

Universidad Cenfotec

Maestría Profesional en Tecnologías de Bases de Datos

Documento final de proyecto de Investigación Aplicada 2

Evaluación de la aplicabilidad de algoritmos predictivos orientados a potenciar el rendimiento académico del estudiantado que ingresa cada año a la Carrera de Informática de la Sede del Atlántico de la Universidad de Costa Rica, con el fin de que sirvan como base para una propuesta institucional de mejora académica mediante los resultados

Elaborado por:

Rolando de Jesús Herrera Sánchez

Diciembre 2021

Declaratoria de derechos de autor

Declaro que el presente proyecto de investigación fue realizado por mi persona Rolando de Jesús Herrera Sánchez, trabajando y fundamentando cada uno de los capítulos establecidos por la Universidad Cenfotec con diferentes fuentes bibliográficas, literatura extraída de fuentes confiables en bases de datos suscritas y respetando los derechos de autor correspondientes.

Se autoriza la reproducción parcial o total de este trabajo posterior al periodo de confidencialidad por tema de datos sensibles que entregó la Universidad de Costa Rica, para ser usado como referencia de trabajos futuros de tipo académico y científico, con la salvedad de siempre respetar e incorporar la referencia respectiva para efectos de derechos de autor.

Agradecimientos

Quiero agradecer a todas las personas que me han permitido llegar a esta etapa final de la maestría, en la cual he logrado adquirir el conocimiento necesario para concluir, de manera exitosa, esta investigación:

Agradezco a cada una de las personas docentes que me han dotado de conocimiento nuevo en cada uno de los cursos que llevé durante este largo recorrido. Especialmente al profesor Luis Naranjo, por todo el conocimiento, guía, orientación y recomendaciones dados en el curso de metodología de la investigación, el cual fue clave para que hoy esté finalizando con éxito este trabajo.

Al profesor Juan Zamora Mora y al profesor Diego Alfaro Bergueiro, los cuales me permitieron iniciar en el increíble mundo de la minería de datos y la inteligencia artificial con el curso de minería de datos del plan de estudios de la maestría y los cursos de aprendizaje automático y profundo de la especialidad en IA. Sin ellos hoy probablemente estaría desarrollando otro tema en otra de las áreas de la tecnología de datos.

A todo el personal de la Universidad Cenfotec por la ayuda brindada al expresar, de manera clara, todos los lineamientos necesarios para concluir con éxito esta investigación.

También quiero agradecer a la Universidad de Costa Rica por permitirme llevar a cabo la investigación en esta gran casa de enseñanza y proveerme de los conjuntos de datos necesarios para el desarrollo de los algoritmos que se implementaron.

Finalmente, quiero agradecer a mi familia por todo el apoyo que me han dado durante estos años en el desarrollo de la maestría y por motivarme a finalizarla de manera satisfactoria.

TRIBUNAL EXAMINADOR

Este proyecto fue aprobado por el Tribunal Examinador de la carrera: **Maestría en Tecnología de Bases de Datos,** requisito para optar por el título de grado de **Maestría,** para el estudiante: **Herrera Sánchez Rolando.**

DIEGO ALONSO ALFARO BERGUEIRO (FIRMA)
Digitally signed by DIEGO ALONSO ALFARO BERGUEIRO (FIRMA)
Date: 2022.03.16 18:11:41 -06'00'

MBD. Diego Alfaro Bergueiro
Tutor

JOSE ALBERTO CABEZAS JAIKEL (FIRMA)
Digitally signed by JOSE ALBERTO CABEZAS JAIKEL (FIRMA)
Date: 2022.03.17 16:07:19 -06'00'

MBD. José Cabezas Jaikel
Lector 1

IGNACIO TREJOS ZELAYA (FIRMA)
Firmado digitalmente por IGNACIO TREJOS ZELAYA (FIRMA)
Fecha: 2022.03.18 10:11:57 -06'00'

M. Sc. Ignacio Trejos Zelaya
Lector 2

San José, Costa Rica, 15 de marzo de 2022

Firmada digitalmente, de conformidad con la Ley de Certificados, Firmas Digitales y Documentos Electrónicos N° 8454, destacando el artículo 9°-

Carta de aprobación del filólogo

Cartago, 30 de marzo de 2022

Los suscritos, Elena Redondo Camacho, mayor, casada, filóloga, incorporada a la Asociación Costarricense de Filólogos con el número de carné 0247, portadora de la cédula de identidad número 3-0447-0799 y, Daniel González Monge, mayor, casado, filólogo, incorporado a la Asociación Costarricense de Filólogos con el número de carné 0245, portador de la cédula de identidad número 1-1345-0416, ambos vecinos de Quebradilla de Cartago, revisamos el trabajo final de graduación que se titula: *Evaluación de la aplicabilidad de algoritmos predictivos orientados a potenciar el rendimiento académico del estudiantado que ingresa cada año a la Carrera de Informática de la Sede del Atlántico de la Universidad de Costa Rica, con el fin de que sirvan como base para una propuesta institucional de mejora académica mediante los resultados*, sustentado por Rolando de Jesús Herrera Sánchez.

Hacemos constar que se corrigieron aspectos de ortografía, redacción, estilo y otros vicios del lenguaje que se pudieron trasladar al texto. A pesar de esto, la originalidad y la validez del contenido son responsabilidad directa del autor.

Esperamos que nuestra participación satisfaga los requerimientos de la Universidad Cenfotec.

X ANA ELENA REDONDO CAMACHO (FIRMA)
Firmado digitalmente por ANA ELENA REDONDO CAMACHO (FIRMA)
Fecha: 2022.03.30 12:03:42 -06'00'

Elena Redondo Camacho
Filóloga - Carné ACFIL n.º 0247

X DANIEL ALBERTO GONZALEZ MONGE (FIRMA)
Firmado digitalmente por DANIEL ALBERTO GONZALEZ MONGE (FIRMA)
Fecha: 2022.03.30 12:02:30 -06'00'

Daniel González Monge
Filólogo - Carné ACFIL n.º 0245

Tabla de contenido

Capítulo I. Introducción 1

1.1. Generalidades 1

1.2. Antecedentes del problema 1

1.3. Definición y descripción del problema 1

1.4. Justificación 2

1.5. Viabilidad 3

1.5.1. Punto de vista técnico. 3

1.5.2. Punto de vista operativo 4

1.5.3. Punto de vista económico. 4

1.6. Objetivos 5

1.6.1. Objetivo general. 5

1.6.2. Objetivos específicos. 6

1.7. Alcances y limitaciones 6

1.7.1. Alcances. 6

1.7.2. Limitaciones. 7

1.8. Marco de referencia organizacional y socioeconómico 7

1.8.1. Historia. 8

1.8.2. Tipo de negocio y mercado meta. 8

1.8.3. Misión, visión y valores. 8

1.8.3.1. Misión. 9

1.8.3.2. Visión. 9

1.8.3.3. Valores. 9

1.8.4. Políticas institucionales. 9

1.9. Estado de la cuestión 10

1.9.1. Planificación de la revisión de literatura existente. 10

1.9.1.1. Foco de la pregunta. 10

1.9.1.2. Amplitud y calidad de la pregunta. 10

1.9.1.2.1. Problema. 11

1.9.1.2.2. Pregunta. 11

1.9.1.2.3. Palabras clave. 11

1.9.1.2.4. Población. 12

1.9.1.2.5. Aplicación. 12

1.9.1.3. Selección de fuentes. 12

1.9.1.3.1. Lista de fuentes. 13

1.9.1.3.2. Evaluación de los índices bibliométricos. 14

Capítulo II. Marco conceptual 19

2.1. Conceptos de contenido 20

2.1.1. Definición de educación. 20

2.1.2. Definición de propuesta institucional. 20

2.2. Conceptos técnicos 22
2.2.1. Definición de minería de datos. 22
2.2.2. Definición de minería de datos educacional. 23
2.2.3. Definición de algoritmo 23
2.2.4. Definición de aprendizaje automático. 23
2.2.4.1. Aprendizaje supervisado 24
2.2.4.1.1. Algoritmos de regresión. 24
2.2.4.1.2. Algoritmos de clasificación. 25
2.3. Regresión logística 26
2.4. Árboles de decisión 29
Capítulo III. Marco metodológico 32
3.1. Tipo de investigación 32
3.2. Alcance investigativo 32
3.3. Enfoque 33
3.4. Diseño 36
3.5. Población y muestreo 38
3.6. Instrumentos de recolección de datos 39
3.7. Técnicas de análisis de información 40
3.8. Estrategia de desarrollo de la propuesta 41
Capítulo IV. Análisis del diagnóstico 43

4.1.1. Identificación de variables que permiten alimentar los algoritmos predictivos que se desarrollan. 44

4.2. Metodología CRISP-DM 46

Capítulo V. Propuesta de solución 51

Capítulo VI. Conclusiones y recomendaciones 60

6.1. Conclusiones 60

6.2. Recomendaciones 62

Referencias 65

Lista de figuras

Figura 1 H-Index..... 14

Figura 2 Cuartiles..... 14

Figura 3 SJR y Citas por documento..... 15

Figura 4 H-Index..... 15

Figura 5 Cuartiles..... 15

Figura 6 SJR y citas por documento 16

Figura 7 H-Index..... 16

Figura 8 SJR y citas por documento 16

Figura 9 Cantidad de veces citado según Google Scholar..... 17

Figura 10 H-Index..... 17

Figura 11 Cantidad de veces citado según Google Scholar..... 18

Figura 12 Cantidad de veces citado según Google Scholar..... 18

Figura 13 Nube de palabras..... 19

Figura 14 Diagrama de Conceptos de contenido 20

Figura 15 Diagrama de conceptos técnicos 22

Figura 16 Función sigmoide usada en el algoritmo de regresión logística 27

Figura 17 Ecuación del modelo de regresión logística..... 27

Figura 18 Flores setosas, versicolor o virginica..... 28

Figura 19 Probabilidad estimada de que una flor sea iris-virginica 28

Figura 20 Límite de decisión lineal..... 29

Figura 21 Árbol de decisión para detectar Spam en correos electrónicos 30

Figura 22 Ontología 34

Figura 23 Diagrama de flujo de Técnicas de Análisis de Información 41

Figura 24 Conjuntos de datos provistos con información de los tres recintos de la Sede 45

Figura 25 Scripts para generar los data sets que alimenten los algoritmos 46

Figura 26 Grado de utilización de las distintas metodologías de minería de datos 47

Figura 27 Modelo CRISP-DM 48

Figura 28 Gráfico de graduados en cuatro años según nota de ingreso a carrera 49

Figura 29 Gráfico de graduados en cuatro años según edad de ingreso a carrera (17 años en adelante) 50

Figura 30 Gráfica de una función sigmoide generada en Python con la librería matplotlib 52

Figura 31 Pruebas hechas al modelo de regresión logística solo con la nota de admisión 53

Figura 32 Matriz de correlaciones con los datos de admisión 54

Figura 33 Matriz de correlaciones aplicando one-hot-encoding en la variable provincia 55

Figura 34 Pruebas hechas al modelo de regresión logística multivariable 55

Figura 35 Resultados de las probabilidades hechas al modelo de regresión logística multivariable 56

Figura 36 Extracto de notas obtenidas en los cursos de Matemática Discreta, Cálculo, Programación I y Programación II por algunos estudiantes 57

Figura 37 Extracto de expediente académico de un estudiante con el carné maquillado para efectos de sensibilidad de los datos 58

Figura 38 Matriz de correlaciones usando las notas de Matemática Discreta,

Cálculo, Programación I y Programación Ii 58

Figura 39 Portal de la Universidad de Costa Rica..... 60

Lista de tablas

Tabla 1 Desglose de salarios 4

Tabla 2 Costo del proyecto aproximado según salario de un docente e investigador con una plaza de interino licenciado en la Universidad de Costa Rica 5

Tabla 3 Costo total del proyecto 5

Tabla 4 Listado de palabras 12

Tabla 5 Estudios que se encontraron ACM 13

Tabla 6 Tipos de tareas según aprendizaje 25

Tabla 7 Criterios de evaluación de la precisión de los algoritmos predictivos 36

Tabla 8 Listas de estudiantes admitidos del 2010 al 2021 43

Tabla 9 Listas de cursos matriculados con su nota por estudiante por periodo de matrícula y por año, del 2010 al 2021 44

Tabla 10 Lista de estudiantes graduados del 2010 al 2021 44

Resumen

El aprovechamiento de los datos que se almacenan todos los años en instituciones de educación ha crecido en los últimos años. Cada vez existen más y mejores herramientas tecnológicas que permiten controlar, de mejor manera, el almacenamiento de datos, en este caso información académica y personal de cada estudiante matriculado. Lamentablemente, aunque el aprovechamiento de estos datos es cada vez mayor, no se saca el provecho que debe. Predecir el rendimiento de un estudiante, detectar comportamientos en carreras universitarias (como la cantidad de estudiantes que suelen reprobar un curso y las causas) permite un mejor aprovechamiento de los recursos que se tienen. El objetivo de este trabajo consiste en determinar la aplicabilidad de las técnicas de minería de datos y aprendizaje automático como base de una eventual propuesta institucional que permita mejorar el rendimiento de las personas estudiantes de Informática de la Sede del Atlántico de la Universidad de Costa Rica mediante la implementación de diferentes modelos de algoritmos predictivos. En pleno auge de la minería de datos y el aprendizaje automático, se pretende aprovechar diferentes técnicas para generar modelos que permitan aprovechar los conjuntos de datos de los que ya dispone la Universidad y predecir comportamientos o detectar patrones en los datos que permitan atacar posibles problemas que existan.

Palabras clave: algoritmos, minería de datos educacional, modelos de aprendizaje automático, predicción, rendimiento estudiantil y modelos.

Capítulo I. Introducción

1.1. Generalidades

Los datos por utilizar son datos reales de estudiantes matriculados en la sede del Atlántico y permiten entrenar los modelos para los algoritmos predictivos. Estos se encargarán de predecir si un estudiante finaliza o no en los cuatro años que establece el plan de estudios de la carrera de Informática Empresarial con base en diferentes escenarios.

1.2. Antecedentes del problema

En la actualidad, no se ha realizado ningún trabajo o investigación dentro de la Universidad de Costa Rica, específicamente en la Sede del Atlántico, que mediante algoritmos predictivos permitan incrementar el rendimiento de las personas estudiantes de la carrera de Informática Empresarial, tampoco se dispone de alguna propuesta institucional que busque potenciar el rendimiento de las personas estudiantes con base en datos históricos o comportamientos que se tienen detectados. Por lo tanto, esto significa que no se dispone tampoco de personal dedicado o que se tenga una empresa externa contratada para resolver lo que la propuesta institucional desarrollará.

1.3. Definición y descripción del problema

Las universidades públicas reciben todo el dinero de las arcas del Estado, por lo tanto, dependen de si el Gobierno realiza un buen manejo de las finanzas año con año y en cada turno. Si se observa detalladamente el panorama actual y las constantes dudas de si el FEES es suficiente, es notorio que se necesita un mejor control monetario y adecuarlo para satisfacer las necesidades básicas que la Universidad tiene con un menor presupuesto.

Este recorte presupuestario influye negativamente y en mayor medida en las sedes regionales, lo que provoca que año con año se cierren grupos, cursos y en algunos casos carreras o que no se tengan los recursos mínimos necesarios

para impartir cursos de calidad, por ejemplo, un laboratorio de cómputo con el equipo actualizado. Si se habla específicamente del cierre de grupos, esto se traduce muchas veces en lo que se conoce como *matrícula insatisfecha*, que significa que varios estudiantes no lograron encontrar campo en cursos que necesitan para avanzar en la carrera. Estos cursos generalmente son de dificultad alta y suelen tener un bajo índice de aprobación.

Debido a lo anterior y si se habla del promedio de tiempo que un estudiante tarda en finalizar su plan de estudios (en su mayoría uno, dos, tres o hasta más años de lo que el plan de estudios indica) se puede ver que hay un efecto de *bola de nieve* que suma a los gastos que la Universidad tiene. Esto porque si el estudiante tiene una beca consumirá espacios en cursos que muchas veces ya ha llevado una, dos o hasta tres veces anteriormente, consumo de recursos de biblioteca, equipo, entre otros.

1.4. Justificación

La Universidad actualmente ya pone en marcha ajustes presupuestarios y recorte de gastos innecesarios. Sin embargo, se pueden reducir costos y mejorar procesos actuales desde ciertos puntos que actualmente no se abordan. Este problema financiero al cual se le hace frente en estos momentos se puede solucionar si las personas estudiantes matriculadas en cada carrera tienen rendimientos académicos de calidad, si son apoyadas en todo el transcurso del plan de estudios, si se bajan los índices de matrícula insatisfecha y, finalmente, si se logra que el tiempo de graduación de cada estudiante sea lo más ajustado a lo que ese plan de estudios indica.

Muchas veces, cuando hay rezago de estudiantes en cursos, se tiene que disponer de un presupuesto extraordinario por cada recinto para nombrar un profesor interino o generar matrículas extraordinarias de cursos en el tercer periodo que se conoce como *verano*, esto sucede año tras año en diferentes cursos, lo que al final se traduce en más cargas sociales para la Universidad, pago de aguinaldos, vacaciones y salario escolar para un docente. Por otro lado, las

personas estudiantes que demoran más años y que necesitan becas representan otro rubro importante del gasto que se puede reducir.

Al lograr mejores rendimientos académicos gana la Universidad, ganan las personas estudiantes al convertirse en profesionales y gana el país en términos de más mano de obra disponible y de calidad. La propuesta institucional que se elaborará con las personas estudiantes de Informática busca que se logren mejores rendimientos, detectar falencias en el estudiantado, crear *estudiaderos*, aprovechar las horas asistente y con esto tener más mano de obra disponible para todas estas empresas tecnológicas que están en el país actualmente y que suelen encontrar dificultades para reclutar.

1.5. Viabilidad

Toda empresa, institución u organización actualmente genera una alta cantidad de datos. Si se parte del supuesto de que en pleno 2021 todo se tiene automatizado o que al menos se utiliza la tecnología para optimizar procesos y agilizar tareas, se puede pensar perfectamente en procesar y analizar los datos que una universidad genera, en este caso específico, la Universidad de Costa Rica. Aunque no se tiene actualmente estudios o casos de aplicación de algoritmos predictivos en el ámbito educacional en la Universidad de Costa Rica y mucho menos en la Sede del Atlántico, este tipo de aplicación de algoritmos en educación ya ha sido objeto de estudio por autores como Asif, Merceron, Ali y Haider (2017) Tsiakmaki, Kostopoulos, Kotsiantis y Ragos (2020) y Paul y K. P. (2015) en el ámbito mundial.

1.5.1. Punto de vista técnico. El autor de esta investigación como profesional, docente, estudiante del posgrado en inteligencia artificial y futuro *master* en tecnologías de datos, cuenta con el conocimiento necesario y experiencia para desarrollar los algoritmos predictivos que mejor se ajusten a los datos que se tienen a disposición. Además, se espera tomar todas las recomendaciones que sugieren los autores que se investiguen dentro del estado de la cuestión, esto con el fin de aplicar todo este conocimiento a la propuesta que se pretende entregar a

la Universidad de Costa Rica.

1.5.2. Punto de vista operativo. Esta investigación no altera el funcionamiento normal de la Universidad, ya que solo se aborda una sede regional y el trabajo por llevar a cabo, así como el uso del conjunto de datos, no detienen el proceso de enseñanza, matrículas y cualquier otro proceso o funciones en departamentos que la Universidad tiene en funcionamiento diariamente.

1.5.3. Punto de vista económico. Este proyecto de investigación busca un ganar-ganar en todo aspecto tanto en el ámbito de la universidad, estudiantes y país, el costo teórico que engloba el desarrollo de este proyecto de investigación, en términos de horas de investigación y todo el equipo necesario (*hardware* y *software*) corren por cuenta del autor actual de este proyecto. Se considera el salario real actual de un docente interino con grado académico de Licenciatura en la Universidad de Costa Rica al cual se le asignan horas para un proyecto de investigación, pero que igual mantiene las tareas académicas de los cursos que suele impartir. El pago por hora es de aproximadamente ₡7.500 para desarrollar un proyecto de investigación en una jornada laboral de 40 h semanales y dedicando 5 h a investigación. Sin embargo, para el desarrollo de esta investigación se trabaja más horas por semana y se desarrolla durante ocho meses, los cuales abarcan los dos seminarios de graduación del plan de estudios de la maestría en datos.

Tabla 1

Desglose de salarios

Salario Anual	**Salario Mensual**	**Salario Diario**	**Salario por hora**
₡14.400.000	₡1.200.000	₡60.000	₡7.500

Tabla 2

Costo del proyecto aproximado según salario de un docente e investigador con una plaza de interino licenciado en la Universidad de Costa Rica

	Horas semanales	**Duración TFG Meses**	**Duración TFG Horas**	**Costo estimado del proyecto**
Rolando Herrera Sánchez	15	8	480	₡3.600.000

Tabla 3

Costo total del proyecto

Rubro	**Monto**
Horas investigador	₡3.600.000 ($6,000)
Formación metodológica investigación	₡962.240 ($594 + $958)
Material bibliográfico	0
Viáticos	₡62.000 ($100)
Total:	₡4.624.240 ($7,500)

Nota. Estos costos se sufragan en su totalidad por la persona investigadora.

1.6. Objetivos

Para la definición de los siguientes objetivos se utilizó la taxonomía jerárquica de Bloom, la cual ha logrado demostrar su eficacia a través de los años, además de su amplio uso en el ámbito de investigaciones en Costa Rica.

1.6.1. Objetivo general. Evaluar la aplicabilidad de algoritmos predictivos orientados a potenciar el rendimiento académico del estudiantado que ingresa cada año a la Carrera de Informática de la Sede del Atlántico de la Universidad de Costa Rica, con el fin de que sirvan como base para una propuesta institucional de

mejora académica mediante los resultados.

1.6.2. Objetivos específicos. A continuación se definen los objetivos específicos de este proyecto.

Nivel 1 de Conocimiento.

Identificar el o los procedimientos que tiene la Universidad actualmente para conocer el rendimiento académico y ayudar al estudiantado durante todo el desarrollo del plan de estudios de la carrera.

Nivel 2 de Comprensión.

Justificar los algoritmos predictivos que ayudarán con la propuesta institucional que permitan adecuar modelos con los datos que se tienen disponibles mediante tecnologías de alto uso.

Nivel 3 de Aplicación.

Utilizar diferentes técnicas de algoritmos predictivos que permitan desarrollar la propuesta institucional en beneficio del estudiantado.

Nivel 4 de Análisis.

Descubrir patrones en los datos mediante un análisis de los resultados de los algoritmos empleados que permitan indagar los factores que intervienen en el rendimiento académico del estudiantado durante su carrera Universitaria.

1.7. Alcances y limitaciones

A continuación, se definen los alcances y limitaciones de este proyecto.

1.7.1. Alcances. Los alcances de este proyecto son:

1. Esta investigación entrega los resultados de aplicar algoritmos de aprendizaje automático que sirvan de base para la propuesta de estrategia institucional por aplicar en la Sede del Atlántico de la Universidad de Costa Rica. La investigación se da a conocer en vida estudiantil, en la asociación de estudiantes de la carrera, la dirección de

la sede y las coordinaciones de carrera según lo acordado al entregar los *data-sets* por parte de los coordinadores de carrera.

2. La presente investigación permite predecir si un estudiante finaliza su plan de estudios en los cuatro años establecidos o no, con base en diferentes escenarios.

3. Se implementan algoritmos predictivos que indiquen cómo es el rendimiento de las personas estudiantes por área de estudio, por ejemplo, el área de Matemáticas o el área de programación.

4. Los algoritmos permiten detectar el comportamiento de la población estudiantil en los diferentes cursos del plan de estudios y con esto atacar puntos débiles que se detecten.

1.7.2. Limitaciones. Las limitaciones de este proyecto son:

1. Este trabajo de investigación utiliza datos solo de estudiantes que cursan la carrera de Informática, por lo que la propuesta de estrategia institucional que se entrega no contempla otras carreras. Además, algunos perfiles de estudiantes diferentes a los que se encuentran en una carrera de Ingeniería no están previstos.

2. No se garantiza que la propuesta institucional tenga el mismo resultado en otras sedes y recintos universitarios.

3. Los algoritmos que se emplean para explotar los datos de los que se dispone son exclusivamente de aprendizaje automático, lo que deja por fuera algoritmos de aprendizaje profundo.

1.8. Marco de referencia organizacional y socioeconómico

La Universidad de Costa Rica es una institución autónoma de cultura superior, constituida por una comunidad de profesores, estudiantes y administrativos que se dedican a la enseñanza, investigación y la acción social. Según el marco estratégico de la institución, la Universidad de Costa Rica propicia

el avance del conocimiento en su máxima expresión y busca responder, de manera efectiva, a las necesidades que generan el desarrollo integral de la sociedad. Esto se hace mediante la formación de profesionales que sirvan como difusores y agentes de cambio.

1.8.1. Historia. La Universidad de Costa Rica se fundó en la administración presidencial del Dr. Rafael Ángel Calderón Guardia (1940-1944), cuando su secretario de instrucción pública, el licenciado Luis Demetrio Tinoco Castro corredactó y presentó al congreso el 15 de junio de 1940 el proyecto de ley para la creación de una nueva universidad, el cual se aprobó inmediatamente. Anteriormente, la casa de enseñanza que se había tenido a lo largo de muchos años había sido la Universidad de Santo Tomás.

Fue el 26 de agosto de 1940 cuando el Dr. Calderón Guardia y el Lic. Tinoco Castro firmaron la Ley n.º 362, que con el nombre de Ley Orgánica de la Universidad de Costa Rica expusieron en su primer artículo: “Se crea la Universidad de Costa Rica, una institución docente y de cultura superior que tendrá por misión cultivar las ciencias, las letras y las bellas artes, difundir su conocimiento y preparar para el ejercicio de las profesiones liberales”. Aparte de esto y en calidad de honroso homenaje a la Universidad de Santo Tomás la Universidad de Costa Rica mantuvo como símbolos al escudo, la divisa y la bandera que había ostentado a través de los años.

1.8.2. Tipo de negocio y mercado meta. La Universidad de Costa Rica es una institución que se dedica la formación de profesionales, proyectos de acción social en beneficio de todos los sectores y regiones del país, la investigación, la meditación y la difusión del conocimiento. Según el artículo 3 del estatuto orgánico de la Universidad, esta debe contribuir con las transformaciones que la sociedad necesita para el logro del bien común, mediante una política dirigida a la consecución de una justicia social, equidad, desarrollo integral, libertad plena y total independencia del pueblo.

1.8.3. Misión, visión y valores. A continuación, se enumera la misión, visión y valores de la Universidad de Costa Rica.

1.8.3.1. Misión. La Universidad de Costa Rica es una institución de educación superior y cultura, autónoma constitucionalmente y democrática, constituida por una comunidad de profesores y profesoras, estudiantes y personal administrativo, dedicada a la enseñanza, la investigación, la acción social, el estudio, la meditación, la creación artística y la difusión del conocimiento.

1.8.3.2. Visión. Fortalecer la excelencia académica mediante el desarrollo y el cultivo permanente de una cultura de calidad, con una articulación estrecha entre docencia, acción social e investigación y por medio de la actualización de los planes de estudio en grado y posgrado en todas sus sedes universitarias, la generación de carreras innovadoras, el mejoramiento continuo y la formación de alto nivel del personal académico y administrativo. Todo esto con el fin de atender, de manera pertinente, las necesidades de la sociedad costarricense y potenciar su liderazgo en el desarrollo de la educación nacional.

1.8.3.3. Valores. Los valores de la Universidad son:

1. Derecho a la educación superior.
2. Excelencia académica e igualdad de oportunidades.
3. Libertad de cátedra.
4. Respeto a la diversidad de etnias y culturas.
5. Respeto a las personas y a la libre expresión.
6. Compromiso con el ambiente.
7. Acción universitaria planificada.

1.8.4. Políticas institucionales. Una de las actividades principales de la Universidad de Costa Rica es invertir recurso humano y recursos económicos para el desarrollo de investigación. Por lo tanto, este trabajo final de graduación tiene el apoyo total de la institución e incluso, según los resultados finales, se pueden llevar a cabo ponencias o publicaciones en revistas para demostrar parte de los resultados.

1.9. Estado de la cuestión

En el ámbito internacional se trabaja en el área de la minería de datos educacional (Educational Data Mining) y hay *papers* que explican qué tipos de algoritmos de minería de datos se adecuan para explotar datos que las instituciones suelen poseer y generar con el tiempo. Con esto se detectan patrones que permiten tomar decisiones para ayudar en el rendimiento de las personas estudiantes que recién ingresan o están en sus primeros bloques del plan de estudios. Por ejemplo Adekitan y Noma-Osaghae (2019) en su *paper* titulado *Minería de datos para predecir el rendimiento académico de las personas estudiantes de primer año en una universidad usando los requerimientos de admisión* utilizaron algoritmos de aprendizaje automático (Naive Bayes, K-NN, 1-NN y Árboles de decisión) desarrollados en *R* para demostrar que la excelencia no solo depende de las pruebas cognitivas y pruebas académicas realizadas a las personas estudiantes en la Universidad, sino que depende de muchos otros factores externos.

1.9.1. Planificación de la revisión de literatura existente. En este punto se formula una pregunta clara que permita profundizar y ampliar el tema por desarrollar, se pretende buscar literatura lo más actualizada posible con el objetivo de conocer cuánto se ha abordado la minería de datos educacional en distintas partes del mundo. Se tiene claro que las universidades, y en general cualquier casa de estudio, están en prácticamente todas las ciudades del planeta.

1.9.1.1. Foco de la pregunta. Esta investigación requiere que la literatura encontrada indique los tipos de algoritmos que se emplean generalmente en la minería de datos educacional y qué tipo de resultados son los que suelen obtenerse con los datos que las mismas instituciones poseen en sus históricos y que generan año tras año. Con esto se pueden tomar decisiones para incrementar el rendimiento de la población estudiantil mediante estrategias institucionales.

1.9.1.2. Amplitud y calidad de la pregunta. Se establece la pregunta de investigación que se desea responder y, a la vez, se hace un listado de conceptos

clave para la búsqueda de literatura.

1.9.1.2.1. Problema. En la actualidad, un estudiante promedio en la Universidad de Costa Rica y, en la mayoría de las universidades públicas no finaliza su plan de estudios en el tiempo que se lo indica. Para un Bachillerato Universitario, en el cual se tarda según el plan de estudios cuatro años, el estudiante termina tardando en promedio de entre 5 y 6 años, con casos en los cuales tardan incluso 7 años o más. El impacto que se tiene es de grandes dimensiones y se traduce en dinero que proviene de las arcas del gobierno que se tiene que invertir por estudiante.

Este gasto se incrementa si el estudiante suele mantener una beca por ser de bajos recursos, por vivir en una zona muy alejada, si hay que darle alimentación o transporte o recursos para gastos en libros o equipo de cómputo. Otro de los impactos que se tiene es que al haber mayor demanda en los cursos, muchos estudiantes se quedan sin campo si el cupo es limitado. La presente investigación se enfoca en los estudios que han realizado a nivel de minería de datos educacional que permita incrementar el rendimiento de las personas estudiantes de primer ingreso y que con esto tengan una mejor promoción año tras año y con el menor rezago posible.

1.9.1.2.2. Pregunta. Se formula la siguiente pregunta de investigación:

¿Cuáles trabajos se realizan en el área de la minería de datos educacional que permitan incrementar el rendimiento de las personas estudiantes para tener mejor promoción año tras año y una conclusión del plan de estudios en la menor cantidad de tiempo posible?

1.9.1.2.3. Palabras clave. A continuación, se muestra una serie de palabras clave que se utilizan para la búsqueda de literatura relacionada con la investigación. Estas se muestran en la Tabla 4.

Tabla 4

Listado de palabras

Palabra	**Equivalente en inglés**
Minería	Mining
Datos	Data
Educacional	Educational
Aprendizaje Automático	Machine Learning
Algoritmo	Algorithm
Estudiante	Student
Rendimiento	Performance
Predicción	Prediction
Técnica	Technique
Clasificación	Clasification

1.9.1.2.4. Población. La población de este proyecto corresponde a las personas estudiantes matriculadas en la carrera de Informática de la Sede del Atlántico (Turrialba, Paraíso y Guápiles) de la Universidad de Costa Rica.

1.9.1.2.5. Aplicación. Esta investigación es de utilidad para las coordinaciones de carrera de todas las sedes regionales de la Universidad de Costa Rica y puede servir como modelo para cualquier otra institución o casa de enseñanza en la cual se busque incrementar el rendimiento de las personas estudiantes en las carreras que se imparten.

1.9.1.3. Selección de fuentes. A continuación, se especifican las fuentes de literatura que se utilizan en esta investigación. Para su selección se utilizaron como referencia principal los índices bibliométricos que ofrece Scimagojr, la popularidad del investigador, la universidad de procedencia y la cantidad de veces

que se citó el documento.

1.9.1.3.1. Lista de fuentes. Las búsquedas se realizaron desde Sibdi que es el sistema de bibliotecas, información y documentación digital que posee la Universidad de Costa Rica.

1. Springer.
2. IEEE Xplore.
3. Science Direct.
4. Elsevier.
5. MDPI Applied Science.

Tabla 5

Estudios que se encontraron ACM

#	Titulo	Autores	Año	URL
1	Analyzing Undergraduate Students 177-194	Asif, Raheela Merceron, Agathe Ali, Syed Abbas Haider, Najmi Ghani	2017	http://dx.doi.org/10.1016/j.compedu.2017.05.007
2	Data mining approach to predicting the performance of first year student in a university using the admission requirements	Adekitan, Aderibigbe Israel Noma-Osaghae, Etinosa	2019	https://*link*-springer-com.ezproxy.sibdi.ucr.ac.cr/article/10.1007/s10639-018-9839-7
3	Implementing AutoML in Educational Data	Tsiakmaki, María	2020	https://www.mdpi.com/2076-3417/10/1/90

	Mining for Prediction Tas	Kostopoulos, Georgios Kotsiantis, Sotiris Ragos, Omiros		
4	Modeling Student Performance in Higher Education Using Data Mining	Istanbullu, Ayhan Guruler, Huseyin	2013	https://link.springer.com/book/10.1007%2F978-3-319-02738-8

1.9.1.3.2. Evaluación de los índices bibliométricos. Guiándose con la Tabla 5, se tienen los siguientes estudios:

- Asif, R.; Merceron, A.; Ali, S. A. y Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007.

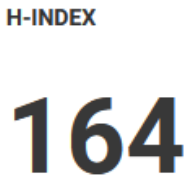


Figura 1

H-Index

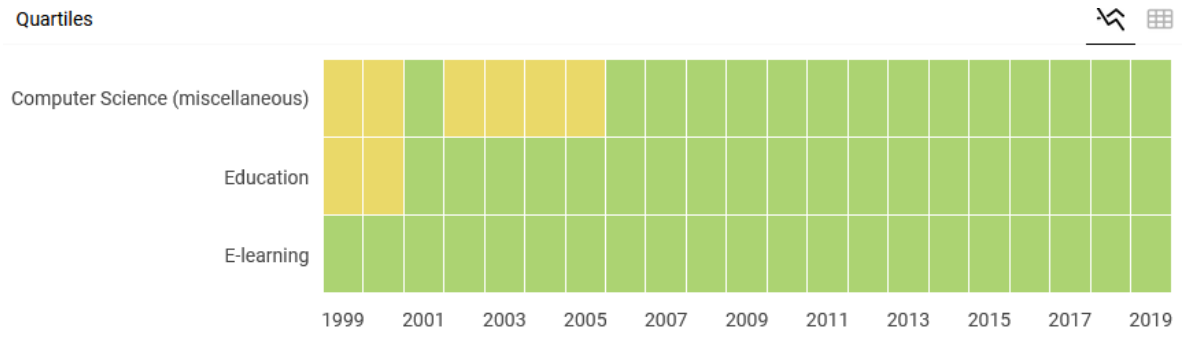


Figura 2

Cuartiles

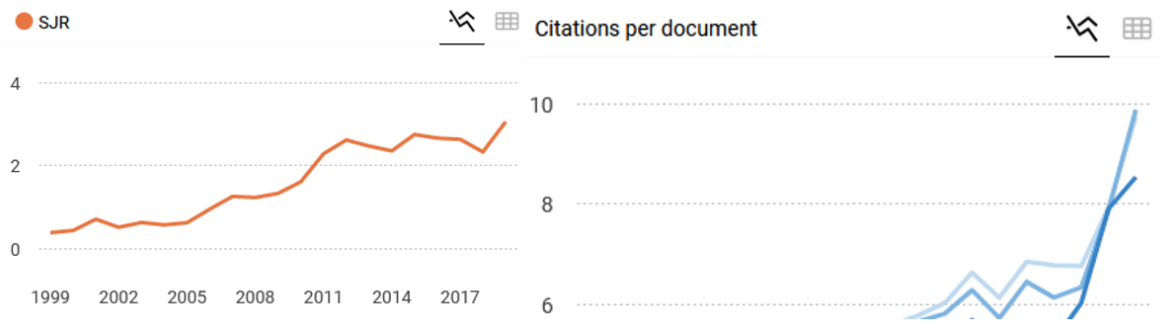


Figura 3

SJR y Citas por documento

- Adekitan, A. I. y Noma-Osaghae, E. (2019). *Data mining approach to predicting the performance of first year student in a university using the admission requirements*. https://doi.org/https://doi-org.ezproxy.sibdí.ucr.ac.cr/10.1007/s10639-018-9839-7.



Figura 4

H-Index

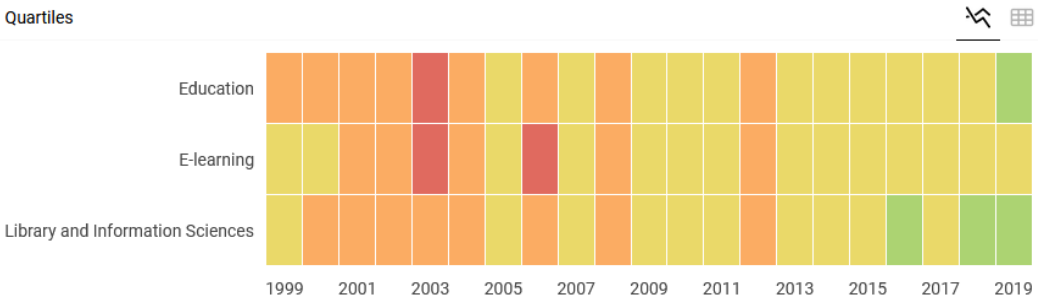


Figura 5

Cuartiles

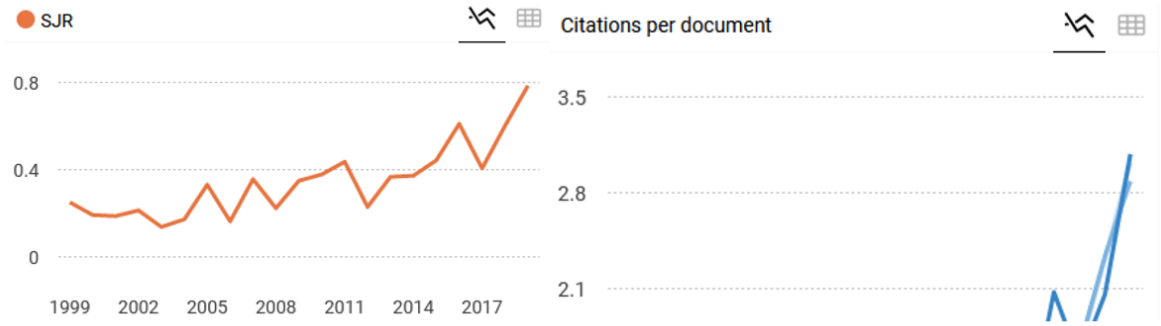


Figura 6

SJR y citas por documento

- Tsiakmaki, M.; Kostopoulos, G.; Kotsiantis, S. y Ragos, O. (2020). Implementing autoML in educational data mining for prediction tasks. *Applied Sciences (Switzerland)*, *10*(1), 1-27. https://doi.org/10.3390/app10010090.



Figura 7

H-Index

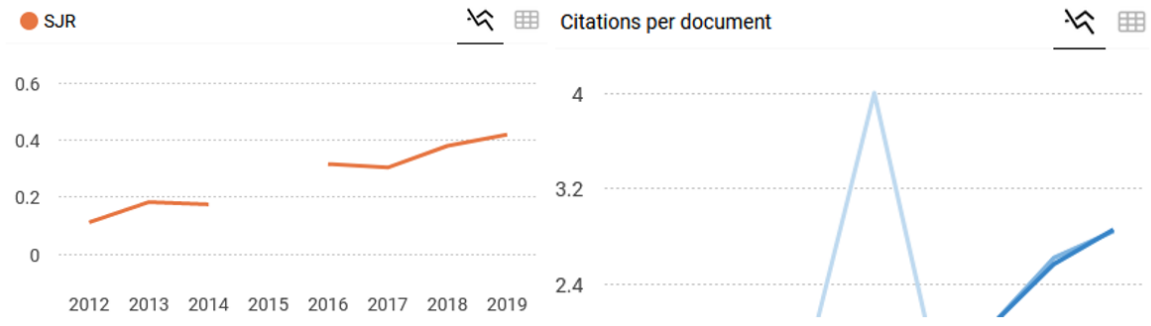


Figura 8

SJR y citas por documento

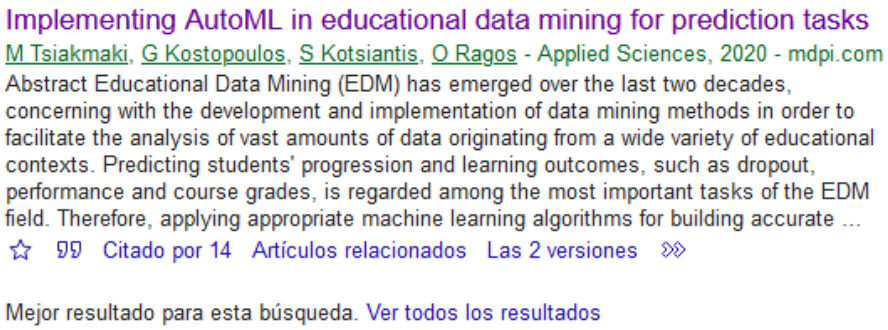


Figura 9

Cantidad de veces citado según Google Scholar

- Istanbullu, A. y Guruler, H. (2014). Educational Data Mining. En A. Peña-Ayala (Ed.), *Educational Data Mining* (pp. 105-124). Springer, Cham. https://doi.org/https://doi-org.ezproxy.sibdi.ucr.ac.cr/10.1007/978-3-319-02738-8_4.

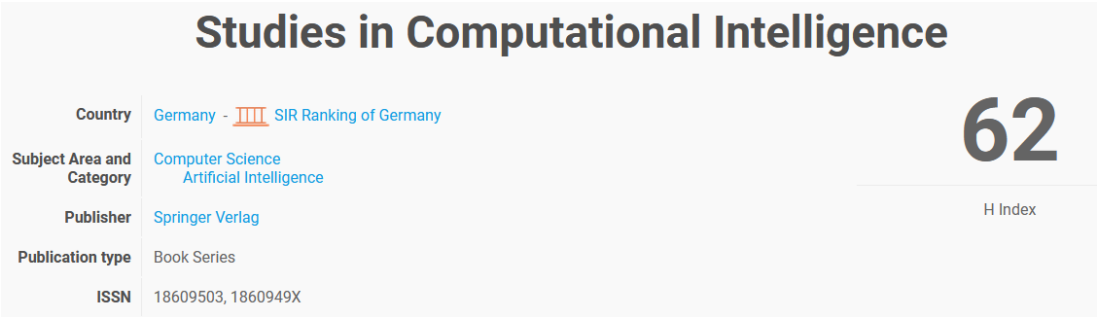


Figura 10

H-Index

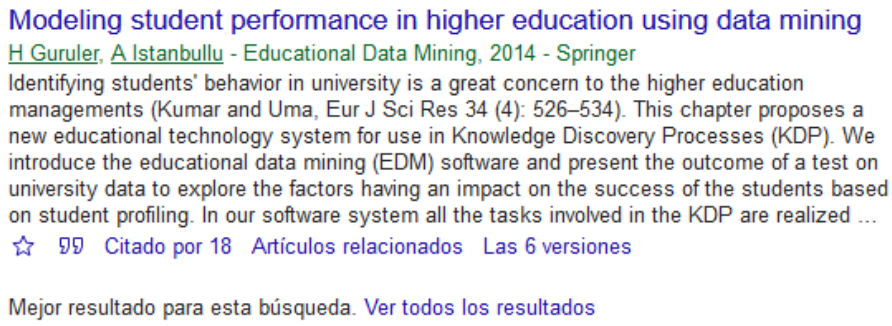


Figura 11

Cantidad de veces citado según Google Scholar

- Devasia, T.; Vinushree, T. P. y Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 91-95. IEEE. https://doi.org/10.1109/SAPIENCE.2016.7684167.

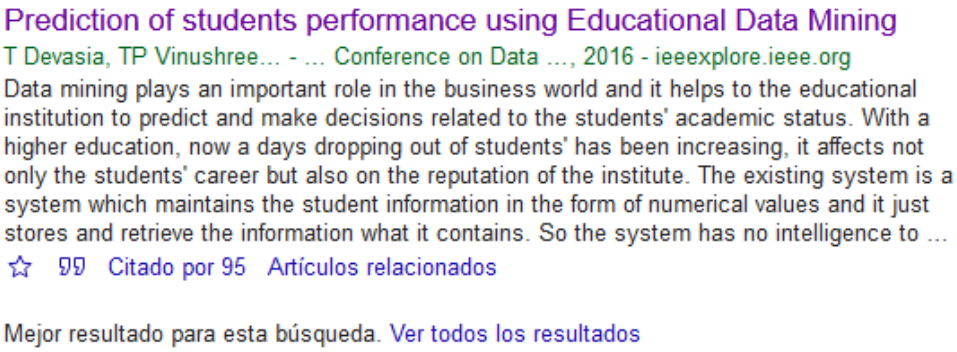


Figura 12

Cantidad de veces citado según Google Scholar

En este caso, a pesar de que los índices bibliométricos del artículo en el Scimagojr no entran en lo esperado, este se citó en 95 ocasiones y su autor está bien posicionado en una universidad de alta investigación.

Capítulo II. Marco conceptual

Se presenta la siguiente nube de conceptos extraídos del estado de la cuestión con la finalidad de identificar cuáles son los conceptos prioritarios en esta investigación.



Figura 13

Nube de palabras

Con base en la nube presentada, los conceptos que se definen van desde lo más general hasta lo específico. De esta manera, es posible aclarar todo lo que se relaciona con el aprendizaje automático o minería de datos del tipo educacional y cómo esto permite dotar de herramientas a las instituciones de enseñanza de un país.

2.1. Conceptos de contenido

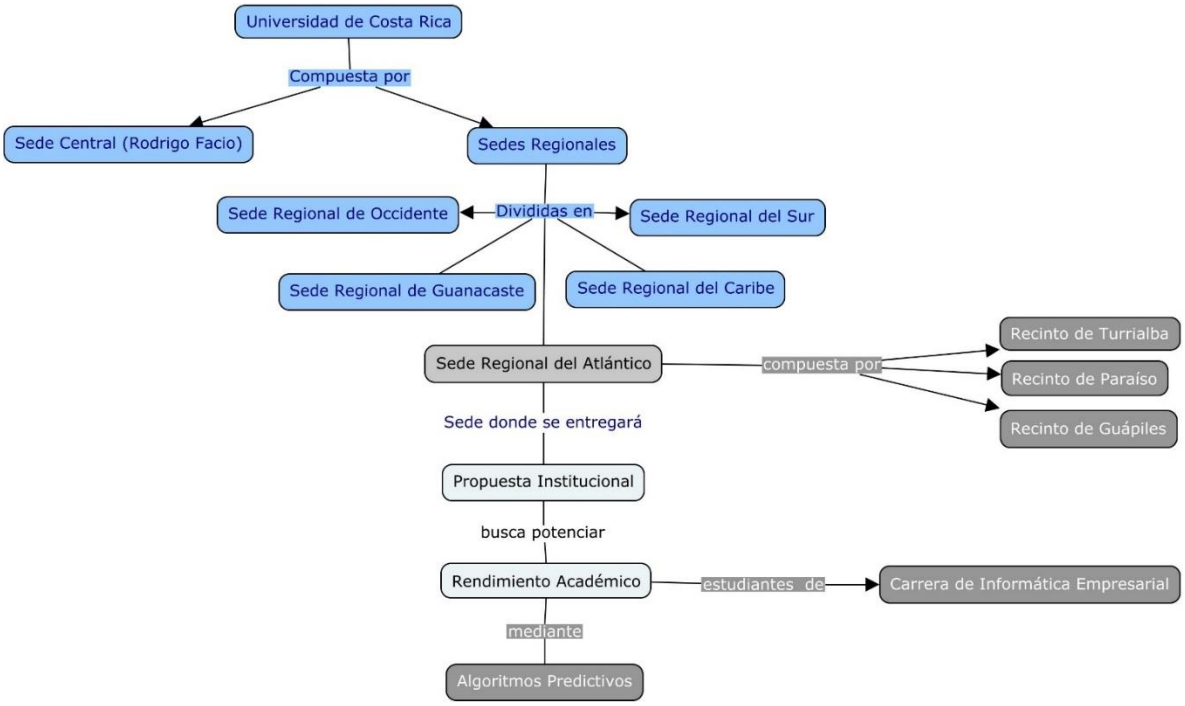


Figura 14

Diagrama de Conceptos de contenido

2.1.1. Definición de educación. La educación se define como:

> Un proceso humano y cultural complejo. La Educación busca la perfección y la seguridad del ser humano, busca asegurarle la libertad al hombre, pero la educación demanda disciplina, sometimiento, conducción y se guía bajo signos de obligatoriedad y a veces de autoritarismo, firmeza y direccionalidad (León, 2007, s. p.).

La educación rige desde que el ser humano está pequeño y recae primero en los padres de familia para posteriormente ir delegándola a las instituciones o casas de enseñanza de cada país. En la actualidad, como lo menciona el autor, la educación busca asegurarle la libertad al ser humano y en Costa Rica se tiene la tranquilidad de que todos pueden educarse de manera gratuita.

2.1.2. Definición de propuesta institucional. La gestión de la calidad educativa es un proceso que se presenta en todos los países donde se tiene educación

pública y gratuita, se pretende que un ente, una organización o un equipo dedicado busque orientar y fortalecer los proyectos educativos de las instituciones para que, de esta manera, la sociedad y el sector productivo de un país reciban profesionales que puedan acoplarse rápido a la fuerza laboral o a emprendimientos propios que, a la vez, generen empleos en una zona geográfica. Como lo indican Arenas Landínez y Jaimes Luna (2008).

> El Estado ha intervenido definiendo de manera participativa, estándares de calidad, es así, como los procesos de acreditación de programas y de instituciones se han posicionado como estrategias primordiales para fomentar y asegurar la calidad en la educación superior, a fin de que las instituciones formadoras rindan cuentas a la sociedad sobre los niveles de calidad con los que están respondiendo a sus requerimientos y sobre el cumplimiento de los objetivos que le han sido trazados en su fundamento misional, visional y funciones. Así mismo [sic], y como resultado necesario de la globalización, la calidad no puede centrarse en una concepción interna del país, sino antes bien, vislumbrar el escenario internacional como referente de comparación en cuanto a índices de calidad y niveles de enseñanza (s. p.).

De acuerdo con lo anterior, una propuesta institucional es aquella que, basada en un estudio previo y con resultados claros que se obtienen de la aplicación de una muestra, puede plantear soluciones o sugerir estrategias para lograr algún resultado en un corto, mediano o largo plazo. En el caso de esta investigación la propuesta que se lleva a cabo está fundamentada en los resultados que se obtienen de la aplicación de los algoritmos que se detallan más adelante y busca dar más calidad a la educación que se imparte en la Sede del Atlántico de la Universidad de Costa Rica, en la carrera de Informática. Sin embargo, no por esto significa que no se pueda adaptar la propuesta y estrategia institucional a otras sedes o carreras.

2.2. Conceptos técnicos

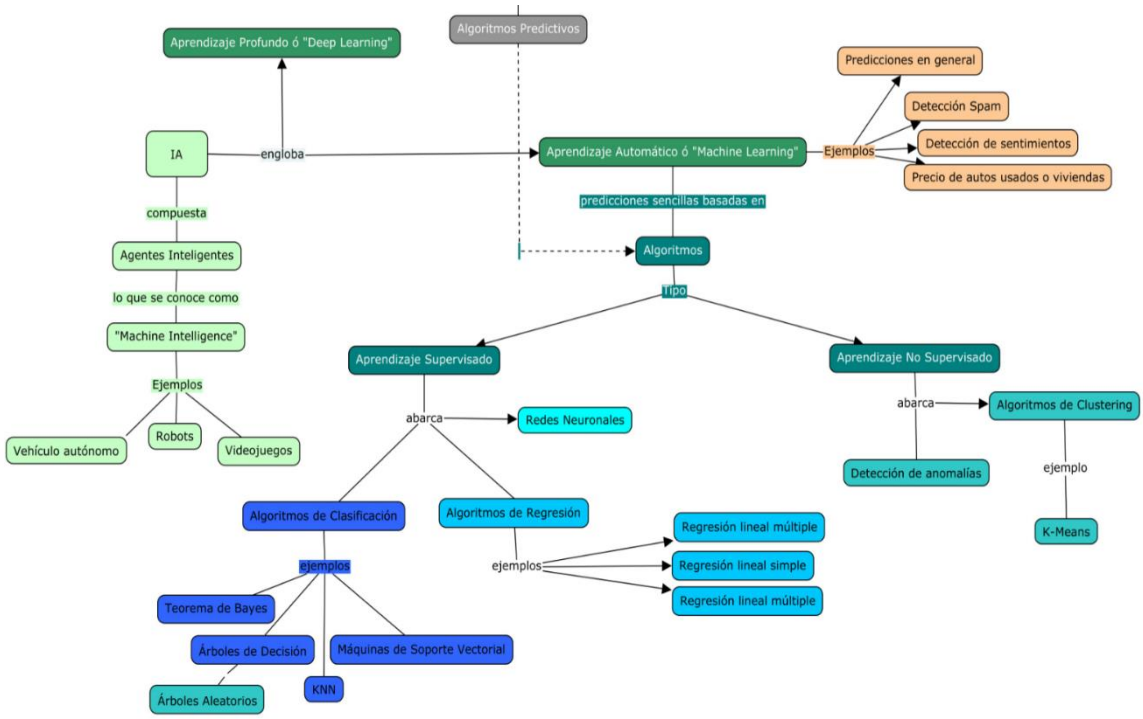


Figura 15

Diagrama de conceptos técnicos

2.2.1. Definición de minería de datos. El concepto de minería de datos es muy amplio y, generalmente, se suele confundir con otros conceptos como inteligencia de negocios, inteligencia artificial, OLAP, aprendizaje automático, aprendizaje profundo, entre otros. El concepto de minería de datos se puede definir como:

> Un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada, teniendo claro que los datos ya no solo están restringidos a tuplas de números o caracteres, si no también imágenes, video, texto, y otros datos numéricos. Como resultado, las herramientas tradicionales de gestión de datos o técnicas estadísticas no son suficientes, y es donde la minería de datos mediante la aplicación de algoritmos específicos permite extraer patrones de los datos (Riquelme, Ruiz y Gilbert, 2006, s. p.).

De acuerdo con esto, estos algoritmos se aplican sobre datos almacenados en las instituciones de enseñanza buscando patrones que no se pueden detectar de otra manera, mucho menos cuando los datos históricos son muchísimos.

2.2.2. Definición de minería de datos educacional. La Minería de datos educacional es: “Una disciplina emergente, preocupada por el desarrollo de métodos para explorar los tipos únicos de datos que provienen de entornos educativos y utilizar esos métodos para comprender mejor a los estudiantes y los entornos en los que aprenden” (Istanbullu y Guruler, 2013, s. p.). Aunque en el ámbito nacional no se suela escuchar que se utilicen herramientas tecnológicas para encontrar patrones en los datos que suelen almacenar año tras año, la realidad es que cada vez se usan más algoritmos predictivos de aprendizaje automático para crear predicciones o análisis de datos para encontrar patrones mediante algoritmos de minería de datos y esta es una tendencia mundial.

2.2.3. Definición de algoritmo. La definición de algoritmo, según Aguilar (2000), es:

> Un algoritmo es un método para resolver un problema, el cual debe ser preciso e indicar el orden de realización de cada paso, debe estar definido, con el fin de obtener el mismo resultado cada vez, y por último debe ser finito, o sea, tener un número finito de pasos (s. p.).

Esta investigación implementa algoritmos predictivos de aprendizaje automático, los cuales se detallan más adelante.

2.2.4. Definición de aprendizaje automático. Los algoritmos que se implementarán en la investigación son del tipo algoritmos predictivos, el significado de aprendizaje automático, según Sandoval Serrano (2018) es:

> Rama de la Inteligencia Artificial que se encarga de generar algoritmos que tienen la capacidad de aprender y no tener que programarlos de manera explícita. El desarrollador no tendrá que sentarse a programar por horas tomando en cuenta todos los escenarios posibles ni todas las excepciones posibles.

> Lo único que hay que hacer es alimentar el algoritmo con un volumen gigantesco de datos para que el algoritmo aprenda y sepa qué hacer en cada uno de estos casos. (s. p.).

Al tomar como experiencia lo aprendido en el módulo 1 del posgrado en inteligencia artificial de la Universidad Cenfotec, aprendizaje automático abarca un conjunto de técnicas que permiten construir *software* capaz de aprender de la experiencia y los datos. Estos abarcan la preparación de datos, desarrollo de modelos y algoritmos supervisados de clasificación y regresión, algoritmos no supervisados de clasificación, reducción de la dimensionalidad, regularización y aplicaciones avanzadas de aprendizaje automático como la detección de anomalías o la elaboración de sistemas recomendadores y procesamiento del lenguaje natural.

2.2.4.1. Aprendizaje supervisado. Aprendizaje supervisado es el tipo de modelo donde los datos de entrada que se tienen permiten determinar la salida, por lo tanto, el algoritmo no tiene que adivinar. Como lo menciona Harrington (2015): “La idea del aprendizaje supervisado es indicarle a la máquina que aprenda de los datos sabiendo cuál es la salida o etiqueta final para cada registro y con esto evitar que tenga que adivinar una etiqueta o variable final” (s. p.).

Los modelos de aprendizaje supervisado permiten tener dos tipos de etiquetas finales, las nominales, por ejemplo, verdadero o falso o un valor dentro de un rango ya definido de etiquetas (reptil, anfibio, mamífero, planta y fungi). El otro caso de clasificación ocurre cuando la etiqueta o variable puede tomar un valor infinito y numérico, por ejemplo, 0.50, 45.340, 1002.745, los cuales se denominan modelos de regresión.

2.2.4.1.1. Algoritmos de regresión. Harrington (2015) muestra los tipos de algoritmos que se pueden llevar a cabo, tanto desde el aprendizaje supervisado como no supervisado:

Tabla 6

Tipos de tareas según aprendizaje

Tareas de aprendizaje supervisado	
Clasificación	**Regresión**
K-vecinos cercanos	Lineal
Naive Bayes	Lineal ponderado localmente
Máquinas de soporte vectorial	Ridge (Cresta)
Árboles de decisión	Lasso
Tareas de aprendizaje no supervisado	
Clustering	**Estimación de densidad**
K-Means	Maximización de expectativas
DBSCAN	Ventana Parzen

Como se indicó en el punto anterior, los algoritmos de regresión pueden tomar valores numéricos y con esto llevar a cabo una predicción. Si se observa gráficamente un modelo de regresión lineal, por ejemplo, se pretende dibujar una recta en un plano de dos dimensiones *x* y *y* que esa recta permita calcular los posibles valores de *y* por los que pasa. Para efectos de esta investigación, no se utilizan, esto porque la idea no es predecir una nota de un curso o un promedio ponderado.

2.2.4.1.2. Algoritmos de clasificación. Lo que buscan los algoritmos de clasificación es dar una etiqueta a un registro o un grupo de registros del conjunto de datos que se estén trabajando. Harrington (2015) lo explica de una manera sencilla al usar películas para ejemplificar estos modelos. Una película pertenece siempre a un género y, aunque se sabe que en una película siempre es posible que existan escenas románticas, de acción, sangrientas o cualquier otra, lo cual haga pensar que la película puede pertenecer a otro género, al hacer un conteo de

cada característica de cada género automáticamente se podría categorizar esa película. Una película romántica tiene en mayor cantidad de besos, mientras que una película de acción contendrá una mayor cantidad de golpes o disparos.

2.3. Regresión logística

El algoritmo de regresión logística es un algoritmo de clasificación que permite etiquetar un valor de manera binaria, por ejemplo, sí o no, 1 o 0. Si se toma como ejemplo el conjunto de datos de esta investigación, lo que se busca es determinar si un estudiante que ingresa el próximo año 2022 durará cuatro años en el plan de estudios, el algoritmo de regresión logística debe ser capaz de responder si es capaz.

Geron (2019) explica que este algoritmo se usa para estimar la probabilidad de que una instancia pertenezca a una clase particular, por ejemplo, si un correo electrónico es *spam* o no. Si la probabilidad es mayor a 50 %, entonces se puede tomar como que la instancia pertenece a una clase llamada positiva etiquetada como *1* y, en caso contrario, pertenecería a una clase negativa etiquetada como *0*, lo que convierte este modelo en un clasificador binario.

Este algoritmo funciona similar al modelo de regresión lineal, sin embargo, hace una suma ponderada de todas las variables que alimentan el modelo junto con el valor *Bias* que se conoce como el error. En lugar de mostrar una ecuación de la recta se usa una función sigmoide que devuelve un valor entre 0 y 1.

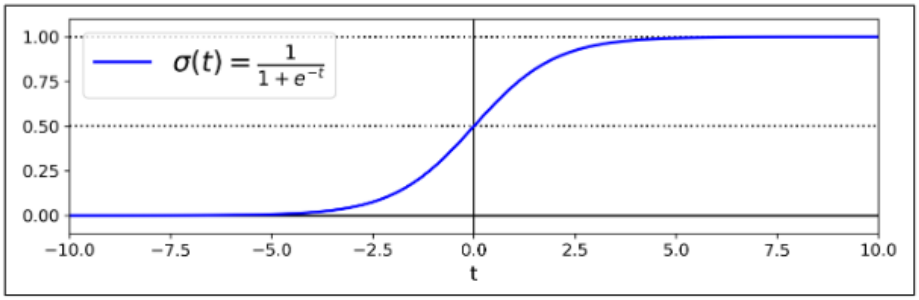


Figura 16

Función sigmoide usada en el algoritmo de regresión logística

Fuente: Zamora (2020).

De acuerdo con Geron (2019): “Una vez que se haya estimado la probabilidad de p = h(x) de que ese x pertenezca a una clase positiva se puede realizar la predicción de manera sencilla como se puede ver en la siguiente fórmula” (s. p.).

$$\hat{y} = \begin{cases} 0 \text{ if } \hat{p} < 0.5 \\ 1 \text{ if } \hat{p} \geq 0.5 \end{cases}$$

Figura 17

Ecuación del modelo de regresión logística

Fuente: Zamora (2020).

Se puede observar en la ecuación anterior que cuando el valor obtenido en la probabilidad es mayor o igual a 0.5 el modelo de regresión logística predecirá 1, mientras que cuando es menor de 0.5 el modelo predecirá 0 o negativo.

A modo de ejemplo, se muestra en las siguientes imágenes la función sigmoide correspondiente a detectar si un pétalo de una flor es *virginica*. Este ejemplo pertenece al famoso conjunto de datos que permiten detectar no solo mediante el ancho y largo de un pétalo, si una flor es *virginica*, sino también si es *setosa* o *versicolor*:

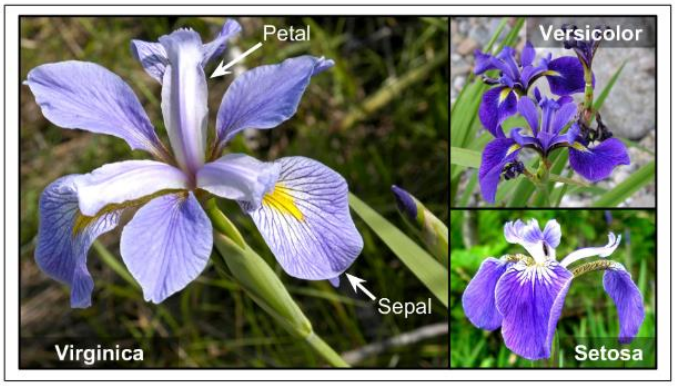


Figura 18

Flores setosas, versicolor o virginica

Fuente: Zamora (2020).

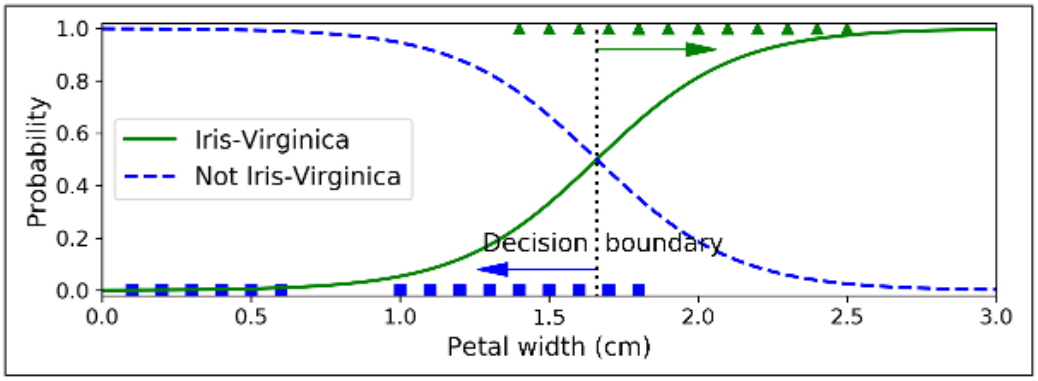


Figura 19

Probabilidad estimada de que una flor sea iris-virginica

Fuente: Zamora (2020).

Con la función que permite detectar si una flor es virginica o no en función del pétalo, se puede observar en la siguiente imagen el límite de decisión lineal que se genera por detrás, pero cuando se tienen dos variables, en este caso el ancho y el largo del pétalo. La imagen muestra un límite que afirma que cuanto más largo y ancho es el pétalo, la probabilidad de que sea iris-virginica es mucho mayor.

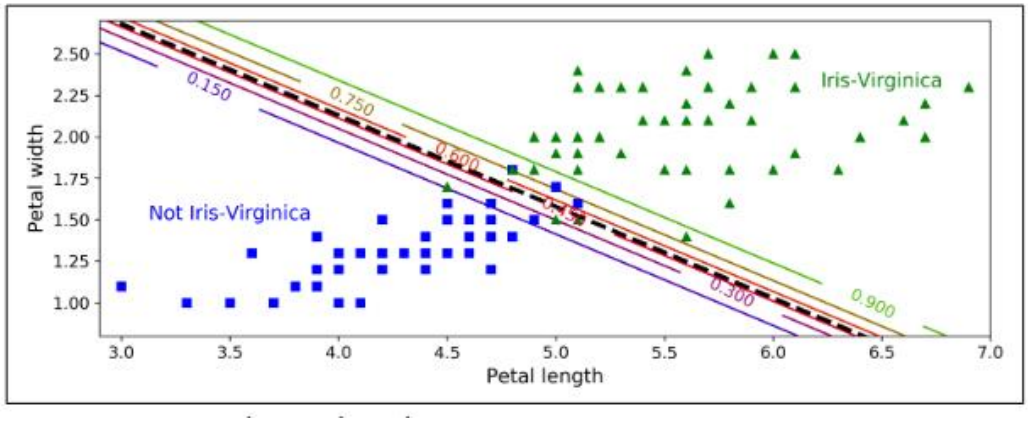


Figura 20

Límite de decisión lineal

Fuente: Zamora (2020).

2.4. Árboles de decisión

La técnica de árboles de decisión es de las más comunes que existen cuando se habla de algoritmos de clasificación y su ventaja recae en que no se tiene que ser un experto en minería de datos o aprendizaje automático para entender cómo funciona. Su fundamento es tomar un conjunto de datos y generar respuestas en función de la variable que se evalúe. Harrington (2015) utiliza el juego *20 preguntas* como ejemplo, en el cual una persona piensa en un objeto y los demás participantes tratan de adivinar con solo 20 preguntas y al recibir solo respuestas de sí o no. Al igual que un árbol de decisión, en este juego las personas que averiguan el objeto dividen todos los objetos que deducen.

Como se puede observar en la Figura 21 la idea de los árboles de decisión es utilizar los datos para generar una serie de bloques de decisión que suelen tener forma de rectángulo y una serie de bloques terminales con forma de óvalo, los cuales contienen las conclusiones a las que se puede llegar. Después se tienen las flechas izquierda o derecha que indican las ramificaciones que pueden tener los bloques de decisión, estas pueden llegar directamente a una conclusión después de validar un bloque de decisión o concatenar otro bloque de decisión para validar más datos.

En la siguiente imagen se puede ver una representación de un árbol de decisión que permite determinar cuando un correo electrónico es *spam* o no.

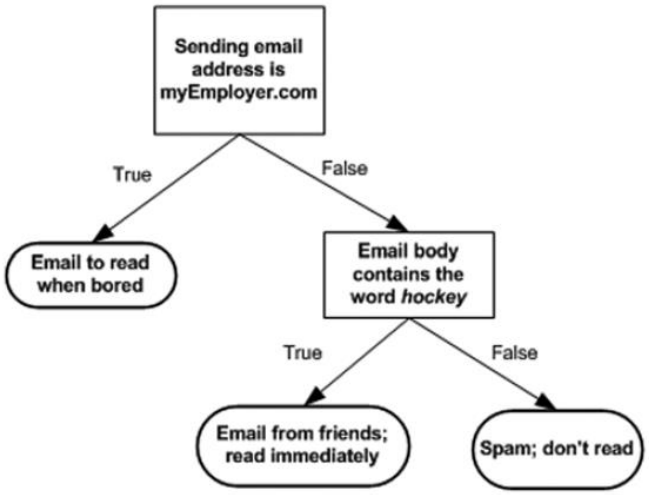


Figura 21

Árbol de decisión para detectar spam en correos electrónicos

Fuente: Harrington (2015).

En la Figura 21 se ven dos bloques de decisión, el primero verifica si el emisor del correo viene de myEmployer.com, lo cual tiene dos ramificaciones, una llega a un bloque de conclusión en forma de óvalo que afirma que en caso afirmativo, es un correo que se puede leer y la otra ramificación conecta otro bloque de decisión que pregunta si el correo contiene la palabra *hockey*, en caso afirmativo permite leer el correo e indicar que es un correo de amigos, mientras que si es falso lo marca como *spam* y manda la advertencia de no leerlo. Aunque es un ejemplo sencillo, la lógica detrás de cualquier otro árbol de decisión es muy similar.

La mayor ventaja de un árbol de decisión es la capacidad de lectura que permite en las personas, por lo que se ajusta a la propuesta que se planea entregar a la Universidad de Costa Rica, ya que no todos son especialistas en Informática y mucho menos en minería de datos. Un árbol de decisión, si se aborda desde la perspectiva de minería de datos educacional, permite observar cómo se llega a bloques de conclusión, los cuales indican si un estudiante tiene un

buen rendimiento o no durante el desarrollo de su plan de estudios o indicar si logrará finalizar en cuatro años o más, mientras que los bloques de decisión pueden generar ramificaciones en función de si un estudiante ha perdido cursos y qué tipo de cursos ha perdido, también preguntar si el estudiante es becado o no o si suele participar de estudiaderos. Los bloques de decisión pueden ser N y todo depende de los datos con los que se alimente el algoritmo.

Harrington (2015) indica que los árboles de decisión tienen las siguientes características:

> Dentro de los pros que un árbol tiene está que son muy baratos a nivel computacional, puede trabajar aún con valores faltantes, son sumamente fáciles de entender para las personas y puede lidiar fácilmente con variables irrelevantes. Del lado de los contras, los árboles de decisión suelen generar mucho *overfitting*, lo cual significa que puede tender a estar sobreentrenado o sobre ajustado, lo que es malo porque tiende a aprender también del ruido o errores de los datos (s. p.).

Capítulo III. Marco metodológico

3.1. Tipo de investigación

Este proyecto busca que la Sede del Atlántico de la Universidad de Costa Rica tenga planes de apoyo y herramientas que potencien el rendimiento académico de sus estudiantes de informática mediante el análisis de los resultados de los algoritmos predictivos. Por lo tanto, esta investigación se aplica debido a que se utilizan herramientas y conocimiento existente del área de aprendizaje supervisado para atender esta necesidad.

Se necesita evaluar los algoritmos que se ajusten mejor a los conjuntos de datos que se tienen en el ámbito institucional, ya que distintos tipos de algoritmos permiten responder a diferentes situaciones. En este caso particular se aplica sobre la Sede del Atlántico de la Universidad de Costa Rica y se busca determinar qué causa que un estudiante finalice su plan de estudios en el tiempo establecido o cuáles factores influyen para que un estudiante tenga mejor rendimiento que otro.

3.2. Alcance investigativo

El alcance es explicativo, ya que más allá de describir la situación actual, se busca establecer la o las causas de los eventos o fenómenos (Barrantes Echavarría, 1999). En este caso es el fenómeno de un bajo rendimiento en estudiantes de informática en la Sede.

Arias Odón (2012) amplía este tipo de investigación al afirmar:

> Esta investigación se encarga de buscar el porqué de los hechos mediante relaciones causa y efecto. Los estudios explicativos pueden ocuparse tanto de la determinación de las causas, como de los efectos, mediante prueba de hipótesis. Sus resultados y conclusiones constituyen el nivel más profundo de conocimientos (s. p.).

Entre los ejemplos que este autor menciona se tiene:

- Indagación de las causas que producen la deserción escolar.
- Investigación de los factores que motivan el bajo rendimiento estudiantil.
- Estudio sobre los efectos del consumo de efedrina en el rendimiento deportivo.

Si se observan los primeros dos ejemplos, se ve que este tipo de investigación calza perfectamente con el área de la minería de datos o aprendizaje automático en el ámbito educacional que se desarrolla.

3.3. Enfoque

El enfoque de esta investigación es alternativo y describe cómo se relacionan los objetivos que se plantearon con las tres dimensiones fundamentales de este enfoque. Según Naranjo Zeledón (2020) el enfoque alternativo:

> Esta propuesta no varía sustancialmente la estructura de un marco metodológico tradicional, pero si implica un cambio importante en la subsección de enfoque, pues libera al investigador para usar cuanto considere adecuado para alcanzar sus objetivos en las subsecciones de diseño e instrumentos de recolección de datos […] agiliza la etapa de planificación de la investigación y concede mayor tiempo para concentrarse en las fases de análisis de hallazgos y redacción de escenarios de propuesta (s. p.).

A continuación, se presentan las tres dimensiones fundamentales de este tipo de enfoque para este trabajo en detalle y la primera dimensión a tratar es la del punto de vista ontológico. El concepto de ontología según Gruber (1993) es: “Una especificación formal y explícita de una conceptualización compartida” (s. p.). Como lo menciona Naranjo Zeledón (2020) esta dimensión debe ser explícita, ya que se necesita especificar de forma consciente todos los conceptos relevantes que conforman la ontología formal mediante un lenguaje de representación formalizado y compartida, ya que se presenta a la comunidad encargada de

evaluarla y usarla, en este caso la Universidad de Costa Rica que recibe la propuesta y Universidad Cenfotec como evaluador.

Este trabajo presenta a continuación una ontología de conceptos como datos, tipos de datos, minería de datos, minería de datos educacional, aprendizaje automático, modelos y algoritmos de predicción para evaluación de rendimientos académicos de estudiantes.

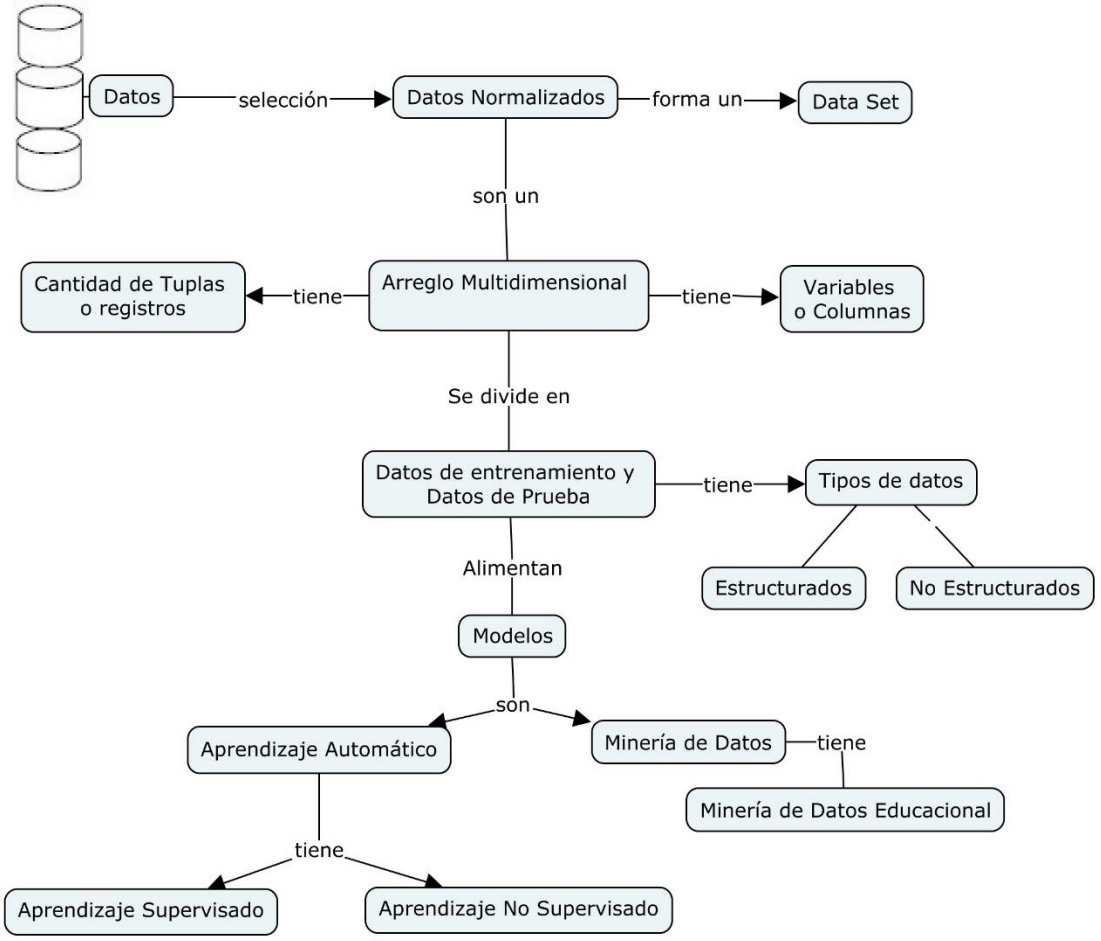


Figura 22

Ontología

Desde la dimensión epistemológica, la persona investigadora tiene una posición de observador, ya que le corresponde mirar los resultados de los algoritmos predictivos que implementa y comparar para que determine cuáles se

ajustan mejor a los conjuntos de datos, después de este proceso puede desarrollar la propuesta institucional con todas las conclusiones que genere.

Por último, desde la dimensión axiológica, como lo menciona Naranjo Zeledón (2020) se pueden formalizar escalas de valores para no utilizar conceptos cuya definición o medición de su valor intrínseco sea muy vaga. En el caso de esta investigación, la cual trabaja con algoritmos predictivos y a partir de lo que mencionan Salazar y Tenorio (2020): “Se determina que desde el punto de vista de la evaluación de precisión de diferentes modelos de minería, no existe un consenso para establecer rangos al clasificar el resultado del algoritmo, como bueno, regular o malo” (s. p.), debido a que todo depende de la calidad de los sets de datos que los alimenten y las variables que se escojan y los procesos de normalización de datos que se realicen, se tiene que usar la probabilidad, el modelo ROC y cómo lo mencionan Salazar y Tenorio (2020) y Jiménez Padilla y Cuevas Covarrubias (2011) se permite ver el rendimiento de un modelo en umbrales de clasificación, en donde un valor de 0 significa que el 100 % de predicción del modelo es incorrecto y falló en predecir correctamente la totalidad de las veces que se utilizó y un valor de 1 significa que el 100 % de predicción del modelo fue correcto.

La curva ROC es la calidad discriminante de S, la cual depende de todas las combinaciones de sensibilidad y especificidad posibles. La curva ROC es la gráfica de la sensibilidad vista como función de la tasa de falsos positivos. Esta sintetiza gráficamente el desempeño potencial de S como un índice de riesgo. El área bajo la curva ROC resume toda esta información en un índice numérico que mide la calidad de S (Jiménez Padilla y Cuevas Covarrubias, 2011).

Tabla 7

Criterios de evaluación de la precisión de los algoritmos predictivos

Evaluación de la precisión del modelo	**Criterio**
Sobresaliente	100 %= Precisión >= 70 % de los casos al si un estudiante finaliza su plan de estudios en el tiempo establecido o si tiene un rendimiento académico aceptable
Aceptable	69 % >= Precisión > 51 % de los casos al predecir si un estudiante finaliza su plan de estudios en el tiempo establecido o si tiene un rendimiento académico aceptable
Deficiente	51 % = Precisión >= 0 % de los casos al predecir si un estudiante finaliza su plan de estudios en el tiempo establecido o si tiene un rendimiento académico aceptable

Fuente: Salazar y Tenorio (2020).

3.4. Diseño

El diseño muestra la concepción o configuración que debe tener el TFG y se relaciona con el enfoque seleccionado. Los diseños en estudios cuantitativos enfatizan en encontrar relaciones entre variables, en tanto que los diseños en estudios cualitativos se concentran en dar sentido y valor a categorías de análisis.

El diseño de esta investigación es experimental *post-test* con grupo de control. Esto significa que entre los conjuntos de datos que se disponen para la elaboración de los modelos de algoritmos predictivos, se usa validación cruzada con el 70 % del *data-set* para entrenar estos modelos y el otro 30 % se usa como prueba de los resultados y medir cuán efectivos son los modelos generados. Esta técnica de validación cruzada no es al azar y se toma con base en los

conocimientos adquiridos en el curso de *aprendizaje automático* y *aprendizaje profundo* del posgrado en inteligencia artificial impartido por el profesor Juan Zamora y de los conocimientos adquiridos en el curso de minería de datos de la Maestría en Tecnologías de Datos de Universidad Cenfotec con el profesor Diego Alfaro Belgueiro. No solo se puede usar 70 %-30 %, también se puede usar 80 %-20 % o 90 %-10 %, pero todo depende del tamaño del *data-set* y la calidad del modelo generado.

Este proceso sigue los siguientes pasos:

- El conjunto de datos que se trabaja proviene del Departamento de Registro y Matrícula y de los expedientes académicos de la Universidad de Costa Rica.
- Para determinar el tiempo que tarda un estudiante en finalizar el plan de estudios se puede usar como referencia el carné, el cual indica el año de ingreso a la Universidad y se verifica la fecha de graduación de este.
- Se usan conjuntos de datos de información personal del estudiante, como lugar de residencia o si dispone de beca o no, además de los datos del expediente académico que tuvo o tiene.
- Se aplica normalización de datos para estandarizar la escala de valores cuantitativos que se tiene y que suele usarse como buena práctica para entrenar el modelo y evitar que este se desempeñe más lento o se vuelva inestable en el momento de generalizar (Bishop, 1995).
- Se implementan los algoritmos de aprendizaje automático que predigan mejor el rendimiento estudiantil o el año de finalización del plan de estudios.
- Se implementan los algoritmos de minería de datos que permitan detectar patrones en los datos finales.
- Se interpretan los resultados y el nivel de confianza de todos los algoritmos.

- Se llevan a cabo las conclusiones.

3.5. Población y muestreo

La población es el punto de partida para recolectar datos. El muestreo, que puede ser probabilístico o no probabilístico, busca conseguir una parte que sea suficientemente representativa de la población para dar validez a los hallazgos.

Levin y Rubin (2004) explican el concepto de población y muestra, el cual afirma: “Una población es un conjunto de todos los elementos que estamos estudiando, acerca de los cuales estamos tratando de sacar conclusiones. Y una muestra es una colección de algunos elementos de la población” (s. p.).

Con los conceptos explicados y al necesitar de un conjunto de datos para entrenar los modelos de aprendizaje automático y generar los conjuntos de entrenamiento y prueba, se procede a explicar las características del conjunto de datos que brinda la Universidad de Costa Rica:

Tamaño: 1200 registros de estudiantes. Este número tiene en cuenta que por año y por recinto hay 40 cupos de matrícula aproximadamente, lo que da un total por año a nivel de Sede de 120 estudiantes de informática nuevos y al obtener los datos de los últimos 12 años da un total de 1200 registros. Se usan 10 años de antigüedad en el conjunto de datos, ya que se puede verificar si las personas estudiantes de 2010, 2011, 2012, 2013, 2014, 2015, 2016 y 2017 lograron finalizar su carrera en el tiempo establecido y se usarán los datos del 2018, 2019, 2020 y 2021 como conjunto de datos de prueba para llevar a cabo predicciones. El manejar los últimos 12 años permite tener una muestra en la que se pueda aplicar la regla de 70-30 para entrenar los modelos.

Atributos: La Universidad entrega los siguientes atributos por estudiante:

- Carné del estudiante.
- Género (hombre-mujer).
- Dirección (provincia, cantón y distrito).

- Nivel de beca (en escala de 1-5).
- Dirección (provincia, cantón y distrito).
- Nota de ingreso a carrera.
- Expediente académico.
    - Reprobó cursos de matemática (sí-no).
    - Reprobó cursos de programación (sí-no).
    - Cantidad de cursos reprobados.
    - Repitió mismo curso más de una vez.
- Año de graduación (2016, 2017, 2018, 2019, 2020, 2021 o 2022).
- Finalizó el plan de estudios en 4 años (sí-no).

3.6. Instrumentos de recolección de datos

En el caso de esta investigación se realiza una recopilación de información que proviene de las coordinaciones de carrera de Informática en los tres recintos que se trabajan y el Departamento de Registro y Matrícula de la Universidad. Esto permite llevar a cabo un análisis documental de todos los datos y lograr:

- Identificar las variables que se ajustan mejor a los modelos de aprendizaje automático o minería de datos que permitan encontrar las causas de un rendimiento en específico o por qué un estudiante logra finalizar el plan de estudios en cuatro años o más.
- Llevar a cabo un proceso de normalización de datos cuando se detecten datos con rangos de valores muy distantes.
- Aplicar técnicas estadísticas robustas como mediana, media o promedio que permitan lidiar con *outliers* (valores aislados) y le quiten peso a observaciones atípicas que alteren el entrenamiento de los modelos.
- Separar los datos en conjuntos de entrenamiento y prueba.

3.7. Técnicas de análisis de información

Con base en la investigación de Salazar y Tenorio (2020) se utiliza un diagrama de flujo que muestre el proceso general de obtención del conjunto de datos, desde las fuentes de la Universidad, para la creación de los conjuntos de datos limpios que permiten alimentar los modelos de minería de datos. Antes de alimentar los algoritmos se realiza un proceso de limpieza y normalización y después se determinará cuán precisos son los algoritmos. Se usan los mismos rangos de valores que utilizan Salazar y Tenorio (2020): “Trabajando con rangos que van de 0 a 50% para un modelo deficiente, de 51% a 69% para un modelo considerado aceptable y por último el rango de 70% a 100% para los modelos considerados sobresalientes” (s. p.), para predecir los rendimientos que pueden tener las personas estudiantes.

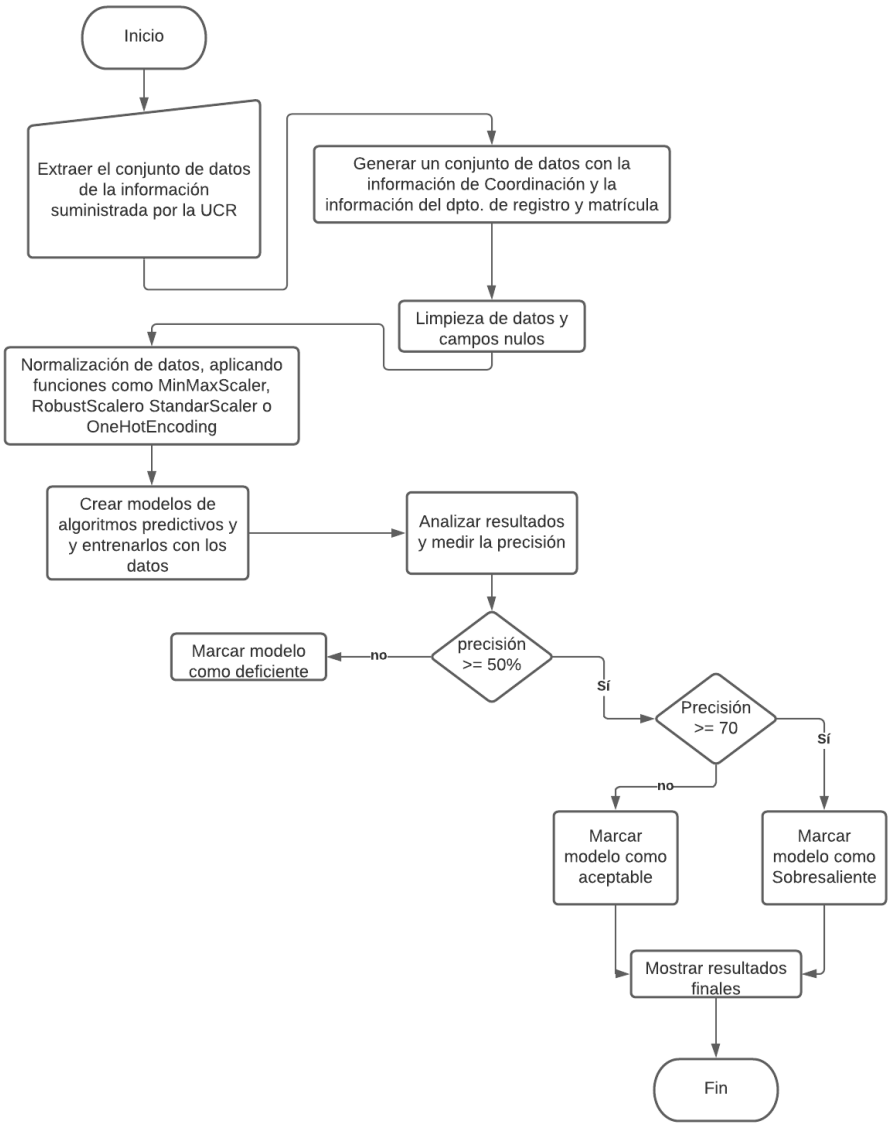


Figura 23

Diagrama de flujo de Técnicas de Análisis de Información

Fuente: Salazar y Tenorio (2020).

3.8. Estrategia de desarrollo de la propuesta

Para esta investigación es necesario programar los algoritmos de aprendizaje automático que permiten indicar si un estudiante, con base en todas las variables establecidas, logrará tener una conclusión del plan de estudios de manera satisfactoria y en el lapso esperado. Es importante aclarar que los datos

en su mayoría son exclusivos de la Universidad de Costa Rica por lo que existe una cláusula de confidencialidad respecto a estos y solo se muestran los resultados en los algoritmos.

El lenguaje de programación por utilizar es Python, en primer lugar, porque es el lenguaje de programación que se usó en los módulos del posgrado en inteligencia artificial con los profesores Juan Zamora y Javier Fabio Porras, el cual recomendaron por encima de R y también por recomendación de Harrington (2015), quien explica por qué es preferible usar Python:

> Este lenguaje es el más utilizado en aprendizaje automático por su sintaxis clara, permite la manipulación de texto de manera extremadamente sencilla y tiene una documentación sumamente amplia, además de un gran desarrollo de algoritmos en organizaciones que se pueden utilizar como guía (s. p.).

Finalmente, el autor menciona que existen librerías probadas a lo largo de varios años que garantizan la velocidad en la implementación de varios algoritmos, por ejemplo, numPy y sciPy. Estos, a la vez, se acoplan a la perfección con librerías como MatPlotLib que permiten graficar en 2D y 3D y mostrar en pantalla prácticamente todos los tipos de gráficos usados comúnmente en el mundo científico.

Capítulo IV. Análisis del diagnóstico

La recopilación de datos para este proyecto se hizo directamente con las personas docentes y coordinadores de la carrera de Informática de los tres recintos, el profesor MGP Álvaro Mena Monge, MAG Carlos Felipe Escalante Solano y el Lic. Nelson Méndez Montero, los cuales, tras una solicitud formal previa, entregaron los datos previstos en el primer punto del diagrama de flujo detallado en la sección 3.7 de esta investigación (se adjunta oficio SA-RP-CIE-46-2021). Estos datos suministrados hay que unificarlos, ya que vienen en *data-sets* separados y divididos por recintos, los *data-sets* se detallan a continuación:

Tabla 8

Listas de estudiantes admitidos del 2010 al 2021

Variable	**Tipo de dato**
Número	Int (identificador único dentro del *data-set*)
Carné	String
Sexo	Int (1 para hombre, 2 para mujer)
Edad Ingreso	Int
Nota Examen admisión	Float
ProvinciaProcedencia	Int (rango de 1 a 7)
CantónProcedencia	Int
DistritoProcedencia	Int
ColegioProcedencia	Int
AñoConcursoACarrera	Int

Tabla 9

Listas de cursos matriculados con su nota por estudiante por periodo de matrícula y por año, del 2010 al 2021

Variable	Tipo de dato
Carné	String
SiglaCurso	String
NombreCurso	String
NotaCurso	Float
PeriodoCurso	Int (1 2 o 3)
AñoCurso	Int

Tabla 10

Lista de estudiantes graduados del 2010 al 2021

Variable	Tipo de dato
AñoIngresoUCR	Int
AñoIngresoCarrera	Int
CicloDerechoGraduación	String
AñoJuramentación	Int
GradoObtenido	String
Duración (en años)	Float

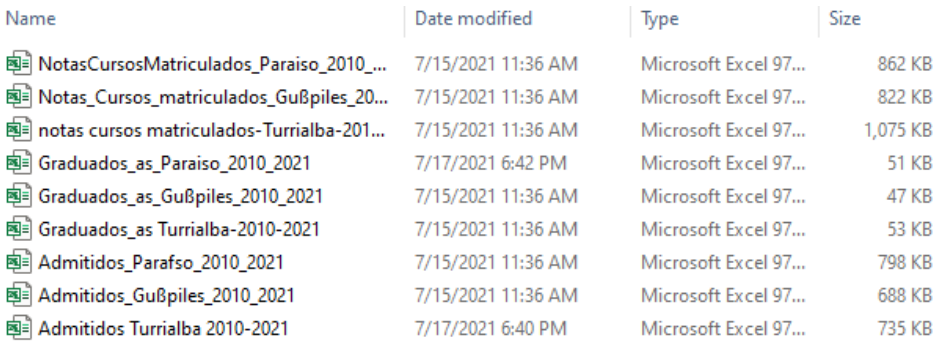
4.1.1. Identificación de variables que permiten alimentar los algoritmos predictivos que se desarrollan. Se desarrolla un algoritmo de regresión logística que indique si un estudiante X logrará concluir el plan de estudios en cuatro años o no. Para este algoritmo de regresión logística se evalúan varios modelos con

distintos tipos de variables, los cuales son:

- Nota del examen de admisión.
- Expediente académico (se puede desglosar por completo o agrupar por áreas: matemática, programación, cursos humanísticos, entre otros y también desglosar por periodos [primer año, segundo año, tercer año]).
- Dirección registrada.
- Colegio de procedencia.

Para efectos del algoritmo de árbol de decisión, se puede generar un único modelo en el cual se evalúen todas las variables y ver cómo es el comportamiento de las personas estudiantes del conjunto de pruebas.

Como se puede observar en el segundo bloque del diagrama de flujo de la sección 3.7 de esta investigación, se necesita generar el conjunto de datos oficial. Para esta tarea se usa SQL Server como herramienta para unificar los conjuntos de datos que entregó la Universidad de Costa Rica. Con el conjunto de datos oficial se cargan como *data-sets* a los *scripts* de Python mediante la librería de Numpy.



Name	Date modified	Type	Size
NotasCursosMatriculados_Paraiso_2010_...	7/15/2021 11:36 AM	Microsoft Excel 97...	862 KB
Notas_Cursos_matriculados_Gußpiles_20...	7/15/2021 11:36 AM	Microsoft Excel 97...	822 KB
notas cursos matriculados-Turrialba-201...	7/15/2021 11:36 AM	Microsoft Excel 97...	1,075 KB
Graduados_as_Paraiso_2010_2021	7/17/2021 6:42 PM	Microsoft Excel 97...	51 KB
Graduados_as_Gußpiles_2010_2021	7/15/2021 11:36 AM	Microsoft Excel 97...	47 KB
Graduados_as Turrialba-2010-2021	7/15/2021 11:36 AM	Microsoft Excel 97...	53 KB
Admitidos_Parafso_2010_2021	7/15/2021 11:36 AM	Microsoft Excel 97...	798 KB
Admitidos_Gußpiles_2010_2021	7/15/2021 11:36 AM	Microsoft Excel 97...	688 KB
Admitidos Turrialba 2010-2021	7/17/2021 6:40 PM	Microsoft Excel 97...	735 KB

Figura 24

Conjuntos de datos provistos con información de los tres recintos de la Sede

Fuente: Universidad de Costa Rica, Sede del Atlántico, Coordinaciones de carrera de Informática.

Esta unificación de datos que se realiza desde la herramienta SQL Server permite exportar posteriormente los datos a documentos de Microsoft Excel que

sirvan como fuente de datos primaria para los modelos por generar. A continuación, se pueden ver los *scripts* generado para la generación de los *data-sets*, se incluyen datos necesarios para el algoritmo de regresión logística:

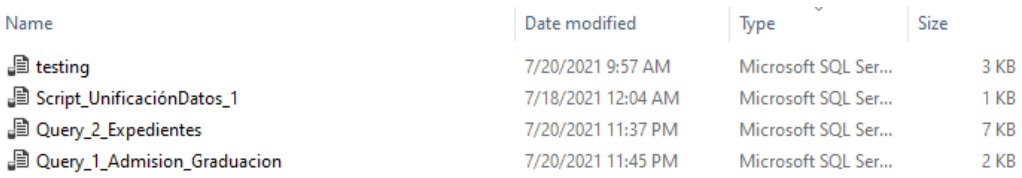


Figura 25

Scripts para generar los data sets que alimenten los algoritmos

4.2. Metodología CRISP-DM

El modelo de minería de datos que se usa en este trabajo se basa en CRISP-DM, ya que es el que más se utiliza cuando se trabaja con proyectos de este tipo. Galán (2015) indica que para implementar una tecnología en un negocio es necesaria una metodología, estos métodos suelen venir de las experiencias propias y también de los procedimientos estándar más conocidos.

En el caso de los proyectos de implementación de minería de datos una de las metodologías que ha tenido más apoyo de las empresas privadas y organismos públicos es CRISP-DM, como se puede observar en la siguiente gráfica (Figura 26). Esta representa el grado de utilización de las principales guías de desarrollo de proyectos de minería de datos según las encuestas realizadas.

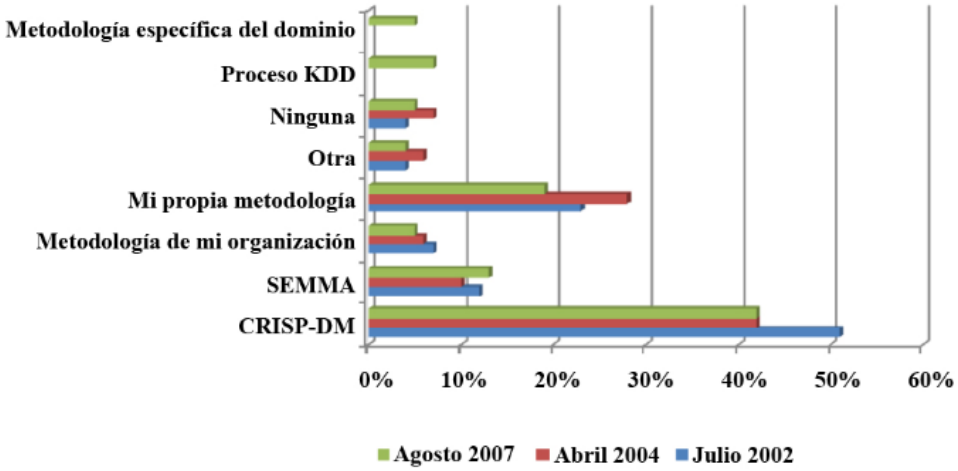


Figura 26

Grado de utilización de las distintas metodologías de minería de datos

Fuente: Galán (2015).

La ventaja de utilizar CRISP-DM es que este modelo es una guía de seis fases, de las cuales algunas son bidireccionales, por lo tanto, se puede regresar a una fase anterior para revisarla y también permite que no se siga un orden fijo, desde la primera hasta la última. Estas fases se observan en la Figura 27:

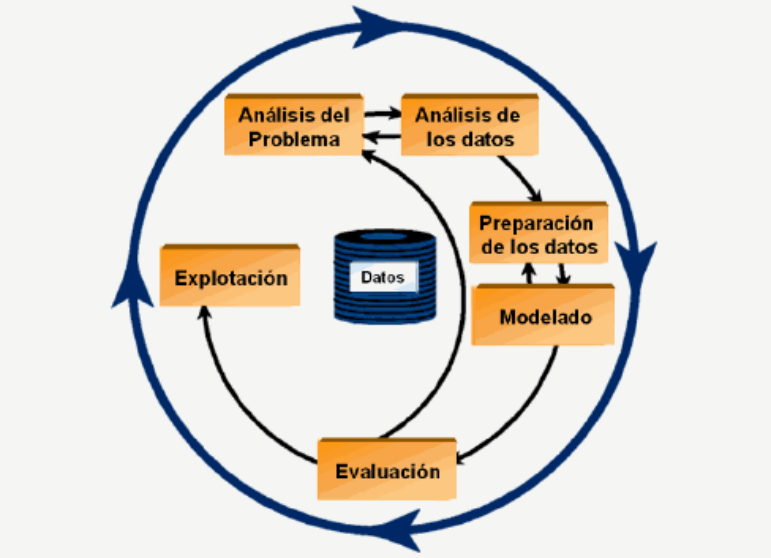


Figura 27

Modelo CRISP-DM

Nota. Tomado de https://blogdaedalus.files.wordpress.com/2013/10/crispdm.gif

En este punto se lleva a cabo un análisis exploratorio de los datos, por lo que se procede con las variables que se tienen del conjunto de datos provisto por la universidad. La primera variable por explorar es la nota de admisión entre el 2010 y el 2021 para ver como se ha comportado el ingreso de estudiantes en términos de sus notas a través de los años, como se puede observar en la Figura 28:

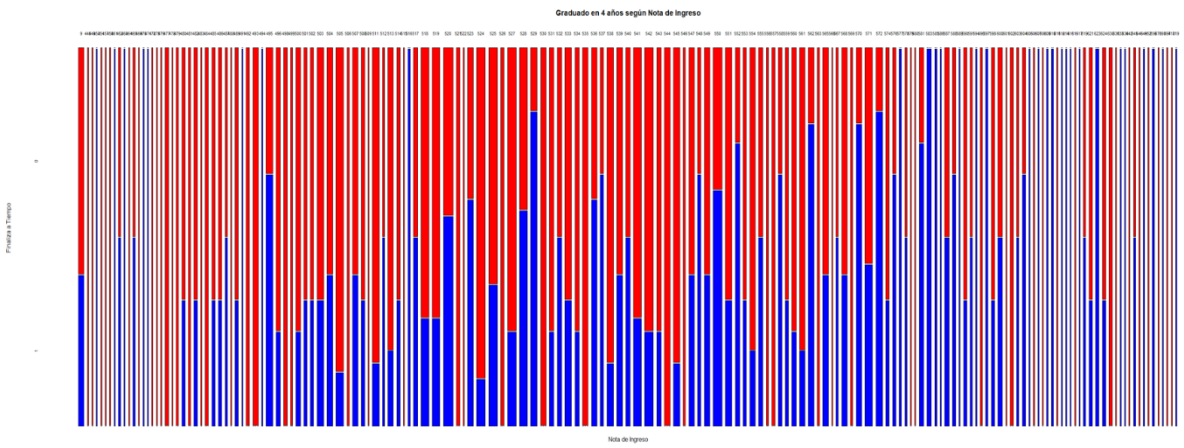


Figura 28

Gráfico de graduados en cuatro años según nota de ingreso a carrera

La Figura 28 muestra todas las personas estudiantes ingresadas en la carrera de Informática, organizadas por nota de admisión de la menor a la mayor de izquierda a derecha. La barra de color rojo corresponde a las personas estudiantes con esa nota específica que no lograron graduarse en 4 años y la barra azul los que sí lo lograron. A pesar de que se puede observar un incremento en el color azul conforme suben las notas de admisión, se puede ver fácilmente que el color rojo siempre existe, por lo que no se puede garantizar que la nota de admisión es un factor determinante cuando de graduarse en el tiempo mínimo se trata.

La siguiente variable por explorar es la edad con la que las personas estudiantes ingresaron a carrera entre el 2010 y 2021 y con esto observar si determina algún comportamiento respecto a graduarse a tiempo a no:

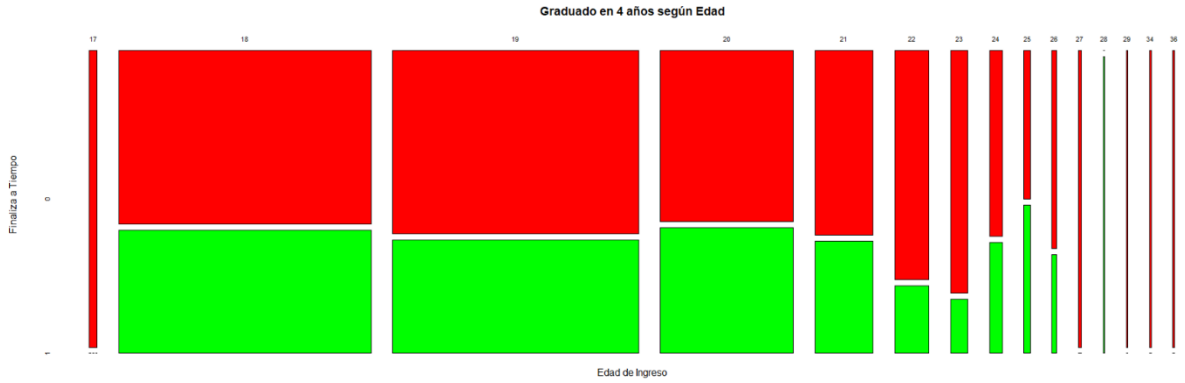


Figura 29

Gráfico de graduados en cuatro años según edad de ingreso a carrera (17 años en adelante)

Capítulo V. Propuesta de solución

El primer escenario que se desarrolla es el del modelo de regresión logística, el cual se propone con dos escenarios, uno es un modelo de regresión de una sola variable y toma la nota del examen de admisión como valor para predecir si el estudiante se graduará en cuatro años o no. Posteriormente, se realiza un modelo de regresión logística multivariable utilizando no solo la nota de examen de admisión, si no también la edad de ingreso del estudiante, el corte que tenía la carrera para primer ingreso, el colegio de procedencia, la provincia, cantón, distrito que registra el estudiante (para las variables categóricas se usa la técnica de *one-hot encoding*, la cual permite representar todas las categorías como un vector, donde todos los valores tienen 0 excepto por la categoría que se quiera representar como verdadera con un 1). Por último, se consideran las notas de los cursos que tiene en su expediente académico.

Este modelo es un modelo de clasificación y cómo se explicó en el marco conceptual, se pretende tener una salida categórica que en este caso es: si finaliza en cuatro años y no finaliza en cuatro años (la carrera de Informática tiene una duración de cuatro años para bachillerato). El primer escenario trabaja con la siguiente función:

f(notaExamenAdmisión)$\mapsto$[0,1], donde 0 es no finaliza y 1 si finaliza.

Al tomar como base los *notebooks* del módulo de *aprendizaje automático* dado por el profesor Juan Zamora, se necesitan varios ingredientes para crear un modelo de regresión logística. Lo primero es tener una función base, en este caso será la función lineal de regresión: $f(x)=B0+B1x$ agregando dos ingredientes adicionales:

1. Una función de enlace o activación que garantiza que la salida está entre 0 y 1, que es lo que se busca para indicar si termina el plan de estudios en cuatro años o no.

2. Una función de error para minimizar, adecuada al tipo de datos binario de la salida.

Zamora (2020), en el *notebook* 1 de la semana tres del módulo de aprendizaje automático, menciona que existe una familia entera de funciones llamadas sigmoides. Se les llama de esa manera porque al graficarlas producen una línea en forma de S. La función logística es una de estas y para efectos del aprendizaje automático se le llama solo función sigmoide. La misma tiene la siguiente definición:

$$sig(x)=1/1+e-x.$$

Y la derivada:

$$sig'(x)=sig(x)(1-sig(x)).$$

Cuando se grafica la función sigmoide se tiene la siguiente gráfica:

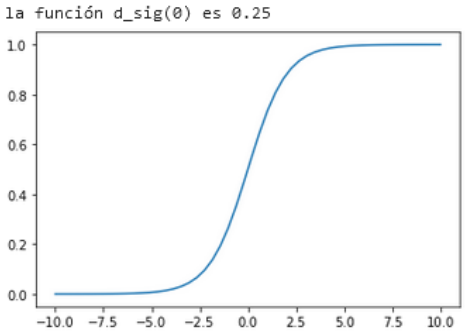


Figura 30

Gráfica de una función sigmoide generada en Python con la librería matplotlib

Fuente: Zamora (2020).

Zamora (2020) indica que el segundo ingrediente para hacer regresión logística es tener una función de error que tome el lugar del MSE y que sea más adecuada para una respuesta binaria. Para esto, se denota la variable Y como un valor original obtenido del *data-set* (siguiente el *data-set* de la UCR, 1=FinalizaATiempo, 0=NoFinalizaATiempo).

Se denota la variable *y*^ como una predicción realizada al modelo. A diferencia de *y* que puede tomar solo los valores 0 o 1, *y*^ puede tomar cualquier valor en el intervalo [0,1]. La idea al final es cuantificar el error de predecir *y*^ cuando el valor real es *y*.

$$\text{bce}(y,\hat{y}) = \begin{cases} -\log(\hat{y}) & \text{si } y = 1 \\ -\log(1-\hat{y}) & \text{si } y = 0 \end{cases}$$

El primer modelo generado es incapaz de predecir si un estudiante se gradúa o no en cuatro años tomando solo como variable la nota de admisión para ingreso a la Universidad, por lo tanto, el algoritmo de regresión logística con una sola variable (nota de admisión) afirma que no importa si un estudiante ingresa con una nota muy alta a carrera o si por el contrario ingreso con una nota baja, con la que apenas le alcanza para concursar según el corte de ingreso para un año en específico. Es importante aclarar que los cortes de ingreso a una carrera varían año tras año, por lo que un año puede haber un corte bastante alto y otro año puede que el corte sea mucho más bajo. Sin embargo, utilizando todos los datos provistos por la universidad, hay estudiantes que se graduaron en cuatro años con notas de ingreso bajas mientras que otros estudiantes que ingresaron con un corte muy alto no lograron finalizar en los cuatro años.

```
# Retorna la etiqueta APROBADO/REPROBADO
def label_finaliza(prob):
    return 'FINALIZA EN 4 AÑOS' if prob > 0.5 else 'NO TERMINA EN 4 AÑOS'

# HACEMOS ALGUNA PRUEBAS ALEATORIAS CON DATOS FICTICIOS DESPUËS DE ENTRENAR EL MODELO DE RL
print('---')
print('Rolando sacó 300 en el examen de admisión: ', label_finaliza(reg_log.predict([[300,1]])))
print('Juan sacó 650 en el examen: '               , label_finaliza(reg_log.predict([[650,3]])))
print('Sofía sacó 525 en el examen: '              , label_finaliza(reg_log.predict([[525,2]])))
print('Natalia 800 en el examen: '                 , label_finaliza(reg_log.predict([[800,2]])))
print('---')
```

Figura 31

Pruebas hechas al modelo de regresión logística solo con la nota de admisión

Se genera el modelo de regresión logística multivariable, sin embargo, es importante observar la matriz de correlaciones entre todas las variables que se tienen para determinar si hay alguna variable que tenga una correlación alta con otra. De esta manera, se puede evitar incluirla en el modelo, ya que no mejorará la

exactitud del modelo. La Figura 32 representa la matriz de correlaciones con los datos provistos por la universidad en términos de admisión y registro:

	EdadIngreso	NotaIngreso	Corte	PLAN_ESTUDIOS_A_TIEMPO
EdadIngreso	1.00	-0.01	-0.05	0.10
NotaIngreso	-0.01	1.00	0.82	0.02
Corte	-0.05	0.82	1.00	-0.09
PLAN_ESTUDIOS_A_TIEMPO	0.10	0.02	-0.09	1.00

Figura 32

Matriz de correlaciones con los datos de admisión

Como se puede observar en la Figura 32 la variable *corte* está estrechamente relacionada con la *notaIngreso* y tiene mucho sentido, ya que el estudiante logrará un ingreso a carrera siempre y cuando le alcance el corte de ese año. De este análisis de variables se omiten automáticamente las variables con alta correlación, ya que no servirán para mejorar el modelo. Por ejemplo, se puede alimentar un modelo con la provincia siempre y cuando esta variable categórica se pase a un arreglo del tipo *one-hot-encoding*, solo para verificar si en términos de ubicación geográfica ciertas provincias tienen un comportamiento favorable una de la otra. Sin embargo, el alimentar también el modelo con el *IdProvincia* no causaría ningún tipo de cambio en los resultados de los modelos generados, como se pudo observar al entrenar el modelo aplicando la variable convertida a *one-hot-encoding*.

	EdadIngreso	NotaIngreso	Corte	PLAN_ESTUDIOS_A_TIEMPO	Provincia_1	Provincia_2	Provincia_3	Provincia_4	Provincia_5	Provincia_6	Provincia_7
EdadIngreso	1.00	-0.01	-0.05	0.10	0.02	-0.04	-0.03	0.16	-0.05	0.02	-0.01
NotaIngreso	-0.01	1.00	0.82	0.02	0.05	-0.03	0.13	-0.15	-0.03	-0.01	-0.12
Corte	-0.05	0.82	1.00	-0.09	0.01	-0.01	0.19	-0.14	-0.01	0.00	-0.18
PLAN_ESTUDIOS_A_TIEMPO	0.10	0.02	-0.09	1.00	0.06	-0.01	-0.08	-0.01	-0.02	0.14	0.02
Provincia_1	0.02	0.05	0.01	0.06	1.00	-0.04	-0.39	-0.04	-0.02	-0.04	-0.15
Provincia_2	-0.04	-0.03	-0.01	-0.01	-0.04	1.00	-0.20	-0.02	-0.01	-0.02	-0.07
Provincia_3	-0.03	0.13	0.19	-0.08	-0.39	-0.20	1.00	-0.19	-0.09	-0.17	-0.71
Provincia_4	0.16	-0.15	-0.14	-0.01	-0.04	-0.02	-0.19	1.00	-0.01	-0.02	-0.07
Provincia_5	-0.05	-0.03	-0.01	-0.02	-0.02	-0.01	-0.09	-0.01	1.00	-0.01	-0.03
Provincia_6	0.02	-0.01	0.00	0.14	-0.04	-0.02	-0.17	-0.02	-0.01	1.00	-0.07
Provincia_7	-0.01	-0.12	-0.18	0.02	-0.15	-0.07	-0.71	-0.07	-0.03	-0.07	1.00

Figura 33

Matriz de correlaciones aplicando one-hot-encoding en la variable provincia

Según lo anterior, eliminar estas variables categóricas evitan saturar el modelo, además de que no impactarán, de manera positiva, en la calidad y exactitud del modelo. Se prueba un segundo modelo, en este caso multivariable, donde todos los resultados son *No logrará finalizar en cuatro años*, al igual que el primer modelo que se llevó a cabo en un principio tomando solo como base la nota de admisión.

```
# HACEMOS ALGUNA PRUEBAS ALEATORIAS CON DATOS FICTICIOS DESPUÉS DE ENTRENAR EL MODELO DE RL
print('---')
print('Rolando de 29 años sacó 300 en el examen de admisión e invicto en mate y progra: ', label_finaliza(reg_log.predict([[300,1,1]])))
print('Juan de 19 años sacó 650 en el examen e invicto en mate y progra: '               , label_finaliza(reg_log.predict([[650,3,1]])))
print('Sofía de 25 años sacó 525 en el examen e invicta en mate y progra: '              , label_finaliza(reg_log.predict([[525,2,1]])))
print('Natalia de 22 años sacó 800 en el examen e invicta en mate y progra: '                 , label_finaliza(reg_log.predict([[800,2,1]])))
print('---')
```

```
433  /  499  clasificados correctamente en aprendizaje
---
Rolando de 29 años sacó 300 en el examen de admisión e invicto en mate y progra:  NO TERMINA EN 4 AÑOS
Juan de 19 años sacó 650 en el examen e invicto en mate y progra:  NO TERMINA EN 4 AÑOS
Sofía de 25 años sacó 525 en el examen e invicta en mate y progra:  NO TERMINA EN 4 AÑOS
Natalia de 22 años sacó 800 en el examen e invicta en mate y progra:  NO TERMINA EN 4 AÑOS
```

Figura 34

Pruebas hechas al modelo de regresión logística multivariable

La razón por la cual en todos los casos evaluados afirma que no se logrará finalizar en cuatro años la carrera no tiene que ver con que esté prediciendo este escenario (falso negativo), sino que más bien la especificidad del modelo es muy baja y, por lo tanto, es poco confiable. Si se toma en cuenta que de los 1057 estudiantes, 740 entrenaron el modelo, se puede ver que el modelo de regresión lineal generado devuelve los siguientes números:

	FALSE	TRUE
0	256	0
1	0	61

Figura 35

Resultados de las probabilidades hechas al modelo de regresión logística multivariable

Con base en los resultados que se pueden ver en los números anteriores, los cuales provienen de llevar a cabo una partición de datos 70/30, donde los datos para probar son 317 estudiantes, se tienen 256 falsos negativos, lo que indica que casi en su totalidad el algoritmo siempre indicará que el estudiante no finalizará a tiempo. Esto es falso si solo se toman como variables de entrenamiento la nota de admisión, la edad del estudiante, el corte de la carrera para un año en específico y las variables categóricas (colegio de procedencia, provincia, cantón o distrito de donde proviene).

Estos mismos resultados los entregó la implementación del algoritmo de árboles de decisión. Esto significa que las variables de entrenamiento son poco significativas y se tiene que trabajar con los expedientes académicos para lograr algún resultado diferente.

El siguiente paso es agregar al *data-set* las notas de cursos del expediente académico que permitan mejorar el modelo y verificar si con base en estas nuevas variables se puede predecir de mejor forma si un estudiante finalizará o no en 4 años. Las variables por agregar son las notas de los cursos de Matemática Discreta, Cálculo y Métodos Numéricos, Programación I y Programación II. El propósito de esto es determinar si estos cursos que suelen tener un nivel de dificultad medio o difícil en el plan de estudios permiten llevar a cabo una predicción de calidad para saber si un estudiante logrará graduarse o no en el tiempo establecido.

Carné	INICIO_DISCRETAS	FIN_CALC	STATUS_MAT	INICIO_P1	FIN_P1	STATUS_PROG	INVICTO	MA0320	NotaDisc	MA0321	NotaCalc	IF2000	NotaP1	IF3000	NotaP2
985141	1998	2030	NOTERMINAPORMATES	1998	1999	VABIENPROGRA	NOINV	MA0320	7,5	MA0321	0	IF2000	7,5	IF3000	7,5
A22533	2005	2030	NOTERMINAPORMATES	2005	2006	VABIENPROGRA	NOINV	MA0320	0,0	MA0321	6,5	IF2000	7,0	IF3000	7,0
A53411	2000	2030	VABIENMATES	2010	2011	VABIENPROGRA	INV	MA0320	0	MA0321	0	IF2000	8,5	IF3000	9,5
A60516	2006	2030	NOTERMINAPORMATES	2006	2007	VABIENPROGRA	NOINV	MA0320	7,5	MA0321	0	IF2000	7,0	IF3000	7,0
A73491	2011	2012	VABIENMATES	2011	2012	VABIENPROGRA	INV	MA0320	8,0	MA0321	7,0	IF2000	8,5	IF3000	7,5
A81116	2008	2009	VABIENMATES	2008	2009	VABIENPROGRA	INV	MA0320	7,0	MA0321	7,0	IF2000	7,0	IF3000	9,5
A82506	2008	2009	VABIENMATES	2008	2015	NOTERMINAPORPROGRA	NOINV	MA0320	5,0	MA0321	7,5	IF2000	4,5	IF3000	7,0
A83787	2008	2012	NOTERMINAPORMATES	2009	2012	NOTERMINAPORPROGRA	NOINV	MA0320	2,0	MA0321	7,0	IF2000	7,5	IF3000	7,5
A85262	2011	2012	VABIENMATES	2012	2015	NOTERMINAPORPROGRA	NOINV	MA0320	7,0	MA0321	8,0	IF2000	2,0	IF3000	7,0
A86024	2009	2009	VABIENMATES	2009	2010	VABIENPROGRA	INV	MA0320	8,0	MA0321	7,5	IF2000	9,0	IF3000	8,0
A86024	2009	2009	VABIENMATES	2009	2010	VABIENPROGRA	INV	MA0320	8,0	MA0321	7,5	IF2000	9,0	IF3000	8,0
A90510	2010	2030	NOTERMINAPORMATES	2012	2030	NOTERMINAPORPROGRA	NOINV	MA0320	2,0	MA0321	4,5	IF2000	5,5	IF3000	3,5
A90782	2009	2009	VABIENMATES	2009	2030	NOTERMINAPORPROGRA	NOINV	MA0320	7,0	MA0321	7,5	IF2000	7,5	IF3000	0
A90859	2011	2011	VABIENMATES	2011	2012	VABIENPROGRA	INV	MA0320	4,0	MA0321	7,0	IF2000	7,5	IF3000	8,0
A91266	2009	2011	NOTERMINAPORMATES	2009	2011	NOTERMINAPORPROGRA	NOINV	MA0320	2,0	MA0321	8,0	IF2000	8,0	IF3000	7,5
A91454	2012	2013	VABIENMATES	2012	2013	VABIENPROGRA	INV	MA0320	8,0	MA0321	7,0	IF2000	8,0	IF3000	7,5
A92801	2010	2013	NOTERMINAPORMATES	2010	2011	VABIENPROGRA	NOINV	MA0320	5,0	MA0321	0	IF2000	9,0	IF3000	9,5
A95021	2010	2013	NOTERMINAPORMATES	2011	2012	VABIENPROGRA	NOINV	MA0320	1,0	MA0321	8,0	IF2000	9,0	IF3000	7,0

Figura 36

Extracto de notas obtenidas en los cursos de Matemática Discreta, Cálculo, Programación I y Programación II por algunos estudiantes

Fuente: Datos provistos por la Universidad de Costa Rica.

La Figura 36 es un pequeño extracto de la lista de estudiantes y sus correspondientes notas en los cursos que se usarán para evaluar los modelos indicados previamente, los datos que se observan fueron extraídos de una consulta SQL que se tuvo que llevar a cabo para generar el conjunto de datos idóneo. Originalmente, los datos provistos por la Universidad venían como se puede observar en la Figura 37:

Carné	Sigla	Nombre del curso	Nota	Pertenece a plan	Período	Año
B700	IF7101	COMPROMISO SOCIAL DE LA INFORMÁTICA	10,0	Si	1	2021
B700	MA0323	MÉTODOS NUMÉRICOS	07,5	Si	1	2021
B700	IF7100	INGENIERÍA DE SOFTWARE	07,0	Si	1	2021
B700	IF3001	ALGORITMOS Y ESTRUCTURAS DE DATOS	09,0	Si	3	2020
B700	MA0322	ÁLGEBRA LINEAL	08,0	Si	2	2020
B700	IF6100	ANÁLISIS Y DISEÑO DE SISTEMAS	08,0	Si	2	2020
B700	IF3001	ALGORITMOS Y ESTRUCTURAS DE DATOS	04,0	Si	1	2020
B700	IF5100	ADMINISTRACIÓN DE BASES DE DATOS	08,0	Si	1	2020
B700	IF4101	LENGUAJES PARA APLICACIONES COMERCIALES	01,0	Si	1	2020
B700	IF4101	LENGUAJES PARA APLICACIONES COMERCIALES	02,5	Si	3	2019
B700	IF4000	ARQUITECTURA DE COMPUTADORES	07,0	Si	2	2019
B700	MA0322	ÁLGEBRA LINEAL	01,0	Si	2	2019
B700	IF4100	FUNDAMENTOS DE BASES DE DATOS	07,0	Si	2	2019
B700	IF3000	PROGRAMACIÓN II	09,5	Si	1	2019
B700	IF3001	ALGORITMOS Y ESTRUCTURAS DE DATOS	05,0	Si	1	2019
B700	MA0321	CÁLCULO DIFERENCIAL E INTEGRAL	07,0	Si	3	2018

Figura 37

Extracto de expediente académico de un estudiante con el carné maquillado para efectos de sensibilidad de los datos

Fuente: Datos provistos por la Universidad de Costa Rica.

Los datos están listados por curso y por carné, por lo que se tuvieron que pivotear ciertos valores para observar los cursos que se necesitan en forma de columna o variable.

	EdadIngreso	NotaIngreso	Corte	INICIO_DISCRETAS	FIN_CALC	INICIO_P1	FIN_P1	PLAN_ESTUDIOS_A_TIEMPO
EdadIngreso	1.00	-0.01	-0.05	-0.25	0.10	-0.23	-0.06	0.10
NotaIngreso	-0.01	1.00	0.82	0.08	0.07	0.08	-0.05	0.02
Corte	-0.05	0.82	1.00	0.18	0.16	0.19	0.03	-0.09
INICIO_DISCRETAS	-0.25	0.08	0.18	1.00	0.30	0.94	0.38	-0.15
FIN_CALC	0.10	0.07	0.16	0.30	1.00	0.34	0.51	-0.22
INICIO_P1	-0.23	0.08	0.19	0.94	0.34	1.00	0.43	-0.18
FIN_P1	-0.06	-0.05	0.03	0.38	0.51	0.43	1.00	-0.24
PLAN_ESTUDIOS_A_TIEMPO	0.10	0.02	-0.09	-0.15	-0.22	-0.18	-0.24	1.00

Figura 38

Matriz de correlaciones usando las notas de Matemática Discreta, Cálculo, Programación I y Programación II

Al evaluar los modelos de regresión logística y árboles de decisión contemplando los cursos del expediente académico se obtuvo un mejor resultado,

esto se debe a que las notas de los cursos permiten saber en términos generales cuando un estudiante va invicto o no en los cursos de la rama de la matemática y que suelen ser los que se pierden en mayor cantidad durante el desarrollo de la carrera.

La malla curricular de la carrera contiene cursos de matemática, desde primer año hasta el cuarto año. Los cursos son Matemática Discreta, Cálculo Diferencial e Integral, Álgebra Lineal, Métodos Numéricos, Economía y Métodos Cuantitativos. Aunado a esto, también se evaluaron cursos de programación I y programación II para ver si son cursos que suelen repercutir en el desarrollo de la carrera de las personas estudiantes.

Con los resultados se mejoró el modelo, sin embargo, la cantidad de estudiantes que no lograron sacar la carrera en 4 años es considerable, de una muestra de 191 registros que llevaban sin perder una sola vez los cursos de matemáticas y que iban invictos en Programación I y Programación II solo 57 lograron sacar la carrera en los 4 años. Esto es un dato muy importante, ya que indica que puede haber muchísimos más factores que provoquen que un estudiante se atrase en completar su plan de estudios. Muchas pueden ser las causas de no graduarse a tiempo incluso cuando van bien en los cursos mencionados, por ejemplo, perder los cursos de redes y comunicaciones de datos; otros estudiantes pueden perder los cursos de bases de datos, los cuales son dos; otros pueden atrasarse solo por el simple hecho de perder cursos de humanidades o cursos de servicio.

Capítulo VI. Conclusiones y recomendaciones

6.1. Conclusiones

Se estableció en un principio que se identificaría el o los procedimientos que tiene la Universidad actualmente para conocer el rendimiento académico y ayudar al estudiantado durante todo el desarrollo del plan de estudios de la carrera. Como conclusión, la Universidad de Costa Rica dispone de un módulo de reportes desde su portal universitario en el cual los coordinadores de carrera tienen acceso y permite sacar listados de estudiantes por sede o recinto y observar el histórico de cursos, fechas de graduación, datos de ingreso en primer año, entre muchos otros datos personales por estudiante.



Figura 39

Portal de la Universidad de Costa Rica

Fuente: Universidad de Costa Rica.

El detalle es que estos reportes se generan en formato de Microsoft Excel, por lo que el procedimiento de análisis se vuelve un poco lento y también se dispone solo de los datos que ya se encuentran cargados previamente en el sistema. No hay manera de que un coordinador de carrera pueda extraer algún

dato que no esté en el portal inmediatamente, por lo que tiene que llamar a la Universidad, específicamente al Centro de Informática para solicitar algún cambio o nuevo dato. Este proceso puede demorar incluso meses, ya que para esta investigación se solicitaron unos datos que no estaban y ese fue el tiempo de respuesta.

Otro detalle es que los datos se generan de manera cruda, según el reporte que se escoja y no se dispone de un *dashboard* o módulo de visualización de estos en el cual se puedan filtrar o llevar a cabo gráficas.

A partir de esto, se plantean las conclusiones específicas por objetivos y se logra justificar el porqué de los algoritmos predictivos que se escogieron (árboles de decisión y regresión logística) y que permitieron adecuar modelos con los datos que se tienen disponibles mediante tecnologías de alto uso. En este caso se generaron modelos en Python que respondían a la pregunta: ¿Se gradúa un estudiante en 4 años o no? Los algoritmos mencionados se utilizan cuando se necesita predecir el resultado de una variable categórica, en este caso *Sí* o *No*.

De esta manera, se concluye que el tercer objetivo que se planteó se cumplió al usar diferentes técnicas de algoritmos predictivos que permitieran desarrollar eventualmente una propuesta institucional en beneficio del estudiantado, aunque los resultados no fueron los que se esperaba. Se concluye que los patrones descubiertos en los datos, mediante el análisis de los resultados de los algoritmos que se emplean y que permiten indagar los factores que intervienen en el rendimiento académico del estudiantado durante su carrera universitaria, son muy pobres. Esto se debe a que el hecho de que un estudiante no se gradúe en 4 años puede tener decenas de causas e ir bien en cursos de matemáticas o en ciertos cursos de programación no garantiza finalizar en 4 años.

Al evaluar los algoritmos de aprendizaje automático y probar los resultados se observa que cualquier estudiante puede no graduarse incluso por el simple hecho de perder un curso como un repertorio o un seminario de realidad nacional o por haber tenido alguna situación imprevista en alguno de los semestres que le impidieran matricular regularmente. El simple hecho de atrasar algún curso

también se traduce en un retraso anual, ya que la mayoría son anuales y no semestrales.

Los algoritmos permitieron observar que la cantidad de estudiantes graduados en 4 años aumenta cuando no se pierde un curso de matemáticas y ningún curso de programación, sin embargo, para el total de estudiantes que se tienen del 2010 al 2021 los resultados no son determinantes. Por esto, solo un bajo porcentaje de estudiantes se gradúa tomando como punto crítico este par de aristas, por ejemplo, aprobar todas las matemáticas, así como Programación I y Programación II.

Como conclusión general, los algoritmos predictivos son una herramienta eficaz que permite, con base en un entrenamiento correcto, predecir respuestas o comportamientos. En este caso lamentablemente los modelos generados no fueron capaces de predecir qué es lo que más influye para que un estudiante se gradúe o no en 4 años utilizando como variables de entrenamiento todos los datos provistos por la universidad. De esta manera, no se puede pensar todavía en llevar a cabo una propuesta institucional que busque mejorar el rendimiento académico de las personas estudiantes.

Es importante destacar que no por esto no se pueden generar recomendaciones o buscar puntos de mejora, ya que, aunque los modelos hayan tenido una exactitud baja, siempre se pueden optimizar o seguir alimentando con muchas más variables. Lo anterior para seguir buscando algún tipo de patrón que logre un mejor índice de aprobación del plan de estudios en el plazo establecido.

6.2. Recomendaciones

A continuación, se presentan algunas de las recomendaciones obtenidas de esta investigación:

- Como primera recomendación se puede solicitar al Departamento de Informática de la Universidad de Costa Rica un plan de mejora del portal universitario que se tiene actualmente para que el módulo de reportes y extracción de datos institucionales sea muchísimo más amigable con el

usuario y una serie de reportes predefinidos que muchas veces ayudan a ciertos departamentos universitarios, pero que quedan debiendo en muchos aspectos cuando llega algún otro tipo de requerimiento. Para efectos de esta investigación, se tuvo que llevar a cabo una solicitud para desplegar los expedientes académicos de las personas estudiantes matriculadas entre 2010 y 2021, ya que los datos que se tenían eran inconsistentes y tenían muchos otros datos faltantes. Esto puede lograrse al agendar una serie de reuniones entre coordinaciones de carrera, escuelas y el equipo encargado de desarrollar estos módulos.

- Respecto al segundo y tercer objetivo que se plantearon en esta investigación, se puede recomendar implementar algún o algunos algoritmos predictivos extra que permitan comparar los resultados entre todos y comprobar si las variables que se implementan generan modelos pobres o si por el contrario alguno reacciona mejor en términos de resultados.

- El cuarto objetivo de esta investigación plantea encontrar patrones en los datos que se utilizan para generar los modelos de predicción, sin embargo, aunque se concluye que no se puede determinar si un estudiante se logrará graduar o no en 4 años con los datos suministrados, se puede recomendar en una eventual propuesta institucional mantener o impulsar todavía más los talleres o *estudiaderos* que permitan mantener un rendimiento académico alto de parte de la mayoría de la población de Informática. Aunque las personas estudiantes que tuvieron un índice bajo de reprobación en cursos de matemática o programación no necesariamente se graduarán en 4 años, tuvieron una probabilidad más alta respecto de los que perdieron algún o algunos cursos de estas áreas.

- Desde el punto de vista técnico se recomienda usar herramientas que permitan el desarrollo rápido y fácil, pero siempre con calidad de algoritmos de minería de datos y aprendizaje automático como Google

Collab y Kaggle cuando no se dispone de licencias o no se puede pagar una herramienta como Pycharm.

- Se recomienda usar todas las librerías existentes, tanto en Python como en *R* que agilicen y garanticen que los algoritmos se desarrollan de manera correcta, sin tener que *experimentar* con código propio para ciertos tipos de cálculos.

Referencias

Adekitan, A. I. y Noma-Osaghae, E. (2019). *Data mining approach to predicting the performance of first year student in a university using the admission requirements*. https://doi.org/https://doi-org.ezproxy.sibdi.ucr.ac.cr/10.1007/s10639-018-9839-7

Aguilar, L. J. (2000). *Programación en C++*. McGraw-Hill. https://www.ingebook.com/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=4366

Arenas Landínez, A. y Jaimes Luna, B. (2008). Calidad y Competencias: Propuesta de un Modelo Educativo en Educación Superior. *Revista UIS Ingenierías, 7*(1), 87-103. https://www.redalyc.org/pdf/5537/553756884007.pdf

Arias Odón, F. G. (2012). *El proyecto de investigación*. https://doi.org/10.2307/j.ctt2050wjh.5

Asif, R.; Merceron, A.; Ali, S. A. y Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education, 113*, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

Barrantes Echavarría, R. (1999). *Investigación un camino al conocimiento*. UNED. https://ebooks.uned.ac.cr/product/investigacin-un-camino-al-conocimiento50118494

Bishop, C. M. (1995). *Building Neural Network for Pattern Recognition*. Clarendon Press. https://doi.org/10.1109/RusAutoCon49822.2020.9208207

Devasia, T.; Vinushree, T. P. y Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 91-95. IEEE.

https://doi.org/10.1109/SAPIENCE.2016.7684167

Galán, V. (2015). *Crisp-Dm a un proyecto de minería*. Universidad Carlos III de Madrid. https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Geron, A. (2019). *Hands-On Machine Learning with ScikitLearn Keras and TensorFlow*. En N. Tache (Ed.), *Hands-On Machine Learning with R*. O'really Publisher. https://doi.org/10.1201/9780367816377

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition, 5*(2), 199-220. https://doi.org/https://doi.org/10.1006/knac.1993.1008

Harrington, P. (2015). Machine Learning in Action. En *Efficient Learning Machines*. Manning. https://doi.org/10.1007/978-1-4302-5990-9_11

Istanbullu, A. y Guruler, H. (2013). Educational Data Mining. En H. Guruler y A. Istanbullu (Eds.), *Educational Data Mining* (pp. 105-124). Springer, Cham. https://doi.org/https://doi-org.ezproxy.sibdi.ucr.ac.cr/10.1007/978-3-319-02738-8_4

Jiménez Padilla, R. y Cuevas Covarrubias, C. (2011). Curvas ROC y Vecinos Cercanos, Propuesta de un Nuevo Algoritmo de Condensación. *Revista de Matemática: Teoría y Aplicaciones, 18*(1), 21. https://doi.org/10.15517/rmta.v18i1.2112

León, A. (2007). Educación. *Red de Revistas Científicas de América Latina, El Caribe, España y Portugal, 11*, 595-604. https://doi.org/10.4272/978-84-9745-427-8.ch9

Levin, R. y Rubin, D. (2004). *Estadística para administración y economía*. Pearson Education.

Naranjo Zeledón, L. (2020). Investigación en Informática: el enfoque alternativo. *Revista Digital Technology Inside by CPIC, 5*, 15. https://cpic-sistemas.or.cr/revista/index.php/technology-inside/article/view/35

Riquelme, J. C.; Ruiz, R. y Gilbert, K. (2006). minería de datos: Conceptos y tendencias. *Revista Iberoamericana de inteligencia artificial, 29*(29), 11-18. https://core.ac.uk/download/pdf/51405945.pdf

Salazar, E. y Tenorio, G. (2020). *Evaluación de la precisión de los modelos de clasificación binaria para la identificación de noticias verdaderas o falsas en Costa Rica*. Universidad Cenfotec.

Sandoval Serrano, L. J. (2018). *Algoritmos de aprendizaje automático para análisis y predicción de datos*. http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Tsiakmaki, M.; Kostopoulos, G.; Kotsiantis, S. y Ragos, O. (2020). Implementing autoML in educational data mining for prediction tasks. *Applied Sciences (Switzerland), 10*(1), 1-27. https://doi.org/10.3390/app10010090

Zamora, J. (2020). *Regresión Logística*.