. ALEKS es un software que utiliza el aprendizaje automático, La máquina inteligente de ALEKS identifica de manera eficiente y exacta cuales temas cada estudiante ha dominado y cuales está listo

para aprender según sus respuestas a un breve número de preguntas que ALEKS elige en función de sus respuestas a todas las preguntas anteriores. ALEKS casi nunca se equivoca al ofrecer al estudiante un tema que está listo para aprender.

Antecedentes conceptuales referenciales

Deserción escolar

El abandono escolar temprano o deserción escolar se refiere a la situación en que un alumno después de un proceso de separación, finalmente abandona los estudios. (Lyche, 2010)

El profesor Stephem Lamb et al. (2016) asevera que las definiciones de deserción escolar varían entre las naciones; en tal razón, las investigaciones deben estar contextualizadas por las convenciones de cada nación.

En 2016 la Comisión Especial de Estadísticas de Educación para el Ecuador, define el término “Abandono Escolar” y establece la fórmula de cálculo de la tasa de Abandono.

“*El Abandono Escolar lo define como el “Número de estudiantes contabilizados al final de un período escolar que abandonan un determinado grado o curso de estudios, expresado como porcentaje del total de estudiantes matriculados al final del mismo grado o curso de estudios y periodo escolar*” (Comisión Especial de Estadísticas de Educación, 2016).

Fórmula de Cálculo:

$$\boldsymbol{TAB} = \frac{Est\ Ab_g^t}{Est\ M_g^t} \times 100$$

Donde:

TAB = Tasa de abandono escolar

$\boldsymbol{Est\ Ab_g^t}$ = Número de estudiantes que abandonan el grado o curso **g** en el periodo escolar **t**.

$\boldsymbol{Est\ M_g^t}$ = Total de estudiantes matriculados en el grado o curso **g** en el periodo escolar **t**.

g = Grado o curso de estudio. (este indicador es aplicable para todos los grados de educación básica y bachillerato)

La deserción estudiantil según (Corzo Salazar, s.f.), significa que un individuo ya no asiste a la escuela y tampoco ha recibido un diploma o su equivalente reconocido.

Según (Viteri Castro & Uquillas Narváez, 2011), la deserción escolar "*se utiliza para referirse a aquellos alumnos que dejan de asistir a clase y quedan fuera del sistema educativo*".

Según (Viteri Castro & Uquillas Narváez, 2011), los alumnos que interrumpen sus estudios y quedan fuera del sistema educativo se conoce como deserción escolar.

"En el plano educativo, se utiliza el término para hablar de aquellos alumnos que abandonan sus estudios por diferentes causas; entendiéndose por estudios a toda educación que se encuentra dentro del sistema educativo impuesto por el gobierno que rija en aquel Estado (primaria, secundaria, universidad, etc.)." (Corzo Salazar, s.f.).

Según, Vásquez et al (2003), se definen tres tipos de deserción estudiantil:

Deserción precoz: Cuando un estudiante abandona un programa de estudios antes de empezar habiendo sido aceptado. (Corzo Salazar, s.f.)

Deserción temprana: Cuando se abandona el programa de estudios durante los primeros cuatro semestres. (Corzo Salazar, s.f.)

Deserción tardía: Desde el quinto semestre en adelante. El enfoque espacial de Vásquez et al (2003) indica que de hecho hay una diferencia entre: **Deserción total:** cuando el alumno abandona por completo un plan educativo y decide no regresar. (Corzo Salazar, s.f.)

Deserción parcial: cuando el alumno hace lo que generalmente se conoce como una baja temporal y cuando se siente seguro regresa al programa educativo para continuar con sus estudios. (Corzo Salazar, s.f.)

En el modelo de evaluación institucional para los Institutos superiores técnicos y tecnológicos 2020, en el indicador 6.1.1 de Acompañamiento a estudiantes remarca en su estándar "*El instituto diseña un proceso de acompañamiento a los estudiantes desde su ingreso (admisión y/o nivelación), que contribuye a su motivación para lograr un adecuado rendimiento académico, desarrollo integral y su permanencia en la institución hasta la culminación de sus estudios*" (CACES, 2020, pág. 76).

Entre las principales razones de deserción estudiantil, se detallan las siguientes: emocionales, comportamentales, económicas, familiares, sociales, convivenciales, de salud, de tiempo, de gestión y planeación del aprendizaje, entre otras; (CACES, 2020).

El objetivo del estándar 6.1.1 claramente indica el interés de que el estudiante culmine sus estudios, para lo cual evalúa la tasa de retención con la siguiente fórmula: (CACES, 2020)

Tasa de retención

$$\mathbf{TR = 100 * \frac{NEM2}{NEM1}}$$

Donde:

TR: tasa de retención

NEM2: Número de estudiantes matriculados en el periodo académico "2"

NEM1: Número de estudiantes matriculados en el periodo académico “1”. (CACES, 2020)

Con esta fórmula implícitamente podemos identificar a los estudiantes desertores, tan solo identificando los estudiantes que iniciaron en un periodo académico y que no constan en el siguiente periodo académico.

Considerando la premisa anterior, y para justificar la realización del presente trabajo de investigación, en conjunto con la Unidad de Bienestar Institucional se analizaron las estadísticas resultantes de la consulta realizada a 19 periodos académicos empezando desde el primer semestre de 2012 hasta el primer parcial de 2021, produciéndose los siguientes resultados que se muestran en la Tabla 1.

Tabla 1

Estudiantes desertores desde el 1S-2012 hasta el 1P-2021

Nivel	Estudiantes desertores	Porcentaje
1	788	35%
2	523	23%
3	329	15%
4	602	27%
5	12	0%

El porcentaje más alto de deserción se concentra en los primeros niveles con un número total de 790 estudiantes y que representa el 35.05% de la población. Estas estadísticas de deserción apalancan el desarrollo del presente trabajo de investigación. Una detección temprana del alumnado en riesgo de deserción facilitará a la Unidad de Bienestar Institucional enfocarse en este segmento y tomar decisiones oportunas para frenar el alto número de deserción que se registran en los primeros niveles.

Modelo Predictivo

El modelado predictivo es utilizado para predecir eventos futuros mediante el análisis de patrones por medio de un conjunto dado de datos de entrada. Es un componente crucial del análisis predictivo, que conlleva un análisis de datos actuales e históricos para pronosticar comportamientos y tendencias futuras. (Lawton & Carew, s.f.)

- El modelado predictivo es uno de los muchos nombres que se refieren al proceso de descubrir relaciones dentro de los datos para predecir algún resultado deseado. Dado que muchos dominios científicos han contribuido a este campo, existen sinónimos para diferentes entidades:
- Los términos sample, data point, observation o instance se refieren a una sola unidad de datos independiente, como un cliente, o un paciente.
- El training set consta de los datos utilizados para desarrollar modelos, mientras que los conjuntos de prueba o validación se utilizan únicamente para evaluar el rendimiento de un conjunto final de modelos candidatos.
- Los predictors, variables independientes, atributos o descriptores son los datos que se utilizan como entrada para la ecuación de predicción.
- El resultado, la variable dependiente, el objetivo, la clase o la respuesta se refieren al evento de resultado o la cantidad que se predice.
- Los datos continuos tienen escalas numéricas naturales. La presión arterial, el costo de un artículo o la cantidad de baños son continuos. Los conteos no pueden ser un número fraccionario, pero aún se tratan como datos continuos.
- Los datos categóricos, también conocidos como datos nominales, de atributos o discretos, toman valores específicos que no tienen escala. El estado del crédito ("bueno" o "malo") o el color ("rojo", "azul", etc.) son ejemplos de estos datos.

- La construcción de modelos, el entrenamiento de modelos y la estimación de parámetros se refieren al proceso de usar datos para determinar los valores de las ecuaciones del modelo. (Kuhn & Johnson, 2013)

Matriz de Variables

De la recopilación efectuada a 19 artículos académicos similares a la deserción estudiantil y de diferentes países, se identificaron las siguientes variables por cada estudio (Anexo 1).

De la revisión bibliográfica realizada en los 19 artículos científicos, en la Tabla 2, se destacan las siguientes variables con mayor frecuencia de ocurrencia.

Tabla 2

Frecuencia de ocurrencia de variables

Variables predictivas	Frecuencia de ocurrencia	Artículos	País
Etnia	3	(CUJI, GAVILANES, & SANCHEZ, 2017)	Ecuador
		(Dursun, 2020)	Estados Unidos
		(Fallb, Vaughna, Robertsb, Kremerc, & Martinezb, 2020)	Estados Unidos
Dirección	4	(NARVÁEZ BARROS & BARRAGÁN REYES, 2015)	Ecuador
		(Dursun, 2020)	Estados Unidos
		(Nikolovski, Stojanov, Chorbev, & Madjarov, 2015)	Macedonia
		(HASAN, 2015)	Bangladesh
Acceso BECA /ayuda financiera recibida	4	(NARVÁEZ BARROS & BARRAGÁN REYES, 2015)	Ecuador
		(Vásquez, 2016)	Chile
		(Díaz Peralta, 2008)	Chile
		(Dursun, 2020)	Estados Unidos
Promedio de las notas de enseñanza media	4	(Vásquez, 2016)	Chile
		(Díaz Peralta, 2008)	Chile
		(Dursun, 2020)	Estados Unidos
		(Nikolovski, Stojanov, Chorbev, & Madjarov, 2015)	Macedonia
Estado Civil	4	(CUJI, GAVILANES, & SANCHEZ, 2017)	Ecuador
		(Amaya Torrado, Barrientos Avendaño, & Heredia Vizcaíno, 2017)	Colombia
		(Díaz Peralta, 2008)	Chile
		(Dursun, 2020)	Estados Unidos
Ingresos	5	(NARVÁEZ BARROS & BARRAGÁN REYES, 2015)	Ecuador

Variables predictivas	Frecuencia de ocurrencia	Artículos	País
		(Amaya Torrado, Barrientos Avendaño, & Heredia Vizcaíno, 2017)	Colombia
		(Vásquez, 2016)	Chile
		(Díaz Peralta, 2008)	Chile
		(Olaya, Vásquez, Maldonado, Miranda, & Verbeke, 2020)	Chile
Puntaje logrado en las secciones de Lenguaje y Comunicación , Matemáticas, Ciencias e Historia	5	(Vásquez, 2016)	Chile
		(Olaya, Vásquez, Maldonado, Miranda, & Verbeke, 2020)	Chile
		(Dursun, 2020)	Estados Unidos
		(Nikolovski, Stojanov, Chorbev, & Madjarov, 2015)	Macedonia
		(HASAN, 2015)	Bangladesh
Nivel de estudio de la madre	6	(Amaya Torrado, Barrientos Avendaño, & Heredia Vizcaíno, 2017)	Colombia
		(Vásquez, 2016)	Chile
		(Díaz Peralta, 2008)	Chile
		(Olaya, Vásquez, Maldonado, Miranda, & Verbeke, 2020)	Chile
		(HASAN, 2015)	Bangladesh
		(Casanova, Cervero, Núñez, Almeida, & Bernardo, 2018)	Portugal
Nivel de estudio del padre	6	(Amaya Torrado, Barrientos Avendaño, & Heredia Vizcaíno, 2017)	Colombia
		(Vásquez, 2016)	Chile
		(Díaz Peralta, 2008)	Chile
		(Olaya, Vásquez, Maldonado, Miranda, & Verbeke, 2020)	Chile
		(HASAN, 2015)	Bangladesh
		(Casanova, Cervero, Núñez, Almeida, & Bernardo, 2018)	Portugal
Edad	6	(CUJI, GAVILANES, & SANCHEZ, 2017)	Ecuador
		(NARVÁEZ BARROS & BARRAGÁN REYES, 2015)	Ecuador
		(Amaya Torrado, Barrientos Avendaño, & Heredia Vizcaíno, 2017)	Colombia
		(Dursun, 2020)	Estados Unidos
		(Casanova, Cervero, Núñez, Almeida, & Bernardo, 2018)	Portugal
		(Fernandez & Silva, 2014)	Ecuador
Género / Sexo	11	(CUJI, GAVILANES, & SANCHEZ, 2017)	Ecuador
		(Amaya Torrado, Barrientos Avendaño, & Heredia Vizcaíno, 2017)	Colombia
		(Vásquez, 2016)	Chile
		(Díaz Peralta, 2008)	Chile
		(Dursun, 2020)	Estados Unidos
		(Fallb, Vaughna, Robertsb, Kremerc, & Martinezb, 2020)	Estados Unidos
		(Nikolovski, Stojanov, Chorbev, & Madjarov, 2015)	Macedonia
		(HASAN, 2015)	Bangladesh
		(Vázquez, Biggio, & García, 2013)	Argentina
		(Casanova, Cervero, Núñez, Almeida, & Bernardo, 2018)	Portugal
		(Fernandez & Silva, 2014)	Ecuador

El Proceso Docente Educativo

El Proceso Docente Educativo es la integración holística y sistémica de la enseñanza y el aprendizaje en todos sus componentes, cualidades, niveles de asimilación, de profundidad y estructural, en sus tres dimensiones: educativa, instructiva y desarrolladora.

Una institución recibe el encargo de educar al hombre para la vida a partir de compromisos sociales, debiendo ser capaz de enfrentarse a nuevas situaciones y problemas que se le presenten y resolverlos en pos de transformar la sociedad.

La tecnología fue adoptada por Instituciones públicas y privadas con el objetivo de obtener buenos resultados en el proceso de enseñanza, facilitando a los docentes contar con herramientas para realizar sus presentaciones y que antes no tenían.

Hoy en día la tecnología juega un papel indiscutible en la formación de futuros profesionales, focalizándose en desarrollar un pensamiento crítico en el estudiante que le permita crear una mejor calidad de vida.

Instituciones públicas y privadas de educación superior deben acoger el uso de herramientas tecnológicas en sus modelos educativos. Aprovechar de estos recursos les permitirá a sus docentes llevar sus clases a un ambiente de magia y fantasía y así facilitar el proceso de enseñanza aprendizaje. (ECURED, s.f.)

Machine Learning en el Proceso Docente Educativo

Recientemente, se ha incrementado el interés en utilizar técnicas de Machine Learning en el estudio educacional, centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educacionales y en el uso de esos métodos para comprender mejor a los estudiantes y el entorno en el que aprenden.

El Machine Learning puede considerarse parte de la inteligencia artificial (IA). El aprendizaje automático es, el proceso de conceder a una máquina o modelo el acceso a los datos y dejar que aprenda por sí misma. En 1959, Arthur Samuel tuvo la brillante idea de que no debíamos enseñar a los ordenadores, sino que podíamos dejarles aprender por sí mismos. Acuñó el término "Machine Learning" para describir su teoría, que ahora es una definición estándar de la capacidad de los ordenadores para aprender de forma autónoma.

Hay muchas otras implementaciones comerciales de Machine Learning, muchas de ellas en el área de la educación. Algunas de las áreas interesantes son:

Predecir el rendimiento de los estudiantes (Una gran aplicación del Machine Learning es predecir el rendimiento de los estudiantes. Al "aprender" sobre cada estudiante, el modelo de aprendizaje automático puede encontrar debilidades y sugerir formas de mejorar, como, por ejemplo, conferencias adicionales o estudiar literatura adicional.

El Machine Learning puede ayudar a crear evaluaciones adaptativas computarizadas. La evaluación basada en el Machine Learning proporciona una retroalimentación constante a los docentes y a los estudiantes sobre cómo aprende el estudiante, el apoyo que necesita y el progreso que está haciendo hacia sus objetivos de aprendizaje).

Mejorar la retención (El Machine Learning, como la analítica de aprendizaje, también ayudará a mejorar los índices de retención. Al identificar a los estudiantes "en riesgo", las escuelas pueden llegar a esos estudiantes y conseguirles la ayuda que necesitan para tener éxito.

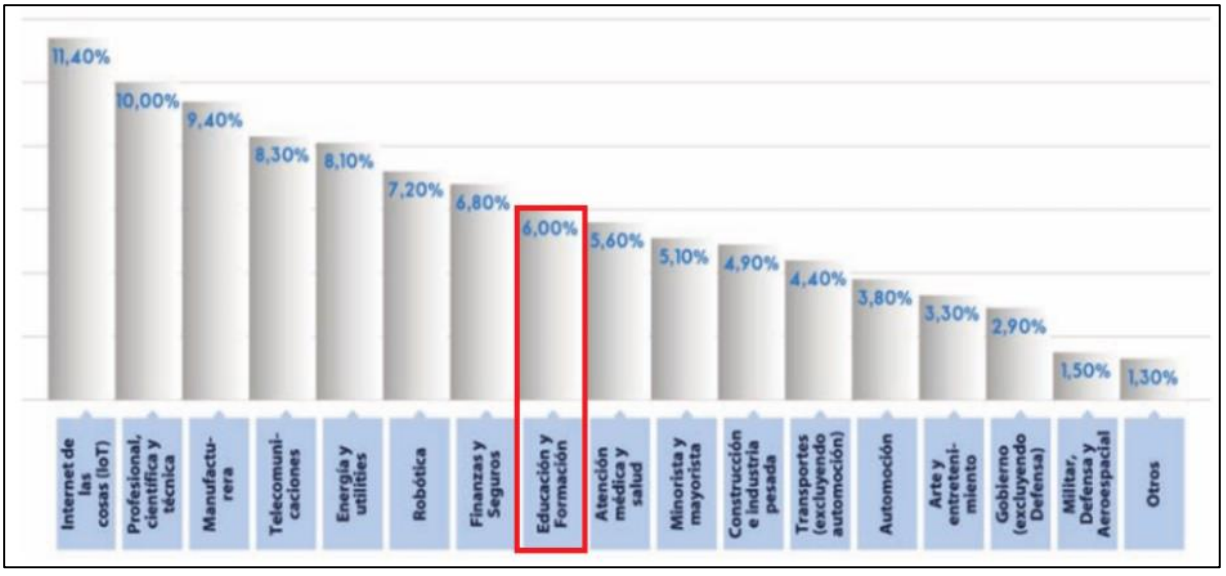
Al identificar tempranamente a los estudiantes "en riesgo", las escuelas pueden detectar y contactar a esos estudiantes y ayudarlos a tener más éxito. La retención de los estudiantes es una parte esencial de muchos sistemas de matriculación. Afecta a casi todos los segmentos de la métrica de la universidad o la escuela: reputación, finanzas, clasificación.

En especial, la retención de estudiantes se ha convertido en una de las cosas más importantes para los directivos de las instituciones de enseñanza superior. Hay pocos estudios que hayan desarrollado modelos para predecir e introducir las razones por las que el número de estudiantes disminuye. El Machine Learning en la educación superior ayuda a prevenir la deserción, a través del aprendizaje continuo de las tecnologías de la información de los patrones que generan los grandes datos o Big Data.

Un estudiante que abandona sus estudios y no logra obtener su grado académico representa una tremenda pérdida potencial para el alumno y un fracaso para la institución de educación superior. Se hace necesario adoptar prácticas y metodologías de apoyo en base a la realidad de la institución, sus datos y los estudios sobre la materia. En la Figura 1 se muestra la presencia de Machine Learning en las distintas industrias.

Figura 1

Presencia del desarrollo de Machine Learning en las distintas industrias 2016



Caracterización tecnológica de las técnicas y métodos Machine Learning

Los tres tipos de aprendizaje automático son: Aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

El objetivo principal del aprendizaje supervisado es aprender un modelo a partir de datos de entrenamiento etiquetados, que permita hacer predicciones sobre datos no vistos o futuros. En este caso, el término supervisado se refiere a un conjunto de muestras donde las señales de salida deseadas (etiquetas) ya se conocen.

Clasificación para predecir las etiquetas de clase

La clasificación es una subcategoría del aprendizaje supervisado en la que el objetivo es predecir las etiquetas de clase categóricas de las nuevas instancias, basándose en observaciones anteriores. Estas etiquetas de clase son valores discretos y no ordenados que pueden entenderse como la pertenencia a un grupo de instancias.

Regresión para predecir resultados continuos

Un segundo tipo de aprendizaje supervisado es la predicción de resultados continuos, que también se denomina análisis de regresión. En el análisis de regresión, nos dan una serie de variables predictoras (explicativas) y una variable de respuesta continua (resultado u objetivo), y tratamos de encontrar una relación entre esas variables que nos permita predecir un resultado.

En el aprendizaje no supervisado, sólo se conocen los datos de entrada y no se dan datos de salida conocidos al algoritmo. Aunque hay muchas aplicaciones exitosas de estos métodos, suelen ser más difíciles de entender y evaluar.

En el aprendizaje no supervisado, se enfrenta con datos no etiquetados o con datos de estructura desconocida. Con las técnicas de aprendizaje no supervisado, se explora la

estructura de los datos para información significativa, sin la guía de una variable de resultado conocida o una función de recompensa.

En el aprendizaje por refuerzo, el objetivo es desarrollar un sistema (agente) que mejore su rendimiento en función de las interacciones con el entorno. Dado que la información sobre el estado actual del entorno suele incluir también una señal de recompensa, podemos considerar el aprendizaje por refuerzo como un campo relacionado con el aprendizaje supervisado. Sin embargo, en el aprendizaje por refuerzo esta retroalimentación no es la etiqueta o el valor correcto de la verdad sobre el terreno, sino una medida de lo bien que la acción fue medida por una función de recompensa. A través de su interacción con el entorno, un agente puede utilizar el aprendizaje por refuerzo para aprender una serie de acciones que maximicen esta recompensa a través de un enfoque exploratorio de ensayo y error o una planificación deliberativa. (Muller & Guido, 2017)

Machine Learning y retención de estudiantes de la Educación Superior

Para Dursun Delen, de la Oklahoma State University, “la retención Institucional forma parte de muchos modelos de gestión de matrícula, por cuanto influye en los rankings de las universidades, la reputación y sus estados financieros”. (Pitzalis, 2020)

Para expertos como él, “*la mejora de la retención Institucional debe comenzar por un profundo conocimiento de los modos de deserción. El comprender esto es el pilar para predecir qué estudiantes están en riesgo de dejar sus estudios, e intervenir de manera apropiada, de modo de retenerles.*" (Pitzalis, 2020)

El profesor Delen estudió una muestra de cinco años de datos institucionales, y desarrolló modelos de análisis para predecir y explicar los motivos de la deserción de los estudiantes de primer año. Concluyó que, en su muestra, las variables educativas y financieras eran los principales predictores de deserciones. (Pitzalis, 2020)

Este tipo de estudios tienen ya varias aplicaciones en la práctica. Con un algoritmo apropiado, sumados a la entrega adecuada de datos, el Machine Learning permite:

Identificar a los estudiantes con alto riesgo de deserción al iniciar un periodo académico.

Identificar los factores de riesgo más recurrentes.

Entrega de informes de las principales causas y factores de riesgo. (Isabel S., 2020)

¿En qué se diferencia la minería de datos del Machine Learning?

Muchos confunden Big Data en la educación con la herramienta Machine Learning.

De acuerdo con SAS, esta última utiliza "muchos algoritmos y técnicas de la minería de datos. La diferencia radica en qué tipo de temas predicen".

La minería de datos detecta patrones e información antes desconocidas.

El Machine Learning reproduce patrones de datos, la integra a otros antecedentes, y automáticamente aplica los resultados a la toma de decisiones y el desarrollo de acciones. (Isabel S., 2020)

Técnicas de Machine Learning

Los **algoritmos de aprendizaje** se dividen en tres categorías:

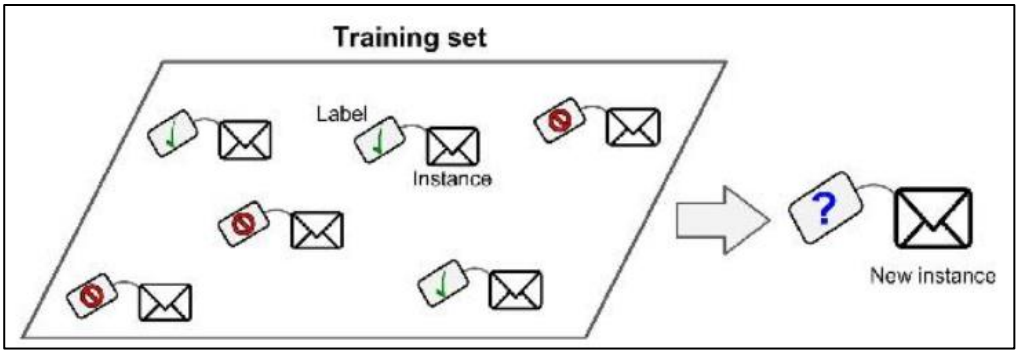
- Aprendizaje supervisado,
- Aprendizaje no supervisado y
- Aprendizaje de refuerzo.

Aprendizaje Supervisado.

En el aprendizaje supervisado, los datos de entrenamiento que se introducen en el algoritmo incluyen las soluciones deseadas, llamadas etiquetas (Figura 2).

Figura 2

Ejemplo conjunto de entrenamiento etiquetado aprendizaje supervisado (spam)

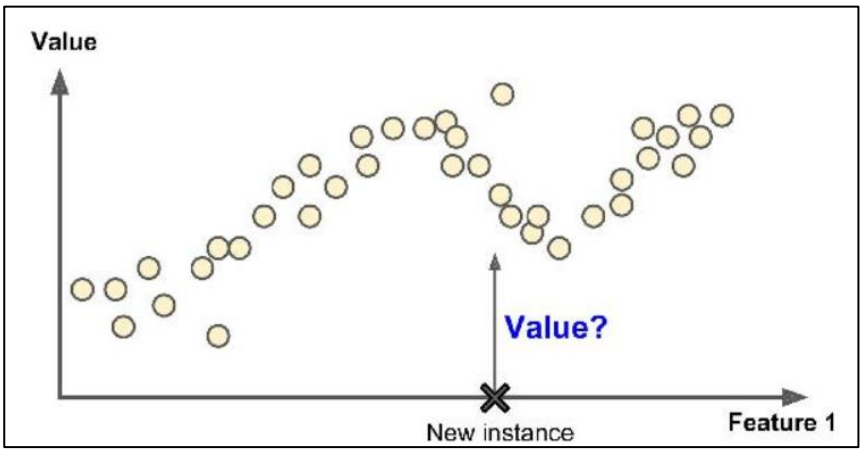


En el aprendizaje supervisado una tarea conocida es la clasificación. Como ejemplo de la clasificación tenemos el conocido spam que se entrena con muchos ejemplos de correos electrónicos de su clase, y debe aprender a clasificar nuevos correos electrónicos.

Otra tarea conocida es predecir un valor numérico objetivo, como el precio de un auto, dado un conjunto de características (kilometraje, edad, marca, etc.) llamados predictores. Este tipo de tarea se denomina regresión (Figura 3).

Figura 3

Regresión



Tener en cuenta que algunos algoritmos de regresión pueden utilizarse también para la clasificación y viceversa. Por ejemplo, la regresión logística se utiliza habitualmente para la clasificación, ya que puede obtener un valor que corresponde a la probabilidad de pertenecer a una clase determinada (por ejemplo 20% de probabilidad de ser spam).

Estos son algunos de los algoritmos de aprendizaje supervisado más importantes:

- Árboles de decisión
- k-Nearest Neighbors
- Bosques aleatorios
- Gradient boosted regression trees (gradient boosting machines)
- Regresión lineal
- Regresión logística
- Máquinas de vectores de apoyo (SVM)
- Redes neuronales

Modelos de Machine Learning de Aprendizaje Supervisado.

Árboles de decisión.

Los árboles de decisión son modelos ampliamente utilizados para tareas de clasificación y regresión. Básicamente, aprenden una jerarquía de preguntas if/else, que conducen a una decisión. Estas preguntas son similares a las preguntas que podría hacer en un juego de 20 preguntas.

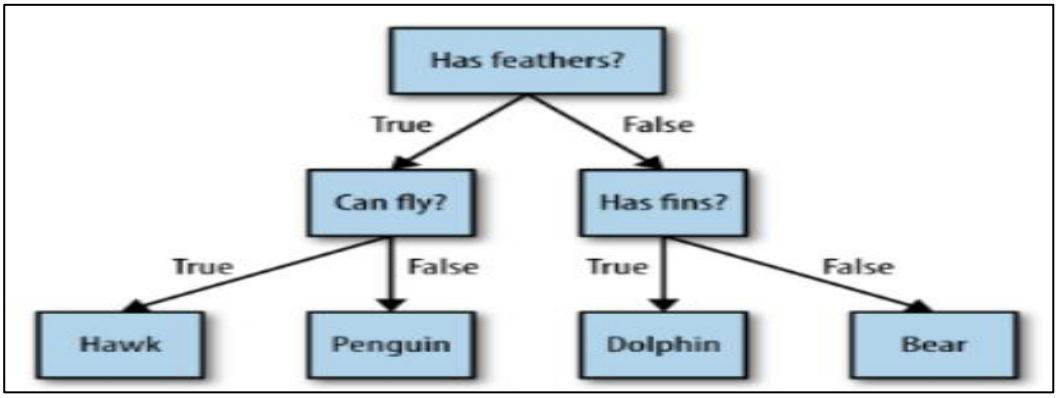
Imagine que quiere distinguir entre los siguientes cuatro animales: osos, halcones, pingüinos y delfines. Su objetivo es llegar a la respuesta correcta preguntando el menor número posible de if/else.

Puede comenzar preguntando si el animal tiene plumas, una pregunta que reduce sus posibles animales a solo dos. Si la respuesta es sí," puede hacer otra pregunta que podría ayudarle a distinguir entre halcones y pingüinos. Por ejemplo, podría preguntar si el animal puede volar. Si el animal no tiene plumas, tus posibles elecciones de animales son delfines y osos, y tendrá que hacer una pregunta para distinguir entre estos dos animales, por ejemplo, preguntando si el animal tiene aletas. (Muller & Guido, 2017)

Esta serie de preguntas se puede expresar como un árbol de decisiones, como se muestra en la Figura 4.

Figura 4

Árbol de decisión



k-Nearest Neighbors.

El clasificador de k-vecinos más cercanos, es fácil de entender. Construir este modelo solo consiste en almacenar el conjunto de entrenamiento. Para hacer una predicción para un nuevo punto de datos, el algoritmo encuentra el punto en el conjunto de entrenamiento que está más cerca del nuevo punto. Luego asigna la etiqueta de este punto de entrenamiento al nuevo punto de datos. La k en k-vecinos más cercanos significa que en lugar de usar solo el vecino más cercano al nuevo punto de datos, podemos considerar cualquier número fijo k de

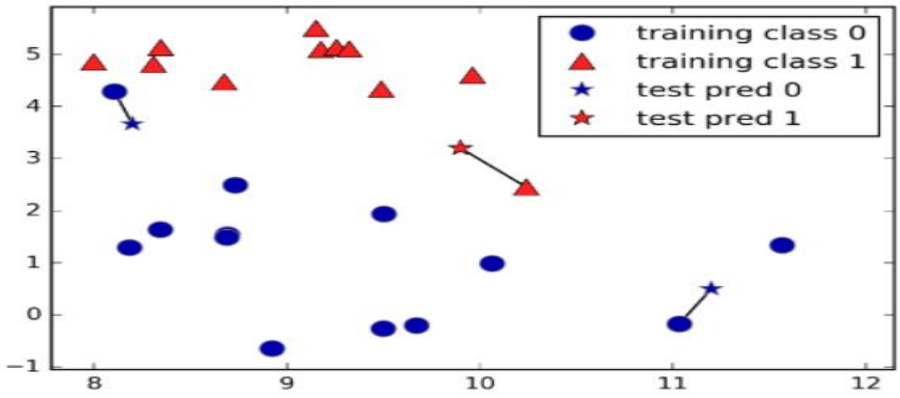
vecinos en el entrenamiento (por ejemplo, los tres o cinco vecinos más cercanos). Luego, podemos hacer una predicción usando la clase mayoritaria entre estos vecinos.

Todos los modelos de aprendizaje automático en scikit-learn se implementan en sus propias clases, que se denominan clases Estimator. El algoritmo de clasificación de k-vecinos más cercanos se implementa en la clase KNeighborsClassifier en el módulo de vecinos. Antes de usar el modelo, necesitamos instanciar la clase en un objeto. Esto es cuando estableceremos cualquier parámetro del modelo. El parámetro más importante de KNeighbors Classifier es el número de vecinos. (Muller & Guido, 2017)

En la figura 5 se muestra un ejemplo del clasificador.

Figura 5

k-Nearest Neighbors



Bosques aleatorios.

Los bosques aleatorios son una forma de abordar este problema. Un bosque aleatorio es esencialmente una colección de árboles de decisión, donde cada árbol es ligeramente diferente de los demás. La idea detrás de los bosques aleatorios es que cada árbol puede hacer una función relativamente buen trabajo de predicción, pero es probable que se sobreajuste en parte de los datos. Si construimos muchos árboles, todos los cuales funcionan

bien y se sobre ajustan de diferentes maneras, podemos reducir la cantidad de sobreajuste promediando sus resultados. Esta reducción en el sobreajuste, al tiempo que conserva el poder predictivo de los árboles, se puede mostrar utilizando matemáticas rigurosas.

Para implementar esta estrategia, necesitamos construir muchos árboles de decisión. Cada árbol debe hacer un trabajo aceptable de predicción del objetivo y también debe ser diferente de los otros árboles. Los bosques aleatorios obtienen su nombre de inyectar aleatoriedad en la construcción del árbol para garantizar que cada árbol sea diferente. Hay dos formas en que los árboles en un bosque aleatorio se aleatorizan: seleccionando los puntos de datos utilizados para construir un árbol y seleccionando las características en cada prueba dividida.

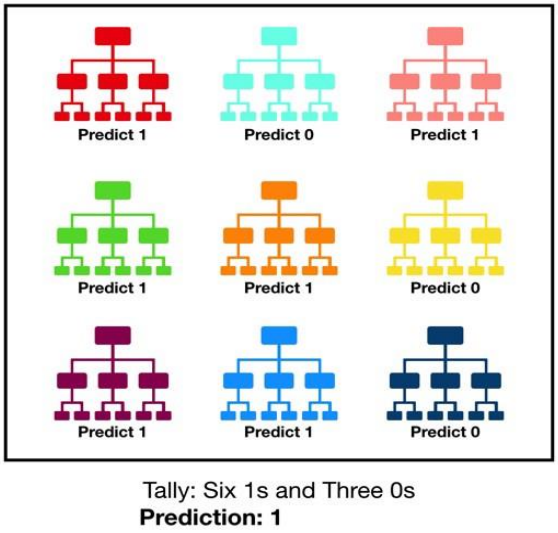
Para construir un modelo de bosque aleatorio, debe decidir la cantidad de árboles que se construirán (el parámetro n_estimators de RandomForestRegressor o RandomForestClassifier).

Digamos que queremos construir 10 árboles. Estos árboles se construirán de forma completamente independiente unos de otros, y el algoritmo hará diferentes opciones aleatorias para cada árbol para asegurarse de que los árboles sean distintos.

Para construir un árbol, primero tomamos lo que se llama una muestra de arranque de nuestros datos. Es decir, a partir de nuestros puntos de datos n_samples, extraemos repetidamente un ejemplo al azar con reemplazo (lo que significa que la misma muestra se puede seleccionar varias veces). Esto creará un DataSet tan grande como el conjunto de datos original, pero faltarán algunos puntos de datos (aproximadamente un tercio) y algunos se repetirán, Ver figura 6. (Muller & Guido, 2017)

Figura 6

Random Forest



Gradient boosted regression trees (gradient boosting machines).

El árbol de regresión potenciado por gradiente es otro método de conjunto que combina múltiples árboles de decisión para crear un modelo más poderoso. A pesar de la "regresión" en el nombre, estos modelos se pueden usar para regresión y clasificación. En contraste con el enfoque de bosque aleatorio, la potenciación de gradientes funciona mediante la construcción de árboles en serie, donde cada árbol intenta corregir los errores del anterior. De forma predeterminada, no hay aleatorización en los árboles de regresión potenciados por gradiente; en su lugar, se utiliza una fuerte poda previa. Los árboles potenciados por gradiente a menudo usan árboles muy poco profundos, de una profundidad de uno a cinco, lo que hace que el modelo sea más pequeño en términos de memoria y hace que las predicciones sean más rápidas.

La idea principal detrás del aumento de gradiente es combinar muchos modelos simples, como árboles poco profundos. Cada árbol solo puede proporcionar buenas predicciones sobre una parte de los datos, por lo que se agregan más y más árboles para

mejorar el rendimiento de forma iterativa. Los árboles potenciados por gradiente son con frecuencia las entradas ganadoras en las competencias de aprendizaje automático y se usan ampliamente en la industria. Por lo general, son un poco más sensibles a la configuración de parámetros que los bosques aleatorios, pero pueden proporcionar una mayor precisión si los parámetros se configuran correctamente.

Si desea aplicar el aumento de gradiente a un problema a gran escala, podría valer la pena examinar el paquete xgboost y su interfaz de Python, que es más fácil de ajustar que la implementación de scikit-learn de aumento de gradiente en muchos conjuntos de datos.

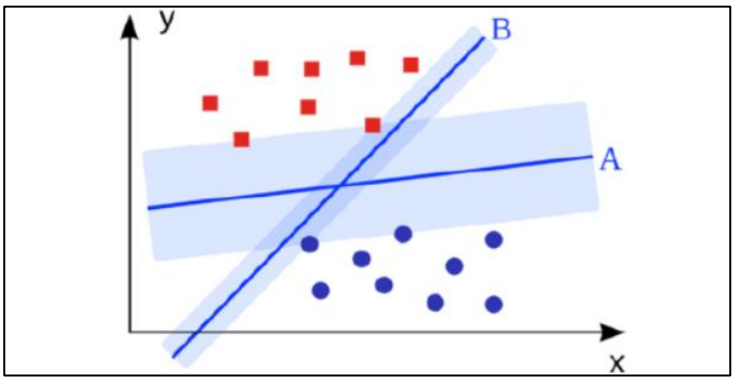
Regresión Logística.

La Regresión Logística (también llamada Regresión Logit) se utiliza habitualmente para estimar la probabilidad de que una instancia pertenezca a una clase determinada (por ejemplo, ¿cuál es la probabilidad de que este correo electrónico sea spam?). Si la probabilidad estimada es superior al 50%, el modelo predice que la instancia pertenece a esa clase (llamada clase positiva, etiquetada como "1"), o bien predice que no pertenece (es decir, pertenece a la clase negativa, etiquetada como "0"). (Géron, 2017)

Esto lo convierte en un clasificador binario. Ver figura 7.

Figura 7

Regresión Logística



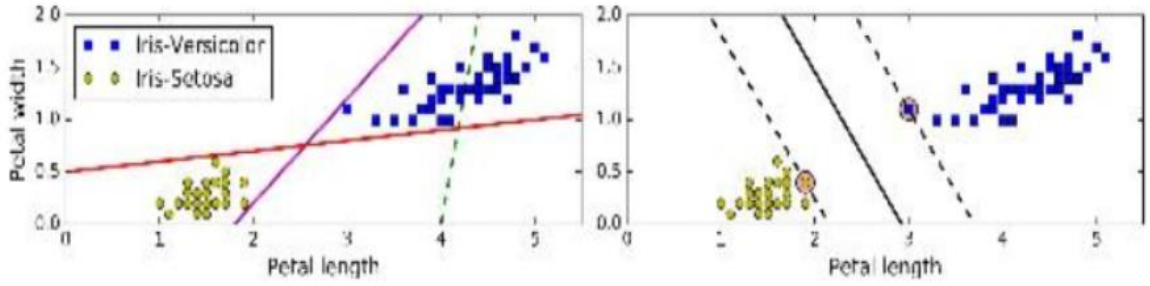
Support Vector Machine (SVM).

El paradigma de la máquina de soporte de vectores (SVM) para el aprendizaje de predictores lineales en espacios de características de alta dimensionalidad. La alta dimensionalidad plantea retos de complejidad muestral y de complejidad computacional.

El paradigma algorítmico de la SVM aborda el reto de la complejidad de la muestra mediante la búsqueda de separadores de gran margen. A grandes rasgos, un semiespacio separa un conjunto de entrenamiento con un gran margen si todos los ejemplos no sólo están en el lado correcto del hiperplano de separación, sino también lejos de él. Restringir el algoritmo a la salida de un separador con un margen grande puede producir una complejidad muestral pequeña incluso si la dimensionalidad del espacio de características es alta. Introducimos el concepto de margen y lo relacionamos con el paradigma de minimización de pérdidas regularizadas, ver Figura 8 (Shalev-Shwartz , 2014)

Figura 8

Support Vector Machine



Redes Neuronales (Deep Learning).

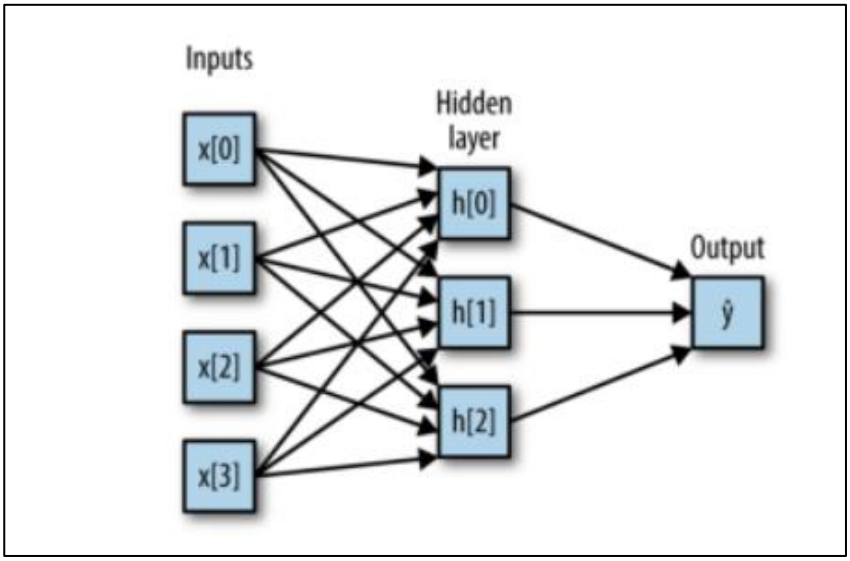
Una familia de algoritmos conocida como redes neuronales ha visto recientemente un resurgimiento bajo el nombre "aprendizaje profundo". Si bien el aprendizaje profundo muestra una gran promesa en muchas máquinas aplicaciones de aprendizaje, los algoritmos de aprendizaje profundo a menudo se adaptan con mucho cuidado a un caso de uso específico.

Existen algunos métodos relativamente simples, perceptrones multicapa para clasificación y regresión, que pueden servir como punto de partida punto para métodos de aprendizaje profundo más involucrados. Los perceptrones multicapa (MLP) son también conocidas como redes neuronales feed-forward (vainilla), o a veces simplemente redes neuronales.

En un MLP, este proceso de cálculo de sumas ponderadas se repite varias veces, primero calcular unidades ocultas que representan un paso de procesamiento intermedio, que son nuevamente combinados usando sumas ponderadas para producir el resultado final, ver figura 9.

Figura 9

Neural Networks

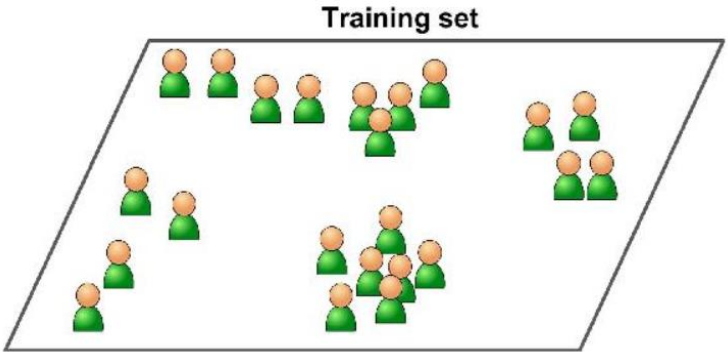


Aprendizaje No Supervisado.

En el aprendizaje no supervisado, como se puede adivinar, los datos de entrenamiento están sin etiquetar (Figura 10). El sistema intenta aprender sin un maestro.

Figura 10

Un conjunto de entrenamiento no etiquetado para aprendizaje no supervisado



Estos son algunos de los algoritmos de aprendizaje no supervisado más importantes:

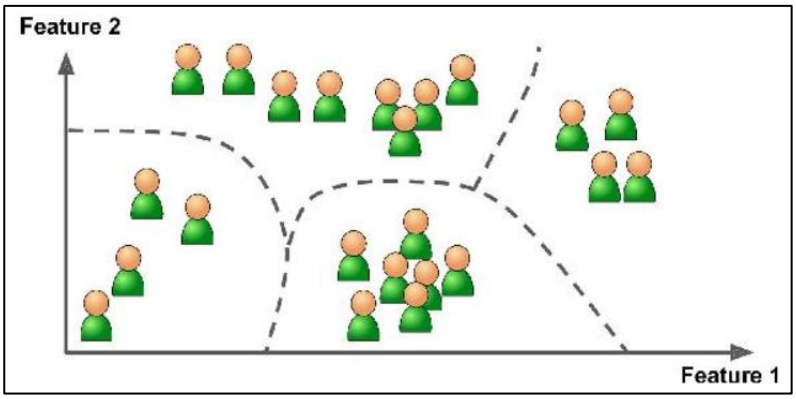
- Clustering
  - k-Means
  - Análisis de clústeres jerárquicos (HCA)
  - Maximización de expectativas
- Visualización y reducción de la dimensionalidad
  - Análisis de componentes principales (PCA)
  - Kernel PCA
  - Incrustación localmente lineal (LLE)
  - Incrustación estocástica de vecinos distribuida en t (t-SNE)
- Aprendizaje de reglas de asociación
  - Apriori
  - Eclat

Por ejemplo, digamos que tiene muchos datos sobre los visitantes de su blog. Es posible que desee ejecutar un algoritmo de agrupación para tratar de detectar grupos de visitantes similares (Figura 11). En ningún momento le dice al algoritmo a qué grupo pertenece

un visitante: el algoritmo encuentra esas conexiones sin su ayuda. Por ejemplo, puede observar que el 40% de sus visitantes son hombres a los que les gustan los cómics y que suelen leer su blog por la noche, mientras que el 20% son jóvenes amantes de la ciencia ficción que lo visitan los fines de semana, etc. Si utiliza un algoritmo de agrupación jerárquica también puede subdividir cada grupo en otros más pequeños. Esto puede ayudarle a orientar sus publicaciones a cada grupo.

Figura 11

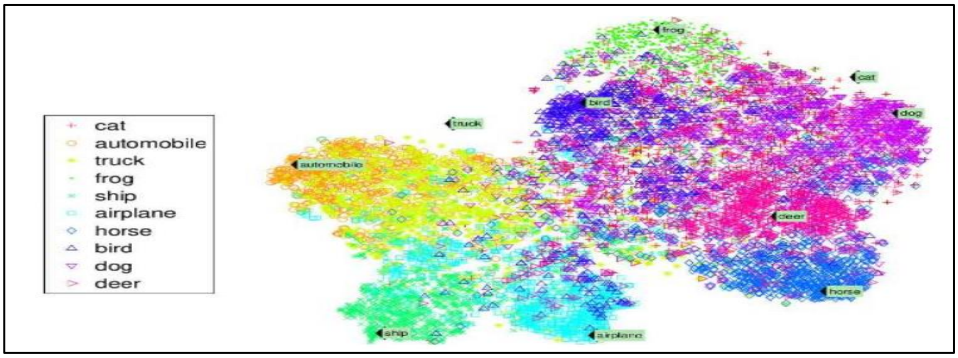
Clustering



Los algoritmos de visualización también son buenos ejemplos de algoritmos de aprendizaje no supervisado: Se les alimenta con un montón de datos complejos y etiquetados datos complejos y etiquetados, y ellos producen una representación 2D o 3D de sus datos que se puede trazar fácilmente (Figura 12). Estos algoritmos intentan reservar toda la estructura que pueden (por ejemplo, tratando de evitar que los clusters separados en el espacio de entrada se superpongan en la visualización), para que se pueda entender cómo están organizados los datos y quizás identificar patrones insospechados.

Figura 12

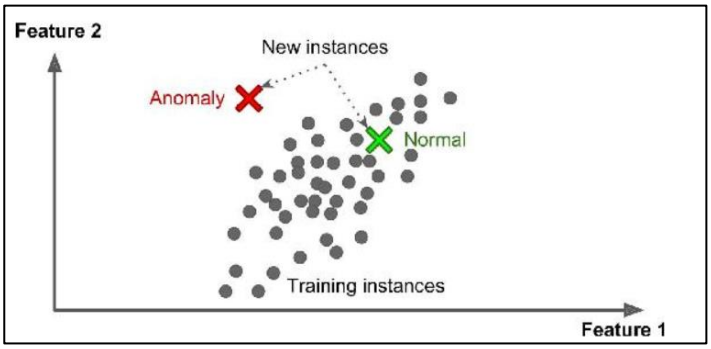
Ejemplo de t-SNE



Otra importante tarea no supervisada es la detección de anomalías: por ejemplo, la detección de transacciones inusuales con tarjetas de crédito para evitar el fraude, la detección de defectos de fabricación o la eliminación automática de valores atípicos de un conjunto de datos antes de alimentar otro algoritmo de aprendizaje. El sistema se entrena con instancias normales y, cuando ve una nueva instancia, puede decir si se parece a una normal o si es normal o si se trata de una anomalía (Figura 13).

Figura 13

Detección de anomalía



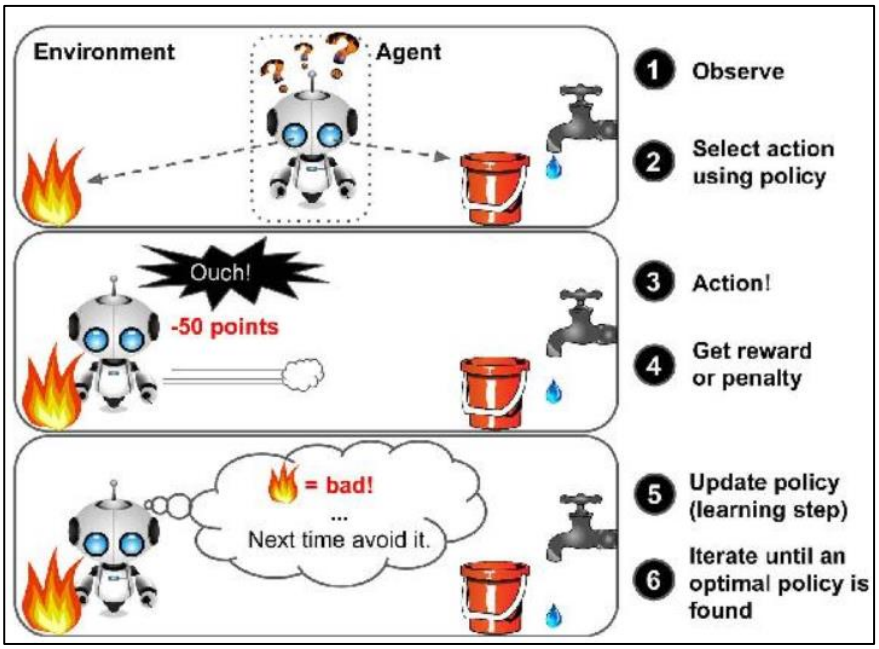
Aprendizaje de Refuerzo.

El sistema de aprendizaje, llamado agente en este contexto, puede observar el entorno, seleccionar y realizar acciones, y obtener recompensas a cambio (o penalizaciones en forma

de recompensas negativas, como en la Figura 14). A continuación, debe aprender por sí mismo cuál es la mejor estrategia, denominada política, para obtener la mayor recompensa a lo largo del tiempo. Una política define qué acción debe elegir el agente cuando se encuentra en una situación determinada.

Figura 14

Aprendizaje de refuerzo



Por ejemplo, muchos robots aplican algoritmos de aprendizaje por refuerzo para aprender a caminar. El programa AlphaGo de DeepMind también es un buen ejemplo de Aprendizaje por Refuerzo: saltó a los titulares en marzo de 2016 cuando venció al campeón mundial Lee Sedol en el juego del Go. Aprendió su política ganadora analizando millones de partidas, y luego jugando muchas partidas contra sí mismo. Hay que tener en cuenta que el aprendizaje se desactivó durante las partidas contra el campeón; AlphaGo sólo aplicaba la política que había aprendido.

Técnicas de evaluación del desempeño Modelos de Machine Learning

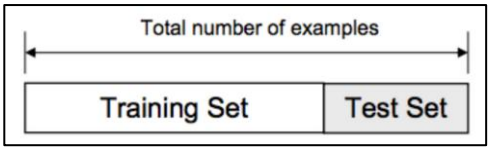
Train_test_split.

Para evaluar los modelos supervisados, hemos dividido nuestro conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba usando la función train_test_split, construimos un modelo en el conjunto de entrenamiento llamando al método fit y lo evaluamos en el conjunto de prueba usando el método score, que para la clasificación calcula la fracción de muestras correctamente clasificadas. (Géron, 2017)

La función train_test_split **es común en todos los modelos de aprendizaje supervisado**. Es la división del DataSet en al menos dos partes: una parte para el entrenamiento *Train* y que representa a la mayor parte del DataSet y que se usa para entrenar el modelo y una parte más pequeña para las pruebas o *Test*, en donde se evalúa el modelo entrenado. Ver figura 15.

Figura 15

Train_test_split



Validación Cruzada (Cross-Validation).

Es una forma más sólida de evaluar el rendimiento de la generalización, analiza los métodos para evaluar el rendimiento de la clasificación y la regresión que van más allá de las medidas predeterminadas de precisión y R^2 proporcionadas por el método score.

La validación cruzada es un método estadístico para evaluar el rendimiento de la generalización que es más estable y completo que usar una división en un conjunto de entrenamiento y otro de prueba.

En la validación cruzada, los datos se dividen repetidamente y se entrenan varios modelos. La versión más utilizada de validación cruzada es la validación cruzada k-fold, donde k es un número especificado por el usuario, generalmente 5 o 10.

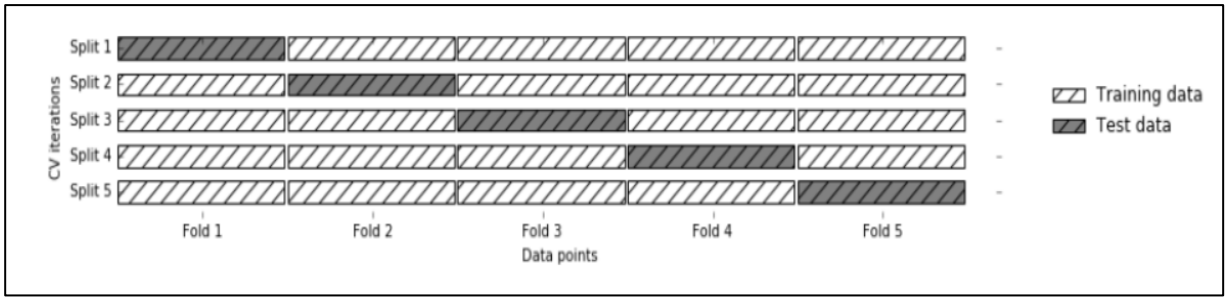
Al realizar una validación cruzada de cinco veces, los datos primero se dividen en cinco partes (aproximadamente) de el mismo tamaño, llamados pliegues. A continuación, se entrena una secuencia de modelos. El primer modelo se entrena usando el primer pliegue como conjunto de prueba, y los pliegues restantes (2–5) se usan como conjunto de entrenamiento.

El modelo se construye utilizando los datos de los pliegues 2 a 5, y luego se evalúa la precisión en el pliegue 1. Luego se construye otro modelo, esta vez usando el pliegue 2 como conjunto de prueba y los datos en los pliegues 1, 3, 4 y 5 como el conjunto de entrenamiento. Este proceso se repite utilizando los pliegues 3, 4 y 5 como conjuntos de prueba.

Para cada una de estas cinco divisiones de datos en conjuntos de entrenamiento y prueba, calculamos la precisión. Al final, hemos recopilado cinco valores de precisión. El proceso se ilustra en la Figura 16 (Muller & Guido, 2017)

Figura 16

Cross Validation



Por lo general, el primer quinto de los datos es el primer pliegue, el segundo quinto de los datos es el segundo pliegue, y así sucesivamente.

La validación cruzada se implementa en scikit-learn mediante la función cross_val_score del módulo model_selection. Los parámetros de la función cross_val_score son el modelo que queremos evaluar, los datos de entrenamiento y los datos de prueba. (Muller & Guido, 2017)

Matriz de Confusión

Una de las formas más completas de representar el resultado de la evaluación de la clasificación binaria es utilizar las matrices de confusión.

Las entradas de la diagonal principal de la matriz de confusión corresponden a las clasificaciones correctas, mientras que las demás entradas nos indican cuántas muestras de una clase se han clasificado erróneamente como otra clase.

Se llama verdaderos positivos a las muestras correctamente clasificadas que pertenecen a la clase positiva y verdaderos negativos a las muestras correctamente clasificadas que pertenecen a la clase negativa. Estos términos suelen abreviarse como FP, FN, TP y TN y conducen a la siguiente interpretación de la matriz de confusión (Figura 17): (Muller & Guido, 2017)

Figura 17

Matriz de Confusión

VALORES REALES	Verdaderos PositivosVP	Falsos Negativos FN
	Falsos Positivos FP	Verdaderos Negativos VN
	VALORES PREDICCION	

VP: El valor real es positivo, y el valor predicho es también es positivo.

VN: El valor real es negativo, y el valor predicho también es negativo.

FN: El valor real es positivo, y el valor predicho es negativo.

FP: El valor real es negativo, y el valor predicho es positivo.

Hay varias formas de resumir la información en la matriz de confusión por medio de las siguientes métricas.

Exactitud (Accuracy): Es el número de predicciones correctas (TP y TN) dividido entre el número de todas las muestras (todas las entradas de la matriz de confusión sumadas)

$$\text{accuracy} = \frac{VP + VN}{VP + FP + FN + VN}$$

Precisión (Precision): La precisión mide cuántas de las muestras predichas como positivas son realmente positivas.

$$\text{precision} = \frac{VP}{VP + FP}$$

Sensibilidad (Recall o Sensitivity): También llamada tasa de verdaderos positivos. Mide cuántas de las muestras positivas son capturadas por las predicciones como positivas.

$$\text{sensibilidad} = \frac{VP}{VP + FN}$$

Especificidad (Especificity): También llamada tasa de verdaderos negativos. Mide cuántas de las muestras negativas son capturadas por las predicciones como negativas.

$$\text{especificidad} = \frac{VN}{VN + FP}$$

Curva ROC

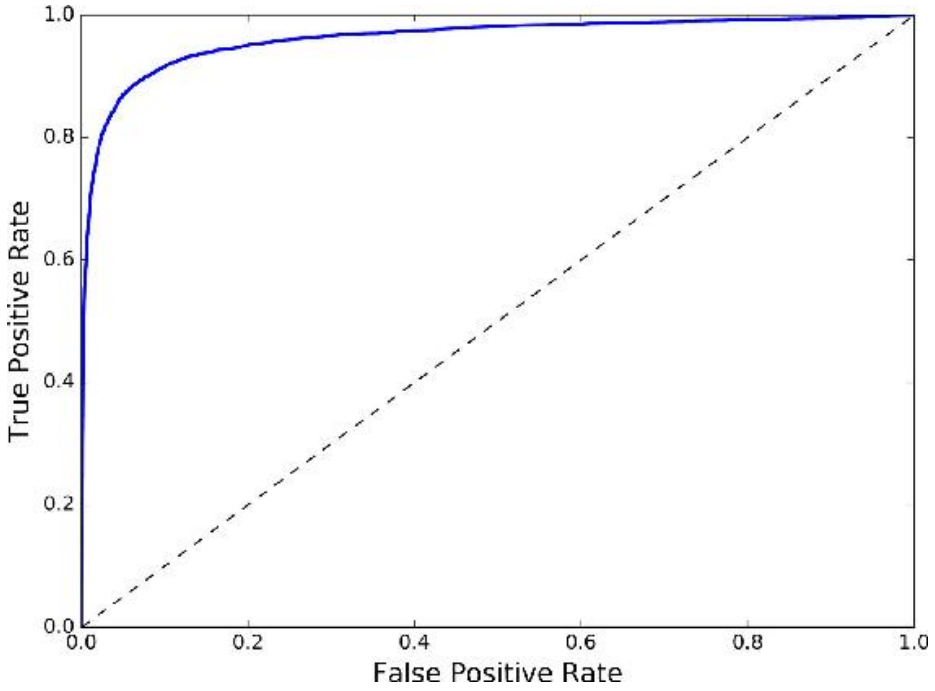
La curva ROC(receiver operating characteristic) es otra herramienta común utilizada con clasificadores binarios. Es muy parecida a la curva de precision/recall, pero en lugar de trazar precision versus recall, la curva ROC traza la tasa de verdaderos positivos frente a la tasa de falsos positivos.

El FPR es la ratio de las instancias negativas que se clasifican incorrectamente como positivos. Es igual a uno menos los verdaderos negativos, que es el ratio de instancias negativas que están correctamente clasificados como negativos.

La curva ROC traza la sensibilidad (recall) versus 1 – especificidad. Para dibujar la curva ROC, previamente se debe calcular el TPR y el FPR para varios valores de umbral, utilizando2.2.1 la función roc_curve(), ver Figura 18. (Géron, 2017)

Figura 18

Curva ROC



La curva ROC ideal está cerca de la parte superior izquierda, para ello se requiere de un clasificador que produzca un alto recall mientras mantiene una baja tasa de falsos positivos. (Muller & Guido, 2017)

Correlación de variables

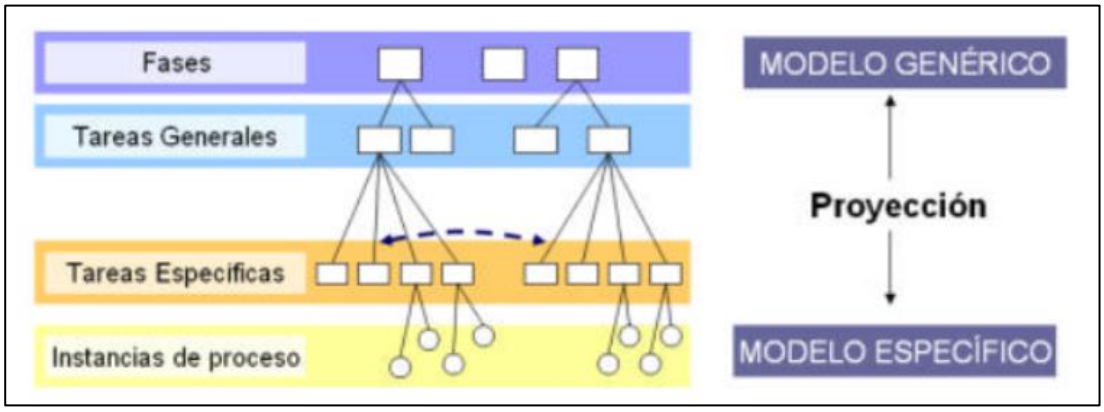
La Correlación es una técnica estadística usada para determinar la relación entre dos o más variables. Se refiere a cuan cerca están dos variables de tener una relación lineal entre sí.

Metodología CRISP-DM

La metodología CRISP-DM (CrossIndustry Standard Process for Data Mining), tiene cuatro niveles de abstracción que van desde el nivel más general hasta los casos más específicos (Figura 19). (Rodríguez Montequín, Álvarez Cabal, Mesa Fernández, & González Valdés, SA)

Figura 19

Modelos de abstracción metodología CRISP-DM



La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. Ver figura 20.

Figura 20

Fases - metodología CRISP-DM

Fase 1. Comprensión del Negocio.

"El primer paso es la Comprensión del Negocio es dar contexto a los objetivos y a los datos para que el ingeniero se haga una idea de la relevancia de los datos en ese modelo de negocio concreto". Se compone de reuniones, encuentros en línea, lectura de documentación, aprendizaje específico sobre el terreno, y una larga lista de formas que ayudan al equipo de desarrollo, a hacer preguntas sobre el contexto relevante" (Rodrigues, 2020).

El producto de este paso es que el equipo de desarrollo entiende el contexto del proyecto. Los objetivos del proyecto deben definirse antes de que éste comience. Por ejemplo, el equipo de desarrollo debería saber ya que el objetivo es aumentar las ventas y, una vez finalizado el paso, entender qué es lo que vende el cliente y cómo lo vende". (Rodrigues, 2020)

Fase 2. Comprensión de los datos.

“El segundo paso es la comprensión de los datos y su objetivo es saber qué se puede esperar y conseguir de los datos. Comprueba la calidad de los datos, en varios términos, como la integridad de los datos, la distribución de los valores o el cumplimiento de la gobernanza de los datos.

Se trata de una parte crucial del proyecto porque define la viabilidad y la fiabilidad de los resultados finales. En este paso, los miembros del equipo hacen una lluvia de ideas sobre cómo extraer el mejor valor de las piezas de información. En caso de que el uso o la relevancia de algún dato no esté claro para el equipo de desarrollo, pueden dar un paso atrás momentáneamente, para entender el negocio y cómo se beneficia de ese dato.

Gracias a este paso el científico de datos ahora sabe, en términos de datos, el resultado debe satisfacer los objetivos del proyecto, qué algoritmo y proceso traen ese resultado, cómo es el estado actual de los datos, y cómo debe ser, para ser útil al algoritmo y proceso involucrados”. (Rodrigues, 2020)

Fase 3. Preparación de la data.

“El tercer paso es la preparación de los datos e implica el proceso de ETLs o ELTs que convierte las piezas de datos en algo útil mediante los algoritmos y el proceso.

A veces las políticas de gobierno de datos no se respetan o no se establecen en una organización, y para dar un verdadero significado a los datos, se convierte en el trabajo de los ingenieros de datos y los científicos de datos para estandarizar la información.

Del mismo modo, algunos algoritmos funcionan mejor bajo ciertos parámetros, alguien no acepta valores no numéricos, otros no funcionan bien con una gran variación de valores. Por otra parte, corresponde al equipo de desarrollo normalizar la información.

La mayoría de los proyectos dedicaron la mayor parte de su tiempo a este paso. Este paso, creo, es la razón por la que existe un perfil informático llamado ingeniero de datos. Como es una tarea que requiere mucho tiempo y que puede resultar realmente compleja cuando se trabaja con grandes cantidades de datos, los departamentos de TI podrían encontrar una ventaja en dedicar recursos a realizar específicamente estas tareas". (Rodrigues, 2020)

Fase 4. Modelado.

"El cuarto paso es la modelización y es el núcleo de cualquier proyecto de aprendizaje automático. Este paso es el responsable de los resultados que deben satisfacer o ayudar a satisfacer los objetivos del proyecto.

Aunque es la parte más glamurosa del proyecto, también es la más corta en tiempo, ya que, si todo lo anterior se hace correctamente, hay poco que ajustar. En el caso de que los resultados sean mejorables, la metodología se establece para dar un paso atrás en la preparación de los datos y mejorar los datos disponibles.

Algunos algoritmos como k-means, clustering jerárquico, series de tiempo, regresión lineal, k-vecinos más cercanos, y muchos varios otros, son las líneas de código de núcleo de este paso en la metodología". (Rodrigues, 2020)

Fase 5. Evaluación.

"El quinto paso es la Evaluación, donde se trata de verificar que los resultados son válidos y correctos. En caso de que los resultados sean erróneos, la metodología permite la revisión de vuelta al primer paso, con el fin de entender por qué los resultados son erróneos.

Normalmente, en un proyecto de ciencia de datos, el científico de datos divide los datos en entrenamiento y prueba. En esta etapa se utilizan los datos de prueba, cuyo objetivo es verificar que el modelo (producto de la etapa de modelización) se ajusta a la realidad.

Dependiendo de la tarea y del contexto, existen diversas técnicas. Por ejemplo, en el contexto del aprendizaje supervisado, con la tarea de clasificar elementos, una forma de verificar los resultados es con la matriz de confusión. Para el aprendizaje no supervisado, hacer la evaluación se hace más difícil, ya que no hay ningún valor estático para separar lo "correcto" de lo "incorrecto", por ejemplo, la tarea de clasificar elementos se evaluaría calculando la distancia inter e intra entre los elementos de uno o varios clusters.

En cualquier caso, es importante especificar alguna fuente de medida de error. Esta medida de error indica al usuario cómo puede confiar en los resultados, ya sea para "seguro que esto funcionará" o "seguro que no". Si de alguna manera la medida de error resulta ser 0 o ninguna para todos los casos, indicaría que el modelo está sobreajustado, y la realidad podría comportarse de forma diferente". (Rodrigues, 2020)

Fase 6. Despliegue.

"La sexta y última etapa es la de Despliegue y consiste en presentar los resultados de forma útil y comprensible, con lo que el proyecto debería alcanzar sus objetivos. Es el único paso que no pertenece a un ciclo.

Dependiendo del usuario final, la forma útil y comprensible puede variar. Por ejemplo, si el usuario final es otra pieza de software, como en el programa del sitio web de ventas que pregunta a su sistema de recomendación qué sugerir para un comprador, una manera útil sería un JSON que lleve la respuesta a una consulta específica. En otro caso, como el de un alto ejecutivo que requiere información proyectada para la toma de decisiones, la mejor manera de presentar los resultados es almacenarlos en una base de datos analítica y presentarlos como un cuadro de mando en una solución de inteligencia empresarial". (Rodrigues, 2020)

Antecedentes contextuales

El Instituto Superior Tecnológico Ramón Barba Naranjo hoy INSTITUTO SUPERIOR TECNOLÓGICO COTOPAXI, de la ciudad de Latacunga, Provincia de Cotopaxi, fue registrado en el Consejo de Educación Superior (CONESUP), bajo el número 05-004 de 19 de octubre de 2000, cómo Instituto Técnico Superior Ramón Barba Naranjo, alcanzando la categoría de Tecnológico mediante acuerdo N°134 del Consejo de Educación Superior, el 08 de septiembre de 2003, y mediante resolución RPC-SO-08-No.140-2017 del 8 de marzo de 2017 se cambia el nombre a INSTITUTO SUPERIOR TECNOLÓGICO COTOPAXI.

El Instituto Superior Tecnológico Cotopaxi es una Institución pública de educación superior acreditada, orientada a la formación integral de profesionales de tercer nivel competentes e innovadores con compromiso ético, social y ambiental que fomentan el desarrollo territorial sostenible.

El Instituto Superior Tecnológico Cotopaxi se encuentra ubicado en la parroquia Tanicuchi Panamericana E35 km.12 vía Latacunga – Quito.

En su oferta académica constan las carreras de:

- Tecnología Superior en Electromecánica
- Tecnología superior en Mantenimiento y Reparación de Motores a Diesel y Gasolina
- Tecnología superior en Mantenimiento Eléctrico y Control Industrial
- Tecnología superior en Logística Multimodal
- Tecnología superior en Floricultura
- Tecnología superior en Desarrollo Infantil Integral.
- Seguridad Penitenciaria
- Centro de Idiomas

El análisis de patrones de deserción aportará principalmente a la Coordinación de Bienestar Institucional, para que oportunamente identifiquen los estudiantes que ingresan a la Institución y que al iniciar su carrera están en riesgo de deserción, para que se pueda diseñar y aplicar un proceso de acompañamiento académico y pedagógico efectivo, que permita disminuir este fenómeno y apoyarlos para que continúen y culminen sus estudios.

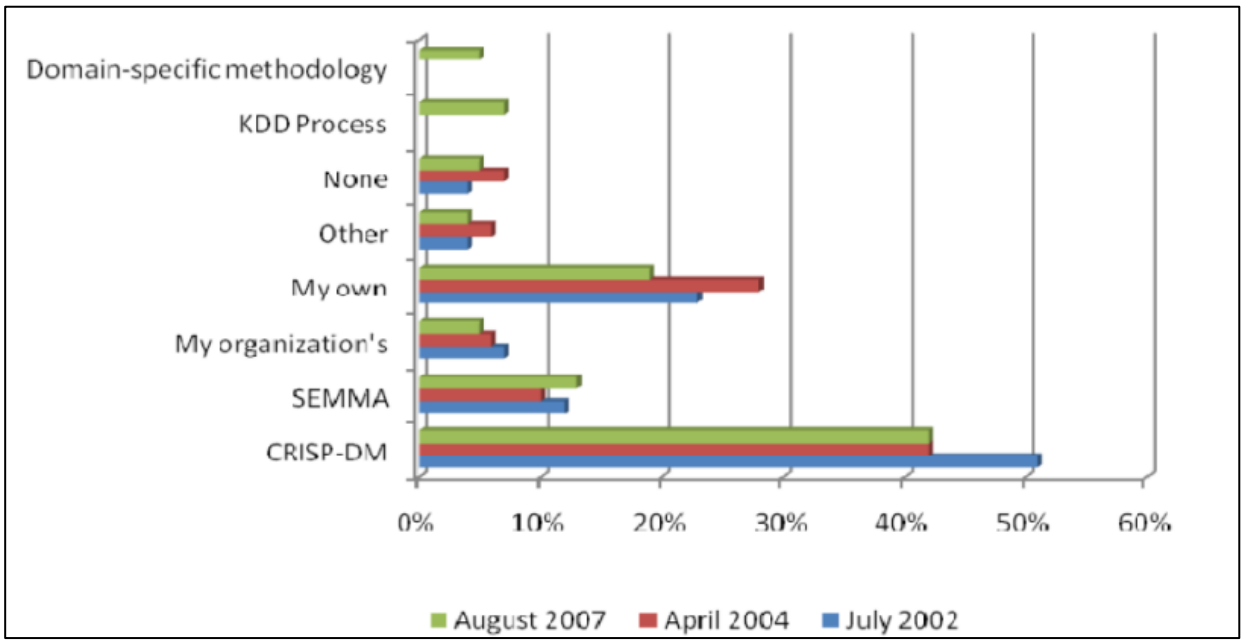
Capítulo III

Desarrollo del modelo inteligente de análisis de datos

CRISP-DM(Cross-Industry Standard Process for Data Mining) es la metodología de desarrollo más usada en proyectos de Data Mining. La empresa kdnuggets.com publicó en el 2007 un ranking de las metodologías más usadas por la industria., ver Figura 21.

Figura 21

Metodologías utilizadas en DataMinig ([kdnuggets, 2007])



Para desarrollar el presente proyecto de investigación se utilizará la metodología CRISP-DM debido a su completitud en todas sus fases que va desde el análisis hasta el despliegue e implementación, y finalmente la presentación de resultados.

Tabla 3

Fase 1: Comprensión del Negocio

Comprensión del negocio	Comprensión de la data	Preparación de la data	Modelado	Evaluación	Despliegue y explotación
Determinar los objetivos empresariales *-Contexto* *-Objetivos Empresariales* *-Criterios de éxito empresarial* **Evaluar la situación inicial** *-Inventario de recursos* *-Requisitos, asunciones y restricciones* *-Riesgos y Contingencias* *-Terminología* *-Costes y Beneficios* **Establecer los objetivos de laminería de datos** *-Objetivos de minería de datos* *-Criterios de éxito de la minería de datos* **Redactar el plan del proyecto** *-Plan del proyecto* *-Evaluación inicial de técnicas y herramientas*	**Recoger Datos iniciales** **-** *Informe* **Describir Datos** -Informe **Explorar Datos** *-Informe* **Verificar la calidad de los datos** -Informe	**Seleccionar la data** *-Razones inclusión/exclusión de datos* **Limpiar la data** *-Informe* **Construir la data** *-Atributos derivados* *-Registros generados* **Ingresar la data** *-Datos combinados* **Formatear la data** *-Datos reformateados*	**Seleccionar técnicas de modelado** *-Técnicas de modelado seleccionados* *-Informe de restricciones de la técnica de modelado* **Diseño de pruebas** *-Prueba de diseño* **Construcción del Modelo** *-Configuración de parámetros* -Modelos -Descripción de los modelos **Evaluar los Modelos** *-Modelos evaluados* -Modelos -Revisión de los parámetros de los modelos	**Evaluar los resultados** *-Evaluación de los resultados de acuerdo con los objetivos del negocio* *-Modelos aprobados* **Revisar el Proceso** *-Informe de la Revisión del Proceso* **Determinar los siguientes pasos** *-Listado de acciones posibles* *-Decisión razonada de cómo proceder*	**Planificar despliegue** *-Plan de despliegue* **Planificar la monitorización y mantenimiento** *-Plan de despliegue* **Planificar la monitorización y mantenimiento** *-Plan de monitorización y mantenimiento* **Redactar el informe final** *-Informe final* *-Presentación y resultados del proyecto* **Revisión final de todo proyecto** *-Documentación de la experiencia adquirida*

Nota. Tomado De “Data Mining. The Crisp-Dm Methodology. The Clem Language And Ibm Spss Modeler (Pág. 16) Por César” Pérez López.

Determinar los objetivos empresariales

Contexto

Se evidencia que no existen estudios relacionados al seguimiento a estudiantes con riesgo de abandono temprano en el Instituto. Con la no existencia de este tipo de información en la Unidad de Bienestar Institucional que permita descubrir algún tipo de patrón por medio del cual se pueda predecir el futuro inmediato de deserción por parte de los estudiantes que ingresan a los primeros niveles del Instituto inicia el presente trabajo de investigación.

Objetivos empresariales

El Instituto Superior Tecnológico Cotopaxi es una Institución de educación superior, que busca explotar la información almacenada en sus bases de datos con el objetivo de obtener información relevante y oportuna por medio del Data Mining que les permita a sus autoridades tomar decisiones. Entre sus objetivos constan los siguientes:

- Disminuir la tasa de deserción del alumnado que ingresa a los primeros niveles del Instituto.
- Potenciar a la Unidad de Bienestar Institucional con una Api de deserción estudiantil que le permita a esta Unidad identificar a los posibles casos de deserción y darles seguimiento para procurar su permanencia en la Institución.
- Incrementar la eficiencia terminal de todas las cohortes que ingresan al Instituto.

Criterios de éxito empresarial

Bajar la tasa de deserción de los estudiantes que ingresan a los primeros niveles del Instituto contribuirá a mejorar el indicador 6.1.1 del Modelo de Evaluación Institucional para Los

Institutos Superiores Técnicos Y Tecnológicos En Proceso De Acreditación 2020, y seguir alcanzando la calidad en la educación.

Evaluar situación inicial

El Instituto Superior Tecnológico Cotopaxi es una institución de educación superior, cuyo propósito es formar profesionales íntegros de tercer nivel con compromiso ético y social que promuevan el desarrollo territorial sostenible.

Con corte al primer parcial de 2021 (1P-2021), el Instituto cuenta con 1239 estudiantes matriculados en todas sus carreras vigentes. El Instituto cuenta con su edificio matriz ubicado en la ciudad de Latacunga, Parroquia Tanicuchí Panamericana E35 km. 12 vía Latacunga – Quito.

Inventario de Recursos

El parque tecnológico del Instituto cuenta con un sistema académico SIGA (Sistema Integrado de Gestión Académica) de SENESCYT. EL motor de base de datos es PostgreSQL. Para desarrollo interno de aplicaciones la Unidad de TICS maneja programas de software libre como PHP, AJAX.

En la Tabla 4, se muestran los equipos que dispone el Instituto.

Tabla 4

Equipos

Cantidad	Equipo	Modelo	Marca	Ubicación
40	Switchs	SG500X-24P	Cisco	Bloque A, B, Talleres, Administrativo
25	AccesPoint	SN	Cisco	Bloque A, B, Talleres, Administrativo, Bar
85	PC's	Hurricane	Hurricane	Bloque A, B, Talleres, Administrativo
0	Servidores	SN	SN	SN
0	Router, Gateways	SN	SN	SN

Requisitos, asunciones y restricciones

Para el desarrollo del presente trabajo es necesario contar con la base de datos con corte al primer periodo de 2021 con toda la información de estudiantes disponibles. Por principio de anonimidad no se presenta información privada de los estudiantes.

Riesgos y Contingencias

El mayor riesgo que se puede tener es no disponer de la data y que esta sea de mala calidad. La copia de la base de datos con corte al primer periodo de 2021 fue entregada por el director de la Unidad de Tics, la misma que fue levantada en un motor de base de datos PostgreSQL.

Se puede experimentar costos adicionales en la realización de este trabajo por parte del investigador siempre y cuando el trabajo a realizar se extienda más del tiempo planificado.

En el periodo de matriculación el estudiante realiza en un formulario web el registro de su información lo que influye directamente en la calidad de la data.

Terminología

Ver en el Anexo 2 el glosario de términos.

Costes y Beneficios

Para el Instituto el proceso de registro de información por parte de los estudiantes al sistema SIGA no genera costes.

Entre los beneficios más significativos para el Instituto destacan el cumplimiento progresivo de los indicadores de retención del alumnado que figuran en el Modelo de Evaluación de la CACES y que periódicamente cada 4 años el Instituyo es sujeto de revisión.

Criterios de éxito de la minería de datos

Según (Géron, 2017) es común usar el % de los datos para entrenamiento y reservar el 20% para pruebas.

Al final, el modelo con el 90% de precisión podría no funcionar.

Redactar el plan del proyecto.

Plan del Proyecto

A continuación, se detalla la planificación del proyecto en todas sus fases con sus respectivos tiempos en la Tabla 5.

Tabla 5

Planificación y seguimiento de la metodología

Fase	Semanas	Personal
1. Comprensión del negocio	3	Director Bienestar, Investigador
2. Comprensión de la data	4	Director TICS, Investigador
3. Preparación de datos	4	Investigador
4. Modelado	4	Investigador
5. Evaluación	2	Director Bienestar, Investigador
6. Implementación o despliegue	1	Departamento de TI, Investigador

Técnicas y herramientas

Para el desarrollo de esta propuesta se considerará las siguientes herramientas y recursos:

Hardware:

- Procesador core I7 de octava generación
- Velocidad del Procesador 2.20GHz o superior
- 32.0 GB de Memoria RAM

- Disco Duro 500 GB estado sólido

Software:

- Python 3.9.6
- Jupyter notebook
- Heroku

Recurso Humano:

- Ing. José Luis Rosero

Extracción de datos:

- PostgreSQL

Técnicas de Modelado:

- Árboles de decisión
- Random Forest
- Redes Neuronales.

Tabla 6

Fase 2: Comprensión de la data

Comprensión del negocio	Comprensión de la data	Preparación de la data	Modelado	Evaluación	Despliegue y explotación
	Recoger Datos iniciales - *Informe* **Describir Datos** -Informe **Explorar Datos** *-Informe* **Verificar la calidad de los datos** -Informe				

Nota. Tomado De “Data Mining. The Crisp-Dm Methodology. The Clem Language And Ibm Spss Modeler (Pág. 16) Por César” Pérez López.

Datos iniciales

Recolección de datos iniciales - Informe

El proceso de recolección inicial de datos empieza con la recolección de información de vistas relacionadas, y el desarrollo de sentencias SQL que permitan extraer toda la información de la base de datos y construir el DataSet. Las sentencias de bases de datos generadas se pueden consultar en el Anexo 3.

Las vistas utilizadas para para la extracción de datos son:

- docente.vw_docente: Vista donde se almacena toda la información del personal docente
- matricula.vw_matriculados: Vista donde se almacena las matrículas que ha tenido el estudiante.

Las tablas utilizadas para para la extracción de datos son:

- administracion.per_persona: Tabla que almacena la ficha de información personal tanto de docentes como de estudiantes.
- matricula.gen_jornada: Tabla catalogo donde se muestran las jornadas vigentes de estudios para todas las carreras.
- matricula.gen_nivel_academico: Tabla catalogo que almacena los diferentes niveles académicos por los que cursa el alumno.
- matricula.ins_asignatura: Tabla que almacena las materias que se dictan por nivel y por carrera con su respectivo número de horas autónomas, docencia, y de práctica.
- matricula.mat_estudiante: Tabla que almacena información complementaria de la matricula del estudiante.

- matricula.mat_ingreso: Tabla catálogo que almacena el destino de los ingresos que produce cuando el estudiante trabaja.
- matricula.mat_matricula_asignatura: Tabla que almacena las notas de los estudiantes por periodo académico.
- matricula.mat_nivel_formacion: Tabla catalogo que almacena el nivel de educación.
- matricula.mat_ocupacion: Tabla catalogo que almacena la ocupación del estudiante.
- matricula.mat_paralelo_periodo: Tabla que almacena información complementaria del docente como carrera, periodo, nivel, asignatura.

Para el desarrollo del modelo se recabaron la siguiente lista de variables que constan en la Tabla 7.

Tabla 7

Lista de variables DataSet

N°	Variable	Tipo de dato	Ejemplos
1	Edad	Entero	18 20
2	Tiene_discapacidad	Categórico	Si No
3	Formación_padre	Categórico	Primaria Secundaria Universidad
4	Formación_madre	Categórico	Primaria Secundaria Universidad
5	Estado_civil	Categórico	Casado Divorciado Soltero Unión de Hecho Viudo
6	Género	Categórico	Femenino Masculino
7	Etnia	Categórico	Indígena Mestizo

N°	Variable	Tipo de dato	Ejemplos
			Otros
8	Región_pais	Categórico	Costa
			Oriente
			Sierra
9	Bono_desarrollo	Categórico	Si
			No
10	Ocupación	Categórico	Solo estudia
			Trabaja
11	Destino_ingresos	Categórico	Financiar sus estudios
			Gastos personales
			No aplica
			Para mantener a su hogar
12	Ingreso_total_hogar	Entero	450
			1000
13	Cantidad_miembros_hogar	Entero	5
			4
14	deserción		1
			0

Las variables que se listan en la Tabla 8 se descartaron para la conformación del DataSet final.

Tabla 8

Listado de variables no seleccionadas

Variables	Motivo
id_persona	no es una variable categórica o numérica representativa
identificacion	no es una variable categórica o numérica representativa
primer_nombre	no es una variable categórica o numérica representativa
segundo_nombre	no es una variable categórica o numérica representativa
primer_apellido	no es una variable categórica o numérica representativa
segundo_apellido	no es una variable categórica o numérica representativa
fecha_nacimiento	no es una variable categórica o numérica representativa
porcentaje_discapacidad	ya se tiene la variable edad
carnet_discapacidad	No es representativa

Variables	Motivo
calle_principal	no es una variable categórica o numérica representativa
numero	no es una variable categórica o numérica representativa
calle_secundaria	no es una variable categórica o numérica representativa
cod_postal	no es una variable categórica o numérica representativa
correo_principal	no es una variable categórica o numérica representativa
correo_secundario	no es una variable categórica o numérica representativa
telefono_domicilio	no es una variable categórica o numérica representativa
telefono_movil	no es una variable categórica o numérica representativa
contacto_emergencia	no es una variable categórica o numérica representativa
telefono_emergencia_domicilio	no es una variable categórica o numérica representativa
telefono_emergencia_movil	no es una variable categórica o numérica representativa
tipo_identificacion	no es una variable categórica o numérica representativa
tipo_sangre	no es una variable categórica o numérica representativa
pueblo	no es una variable categórica o numérica representativa
tipo_discapacidad	no es una variable categórica o numérica representativa
efermedad	no es una variable categórica o numérica representativa
pais_nacimiento	no es una variable categórica o numérica representativa
provincia_nacimiento	no es una variable categórica o numérica representativa
pais_residencia	no es una variable categórica o numérica representativa
provincia_residencia	no es una variable categórica o numérica representativa
parentesco	no es una variable categórica o numérica representativa
etniapp	no es una variable categórica o numérica representativa
siglas_carrera	no es una variable categórica o numérica representativa
maxima_nota_primero	no es una variable categórica o numérica representativa

Variables	Motivo
minima_nota_primero	no es una variable categórica o numérica representativa
promedio_primero	no es una variable categórica o numérica representativa
desviacion_notas_primero	no es una variable categórica o numérica representativa
asistencia_promedio_primero	no es una variable categórica o numérica representativa
nro_materias_primero	no es una variable categórica o numérica representativa
nro_materias_reprobados_primero	no es una variable categórica o numérica representativa
min_periodo_primero	no es una variable categórica o numérica representativa
reingreso	no es una variable categórica o numérica representativa
gratuidad	no es una variable categórica o numérica representativa
id_ocupacion	no es una variable categórica o numérica representativa
id_ingreso	no es una variable categórica o numérica representativa
id_nivel_formacion_padre	no es una variable categórica o numérica representativa
id_nivel_formacion_madre	no es una variable categórica o numérica representativa
anio_titulo_colegio	no es una variable categórica o numérica representativa
internet	no es una variable categórica o numérica representativa
periodo_graduacion	no es una variable categórica o numérica representativa
dato	no es una variable categórica o numérica representativa

Describir Datos

Informe

A continuación, se procede a describir las variables de entrada y las variables de salida:

Variables de entrada:

- Edad: Número de años que tiene el estudiante
- Tiene_discapacidad: Indica si el alumno tiene alguna discapacidad

- Formación_padre: indica el grado de educación que tiene el padre del alumno.
- Formación_madre: indica el grado de educación que tiene la madre del alumno.
- Estado_civil: indica el estado civil del alumno
- Género: indica el sexo del alumno
- Etnia: indica la etnia con la que se identifica el alumno
- Región país: indica la región de donde viene el estudiante
- Bono_desarrollo: indica si el alumno recibe o no el bono de desarrollo
- Ocupación: indica si el estudiante trabajo o solo estudia
- Destino_ ingresos: si recibe algún ingreso indica el destino de este.
- Ingreso_total_hogar: indica los ingresos de su hogar
- Cantidad_miembros_hogar: indica el número de miembros que conforman el núcleo familiar.

Variable de salida:

- La variable de salida representa el resultado de la predicción del modelo.

Explorar datos

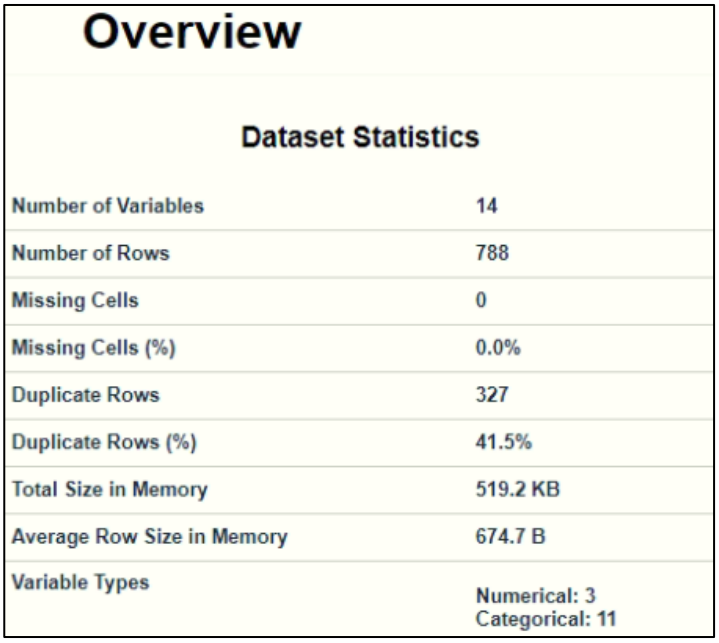
Informe

Se realizó el EDA(Análisis exploratorio de datos) a las 14 variables presentes en el DataSet, con el propósito de obtener estadísticas particulares de cada una en relación a la variable dependiente y poder identificar su tendencia.

En la Figura 22, se muestra el resumen de todas las variables que intervienen en el presente trabajo.

Figura 22

Estadísticas del DataSet



Overview

Dataset Statistics

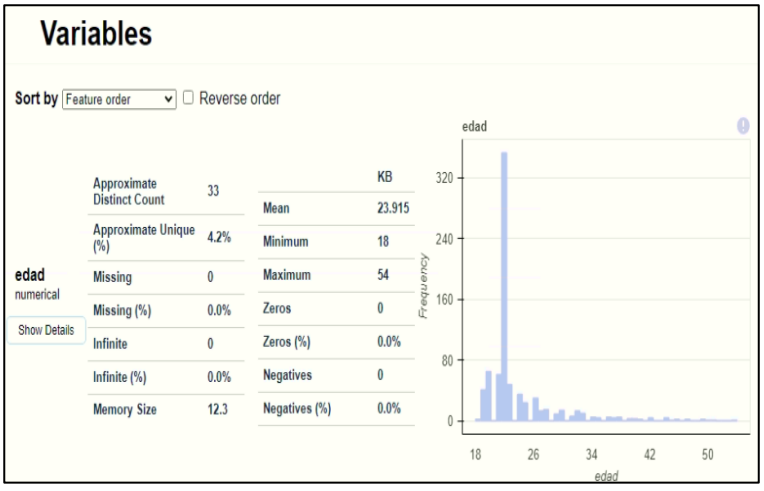
Number of Variables	14
Number of Rows	788
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	327
Duplicate Rows (%)	41.5%
Total Size in Memory	519.2 KB
Average Row Size in Memory	674.7 B
Variable Types	Numerical: 3 Categorical: 11

A continuación, se muestra un resumen del EDA por cada variable.

Edad.- El 44.05% de estudiantes a la edad promedio de 23 años deserta. (Ver Figura 23)

Figura 23

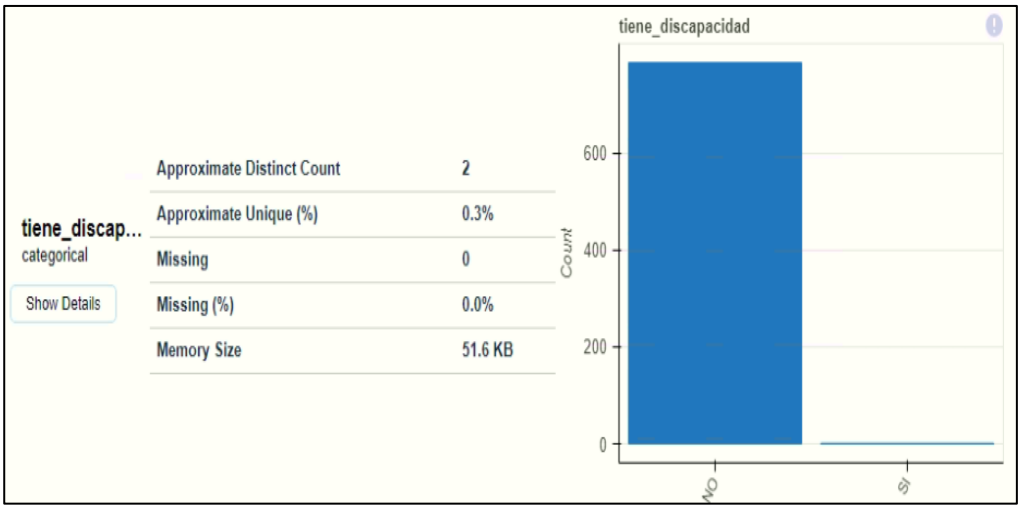
Exploración de edad



Tiene_discapacidad.- El 99.87% de estudiantes No registra discapacidad. (Ver Figura 24)

Figura 24

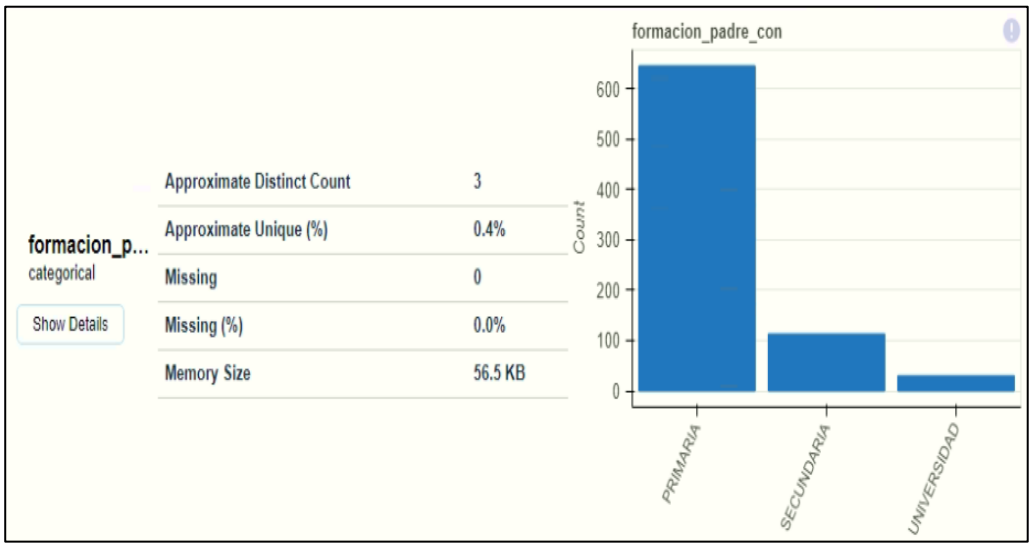
Exploración de variable tiene discapacidad



Formación_padre.- El 81.7% de los padres de familia de los estudiantes tiene formación de primaria. (Ver Figura 25)

Figura 25

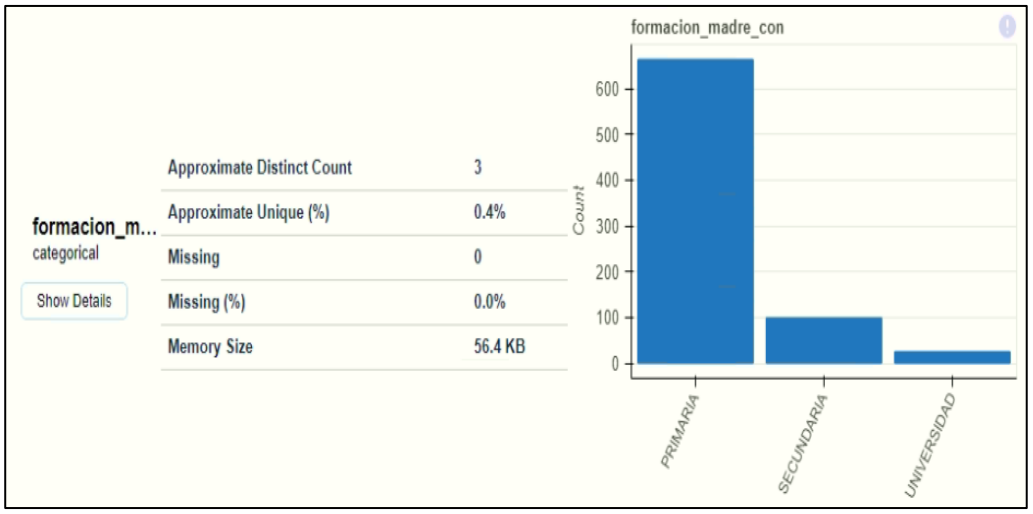
Exploración de variable formación_padre



Formación_madre.- El 84.7% de las madres de familia de los estudiantes tiene formación de primaria. (Ver Figura 26)

Figura 26

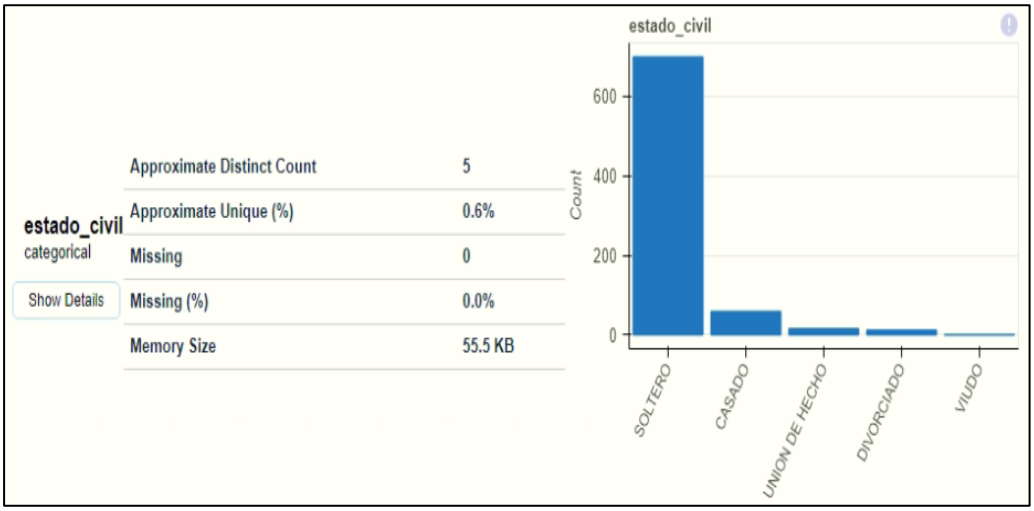
Exploración de variable formación_madre



Estado_civil.- El 88.86% de estudiantes son de estado civil soltero. (Ver Figura 27)

Figura 27

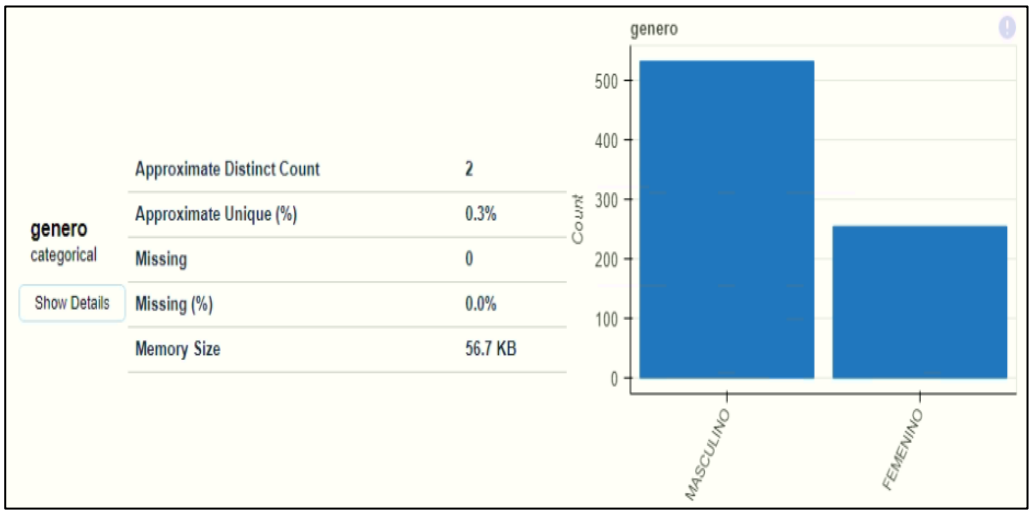
Exploración de variable estado_civil



Género.- El 67.72% de estudiantes son de género masculino. (Ver Figura 28)

Figura 28

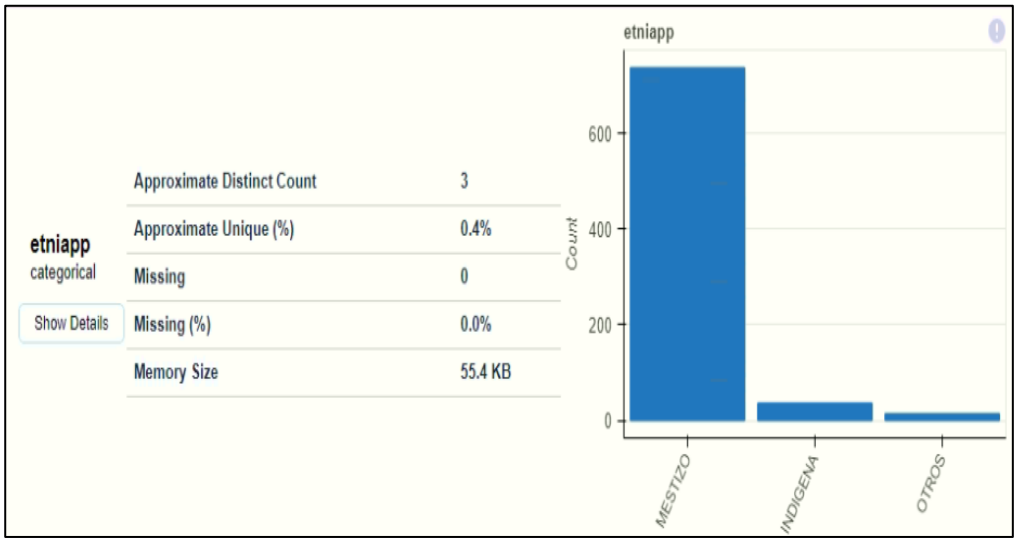
Exploración de variable género



Etnia .- El 93.54% de estudiantes son de etnia Mestiza. (Ver Figura 29)

Figura 29

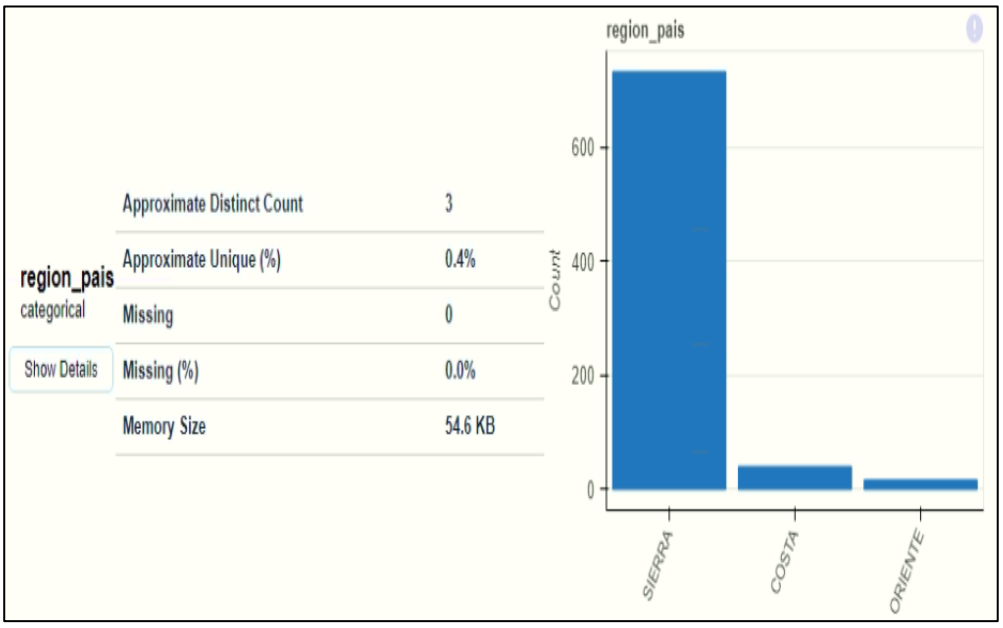
Exploración de variable etnia



Region_pais.- El 92.91% de estudiantes son de la región Sierra. (Ver Figura 30)

Figura 30

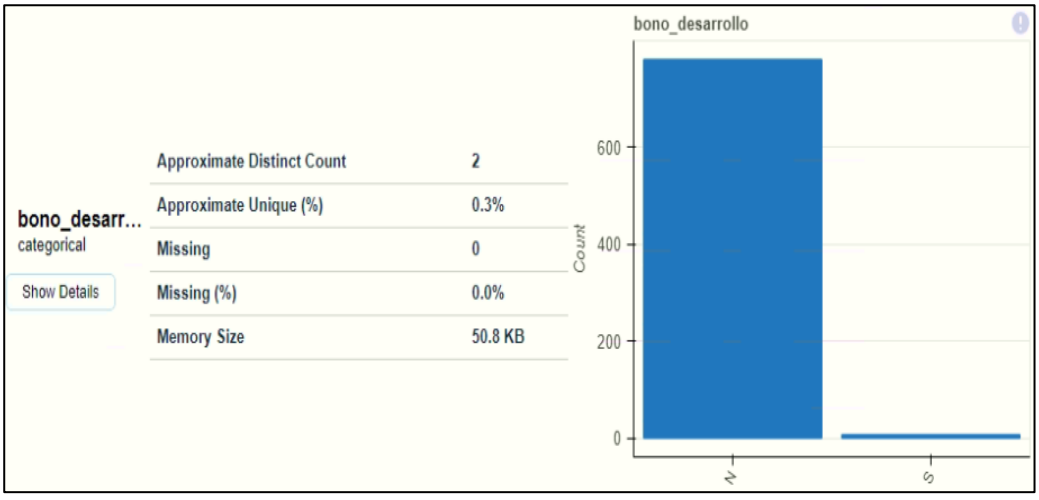
Exploración de variable region_pais



Bono de desarrollo. - El 99.11% de estudiantes No reciben el bono de desarrollo del Gobierno Nacional. (Ver Figura 31)

Figura 31

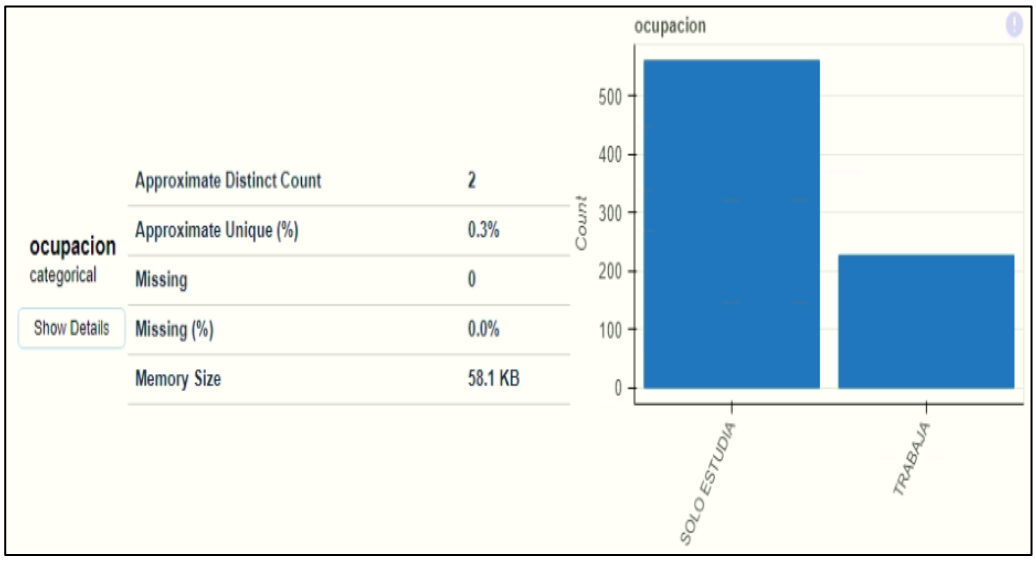
Exploración de variable bono_desarrollo



Ocupación. - El 71.26% de estudiantes se dedica Solo a estudiar. (Ver Figura 32)

Figura 32

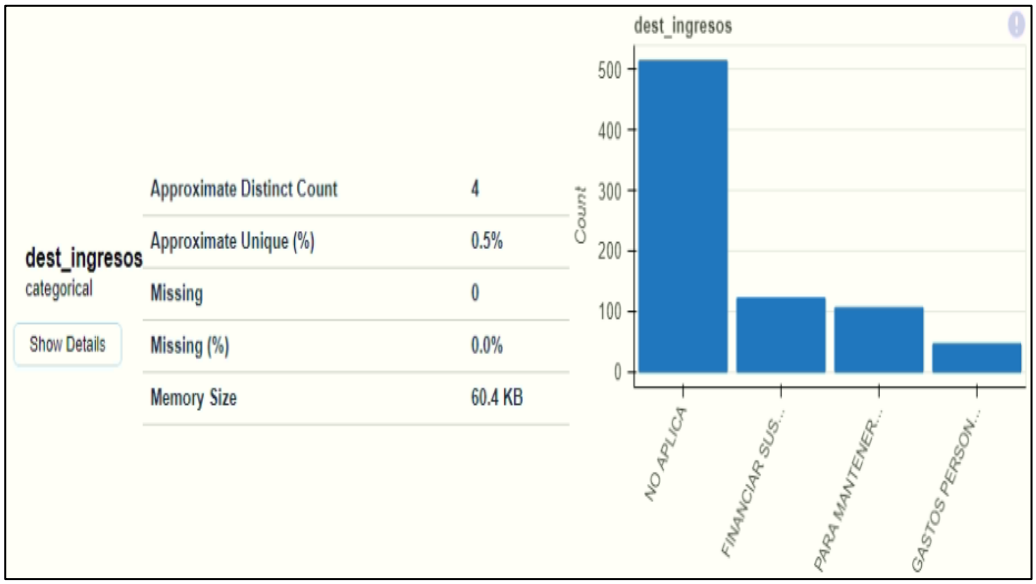
Exploración de variable ocupación



Destino de los ingresos. - El 65.32% de estudiantes no trabaja. (Ver Figura 33)

Figura 33

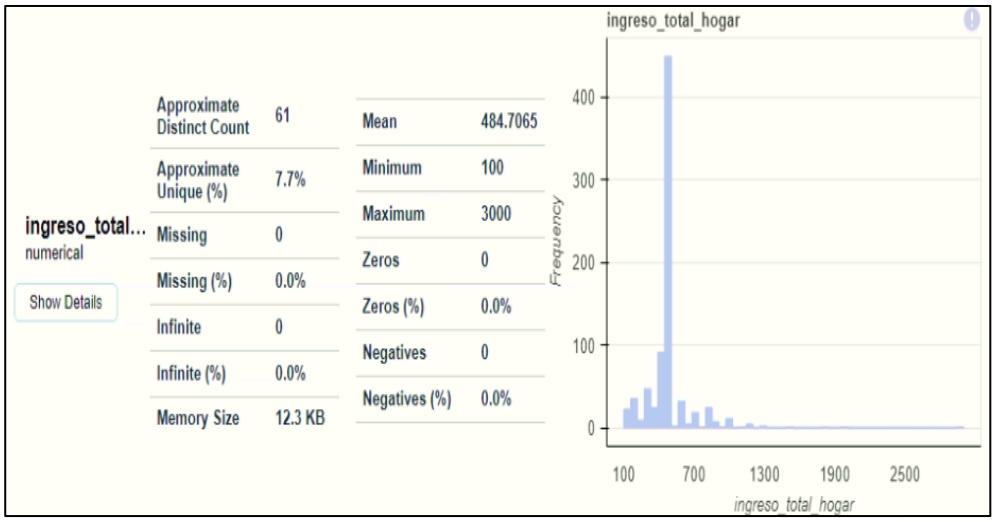
Exploración de variable dest_ingresos



Ingreso Total del Hogar. - El 47.22% de estudiantes registra en su hogar un ingreso promedio de 483.85 dólares. (Ver Figura 34)

Figura 34

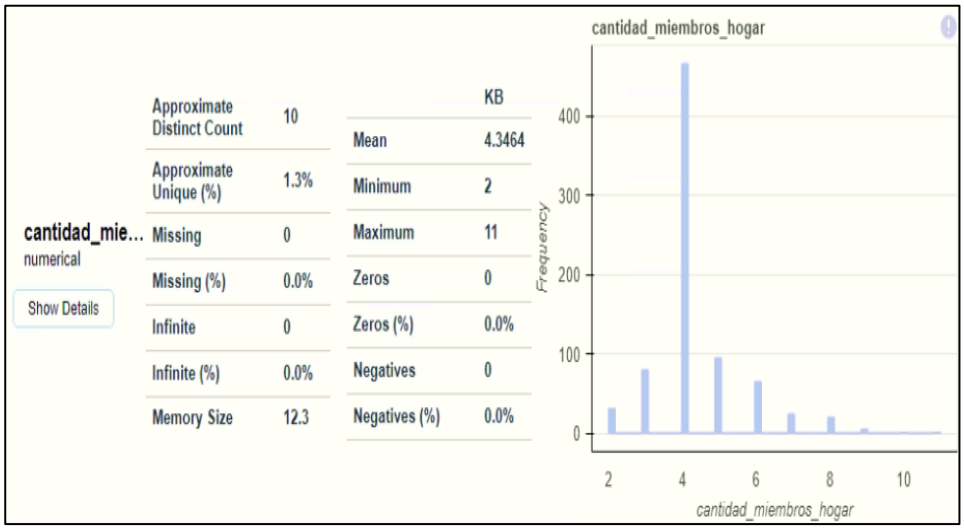
Exploración de variable ingreso_total_hogar



Cantidad de Miembros del Hogar. - El 58.99% de estudiantes registra en su hogar un promedio de 4 miembros. (Ver Figura 35)

Figura 35

Exploración de variable cantidad_miembros_hogar



Calidad de los Datos

Informe

La ISO/IEC 25012 es un modelo de calidad para aquellos datos que están representados en un formato estructurado dentro de un sistema informático, por medio de sus características se puede definir la calidad de la data.

La calidad de los datos está compuesta por las siguientes características:

Accesibilidad (AC).- El acceso a los datos se da en un grado especifico (Calebrese, Esponda, Pasini, Boracchia, & Pesado, 2019).

Conformidad (CO).- Los datos cumplen con normativas, convenciones y estándares. (Calebrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

Confidencialidad (CF).- Los datos son accedidos por usuarios que cuentan con permisos de autorización. (Calebrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

Eficiencia (EF). - El análisis de los datos se espera conforme con sus niveles de rendimiento. (Calebrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

Precisión (PR). -. Debe existir exactitud en los datos (Calebrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

Trazabilidad (TZ). - Se analizan los datos proporcionados para ser almacenados en un registro. (Calebrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

Comprensibilidad (CP). – Los datos se expresan mediante símbolos y lenguajes apropiados para posterior ser leido por otros usuarios. (Calebrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

La ponderación de cada variable de acuerdo con la característica de validez de datos se puede ver en la Tabla 9.

Tabla 9

Comprobación de datos

VARIABLES/FEATURES	AC	CO	CF	EF	PR	TZ	CP	TOTAL
Edad	1	1	1	0	1	1	1	6
Tiene_discapacidad	1	1	1	1	0	1	1	7
Formación_padre	1	0	1	1	0	1	1	5
Formación_madre	1	1	1	0	0	1	1	5
Estado_civil	1	1	1	1	1	1	1	7
Género	1	1	1	0	1	1	1	6
Etnia	1	1	1	1	1	1	1	7
Región_pais	1	1	1	0	1	1	1	6
Bono_desarrollo	1	1	1	1	1	1	1	7
Ocupacion	1	1	1	1	1	1	1	7
Destino_ingresos	1	1	1	1	1	1	1	7
Ingreso_total_hogar	1	1	1	1	1	1	1	7
Cantidad_miembros_hogar	1	1	1	0	0	1	1	5

Realizada la comprobación de las variables, se determina que los índices de Eficiencia y Precisión deben mejorar.

Tabla 10

Fase 3: Preparación de la data

Comprensión del negocio	Comprensión de la data	Preparación de la data	Modelado	Evaluación	Despliegue y explotación
		Seleccionar la data *-Razones inclusión/exclusión de datos* **Limpiar la data** *-Informe* **Construir la data** *-Atributos derivados* *-Registros generados* **Ingresar la data** *-Datos combinados* **Formatear la data** *-Datos reformateados*			

Nota. Tomado De “Data Mining. The Crisp-Dm Methodology. The Clem Language And Ibm Spss Modeler (Pág. 16) Por César” Pérez López.

Seleccionar la data

Razones de inclusión/exclusión de datos

De la revisión de variables realizada a la base de datos, se forma el DataSet final compuesta por 14 variables con un número total de 790 registros correspondiente a los primeros niveles de todas las carreras que van desde el primer semestre de 2012 hasta el primer periodo de 2021.

Limpiar la data

Informe

De la revisión realizada a todas las variables se evidencia que existe información almacenada con valores nulos, información errada que el estudiante ingresa al sistema SIGA cuando ingresa su información personal en la ficha de datos.

En este paso de la metodología se procede a la limpieza de la data almacenada en cada una de las variables que presentan inconsistencias.

Edad. - Mediante Quipux, Rectorado envió una solicitud de información a la Dirección General de Registro Civil, Identificación y Cedulación solicitando se nos facilite información personal de los estudiantes que no tienen registrado su fecha de nacimiento en el sistema SIGA para poder realizar el cálculo de la edad, pero la respuesta por parte de la entidad de Registro fue negativa según Oficio Nro. DIGERCIC-CZ3.OT05-2021-1138-O que consta en el Anexo 5. Ante esto se actualizó la variable fecha_nacimiento con fechas randomicas correspondientes a la media de edad de los estudiantes.

Tiene discapacidad. - Este campo fue actualizado con NO cuando el campo porcentaje_discapacidad tenía nulos. Se actualizó con SI cuando el campo porcentaje_discapacidad era diferente de nulo.

Estado civil. – Los valores nulos fueron actualizados **con** Soltero.

País de nacimiento. – Los valores nulos fueron actualizados con Ecuador, los países diferentes a Ecuador fueron actualizados como Otros.

Provincia de nacimiento. – Los valores nulos fueron actualizados con Cotopaxi.

Bono de desarrollo. – Los valores nulos fueron actualizados con N.

Ocupación. – Los valores nulos fueron actualizados a solo estudia. Cuando estaba ingresado trabaja y estudia, se actualizó a trabaja.

Destino de los ingresos. – Los valores nulos fueron actualizados a no aplica.

Formación del padre. – Los valores nulos fueron actualizados a primaria.

Formación de la madre. – Los valores nulos fueron actualizados a primaria.

Ingreso total del hogar. – Los valores nulos y los valores ingresados menores a 100 dólares fueron actualizados con la media de este campo.

Cantidad de miembros del hogar. – Los valores nulos fueron actualizados con la mediana de este campo.

Internet. – Valores nulos y los valores ingresados con una X, fueron actualizados a N.

Género. – Los valores nulos fueron actualizados a masculino por ser los más representativos cuantitativamente hablando en este campo.

Construir la data

Atributos derivados

Transformar los datos tipo objeto a categóricos facilitará su entendimiento y entrenamiento del algoritmo.

El DataSet cuenta con tres variables numéricas, y 11 variables categóricas.

En la Tabla 11, se detalla estas variables y se visualiza su contenido.

Tabla 11

Ponderación numérica

Variables	Rango	Descripción
Edad	18 a 54	Edad del estudiante
Tiene discapacidad		No Si
Formación del Padre		Primaria Secundaria Universidad
Formación de la Madre		Primaria Secundaria Universidad
Estado civil		Casado Divorciado Soltero Unión de Hecho Viudo
Género		Femenino Masculino
Etnia		Indígena Mestizo Otros
Región país		Costa Oriente Sierra
Bono desarrollo		N S
Ocupación		Solo estudia Trabaja
Destino ingresos		Financiar sus estudios Gastos personales No aplica Para mantener a su hogar
Ingreso total del hogar	100 – 3000	
Cantidad miembros del Hogar	2 - 11	

Registros generados

Con la conversión de las variables categóricas a numéricas, quedan todas las variables del mismo tipo y con ello se normaliza las variables para obtener un mejor desempeño del modelo. Una vez que todas las variables están en formato numérico, se puede proceder a la siguiente fase.

Integrar datos

Datos combinados

No se realiza la integración con otras bases de datos. Se utiliza únicamente la información almacenada en la base de datos y extraída mediante el motor de base de datos PostgreSQL.

Formatear la data

Datos reformateados

Una vez que el DataSet se encuentra normalizado todas sus variables a numéricas, como se muestra a continuación en la Figura 36.

Figura 36

Muestra de las variables normalizadas

```
raw_data2.head()
```

	edad	tiene_discapacidad	formacion_padre	formacion_madre	estado_civil	genero	etnia	region	bono_desarrollo	ocupacio
0	24	0	1	1	3	2	2	3	0	
1	44	0	1	1	3	1	2	3	0	
2	22	0	2	2	3	2	2	3	0	
3	21	0	1	1	3	2	2	3	0	
4	21	0	1	2	3	2	2	3	0	

Tabla 12

Fase 4: Modelado

Comprensión del negocio	Comprensión de la data	Preparación de la data	Modelado	Evaluación	Despliegue y explotación
			Seleccionar técnicas de modelado *-Técnicas de modelado seleccionadas* *-Informe de restricciones de la técnica de modelado* **Diseñar las pruebas del modelo** *-Pruebas del modelo* **Construir los modelos** *--Configuración de parámetros* *-Modelos* -Descripción de los modelos **Evaluar los modelos** *-Modelos evaluados* *-Revisión de los parámetros de los modelos*		

Nota: Tomado De “Data Mining. The Crisp-Dm Methodology. The Clem Language And Ibm Spss Modeler (Pág. 16) Por César” Pérez López.

Seleccionar técnicas de modelado

Técnicas de modelado seleccionado

En base a la revisión bibliográfica realizada al inicio de este trabajo de investigación para determinar las variables con mayor recurrencia, también se identificaron los modelos más usados para realizar predicciones en el ámbito de deserción estudiantil. Los modelos seleccionados son los siguientes:

- Árboles de decisión
- Random Forest
- Redes Neuronales

Los tres modelos serán desarrollados y evaluados en Python. En la evaluación de los modelos se utilizarán métricas de validación y así determinar cuál es el modelo que tiene un mejor ranking por modelo.

Informe de restricciones de la técnica de modelado

Conforme a la revisión bibliográfica realizada y al estudio y análisis de cada modelo a desarrollar se puede determinar que no se presentan restricciones para su desarrollo.

Diseñar las pruebas del modelo

Pruebas del modelo

Antes de desarrollar y entrenar a cada modelo se debe definir un plan de pruebas que permita dividir el set de datos en entrenamiento y en pruebas. Seguidamente para evaluar a cada modelo se deben tomar en cuenta las métricas de validación descritas en el Capítulo dos de la matriz de confusión y determinar cuáles son las métricas con mejores indicadores de precisión de cada modelo.

Para este trabajo de investigación el plan de pruebas inicia definiendo el porcentaje de datos a utilizar en el entrenamiento y el porcentaje de datos a utilizar en el entrenamiento.

Primeramente, se divide el set de datos en 80% de los datos para entrenamiento y se reserva el 20% para pruebas.

Seguidamente para la evaluación de modelos se utilizará las siguientes técnicas de evaluación de la Matriz de Confusión:

- Exactitud
- Precision
- Sensibilidad
- Especificidad
- Tasa de error
- F1 Score

Finalmente se evalúa cada modelo con Cross Validation o validación cruzada. La validación cruzada es un método estadístico para evaluar el rendimiento de la generalizado que es más estable y exhaustivo que el uso de una división en un conjunto de entrenamiento y otro de prueba.

Construir los modelos

Configuración de parámetros

Árboles de decisión.

Para la construcción del árbol de decisión se establece el criterio a entropía y una profundidad máxima de 5.

Random Forest.

Para la construcción de modelo Random forest se establece el criterio de entropía, una profundidad máxima de 5, el número de árboles en el bosque de 100, max_features por defecto sqrt para buscar el mejor Split, muestras bootstrap por defecto en verdadero al construir los árboles, el número de muestras a extraer de X max_samples = 2/3, oob_score=True Si se utilizan muestras fuera de la bolsa para estimar la puntuación de generalización. Sólo está disponible si bootstrap=True.

Red Neuronal.

Este modelo contiene los siguientes parámetros para Multi Layer Perceptron; el número de capas ocultas de (10,), número máximo de iteraciones de 500, un optimizador de solución basado en lbfgs para pequeños DataSets.

Modelos

Árboles de decisión

Figura 37

Importa las librerías necesarias

```
# Paquetes / Librerias
import os
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
# Datetime lib
from pandas import to_datetime
import itertools
import datetime
import warnings
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import plot_roc_curve
from scikitplot.metrics import plot_roc_curve
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, r2_score, precision_score, recall_score, f1_score
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import KFold
#BDD
from sqlalchemy import create_engine

warnings.filterwarnings('ignore')

%matplotlib inline

# Cambiar a float
np.set_printoptions(formatter={'float_kind':'{:f}'.format})

# Incrementa el tamaño de los plots
sns.set(rc={'figure.figsize':(8,6)})
```

Figura 38

Conexión a la base de datos

```
engine = create_engine('postgresql://postgres:sql@localhost:5432/Cotopaxi')
raw_data = pd.read_sql("select * from modelosf_primero", engine)
```

Figura 39

Verificación que no haya valores nulos

```
raw_data.isnull().sum()
```

```
edad                    0
ingreso_total_hogar     0
cantidad_miembros_hogar 0
tiene_discapacidad      0
formacion_padre         0
formacion_madre         0
estado_civil            0
genero                  0
etnia                   0
region                  0
bono_desarrollo         0
ocupacion               0
destino_ingresos        0
desercion               0
dtype: int64
```

Figura 40

Conversión de tipo de variables de object a category

```
# de object a entero

raw_data["tiene_discapacidad"]=raw_data["tiene_discapacidad"].astype("int64")
raw_data["formacion_padre"]=raw_data["formacion_padre"].astype("int64")
raw_data["formacion_madre"]=raw_data["formacion_madre"].astype("int64")
raw_data["estado_civil"]=raw_data["estado_civil"].astype("int64")
raw_data["genero"]=raw_data["genero"].astype("int64")
raw_data["etnia"]=raw_data["etnia"].astype("int64")
raw_data["region"]=raw_data["region"].astype("int64")
raw_data["bono_desarrollo"]=raw_data["bono_desarrollo"].astype("int64")
raw_data["ocupacion"]=raw_data["ocupacion"].astype("int64")
raw_data["destino_ingresos"]=raw_data["destino_ingresos"].astype("int64")


raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2420 entries, 0 to 2419
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   edad                    2420 non-null   int64
 1   ingreso_total_hogar     2420 non-null   float64
 2   cantidad_miembros_hogar 2420 non-null   float64
 3   tiene_discapacidad      2420 non-null   int64
 4   formacion_padre         2420 non-null   int64
 5   formacion_madre         2420 non-null   int64
 6   estado_civil            2420 non-null   int64
 7   genero                  2420 non-null   int64
 8   etnia                   2420 non-null   int64
 9   region                  2420 non-null   int64
 10  bono_desarrollo         2420 non-null   int64
 11  ocupacion               2420 non-null   int64
 12  destino_ingresos        2420 non-null   int64
 13  desercion               2420 non-null   int64
dtypes: float64(2), int64(12)
memory usage: 264.8 KB
```

Figura 41

Limitando la data

```
# Limitando la data
raw_data2 = raw_data[['edad', 'tiene_discapacidad',
       'formacion_padre', 'formacion_madre', 'estado_civil', 'genero',
       'etnia', 'region', 'bono_desarrollo', 'ocupacion', 'destino_ingresos',
       'ingreso_total_hogar', 'cantidad_miembros_hogar', 'desercion']]
```

Se elimina el campo que no se necesita en el entrenamiento, y se divide la data en 80% de entrenamiento y 20% pruebas. Los datos de “X” y “y” se utilizarán en los tres modelos.

Figura 42

Eliminación de campo

```
X = raw_data2.drop('desercion', axis=1).values
y = raw_data2['desercion'].values
print('X shape: {}'.format(np.shape(X)))
print('y shape: {}'.format(np.shape(y)))

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
X shape: (2420, 13)
y shape: (2420,)
```

Corriendo el árbol de decisión con los criterios indicados en la fase de modelado

Figura 43

Corriendo el árbol de decisión

```
dt = tree.DecisionTreeClassifier(criterion='entropy', max_depth=5)
```

Se entrena el modelo utilizando fit con los datos de entrenamiento de “X” y de “y”

Figura 44

Entrenando el modelo

```
dt.fit(X_train, y_train)
```

Se hace una predicción con el método predict y los datos de test.

Figura 45

Predicción

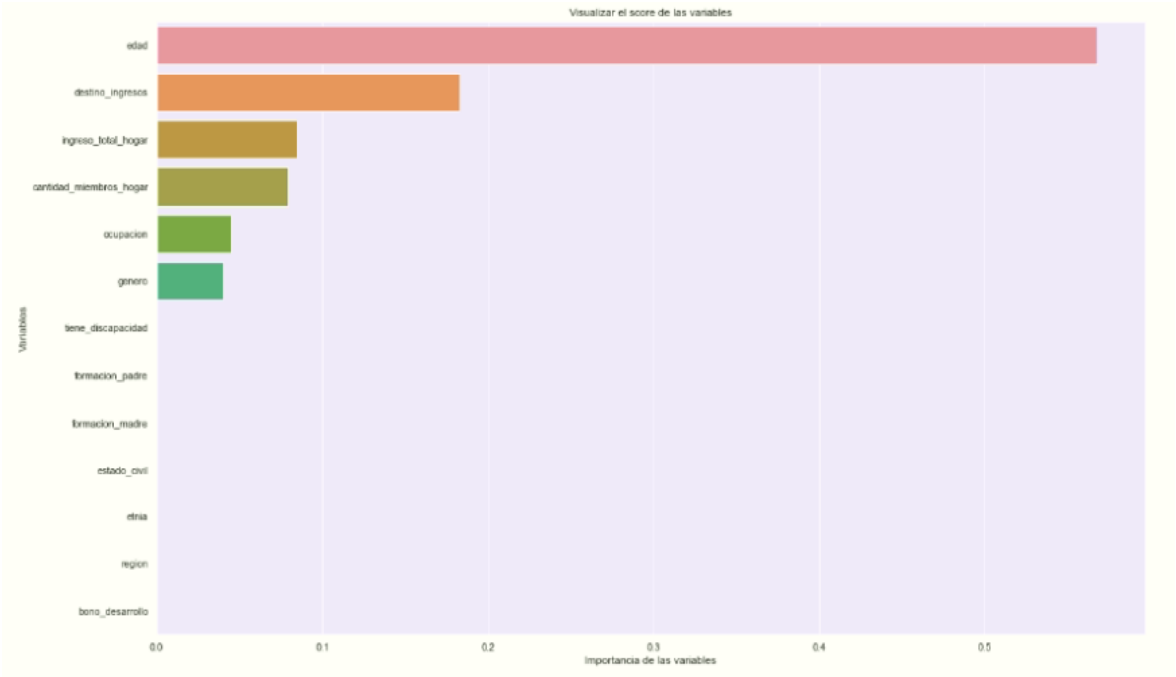
```
##  Realiza la predicción
y_pren = dt.predict(X_test)
```

Realizada la predicción se obtiene el ranking de las variables más importantes para el modelo.

Figura 46

Obtención del ranking de variables

```
#score de las variables
f, ax = plt.subplots(figsize=(20, 12))  #30,24
ax = sns.barplot(x=feature_scores, y=feature_scores.index)
ax.set_title("Visualizar el score de las variables")
ax.set_yticklabels(feature_scores.index)
ax.set_xlabel("Importancia de las variables")
ax.set_ylabel("Variables")
plt.show()
```



Random Forest.

Corriendo el modelo de random forest con los criterios indicados en la fase de modelado

Figura 47

Corriendo el modelo de random forest

```
rf = RandomForestClassifier(n_estimators=100,
                            criterion='entropy',
                            max_depth=5,
                            max_features="sqrt",
                            bootstrap=True,
                            max_samples= 2/3,
                            oob_score=True)
```

Se entrena el modelo utilizando fit con los datos de entrenamiento de “X” y de “y”

Figura 48

Entrena del modelo utilizando fit

```
rf.fit(X_train, y_train)
```

Se hace una predicción con el método predict y los datos de test.

Figura 49

Predicción con el método predict

```
##  Prediccion
prediction_test = rf.predict(X=X_test)
```

Realizada la predicción se obtiene el ranking de las variables más importantes para el modelo.

Figura 50

Ranking de las variables más importantes para el modelo

```
# Grafico de barras con seaborn

f, ax = plt.subplots(figsize=(20, 12))  #30,24
ax = sns.barplot(x=feature_scores, y=feature_scores.index)
ax.set_title("Visualize feature scores of the features")
ax.set_yticklabels(feature_scores.index)
ax.set_xlabel("Feature importance importances")
ax.set_ylabel("Features")
plt.show()
```

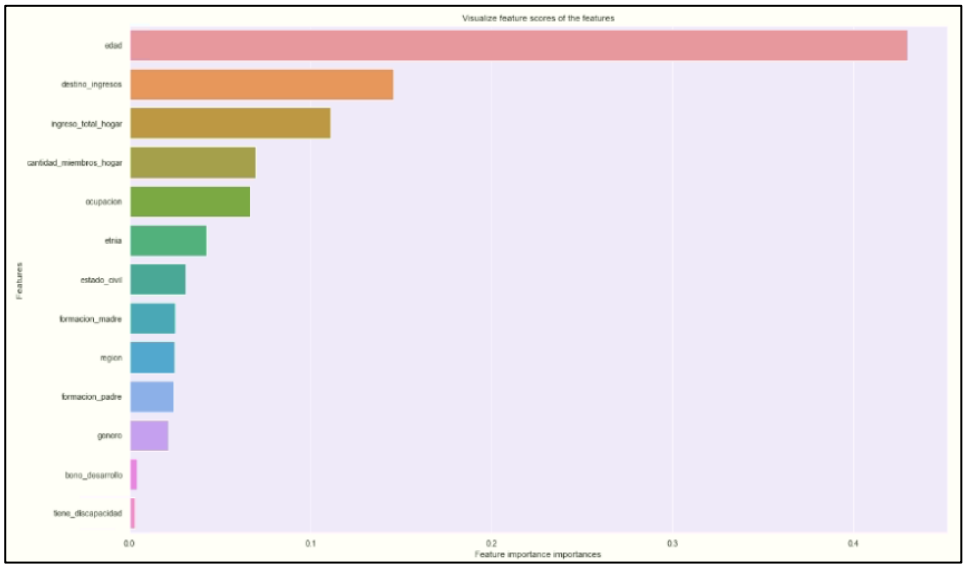


Figura 51

Redes Neuronales. Generar y entrenar el modelo

```
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test  = scaler.transform(X_test)

mlp=MLPClassifier(hidden_layer_sizes=(10,), max_iter=500, solver='lbfgs')

mlp.fit(X_train, y_train)
```

Se realiza una predicción con los datos de test.

Figura 52

Predicción con los datos de test

```
y_pred_nn = mlp.predict(X_test)
```

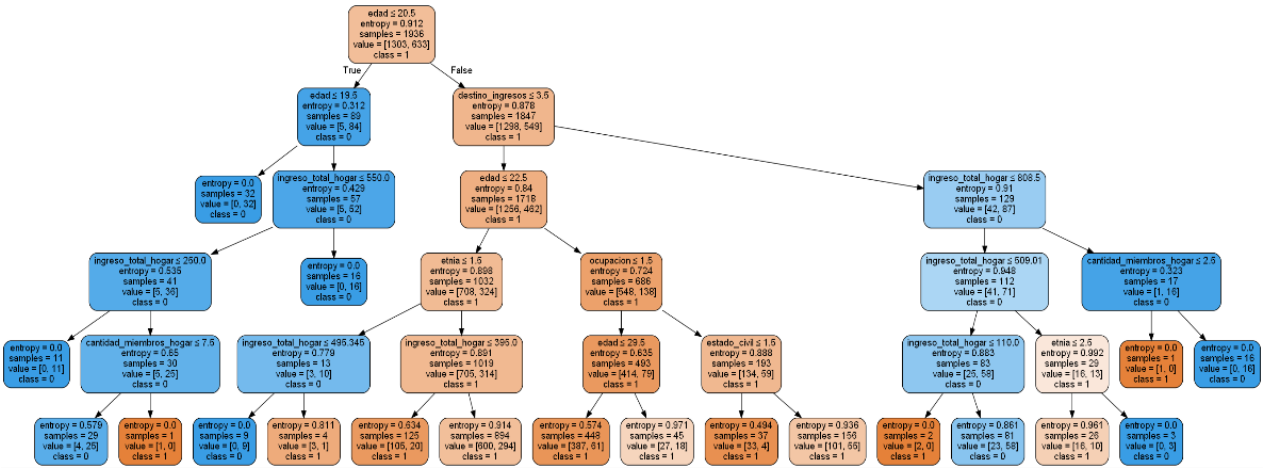
Descripción de los modelos

Árboles de decisión.

Inicia el desarrollo de este modelo con la importación de las librerías requeridas, se procede a conectar la base de datos, la división del set de datos en 80% para entrenamiento y el 20% para pruebas, seguidamente se corre el modelo de árbol de decisión con los criterios indicados en la fase de modelado, se entrena el modelo utilizando le método fit con los datos de entrenamiento de “X” y de “y” y finalmente se hace una predicción con el método predict y los datos de test. En la figura 53 se muestra el árbol generado.

Figura 53

Árbol de decisión



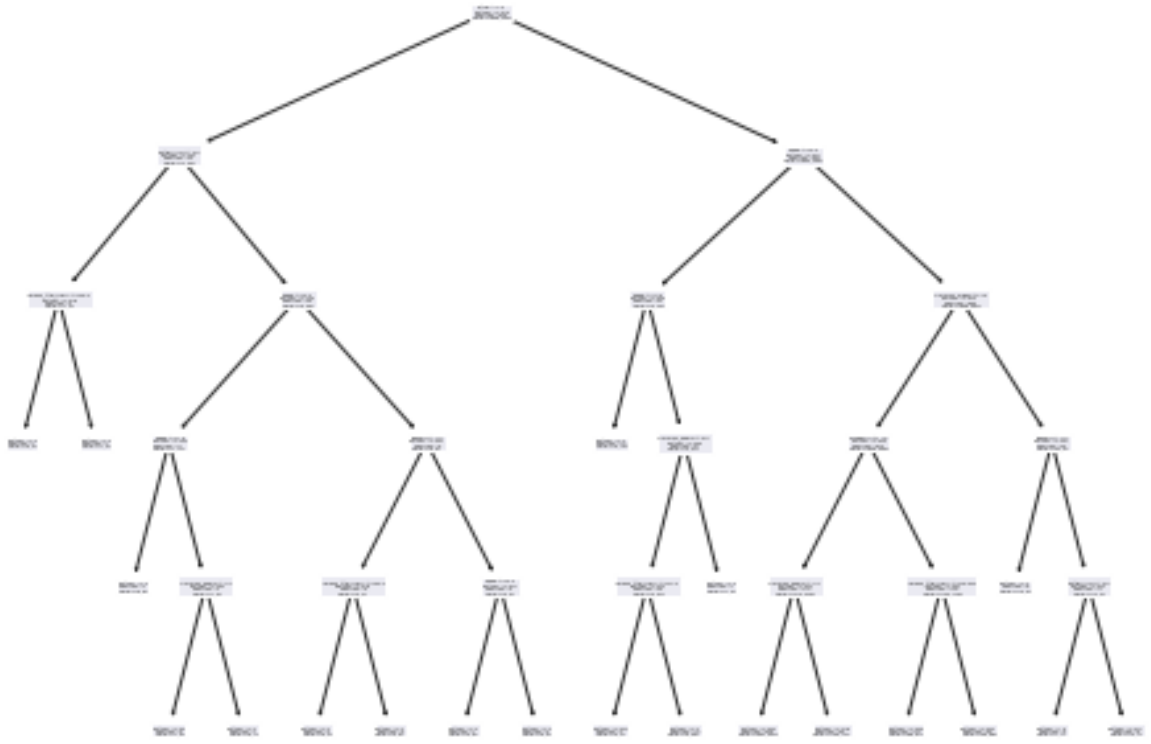
Random Forest.

Ya importadas las librerías y realizada la conexión a la base de datos se corre el modelo random forest con los criterios indicados en la fase de modelado, se entrena el modelo utilizando el método fit con los datos de entrenamiento de “X” y de “y”.

Finalmente se hace una predicción con el método predict y los datos de test. En la figura 54 se muestra el bosque generado.

Figura 54

Random Forest



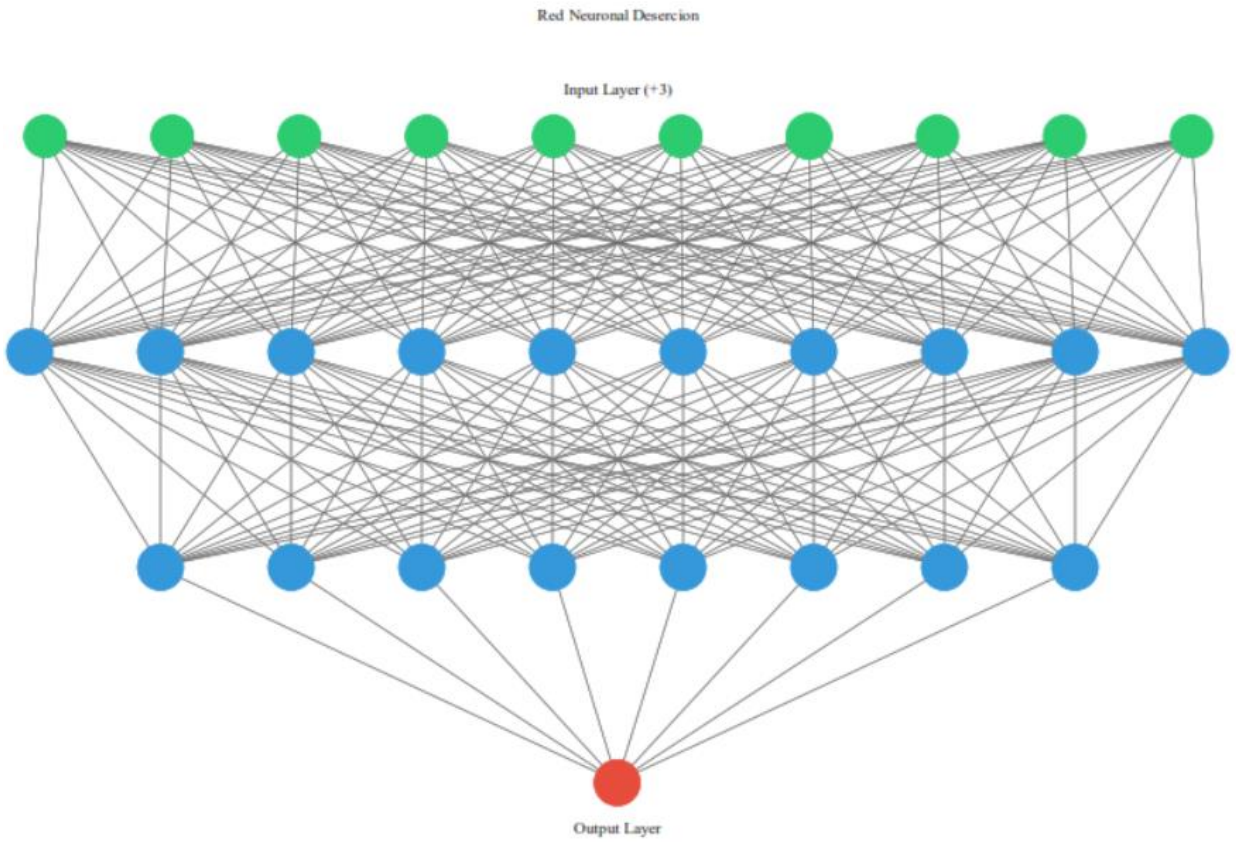
Redes Neuronales.

Para el último modelo ocupamos las librerías ya importadas junto con la conexión a la base de datos, se genera el modelo con los parámetros definidos en la fase de modelado y se

entrena. Finalmente se realiza una predicción con los datos de test y se obtiene la gráfica de la red neuronal. Ver figura 55.

Figura 55

Redes Neuronales



Evaluar los modelos

Modelos evaluados

- Modelo de Árboles de decisión

Para realizar la validación de cualquier modelo se utiliza la matriz de confusión con datos reales, para ello en la sección de librerías se importó confusión_matrix de la librería sklearn.

Figura 56

Validación

```
##  Realiza la predicción
y_pren = dt.predict(X_test)

# Graficar la matriz de confusion
cm = confusion_matrix(y_test, y_pren)
cm_norm = cm/cm.sum(axis=1)[:, np.newaxis]
plt.figure()
plot_confusion_matrix(cm_norm, classes=dt.classes_, title='Matriz de Confusion')
print('Matriz de Confusión')
print(cm)
X_test.shape
```

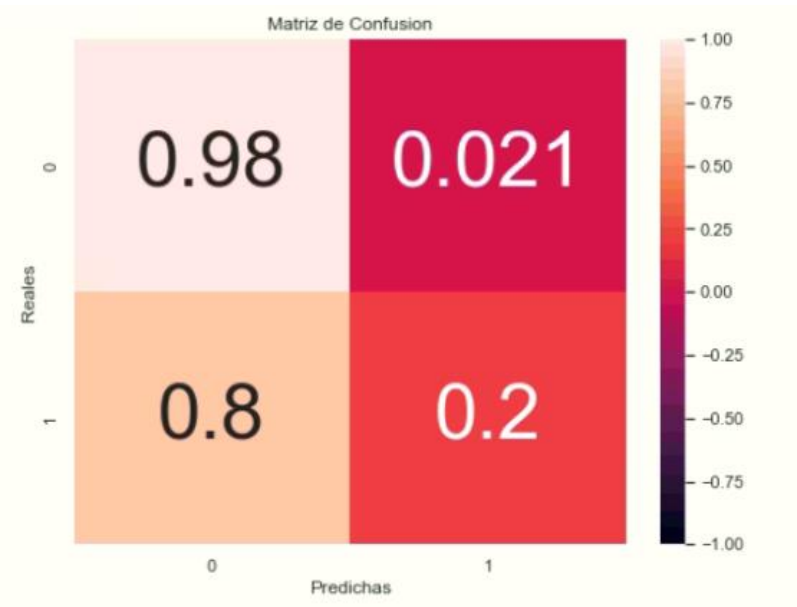
```
Matriz de Confusión
[[319   7]
 [126  32]]

(484, 13)
```

A continuación, en la Figura 57, se muestra los resultados arrojados de la matriz de confusión:

Figura 57

Matriz de confusión – Árboles de decisión



351 registros fueron clasificados correctamente y 133 fueron clasificados erróneamente, además se puede destacar lo siguiente:

319 alumnos que representan el 98% fueron clasificados de forma correcta que no desertan.

32 alumnos que representan el 20% fueron clasificados de forma correcta como desertores.

126 alumnos que representan el 80% no desertan, pero fueron clasificados erróneamente como alumnos que desertan.

7 alumnos que representan el 2.1% desertan, pero fueron clasificados erróneamente como no desertores.

Seguidamente se calculan las métricas de precisión.

Exactitud:

$$AC = \frac{VP + VN}{VP + FP + FN + VN}$$

$$AC = \frac{319 + 32}{319 + 126 + 7 + 32} = 0.7252$$

Precisión:

$$P = \frac{VP}{VP + FP}$$

$$P = \frac{319}{319 + 126} = 0.7169$$

Sensibilidad:

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

$$TP = \frac{319}{319 + 7} = 0.9785$$

Especificidad:

$$TN = \frac{VN}{VN + FP}$$

$$TN = \frac{32}{32 + 126} = 0.2025$$

Tasa de error:

$$Tasa\ de\ error = \frac{FP + FN}{FP + FN + VP + VN}$$

$$Tasa\ de\ error = \frac{126 + 7}{126 + 7 + 319 + 32} = 0.2748$$

F1 Score:

$$F1\ Score = \frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad}$$

$$F1\ Score = \frac{2 * 0.7169 * 0.9785}{0.7169 + 0.9785} = 0.8274$$

Los resultados arrojados en la matriz de error son óptimos con un 72.52% de exactitud y una tasa de error 27.48%.

La métrica F1 score es del 82.74% lo que se considera aceptable.

- Modelo Random Forest

Se realiza la predicción, y se ejecuta la matriz de confusión.

Figura 58

Ejecución de matriz de confusión

```
##  Prediccion
prediction_test = rf.predict(X=X_test)

# Matriz de Confusión
cm_rf = confusion_matrix(y_test, prediction_test)
cm_norm_rf = cm_rf/cm_rf.sum(axis=1)[:, np.newaxis]
plt.figure()
plot_confusion_matrix(cm_norm_rf, classes=rf.classes_, title='Matriz de Confusión')
print('Matriz de confusion')
print(cm_rf)
X_test.shape
```

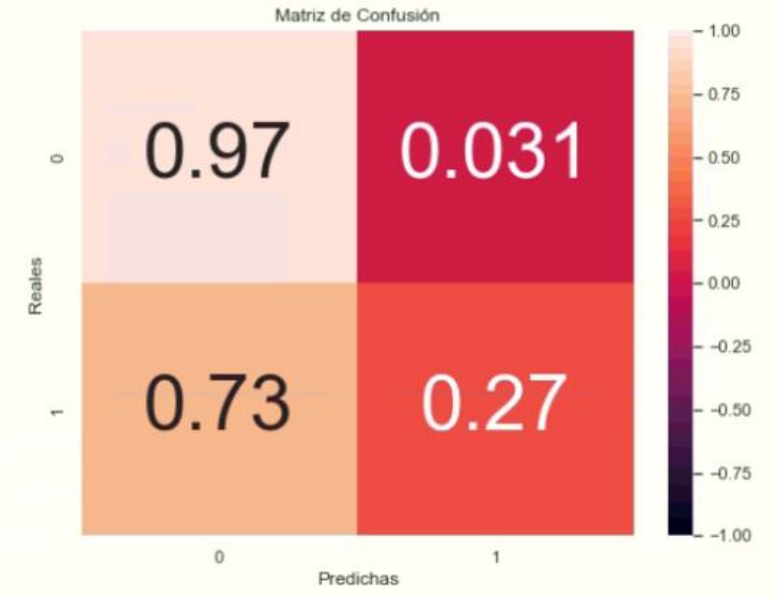
```
Matriz de confusion
[[316  10]
 [116  42]]

(484, 13)
```

A continuación, en la Figura 59, se muestra los resultados arrojados de la matriz de confusión del modelo:

Figura 59

Matriz de Confusión – Random Forest



358 registros fueron clasificados correctamente y 126 fueron clasificados erróneamente, además se puede destacar lo siguiente:

316 alumnos que representan el 97% fueron clasificados de forma correcta que no desertan.

42 alumnos que representan el 27% fueron clasificados de forma correcta como desertores.

116 alumnos que representan el 73% no desertan, pero fueron clasificados erróneamente como alumnos que desertan.

10 alumnos que representan el 3.1% desertan, pero fueron clasificados erróneamente como no desertores.

Seguidamente se calculan las métricas de precisión.

Exactitud:

$$AC = \frac{VP + VN}{VP + FP + FN + VN}$$

$$AC = \frac{316 + 42}{316 + 116 + 10 + 42} = 0.7397$$

Precisión:

$$P = \frac{VP}{VP + FP}$$

$$P = \frac{316}{316 + 116} = 0.7315$$

Sensibilidad:

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

$$TP = \frac{316}{316 + 10} = 0.9693$$

Especificidad:

$$TN = \frac{VN}{VN + FP}$$

$$TN = \frac{42}{42 + 116} = 0.2658$$

Tasa de error:

$$Tasa\ de\ error = \frac{FP + FN}{FP + FN + VP + VN}$$

$$Tasa\ de\ error = \frac{116 + 10}{116 + 10 + 316 + 42} = 0.2603$$

F1 Score:

$$F1\ Score = \frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad}$$

$$F1\ Score = \frac{2 * 0.7315 * 0.9693}{0.7315 + 0.9693} = 0.8337$$

Los resultados arrojados en la matriz de error son óptimos con un 73.97% de exactitud y una tasa de error del 26%.

La métrica F1 score es del 82.74% lo que se considera aceptable.

Redes Neuronales.

Se realiza la predicción, y se ejecuta la matriz de confusión.

Figura 60

Ejecución de matriz de confusión

```
mlp.fit(X_train, y_train)

y_pred_nn = mlp.predict(X_test)

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_nn))
```

```
#matriz de confusion
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix

cm_nn = confusion_matrix(y_test, y_pred_nn)
cm_norm_nn = cm_nn/cm_nn.sum(axis=1)[:, np.newaxis]
plt.figure()
plot_confusion_matrix1(cm_norm_nn, classes=rf.classes_, title='Matriz de Confusión')
print(cm_nn)
X_test.shape
```

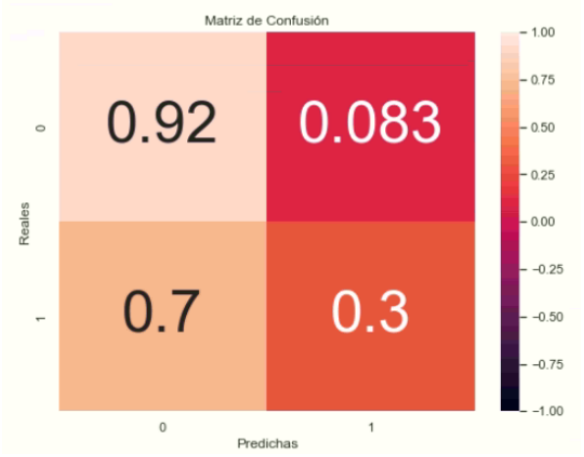
```
[[299  27]
 [110  48]]

(484, 13)
```

A continuación, en la Figura 61, se muestra los resultados arrojados de la matriz de confusión del modelo:

Figura 61

Matriz de confusión – Redes neuronales



347 registros fueron clasificados correctamente y 137 fueron clasificados erróneamente, además se puede destacar lo siguiente:

299 alumnos que representan el 92% fueron clasificados de forma correcta que no desertan.

48 alumnos que representan el 30% fueron clasificados de forma correcta como desertores.

110 alumnos que representan el 70% no desertan, pero fueron clasificados erróneamente como alumnos que desertan.

27 alumnos que representan el 8.3% desertan, pero fueron clasificados erróneamente como no desertores.

Seguidamente se calculan las métricas de precisión.

Exactitud:

$$AC = \frac{VP + VN}{VP + FP + FN + VN}$$

$$AC = \frac{299 + 48}{299 + 110 + 27 + 48} = 0.7169$$

Precisión:

$$P = \frac{VP}{VP + FP}$$

$$P = \frac{299}{299 + 110} = 0.7311$$

Sensibilidad:

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

$$TP = \frac{299}{299 + 27} = 0.9172$$

Especificidad:

$$TN = \frac{VN}{VN + FP}$$

$$TN = \frac{48}{48 + 110} = 0.3028$$

Tasa de error:

$$Tasa\ de\ error = \frac{FP + FN}{FP + FN + VP + VN}$$

$$Tasa\ de\ error = \frac{110 + 27}{110 + 27 + 299 + 48} = 0.2831$$

F1 Score:

$$F1\ Score = \frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad}$$

$$F1\ Score = \frac{2 * 0.7311 * 0.9172}{0.7311 + 0.9172} = 0.8136$$

La matriz de error indica de manera general que el grado de clasificación tiene un 72% de exactitud y una tasa de error o de clasificación incorrecta del 28%. La métrica F1 Score es del 81.36% lo que se considera aceptable. También el modelo indica que clasifica los casos positivos con una probabilidad del 91.72% y los casos negativos con una probabilidad del 30.28%.

Tabla 13

Fase 5. Evaluación

Comprensión del negocio	Comprensión de la data	Preparación de la data	Modelado	Evaluación	Despliegue y explotación
				Evaluar los resultados *-Evaluación de los resultados de acuerdo con los objetivos del negocio* *-Modelos aprobados* **Revisar el proceso** *-Informe de la revisión del proceso* **Determinar los siguientes pasos** *-Listado de acciones posibles* *-Decisión razonada de cómo proceder*	

Nota. Tomado De “Data Mining. The Crisp-Dm Methodology. The Clem Language And Ibm Spss Modeler (Pág. 16) Por César” Pérez López.

Evaluar los resultados

Evaluación de resultados

En el presente trabajo de investigación se desarrollaron tres modelos de clasificación: Árboles de decisión, Random Forest y Redes Neuronales.

En la tabla 14 se muestran los resultados de las métricas aplicadas con los datos de prueba para cada modelo.

Tabla 14

Resultados de las métricas

Modelo/Métrica	Exactitud	Precisión	Sensibilidad	Especificidad	Tasa de Error	Vp	Vn	Fp	Fn	Roc	F1 Score
ML1. Árboles de decisión	0.7252	0.7169	0.9785	0.2025	0.2748	319	32	126	7	0.6208	0.8274
ML2. Random Forest	0.7397	0.7315	0.9693	0.2658	0.2603	316	42	116	10	0.6176	0.8337
ML3. Redes Neuronales	0.7169	0.7311	0.9172	0.3038	0.2831	291	39	128	27	0.6104	0.8136

Premisa: El resultado de las pruebas fueron realizadas con el modelo entrenado y el resultado de las métricas se obtuvo de la data del test.

La métrica de exactitud no funciona bien cuando las clases están desbalanceadas como es en este caso, se tienen 1632 estudiantes que no tienen problemas de deserción que son la mayoría y que representan el 72% de la población de estudiantes. 788 estudiantes son desertores y representan el 35% de la población de estudiantes que desertan en los primeros niveles.

En este caso al estar las clases desbalanceadas, es mucho mejor usar las métricas de precisión (precision) y sensibilidad (recall) que dan una mejor idea de la calidad del modelo.

Análisis del resultado de predicción de los 3 modelos:

- Según el modelo ML2, el 73% de estudiantes no tiene riesgo de deserción. La tasa de error es del 26% según el modelo ML2 y que representa a los estudiantes que desertan.
- El modelo ML1 y ML2 son capaces de identificar en un 97% y 96% a los estudiantes de los primeros niveles que desertarán.
- Al ingresar los datos de las variables de entrada y proceder a ejecutar el modelo en un tiempo corto se obtienen los resultados del alumno si deserta o no.

Modelos aprobados

En el presente trabajo de investigación la métrica que nos ayudará a predecir los estudiantes con riesgo de deserción temprana es la sensibilidad (recall). La métrica de sensibilidad informa sobre la cantidad de estudiantes que el modelo de machine learning es capaz de identificar. La métrica de sensibilidad (recall) es la respuesta a la pregunta ¿qué porcentaje de estudiantes matriculados en los primeros niveles que tienen un riesgo de deserción somos capaces de identificar?

Con una baja precisión y un alto recall, el modelo seleccionado detecta bien la clase, pero también incluye muestras de la otra clase.

Las métricas de precisión (precision) y exactitud (recall) nos dan un 73% de acierto y una tasa de error del 26%, pero al tener datos desbalanceados las métricas de exactitud no son muy confiables. Ante ello la métrica F1 Score es una métrica que resume la precisión y sensibilidad en una sola métrica y es de gran utilidad cuando las clases son desbalanceadas.

Con estos resultados, el modelo Random Forest de Machine Learning da un 96% de sensibilidad y un F1 Score del 83% lo que se considera al modelo como aceptable y se constituye en el modelo ganador para su implementación.

Revisar el proceso

Informe de la revisión del proceso

El proceso de construcción de los tres modelos de machine Learning se realizó conforme a lo planificado. Las fases 1 y 2 de comprensión del negocio y comprensión de la data, fueron las etapas de mayor exigencia y demanda de tiempo, debido sobre todo a la falta de información en campos de ingreso obligatorio como es la fecha de nacimiento.

Las inconsistencias encontradas en la data se deben a la falta de controles de validación en campos de ingreso obligatorio en el sistema académico utilizado por la institución desde sus orígenes. Estas inconsistencias fueron migradas al nuevo sistema informático SIGA de la SENESCYT, por lo que el sistema actual heredó estas inconsistencias, las cuales fueron gestionadas para mejorar la calidad de la data como se indicó en el apartado limpieza de datos de la fase 3 Preparación de la data.

Determinar los siguientes pasos a ejecutar

El siguiente paso es iniciar la ejecución la etapa de implantación.

Tabla 15

Fase 6. Despliegue y explotación

Comprensión del negocio	Comprensión de la data	Preparación de la data	Modelado	Evaluación	Despliegue y explotación
					Planificar despliegue *-Plan de despliegue* **Planificar la monitorización y mantenimiento** **-***Plan de monitorización y mantenimiento* **Redactar el informe final** **-***Informe final* *-Presentación de resultados del proyecto* **Hacer una revisión final de todo el proyecto** *-Documentación de la experiencia adquirida.*

Nota. Tomado De “Data Mining. The Crisp-Dm Methodology. The Clem Language And Ibm Spss Modeler (Pág. 16) Por César” Pérez López.

Planificar despliegue

Plan de despliegue

El proceso de implementación del modelo en el Instituto Superior Tecnológico Cotopaxi se lo realiza en 4 fases.

Fase 1: En la primera fase se exporta o se serializa el modelo entrenado para poder ser utilizado mediante una Api Web que es una interfaz web para páginas y aplicaciones web.

Fase 2: En la segunda fase se desarrolla una API web en el framework Flask que permite la creación de aplicaciones web en Python. Seguidamente el proyecto se lo envía a la nube mediante la plataforma de servicio Heroku. Por medio de esta aplicación los encargados en la Unidad de Bienestar Institucional podrán realizar la predicción de los alumnos matriculados en los primeros niveles del Instituto que podrían estar en riesgo de deserción.

Fase 3: Capacitar a los encargados en la Unidad de Bienestar Institucional sobre el uso de la API web. La Unidad de Bienestar Institucional es la encargada de brindar el apoyo y acompañamiento a los estudiantes que presenten algún tipo de impedimento en cursar normalmente sus estudios en el Instituto.

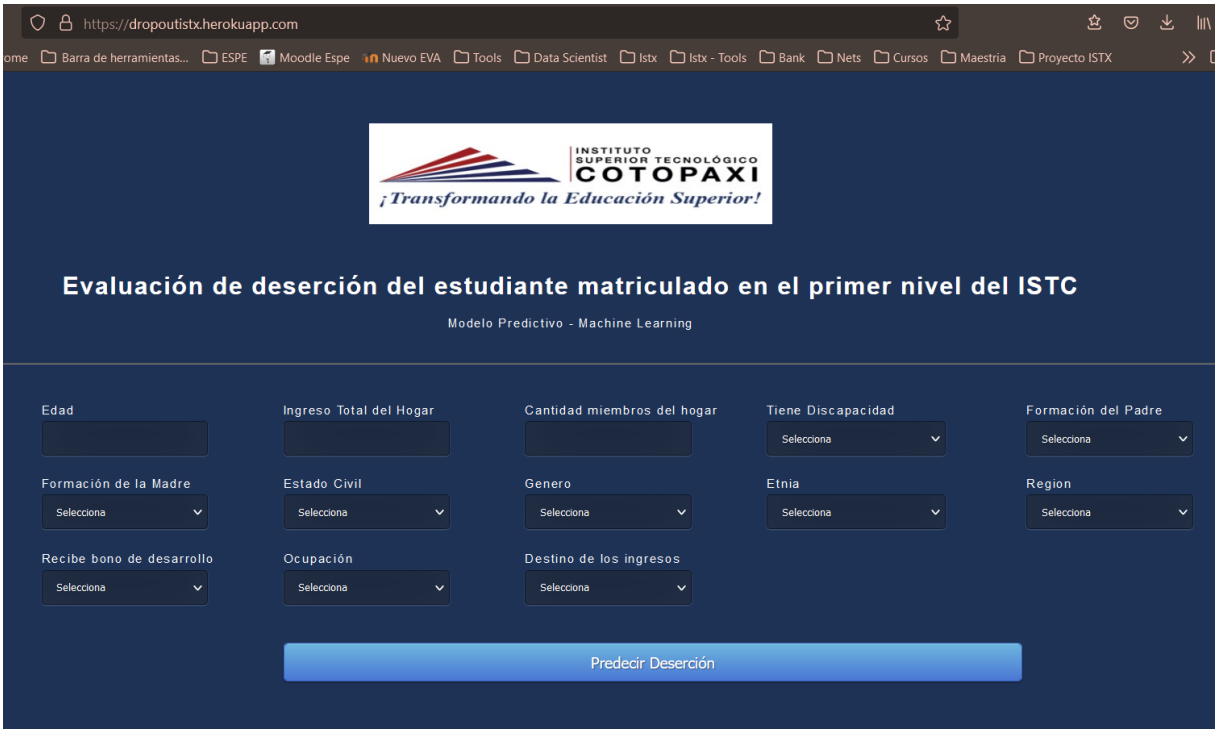
Fase 4: Realizar una entrevista al director de la Unidad de Bienestar Institucional para conocer su nivel de satisfacción al incorporar a su Unidad una aplicación web mediante el uso de Machine Learning.

Resultado del despliegue

Se accede al siguiente link creado https://dropoutistx.herokuapp.com/ en el cual visualizó la siguiente pantalla, ver figura 62

Figura 62

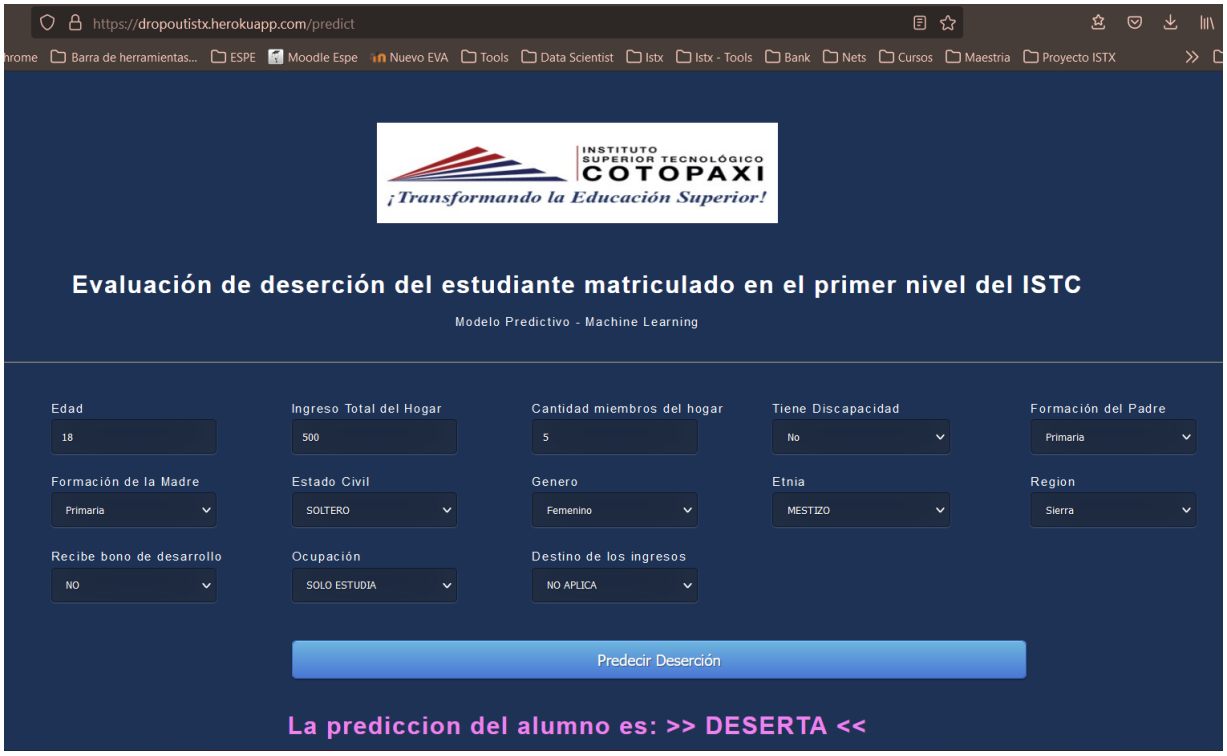
Despliegue



Se realizan pruebas de ingreso de información al API, produciéndose exitosamente la predicción de deserción, ver figura 63

Figura 63

Predicción



Monitorización y mantenimiento

Plan de monitorización y mantenimiento

Lo más importante en una implementación de un modelo de Machine Learning es el de contar con una data de alta calidad que favorezca a su correcto funcionamiento.

Para ello se han definido los siguientes procesos:

- Realizar un análisis exploratorio de datos de manera trimestral por medio de una extracción de data actualizada.

- Generar el modelo con los datos extraídos considerando el (80% de la data para entrenamiento y el 20% de la data para pruebas).
- Exportar y serializar el modelo para ser usado en la API web.
- Generación de una bitácora con las mejoras y actualizaciones que se realice al modelo e irlas versionando.

Informe final

Ver Anexo 9

Presentación de resultados del proyecto

Culminada las etapas de la metodología CRISP-DM, se presenta los resultados obtenidos de este trabajo de investigación al director de la Unidad de Bienestar Institucional, obteniendo su aceptación y satisfacción del trabajo entregado, y la carta de conformidad del trabajo realizado por parte de la máxima autoridad del Instituto Superior Tecnológico Cotopaxi. Ver anexo 8.

Revisión final de todo el proyecto

Documentación de la experiencia adquirida

La metodología CRISP-DM facilita la identificación de incidencias en cada una de sus fases, con lo que se podrá crear una bitácora de hallazgos y documentarlos para consideración en trabajos futuros.

El éxito en todo proyecto de Data Mining siempre se basará en la calidad de la data almacenada en sus variables, así mismo, que el sistema académico cuente con más variables descriptivas o numéricas en relación al alumno creará un espectro más grande de posibilidades para crear más y nuevos patrones de relación entre ellas.

En las primeras fases del proyecto es muy importante contar con la participación activa de personas con experiencia, conocimiento del tema, y que tengan relación directa con los procesos de calidad en el campo educativo institucional.

Capítulo IV

Discusión de resultados

Al tener las clases desbalanceadas, la métrica F1-Score resume la precisión y sensibilidad, siendo el modelo de Random Forest el que genera un mayor score 83.37% en comparación al resto de modelos.

Para verificar el cumplimiento o no de la hipótesis planteada y del planteamiento del problema, se desarrolló una API, la cual permitió al usuario final ingresar los datos de todas las variables que forman parte del proceso de predicción, una vez ingresadas las variables se procesa el modelo entrenado, y éste a su vez muestra como resultado si el estudiante de primer nivel desertará o no a sus estudios.

Para la revisión de los resultados se realizó la capacitación ver Anexo 6 al director de la Unidad de Bienestar Institucional en su puesto de trabajo en donde se le desplegó el API desarrollada. Posterior a ello se realizó la entrevista de los resultados obtenidos ver Anexo 7.

Los resultados de la capacitación y de la entrevista indica que, con esta implementación, se logrará dar atención prioritaria a los estudiantes que presenten riesgo de deserción en una etapa temprana y poder tomar los correctivos necesarios del caso y con ello mejorar los indicadores de deserción en los primeros niveles del Instituto.

Capítulo V

Conclusiones y recomendaciones

Conclusiones

- Se acepta la hipótesis planteada Ho al desarrollar un modelo inteligente de análisis de datos, se genera una predicción de los alumnos con riesgo de deserción en etapas tempranas al iniciar sus estudios en los primeros niveles en el Instituto Superior Tecnológico Cotopaxi con lo cual se da cumplimiento al objetivo general.
- Las variables de causalidad de mayor incidencia identificadas para predecir la deserción estudiantil en los primeros niveles son: edad, destino ingresos, y cantidad de miembros en la familia.
- La información proporcionada por parte de Instituto Superior Tecnológico Cotopaxi sirvió de base fundamental para la construcción del DataSet que cumplió con los estándares requeridos para ser utilizado en los modelos seleccionados para realizar la predicción de deserción.
- Se validaron los tres modelos de clasificación propuestos con sus respectivas métricas, siendo el modelo de bosques aleatorios el que registró un ranking superior en sus métricas respecto a los otros modelos.
- La metodología CRISP-DM por su completitud en todas sus fases apalancó el desarrollo del presente trabajo de investigación, además que facilitó la resolución de los objetivos e hipótesis planteada.
- Las técnicas de clasificación binaria de Machine Learning son ideales para su aplicación en un conjunto de datos y poder predecir posibles eventos futuros.
- La API implementada en el Instituto Superior Tecnológico Cotopaxi para la evaluación de deserción de los estudiantes matriculados en primer nivel contribuye

a detectar los posibles casos de deserción mediante la emisión de una alerta, realizar por parte de la Unidad de Bienestar Institucional un seguimiento, monitoreo e intervención en los casos identificados para prevenir la deserción y mejorar los indicadores de retención estudiantil.

Recomendaciones

- Se recomienda la incorporación de más variables numéricas y categóricas al sistema académico del Instituto, dichas variables se encuentran disponibles en el punto 3 de la guía de registro de Institutos públicos (GRIPP) de la SENESCYT y fijarlas en la base de datos de ingreso obligatorio, con ello se podría actualizar el DataSet en el modelo de predicción.
- Se recomienda validar y depurar el DataSet antes de ser utilizado en un algoritmo de clasificación para evitar alteraciones en los scores de sus métricas de validación.
- En el ciclo de CRISP-DM, se recomienda destinar el mayor tiempo posible a las primeras fases de la metodología para garantizar data de calidad en la implementación de los modelos.
- Se sugiere realizar validaciones periódicas a la base de datos del Instituto con el objetivo de mantener data de calidad para futuros proyectos de Machine Learning.
- Se sugiere que la Unidad de Tics emita reportes de control de los datos del estudiante y enviarlos a la Unidad de Bienestar Institucional o Coordinación Estratégica para su gestión y actualización de la data del estudiante.
- Se recomienda la incorporación del modelo desarrollado al sistema SIGA, para que las predicciones se ejecuten a todo un conjunto de datos en específico y no de manera individual.
- Como trabajos futuros se recomienda trabajar sobre el modelo desarrollado, unificando la información del estudiante con sus calificaciones y asistencias al terminar el primer parcial de un periodo académico.

Bibliografía

Amaya Torrado, Y. K., Barrientos Avendaño, E., & Heredia Vizcaíno, D. J. (2017). Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. 14.

American, S. (s.f.). "How the Computer Beat the Go Master" .

Banker, A. (s.f.). "How PayPal Is Taking a Chance on AI to Fight Fraud" .

Bradley, S., & Migali, G. (2018). *The effects of the 2006 tuition fee reform and the Great Recession.*

Burgos, C., Campanario, M., de la Peña, D., Lara, J., Lizcano, D., & Martínez, M. (2017). *Data mining for modeling students' performance: A tutoring.*

CACES. (2020). MODELO DE EVALUACIÓN INSTITUCIONAL PARA LOS INSTITUTOS SUPERIORES TÉCNICOS Y TECNOLÓGICOS EN PROCESO DE ACREDITACIÓN 2020. www.caces.gob.ec.

Calebrese, J., Esponda, S., Pasini, A., Boracchia, M., & Pesado, P. (2019). *Guía para evaluar calidad de datos basada en .*

Casanova, J., Cervero, A., Núñez, J., Almeida, L., & Bernardo, A. (2018). *Factors that determine the persistence and dropout of university.*

Chopra, R. (s.f.). Advanced Computer Architecture.

Chung, J., & Lee, S. (2018). *Dropout early warning systems for high school students using.*

Corso Salazar, C. (s.f.). *Deserción escolar*. Obtenido de https://www.uaeh.edu.mx/scige/boletin/prepa3/n8/p1.html#refe1

CUJI, B., GAVILANES, W., & SANCHEZ, R. (2017). Modelo predictivo de deserción estudiantil basado. *ESPACIOS*, 9.

Department of Engineering: Computer Science, S. U. (The 1940's to the 1970's). *Neural Networks History: The 1940's to the 1970's*.

Department of Psychology, U. o. (s.f.). "Artificial Neural Networks Technology".

Díaz Peralta, C. (2008). *Modelo Conceptual para la deserción estudiantil universitaria Chilena.* Concepción.

Dursun, D. (2020). *A comparative analysis of machine learning techniques for student retention management.*

ECURED. (s.f.). *ECURED*. Obtenido de https://www.ecured.cu/Proceso_Docente_Educativo

Fallb, A.-M., Vaughna, M., Robertsb, G., Kremerc, K., & Martinezb, L. (2020). *Preliminary validation of the dropout risk inventory for middle and high.*

Fernandez, X., & Silva, E. (2014). *Deserción estudiantil universitaria en el primer semestre. El caso de una.*

Forbes. (s.f.). "10 Ways Machine Learning Impacts Customer Experience".

Forbes. (s.f.). "A Short History of Machine Learning -- Every Manager Should Read".

Forbes. (s.f.). "Six Novel Machine Learning Applications" .

Fortune. (s.f.). "Can Artificial Intelligence Silence Internet Trolls?" .

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and Tensor Flow.* Sebastopol, CA 95472: O'REILLY.

Gizmodo. (s.f.). "Google's Artificial Brain Loves to Watch Cat Videos" .

Graduate School of Stanford Business, S. U. (s.f.). "Predictive Data Can Reduce Emergency Room Wait Times" .

Guzmán Castillo, S., Korner, F., Pantoja-García, J., Nieto-Ramos, L., Gómez-Charris, Y., Castro-Sarmiento, A., & Romero-Conrado, A. (2020). *Implementation of a Predictive Information System for University Dropout Prevention.*

HASAN, M. (2015). *PREDICTING STUDENT PERFORMANCE TO REDUCE DROPOUT USING J48 DECISION TREE ALGORITHM.*

Home of the Oncology, J. (s.f.). Computer Technology Helps Radiologists Spot Overlooked Small Breast Cancers Cancer Network.

Insider, B. (s.f.). "This algorithm could help predict ISIS' future moves" .

Isabel S. (2020). Cómo el Machine Learning mejora la retención estudiantil.

Kubrick. (s.f.). The New Yorker.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling.* New York: Springer.

Kurenkov, A. (s.f.). "A 'Brief' History of Neural Nets and Deep Learning, Part 4".

Lawton, G., & Carew, J. (s.f.). *Predictive modeling*. Obtenido de https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling

Lyche, C. (9 de 11 de 2010). *Taking on the Completion Challenge: A literature Review on Policies to Prevent Dopout and Early School Leaving*. Obtenido de http://dx.doi.org/10.1787/5km4m2t59cmr-en

Marvin Minsky, p. i. (s.f.). The Tech, Massachusetts Institute of Technology.

Merlino, A., Ayllón, S., & Escanés, G. (2013). *VARIABLES QUE INFLUYEN EN LA DESERCIÓN DE ESTUDIANTES.*

Muller, A., & Guido, S. (2017). *Introduction to Machine Learning with Python.* 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'REILLY.

NARVÁEZ BARROS, M. A., & BARRAGÁN REYES, G. E. (2015). ANÁLISIS SOBRE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD POLITÉCNICA SALESIANA, SEDE GUAYAQUIL: CASO DE LAS CARRERAS DE ADMINISTRACIÓN DE EMPRESAS Y CONTABILIDAD Y AUDITORÍA PERIODO DE APLICACIÓN 2007 - 2012. Guayaquil, Guayas, Ecuador.

Networks, L. A. (s.f.). "Perceptron and Adaline" .

Nikolovski, V., Stojanov, R., Chorbev, I., & Madjarov, G. (2015). *Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education.*

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). *Uplift Modeling for preventing student dropout in higher education.*

Pacheco. (2015). *La primera evaluación de la universidad ecuatoriana.* Quito: Consejo de Educación Superior.

Pitzalis, L. (23 de 04 de 2020). *Machine Learning para reducir el abando escolar*. Obtenido de https://www.linkedin.com/pulse/machine-laerning-para-reducir-el-abando-escolar-lidia-pitzalis/?originalSubdomain=es

Ramirez, R. (2016). *UNIVERSIDAD URGENTE para una sociedad emancipada.* Quito: SENESCYT-IESALC.

raona. (2017). *Los 10 Algoritmos esenciales en Machine Learning*. Obtenido de https://www.raona.com/los-10-algoritmos-esenciales-machine-learning/

Review, T. (s.f.). "AI Has Beaten Humans at Lip-reading".

Rodríguez Montequín, M., Álvarez Cabal, J., Mesa Fernández, J., & González Valdés, A. (SA). METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE DATA MINING.

Runner, B. (s.f.). IMDb.

Samuel, A. (s.f.). *Wikipedia*.

Samuel, A. (From 1949 through the late 1960s). Pioneer in Machine Learning, Stanford University Infolab.

Samuel, A. L. (s.f.). IEEE Computer Society.

Samuel, A. (s.f.). Pioneer Researcher In Computer Science The New York Times.

Seeker. (s.f.). Artificially Intelligent Investors Rack Up Massive Returns in Stock Market Study.

Shalev-Shwartz , S. (2014). *Understanding Machine Learnig.* Cambridge University Press.

Spady. (1971). Dropouts from higher education: toward an empirical model. Interchance 2, 38-62. http://dx.doi. org/10.1007/BF02282469 .

Stanford, U. (s.f.). Stanford Cart.

Time. (s.f.). "Interview with Eugene Goostman, the Fake Kid Who Passed the Turing Test".

Times, T. N. (s.f.). "Computer Wins on ‘Jeopardy!’: Trivial, It’s Not" .

Times, T. N. (s.f.). Learning, Then Talking.

Vásquez, J. (2016). Modelo predictivo para estimar la deserción de estudiantes en una Institución de Educación Superior. 156.

Vázquez, S., Biggio, M., & García, S. (2013). *Relaciones entre rendimiento académico, competencia.*

VentureBeat. (s.f.). "The North Face to launch insanely smart Watson-powered mobile shopping app next month" .

Viteri Castro, D., & Uquillas Narváez, M. (2011). *Estudio sobre la deserción estudiantil en la Pontificia Universidad Católica del Ecuador - Matriz, en los niveles 1ro, 2do, y 3ro de*

todas las Facultades y Escuelas del primer semestre del año académico 2007-2008. Quito.

Week, I. (s.f.). "11 Cool Ways to Use Machine Learning" .

Wikipedia. (s.f.). "Geoffrey Hinton" .

Wikipedia. (s.f.). Deep Blue (chess computer).

Wired. (s.f.). "Bellkor’s Pragmatic Chaos Wins $1 Million Netflix Prize by Mere Minutes".

Wired. (s.f.). "Google’s Artificial Brain Learns to Find Cat Videos" .

Wired. (s.f.). "How the Netflix Prize Was Won".

Wired. (s.f.). "Meet the Man Google Hired to Make AI a Reality".

Wired. (s.f.). "Now Anyone Can Deploy Google's Troll-Fighting AI" .

Wired. (s.f.). Autonomous Cars Through the Ages.

Anexos