



UNIVERSIDAD POPULAR AUTÓNOMA DEL ESTADO DE PUEBLA

DECANATURA DE INGENIERÍAS

Aplicación web para la detección oportuna de la deserción universitaria aplicando ciencia de datos

Trabajo práctico que presenta

Luis Andrés Méndez Ortega

Matrícula: 17970007

Para obtener el grado de
Maestro en Ciencia de Datos e
Inteligencia de Negocios

Dra. Argelia Berenice
Urbina Nájera
Directora de Trabajo Práctico

Dr. Raúl Cruz Barbosa
Co-Director de Trabajo Práctico

Mayo 2020



UPAEP – Secretaría General

Dirección General de Apoyos Académicos

Dirección del Centro de Recursos para el Aprendizaje y la Investigación.

Biblioteca Central - **Karol Wojtyla**

Tesis Digitales Restricciones de uso:

DERECHOS RESERVADOS ©

PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de textos, imágenes, gráficas, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente de donde la obtuvo mencionando el autor o autores involucrados en el documento.

Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Contenido

Carta de liberación del comité	1
Carta de liberación de Trabajo Práctico	2
Agradecimientos	3
Resumen	4
Introducción	6
1 Propósito y organización	7
1.1 Planteamiento del problema	7
1.2 Justificación	9
1.3 Objetivos del trabajo práctico.....	11
1.4 Alcances y limitaciones.....	11
1.5 Propuesta metodológica	12
1.5.1 Comprensión.....	13
1.5.2 Modelado	14
1.5.3 Despliegue en web.....	18
1.5.4 Validación	19
1.5.5 Herramientas de software	20
1.6 Organización del proyecto	21
2 Estado del arte y marco teórico.....	22
2.1 Deserción escolar	22
2.2 Minería de datos.....	25
2.2.1 Proceso KDD	29
2.2.2 Selección de atributos	30
2.2.3 Evaluación de desempeño	30
2.2.4 Árboles de decisión.....	32
2.2.5 Redes neuronales artificiales	32

2.3 Modelo de prototipos	32
2.3.1 Evaluación de usabilidad.....	34
3 Sistema embebido	35
3.1 Descripción de la propuesta	35
3.1.1 Requisitos funcionales y no funcionales	37
3.1.2 Producto esperado	39
4 Prototipo	40
4.1 Interfaz inicial	42
4.2 Instrumento de evaluación.....	43
4.2 Presentación de resultados	46
4.3 Ajustes al prototipo	49
5 Resultados	51
5.1 Conjunto de datos utilizado.....	51
5.2 Atributos explicativos	60
5.3 Modelos de clasificación	65
5.4 Implementación	76
5.5 Evaluación	83
6 Conclusiones y trabajo futuro	85
Referencias	88
Anexo 1 Ejemplo de instrumento tipo A 1.1.....	94
Anexo 2 Ejemplo de instrumento tipo B 1.1.....	96
Anexo 3 Ejemplo de instrumento tipo C 1.1	97
Anexo 4 Ejemplo de instrumento tipo D 1.1	101
Anexo 5 Reglas generadas por árbol de decisión	102
Anexo 5 Manual de usuario.....	103

Ciudad de Puebla, a 21 de mayo de 2020

Dra. Ma. del Rocío Guadalupe Morales Salgado

Directora de Programa Educativo

Maestría en Ciencia de Datos e Inteligencia de Negocios

UPAEP

PRESENTE

Asunto: Aprobación para digitalización
de trabajo práctico

Los que suscriben la presente, en calidad de comité asignado al **trabajo práctico** del Lic. **Luis Andrés Méndez Ortega**, estudiante de la Maestría en **Ciencia de Datos e Inteligencia de Negocios** con número de matrícula **17970007**, le informamos que hemos revisado el trabajo práctico titulado **Aplicación web para la detección oportuna de la deserción universitaria aplicando ciencia de datos**. Como resultado de esta revisión consideramos que el trabajo práctico cumple con los requisitos necesarios para ser defendido con el propósito de obtener el grado de Maestro en **Ciencia de Datos e Inteligencia de Negocios**, por lo que aprobamos que se realice la digitalización correspondiente del documento y por ende la programación del examen de grado.

La saludamos afectuosamente, anticipando nuestro agradecimiento por la atención a la presente.

Cordialmente,


Dra. Argelia Berenice Urbina Nájera
Directora de trabajo práctico


Dr. Raúl Cruz Barbosa
Co-director de trabajo práctico


Mtra. Mónica Sissy Vela Larios
Asesor


POSGRADOS
17 SUR 901
Barrio de Santiago
Puebla, Pue. México
C.P. 72410

Tel.: 01 (222) 229 9400
Fax: 01 (222) 2325251
01 800 224 2200
www.upaep.mx

U
P
A
E
P

Carta de liberación de Trabajo Práctico

Carta de Liberación de Trabajo Práctico

H. Puebla de Zaragoza, México, a 31 de mayo del 2020

Dr. Argelia Berenice Urbina Nájera
Dr. Raúl Cruz Barbosa
Mtra. Mónica Sissy Vela Larios
Comité de trabajo práctico
UPAEP
PRESENTE

Por medio de la presente, se informa que el alumno **Luis Andrés Méndez Ortega Nombre**, estudiante de la Maestría en Ciencia de Datos e Inteligencia de Negocios, con número de matrícula 17970007; ha concluido satisfactoriamente el proyecto titulado:

**“Aplicación web para la detección oportuna de la deserción
universitaria aplicando ciencia de datos”**

el cual desarrolló en el departamento, mismo que satisface los requisitos solicitados y cuya implementación ha contribuido a mejorar el seguimiento estudiantil.

Sin otro particular, me despido, quedando a su disposición para cualquier aclaración.

Atentamente



Lic. Gabriela Guadalupe Rentería Montemayor
Jefe de área de Apoyo y Seguimiento al Estudiante
Universidad Popular Autónoma del Estado de Puebla

Agradecimientos

A mi esposa, a mi hija

A mis padres y hermanos y hermanas

A mi Jefa, compañeros y compañeras de trabajo

A mi directora de trabajo práctico y co-director, así como profesores

A la UPAEP

Resumen

La deserción de un periodo a otro es un desafío que las instituciones de educación superior tienen que enfrentar tratando de identificar aquellas causas que lo originan. En este sentido UNIVR, una Universidad de Educación Superior en México ha realizado un esfuerzo para afrontar este reto implementando aprendizaje computacional y análisis histórico de datos de tres decanatos académicos de licenciatura para pronosticar posibles deserciones. La información analizada incluyó datos de estudiantes del 2010 al 2019, recolectando un total de 51,497 instancias con 26 atributos. Se implementaron *redes neuronales artificiales* y *árboles de decisión* como algoritmos de clasificación, también, se usaron algoritmos de selección de atributos y métodos de re-muestreo para el desbalanceo de clase. Se propuso e implementó un sistema embebido en la plataforma de *Inteligencia de Negocio* de UNIVR para que el departamento de seguimiento estudiantil tenga una herramienta web que le permita monitorear de una manera rápida, efectiva y única el estatus académico del estudiante con un enfoque predictivo. En el desarrollo de la aplicación se utilizó el *modelo de prototipos* y *evaluación heurística* y *recorrido cognitivo* como metodologías para medir usabilidad y funcionalidad. La arquitectura propuesta permite de manera *off-line* mantener los modelos actualizados una vez al periodo escolar y la clasificación de las predicciones diariamente. El modelo con mejor desempeño fue el de árboles de decisión con un coeficiente de correlación Matthew del 87.43% vs 53.39% de redes neuronales. Además, la técnica de selección de atributos por Envoltura, dio los mejores atributos *con random-forest*. Aunque solo se utilizaron datos principalmente académicos e históricos, se obtuvieron resultados satisfactorios.

Palabras clave

Pronóstico, arboles de decisión, atributos, balance

Abstract

Dropout from one period to another is a challenge that higher education institutions have to face the treatment of identifying the causes that originate it. In this sense, UNIVR, a University of Higher Education in Mexico, has made an effort to face this challenge, implementing computational learning and historical analysis of data from three undergraduate academic deanships to predict possible dropouts. The analyzed information included student data from 2010 to 2019, collecting 51,497 total instances with 26 attributes. Artificial neural networks and decision trees implemented as classification algorithms, also, attribute selection algorithms and resampling methods used for class imbalance. A system embedded in the UNIVR Business Intelligence platform was proposed and implemented so that the students monitoring department has a web tool that allows it to quickly, effectively and uniquely monitor the student's academic status with a predictive approach. In the development of the application, the prototype model and heuristic evaluation and cognitive journey used as methodologies to measure usability and functionality. The proposed architecture allows off-line mode to keep the models updated once a school period, and the classification of the predictions daily. The best performing model was the decision tree model with a Matthew correlation coefficient of 87.43% vs. 53.39% for neural networks. In addition, the attribute selection technique by wrapper gives the best attributes with random-forest. Although only mainly academic and historical data were used obtained satisfactory results.

Keywords

Forecast, decision tree, attributes, balance

Introducción

En la actualidad, la OCDE (Organización para la Cooperación y el Desarrollo Económicos) indica que la sociedad mexicana tiene el gran reto de disminuir la deserción ya que se encuentra en los primeros países con esta problemática (Díaz, 2012).

De acuerdo con Alonso, Blanco, Fernández y Solís (2014), disminuir la deserción estudiantil en la educación superior, impacta de manera positiva al fomentar una sociedad mejor preparada para enfrentar los retos globales en los que estamos inmersos, y mejorar la calidad de vida de las personas, obtener mejores empleos, salarios, posibilidades de crecimiento intelectual, etc. Los beneficios son para todos y, por lo tanto, no debe desestimarse cualquier esfuerzo orientado a reducir el abandono escolar.

La literatura sobre deserción académica universitaria, menciona que los primeros esfuerzos fueron encaminados a explicar los factores que desencadenan la deserción, sin embargo, en más recientes investigaciones, se sugiere aplicar los últimos avances en tecnologías de la información y comunicación (TIC's) y también en ciencia de datos para afrontar esta problemática (Ramírez, Espinosa y Millán, 2016). De esta manera, no solo se podrán detectar factores explicativos, si no también crear modelos predictivos para hacer pronósticos y tomar mejores decisiones.

Con el objetivo de contribuir a reducir la deserción Universitaria en una institución privada en el estado de Puebla (UNIVER), se propone el uso de algoritmos de aprendizaje computacional para crear un modelo embebido en su plataforma de inteligencia de negocios. Esta herramienta se va a sumar a las acciones encaminadas a ayudar a reducir el índice de deserción en sus programas de licenciatura de Ingeniería, Ciencias Sociales y Administrativas.

Para lograr este objetivo se propone la creación y utilización de modelos de árboles de decisión o redes neuronales los cuales han sido utilizados para tratar de pronosticar deserción con resultados prometedores (Ramírez y Grandón, 2018) y (Chinkes, 2017). Además, la implementación de una interfaz web puede permitir consultar la información clasificada y mostrar el análisis en forma de lista o expediente (por alumno), para su uso en el área de seguimiento estudiantil. De esta manera el usuario podrá tener la información una manera rápida, en línea y sin necesidad de estar consultando diferentes fuentes. Todo este análisis deberá procesarse de forma automática y mantenerse actualizada al día, considerando las políticas tecnológicas y de información de la institución.

1 Propósito y organización

De acuerdo al estudio realizado por López y Beltrán (2017), la deserción académica es el resultado de diversos factores como: la situación personal, calidad educativa, instalaciones, factores socio-culturales y económicos. Esto provoca que el estudiante tenga que interrumpir sus estudios y por lo tanto repercute en su vida académica, en la institución y también en la sociedad. Tinto (1989) sentó las bases para tratar de enfrentar esta problemática, y sugiere analizar los factores en perspectivas personales, institucionales y estatales.

Por otra parte, la Secretaría de Educación Pública (2016), en México, indica que la deserción académica es una problemática que las instituciones de educación superior deben afrontar para evitar que los jóvenes interrumpan sus estudios y mejorar el impacto social, económico y cultural del país.

Es por ello que, en este trabajo se busca la forma de identificar la posible deserción de un alumno antes de que suceda con el desarrollo de una aplicación que utilice algoritmos de aprendizaje computacional implementados en una universidad privada del estado de Puebla, la cual tiene actualmente un índice de deserción escolar del 20%.

El trabajo se apoya de redes neuronales y árboles de decisión ya que han sido utilizados en el análisis de factores que influyen en la deserción universitaria para crear modelos predictivos con resultados satisfactorios (Chinkes, 2017) y árboles de decisión (Ramírez y Grandón, 2018).

A continuación, se explica el problema, su justificación, objetivo general y los objetivos específicos, así, como también algunas limitantes y se expone el programa de actividades.

1.1 Planteamiento del problema

De acuerdo a Díaz (2012), en México el 38% de las personas que pueden acceder a estudios de educación superior **no egresan**¹, razón por la cual, la OCDE (Organización para la Cooperación y el Desarrollo Económicos) ubica a México y a Turquía (mismo porcentaje) como los países con un problema grave en abandono. Este porcentaje contrasta con el de países como Alemania y Finlandia donde tienen un 4.03% y 0.45% respectivamente.

¹ En su tiempo normal de duración del programa de estudios

Adicionalmente, los efectos de la deserción escolar se reflejan en una desigualdad laboral y social, porque la probabilidad de encontrar mejores empleos con mayores privilegios resulta permeada por este problema (Alonso et al., 2014). Según datos de la OCDE del 2018 el 85% de las personas con estudios de educación superior (25 a 64 años) tienen empleo contra el 75.2% de los que solo tienen educación medio superior, lo que deja ver que existen mayores oportunidades laborales para quienes avanzan y terminan sus estudios académicos. En el caso de México, esta proporción es de 80% para los que tienen estudios superiores y del 70.6% para estudios inferiores, encontrándose así por debajo de la media de la OCDE. Además, según el reporte de Educación Superior en México (2018), por la OCDE, la contratación de jóvenes entre 25-34 años con educación superior es del 80.7%, el cual es menor al 84.1% del promedio de los demás países miembros.

Rodríguez y Hernández (2008), mencionan que algunas causas de la deserción escolar universitaria giran en torno a la conjunción de factores socio-económicos, desempeño escolar e integración de los estudiantes, sin embargo, Tinto (1989), señala que es un tema complejo y recomienda definir la deserción en términos de perspectivas, como la del estudiante, institucional y estatal y/o nacional y así, encontrar los factores que mejor expliquen el problema.

En un contexto nacional, el Gobierno Mexicano implementó una reforma educativa en el 2012 haciendo obligatoria la educación media superior, dentro de sus objetivos estuvo contrarrestar los altos niveles de deserción (Secretaría de Educación Pública de Educación Media Superior, 2012), lo cual repercute positivamente en la educación superior.

Según el reporte de Principales Cifras del Sistema Educativo Nacional emitido por la Secretaría de Educación Pública en 2016, indica que a nivel nacional el índice de abandono en educación superior fue de 6.9% en el ciclo escolar 2014-2015 y de 6.8% para el 2015-2016 (modalidad escolarizada). Para el estado de Puebla el índice de abandono fue de 10.7% para el 2014-2015 y de 10.5% en el 2015-2016 ubicando al estado en el octavo lugar con mayor porcentaje de abandono, de esta manera, se encuentra en iguales condiciones con Zacatecas y sólo por debajo de Tabasco 10.8%, Quintana Roo 11.1%, Sonora 11.1%, Veracruz 11.5%, Campeche 11.8% y Baja California Sur 11.8% (Secretaría de Educación Pública, 2016). En este sentido en Puebla, aunque redujo del 10.7 % al 10.5%, no llegó a la media nacional del 6.8 % en 2016.

Dentro de este escenario, UNIVR es una universidad privada en Puebla que, como cualquier institución de educación superior, tiene que afrontar esta realidad. Presenta índices de deserción del 20% en los decanatos de *Ingenierías*, *Ciencias Sociales* y *Administrativas*, lo cual motiva a la

universidad a dar seguimiento de los alumnos por medio de un departamento para el apoyo del estudiante y a la creación de indicadores de deserción escolar. Dentro de las acciones remediales ha implementado indicadores con diferentes ópticas como periodos por cohorte, decanatos y por programa académico, además ha utilizado encuestas de baja para obtener información sobre las posibles causas; todo esto, con el fin de detectar a los estudiantes en riesgo y así aplicar planes de acompañamiento académico. Es entonces que, se desea implementar estrategias automáticas y especializadas para detectar comportamientos de abandono, con el fin de tomar acciones preventivas y no correctivas como sucede actualmente.

En este sentido, algunos autores como Saavedra, Sarmiento y Rojas (2016) señalan que es posible utilizar las TIC's (Tecnologías de la información y la comunicación) para contribuir a la reducción de la deserción universitaria optimizando recursos y reduciendo gastos (Ramírez y Grandón, 2018).

1.2 Justificación

El proyecto se centra en contribuir a disminuir la deserción académica para fomentar en los alumnos la culminación de sus estudios de educación superior y así, aportar egresados con mejor preparación académica y competencias laborales.

En un sentido de modelo de negocios, las instituciones privadas deben aplicar lo último en tecnología y a contribuir a mejorarla, utilizar modelos de Inteligencia de Negocios y explotar la información de una manera ordenada, precisa y efectiva, así, se pueden obtener ventajas como, la mejora de la administración, monitoreo del desempeño con tableros, y la aplicación de minería de datos (Ahmed y Kumar, 2015).

Un sistema automatizado para la predicción de la deserción, puede permitir análisis diarios, siendo afectado por los cambios que ocurren dentro de la vida estudiantil, en pocas palabras, se puede ajustar al comportamiento de las acciones académicas de los alumnos: calificaciones, asistencias, reprobación, etc.

El sistema también puede contribuir al crecimiento de otras áreas como la de Inteligencia de Negocios Institucional, apoyando proyectos con ciencia de datos y a la explotación de información histórica estudiante.

UNIVER cuenta con indicadores estratégicos, dentro de los cuales se encuentra la reducción de los índices de deserción. Por este motivo, cualquier esfuerzo que implique mejorar estos números son bienvenidos a nivel Dirección Académica.

Con las acciones que se puedan tomar partiendo del sistema, más estudiantes con perfiles de riesgo podrían sentirse acompañados por la institución y percibir un interés por que termine sus estudios, permitiendo así, evitar acciones remediales.

Desde una perspectiva metodológica, los algoritmos de árboles de decisión y redes neuronales están bien fundamentados y se sabe que tienen una variedad de aplicaciones para resolver problemas y entender comportamientos (Wu, 2006). Por otro lado, existen herramientas libres y especializadas en procesamiento de datos utilizando técnicas de aprendizaje computacional, las cuales se utilizan en este proyecto ya que no exige recursos adicionales que afecten a la institución.

Partiendo de estos puntos, UNIVR se puede apoyar del sistema para tomar decisiones basadas en las características de su funcionamiento académico y su población; buscar incrementar el prestigio de la universidad, mejorar programas educativos, mejorar apoyos económicos a estudiantes y empleados, etc.

Como consecuencia, es posible afectar positivamente a la productividad y competitividad de la sociedad, incluso aportar a reducir la brecha competitiva con los países desarrollados (Organización para la Cooperación y el Desarrollo Económicos, 2019).

Si se analiza en un sentido económico y social, trasciende de manera positiva el hecho de que un individuo tenga mejores esperanzas de tener un empleo bien remunerado o alguna otra fuente de ingresos, cambiando así su estilo de vida y propiciando entornos sociales más saludables y seguros, ejemplo de este cambio es Corea del Sur, ya que fijó su objetivo en mejorar la educación y actualmente es un modelo de desarrollo económico. Si la población interrumpe sus estudios, puede existir un desaprovechamiento económico de los recursos que el gobierno destina a la educación y preparación profesional de sus habitantes, por el contrario, una sociedad preparada contribuye mejor a su entorno, incluso a la vida política de su país creando entornos democráticos deseables (Banco Mundial, 2018).

1.3 Objetivos del trabajo práctico

Objetivo general

Implementar una aplicación web embebida basada en algoritmos de aprendizaje computacional que anticipe la deserción universitaria.

Objetivos específicos

- Realizar un análisis descriptivo para el pre-procesamiento de datos con la información histórica de educación superior disponible en el almacén de datos de UNIVR: incluye decanatos de ingenierías, ciencias sociales y administrativas.
- Analizar las variables que la literatura recomienda y aplicar algoritmos de selección de atributos para identificar aquellas más relevantes.
- Aplicar árboles de decisión y redes neuronales para obtener un modelo predictivo que identifique una deserción temprana
- Desarrollar una aplicación web de consulta para pronosticar una probabilidad de deserción de manera temprana
- Realizar pruebas de usabilidad para mejorar la experiencia del usuario con encuestas *in situ* a través de una evaluación heurística y recorrido cognitivo.

1.4 Alcances y limitaciones

A continuación, se mencionan alcances y limitaciones, teniendo en cuenta que pueden cambiar durante el desarrollo del proyecto, como incrementar el número de programas educativos involucrados o rangos en los históricos.

Alcances

- Los datos serán obtenidos del almacén de datos con el que cuenta la UNIVR, desde el ciclo escolar 2010 al 2019, considerando las variables: promedio ingreso, promedio general, semestre, asistencias, faltas, créditos, avance por créditos, reprobación, reincidencia de reprobación, asignaturas inscritas, edad, género, estatus civil, foráneo, nuevo ingreso y programa académico.
- Se utilizarán bajas temporales y definitivas.

- Se usarán histórico de 25 programas usados pertenecientes a los Decanatos de Ciencias Administrativas, sociales e ingenierías.
- Se van a separar datos del periodo otoño 2018 hasta primavera 2020 para validar el modelo.

Limitaciones

- Se considerará información institucional, sin agregar variables ajenas al ERP (por sus siglas en inglés, *enterprise resource planning*) institucional.
- Se contemplarán árboles de decisión y redes neuronales como algoritmos de aprendizaje computacional.
- Se restringe a usar datos de los programas autorizados del decanato de ciencias administrativas, sociales e ingenierías (por temas de seguridad de datos).
- Se utilizará software que sea compatible con la plataforma de inteligencia de negocios de la universidad, como R, Python, o conectores a plataformas de análisis de datos (pueden surgir restricciones al momento de desarrollar la aplicación). Adicionalmente se auxiliará de herramientas como Weka.

1.5 Propuesta metodológica

La elección de una metodología puede permitir seguir de manera sistemática todo el proceso de desarrollo del proyecto, razón por la cual, se analizaron proyectos similares para encontrar aquellas fases metodológicas que se puedan aplicar al entorno académico.

Las metodologías en la cual se basa este proyecto fueron KDD, SEMMA, CATALYST y CRISP-DM, las cuales han resultado ser las más discutidas en ejercicios de minería de datos (Monie (2013), Cendejas et al., (2017)).

Estas metodologías tienen fases en común y en algunas dejan la definición de las tareas al usuario, razón por la cual, se puede seleccionar aquellas etapas que se adecuen al entorno actual del proyecto:

- Se cuenta con un almacén de datos (modelo multidimensional).
- Se dispone de un sistema de Inteligencia de Negocios (plataforma web).
- Existe un proceso de integración de datos (ETL) el cual actualiza la información del modelo una vez por día en un horario nocturno.

En consecuencia, en la Figura 1 se presenta la metodología a implementar donde se retoman algunas fases, para formar una propia.

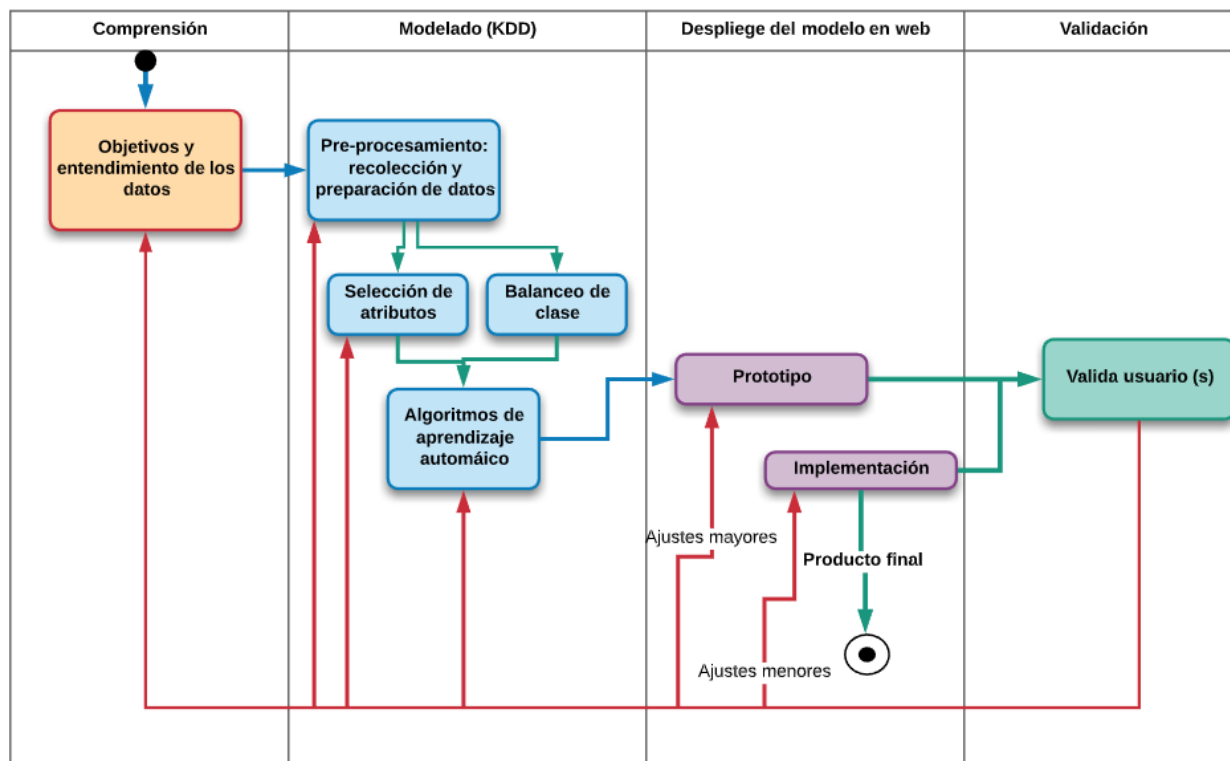


Figura 1. Metodología propuesta

1.5.1 Comprensión

Esta fase está encaminada a definir los **objetivos** que tendrán que ser alcanzados con la metodología planteada, los cuales deberán estar sustentados con los **datos** disponibles y los **requerimientos** del usuario.

Para conocer los requerimientos del usuario, se proponen el uso de algunas entrevistas en las juntas que existen entre las áreas de Inteligencia de Negocios y el de Seguimiento de Estudiantes para recabar información de manera directa con los integrantes del área y conocer la forma en que trabajan, sus necesidades y preocupaciones.

En cuanto a los datos, según Han, Kamber y Pei (2012), recomiendan obtener medidas de tendencia central, algunas de dispersión de los datos y utilizar una representación visual básica de los datos para un mejor entendimiento de los datos.

1.5.2 Modelado

Pre-procesamiento

En este paso se deben de extraer los datos del almacén de datos institucional y otras fuentes como Salesforce y archivos Excel, ver Figura 2. Esta información ya se encuentra limpia y está automatizado por un sistema de Integración de datos, en el cual, por medio de *mapas de proceso* realiza las consultas, validaciones y transformaciones en un ambiente intermedio entre la extracción y la carga de los datos. Todos los datos finalmente se encuentran en tablas de dimensiones y hechos, siguiendo los estándares marcados en la literatura (Kimball, 2008). Se utilizan bases de datos relacionales como Oracle 11g, MySQL y MSQl Server.

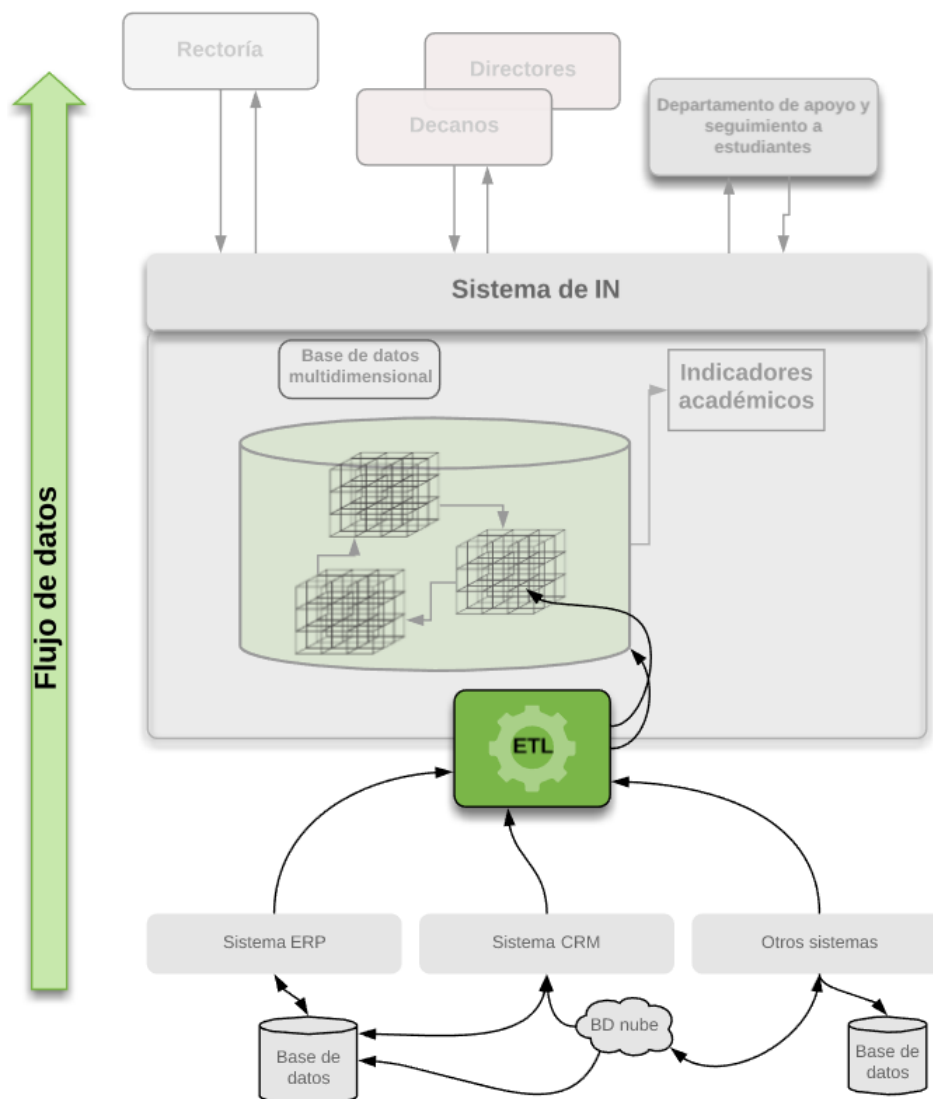


Figura 2. Recolección de datos dentro de la arquitectura tecnológica institucional.

Para la realización de este proyecto, se autorizaron datos de tres decanatos, distribuidos en 25 programas, ver Tabla 1.

Tabla 1. Número de instancias por decanato académico.

<i>Decanato</i>	<i># Programas</i>	<i>Número registros</i>	<i>% por programa</i>
<i>Ingenierías</i>	10	19,174	37.23 %
<i>Ciencias Sociales</i>	6	11,092	21.54 %
<i>Ciencias Administrativas</i>	9	21,231	41.23 %
<i>TOTAL</i>	26	51,497	100%

Cada registro corresponde a información recabada a nivel de alumno y periodo, la cual ha sido recabada desde el otoño 2010 al otoño 2019: periodo escolar, promedio al periodo, promedio al inicio de inscripción, edad del alumno, faltas acumuladas, asistencias acumuladas, género, número de semestre, total de asignaturas reprobadas, asignaturas reprobadas una vez, asignaturas reprobadas dos veces, asignaturas reprobadas más de tres veces, número de periodos inscritos, créditos cursados, avance según créditos, asignaturas inscritas en periodo, faltas, estatus civil, nuevo ingreso o no, foráneo, asignaturas reprobadas en primer parcial, segundo parcial y tercer parcial. En total se cuenta con 26 atributos más la clase *Desertor*.

Los datos están etiquetados por la clase *Desertor*, de los cuales *Desertor(S)* son 6,282 y *Desertor(N)* 45,215, por lo tanto, se tiene una clase desbalanceada. Este tipo de clases son un problema para los modelos predictivos, ya que pueden conducir a resultados erróneos en la predicción, es decir, la clase con un mayor número de muestras obtiene un mejor desempeño de predicción que la clase con pocas muestras, la cual, muchas veces es la de mayor interés (Arnejo, H., 2016). Entonces, como se mencionó anteriormente, la clase *Desertor(S)* representa el 12.19% del total de muestras, exponiendo así un problema de balanceo de clase. Para resolverlo, se utilizará las técnicas de *oversample* y *undersample*, ya que como lo menciona Arnejo (2016), son fáciles de implementar y se obtienen excelentes resultados. La propuesta es utilizar *oversample* para la clase *Desertor(S)*, y *undersample* para la clase *Desertor(N)*, ver Figura 3. De esta manera se obtendrá un nuevo conjunto de datos con la clase balanceada, la cual puede pasar a la fase de entrenamiento.

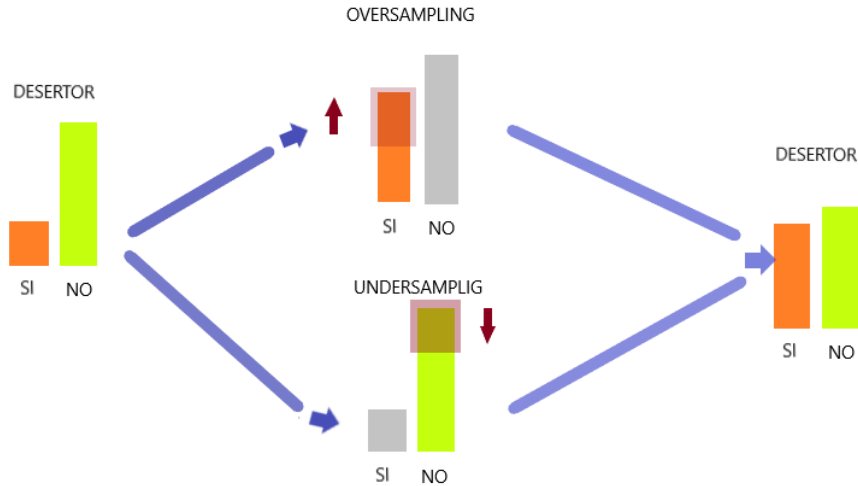


Figura 3. Técnicas de balanceo de clase.

Adicionalmente se contará con datos de primavera 2020 sin etiquetar, con el objetivo de poder pronosticar su deserción con el modelo final. Este conjunto de datos tiene 2,244 registros.

Estas tareas se piensan realizar con procedimientos almacenados en base de datos en conjunción con *scripts* en *R*, lo cual va a permitir trasladar datos, filtrar, validar, y transformar.

Selección de atributos con algoritmos de aprendizaje computacional

Se aplicará *algoritmos de selección de atributos*, algunos de los cuales fueron comparados en el trabajo realizado por Ruiz., R, Aguilar, J. y Riquelme, J. (2002), dichos algoritmos son: X^2 , Relief y SOAP (Selection of Attributes by Projections, por sus siglas en inglés). En este caso, se tienen 26 posibles atributos y según los autores, los beneficios de utilizar algoritmos son:

- Incrementa la precisión.
- Disminuye el *overtraining* ya que elimina datos con menor significancia.
- Aumenta la velocidad de entrenamiento.

Para estos análisis se tiene *R* con conexión directa a la base de datos multidimensional. Por lo tanto, para obtener los mejores atributos se propone utilizar algunos de los métodos disponibles en *R*:

- *Selección por subsección*- con CFS (Correlation Feature Selección, por sus siglas en inglés)
- *Filtro*: *chi_cuadrada* y *gain information*.

- *Envoltura*: con random forest.

Modelado predictivo

En la fase de modelado, se van a utilizar árboles de decisión y redes neuronales artificiales como algoritmos para la generación de los modelos predictivos. La Figura 4 detalla visualmente el proceso del modelado. Para trabajar con estos algoritmos se utilizará la herramienta RStudio 1.1.463 con R en su versión 3.5.1. De esta manera se podrán generar los scripts y manipular los datos desde la misma plataforma.

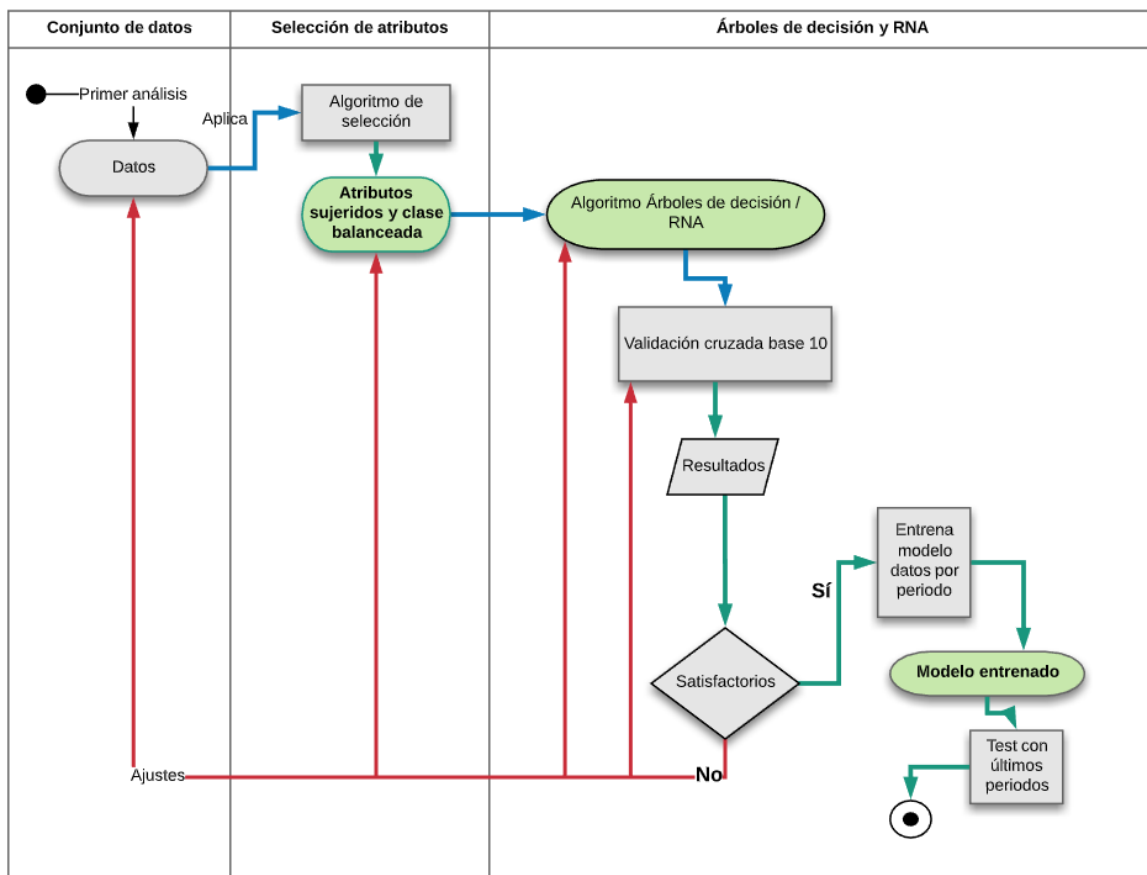


Figura 4. Detalle de la fase de modelado.

Las librerías a utilizar serán:

- Para **árboles de decisión** se va a utilizar el algoritmo C5, el cual se encuentra en la librería C50 en R.

- En **Redes neuronales artificiales** se va a utilizar el paquete *neuralnet* para implementar las RNA con R, con el algoritmo de *perceptrón multicapa* con topologías de 1 y 2 capas con diferentes números de neuronas. Otra librería opcional es *nnet* la cual utiliza redes *Feed-Forward*.

Entrenamiento y pruebas

Para el entrenamiento y pruebas se van a utilizar ejecuciones con diferentes proporciones en número de instancias, con el afán de comparar sus métricas de y encontrar el modelo con mejor desempeño:

1. Cuando sea necesario, se proponen con el 70%, 80% y 90% de los datos disponibles.
2. Una limitante en los entrenamientos son los tiempos de ejecución, ya que no deberían pasar los 20 minutos.
3. Para las evaluaciones iniciales de los modelos se propone usar Validación cruzada base 10.

1.5.3 Despliegue en web

En esta fase se diseña y propone la implementación de la aplicación web que va a interactuar con el modelo predictivo y el usuario del departamento de seguimiento de alumnos.

Para la realización de la aplicación se propone utilizar una metodología centrada en el usuario por medio de un prototipo de alta fidelidad, lo cual va a permitir enfocarse en las observaciones del usuario en un proceso iterativo. Este proceso se basa en la metodología de “recorrido cognitivo” en la cual se involucran los usuarios en la evaluación (Granollers, Perdrix y Lorés, 2004). Además, actualmente es utilizada en metodologías de UX (por sus siglas en inglés User eXperience).

Superado la fase del prototipado, se procederá a la evaluación del mismo con el instrumento adecuado aplicado a los usuarios expertos y el usuario finales, si surgen observaciones, se retornará a la fase del prototipo, si no pasa a la siguiente etapa de implementación, si en este paso existen observaciones, entonces se regresa a la fase de prototipado y se itera de nuevo hasta que ya no existan cambios o se alcance las fechas límite del proyecto, lo que suceda primero, Figura 5.

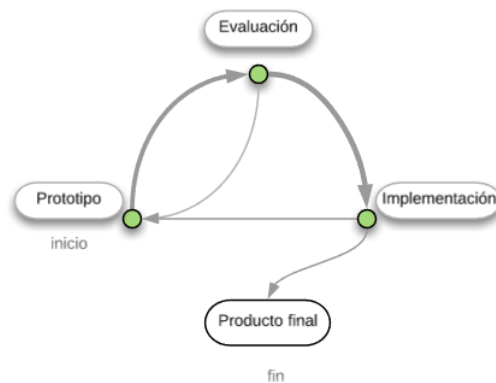


Figura 5. Proceso iterativo para el despliegue web.

Con este prototipo, se podrá realizar el proceso de evaluación por Recorrido Cognitivo, es decir, mostrar al usuario, evaluar su comportamiento y/o experiencia en un proceso iterativo.

Es importante mencionar que se hará una simulación de resultados predictivos, para efectos de evaluación de la aplicación y su interacción con el usuario.

Se crearán las interfaces web basándose en el prototipo aprobado, las cuales, se conectarán al modelo predictivo para realizar la consulta de los datos. En esta fase se espera que la implementación sufra el mínimo de cambios con el menor impacto. La información utilizada será la de los 3 decanatos académicos mencionados en los objetivos del proyecto.

La implementación se deberá hacer desde plataforma de IN institucional, cumpliendo así con las políticas del departamento de tecnologías de información la cual autoriza la creación de servicios web institucionales. Dicha herramienta tiene su entorno de desarrollo en web con *Asp.Net*, *Html5* y *JavaScript*. R permitirá crear los modelos y procesar los datos con conexión directa al almacén de datos. La seguridad estará administrada por la misma plataforma de IN, (accesos, timeout, contraseñas).

Al igual que en el prototipo, se aplicarán las evaluaciones para encontrar problemas de diseño y funcionalidad.

1.5.4 Validación

La evaluación ocurrirá dentro de la etapa final del proyecto, en ella participaran usuarios expertos y finales. Existirán dos momentos de validación:

- *Evaluación del prototipo*: este proceso permitirá realizar cambios al diseño antes de hacer alguna implementación, lo cual acelerará y simplificará el desarrollo. Se aplicará la evaluación con usuarios expertos y finales. Se buscarán sesiones de 20 – 30 minutos para su revisión con el prototipo en línea y acceso desde el pc de los participantes. Es un proceso iterativo.
- *Evaluación de la implementación*: esta fase servirá para refinar la aplicación y obtener el producto final. Se espera que los cambios que llegasen a surgir sean mínimos y que las sesiones de revisión duren menos tiempo: 15-20 minutos.

La metodología de evaluación de Heurística se utilizará principalmente para detectar problemas de diseño en la interfaz web, de esta manera, se podrá ofrecer un producto de calidad. Se complementará con el método de Recorrido Cognitivo con el objetivo de que el usuario se involucre en las decisiones y así, asegurar que se cumplan las expectativas del usuario.

1.5.5 Herramientas de software

Para cada fase es necesario tener diferentes herramientas de software que permitirán ejecutar las tareas correspondientes, desde conectores para la extracción de datos, hasta el software que ejecuta los algoritmos de aprendizaje computacional. En la Tabla 2, se muestran las herramientas dependiendo de la fase metodológica. El objetivo es aprovechar aquellas herramientas libres y utilizar los recursos que la institución facilita para la realización de este proyecto.

Tabla 2. Herramientas de software.

Fase	Herramientas de software
<i>Recolección y preparación de datos</i>	Base de datos Oracle 11g. R 3.2.5 y R Studio 1.1.463 Excel 2016 Windows server 2012
<i>Selección de atributos</i>	R 3.2.5 / R Studio 1.1.463 Conectores a base de datos R a Oracle.
<i>Modelado</i>	R 3.2.5 / R Studio 1.1.463 Árboles de decisión con R y librería neuralnet Conectores a base de datos R a Oracle. Windows server 2012 Weka 3.9 (apoyo para probar algoritmos)
<i>Despliegue web</i>	Lenguaje programación Asp.Net Conectores a Oracle 11g Dossier 2019 (prototipo)

Se propone utilizar R porque es una plataforma robusta para el análisis de datos y se complementa con librerías para diferentes funciones: para conexión a base de datos, algoritmos de clasificación, transformaciones, análisis estadísticos, etc. Además, R permite evitar temas de restricciones de licenciamiento como con RapidMiner.

1.6 Organización del proyecto

Para alcanzar los objetivos planteados anteriormente se propone el siguiente programa de actividades. Se observa que se divide en tres etapas, cada una tiene una duración de cuatro meses:

- La etapa 1 corresponde a la conceptualización y delimitación del proyecto.
- En la etapa 2 se va a plantear la metodología a seguir, y a definir el prototipo, según objetivos y requerimientos del usuario.
- Adicionalmente, se realizará el pre-procesamiento de los datos, identificación de las variables utilizando algoritmos de selección de atributos, balanceo de clase y aplicación del modelado con árboles de decisión y redes neuronales.
- En la etapa 3 se va a implementar la aplicación web, según el prototipo y se realizarán pruebas finales, después se obtendrán las conclusiones y se definirá el trabajo futuro.

Ver Figura 6.

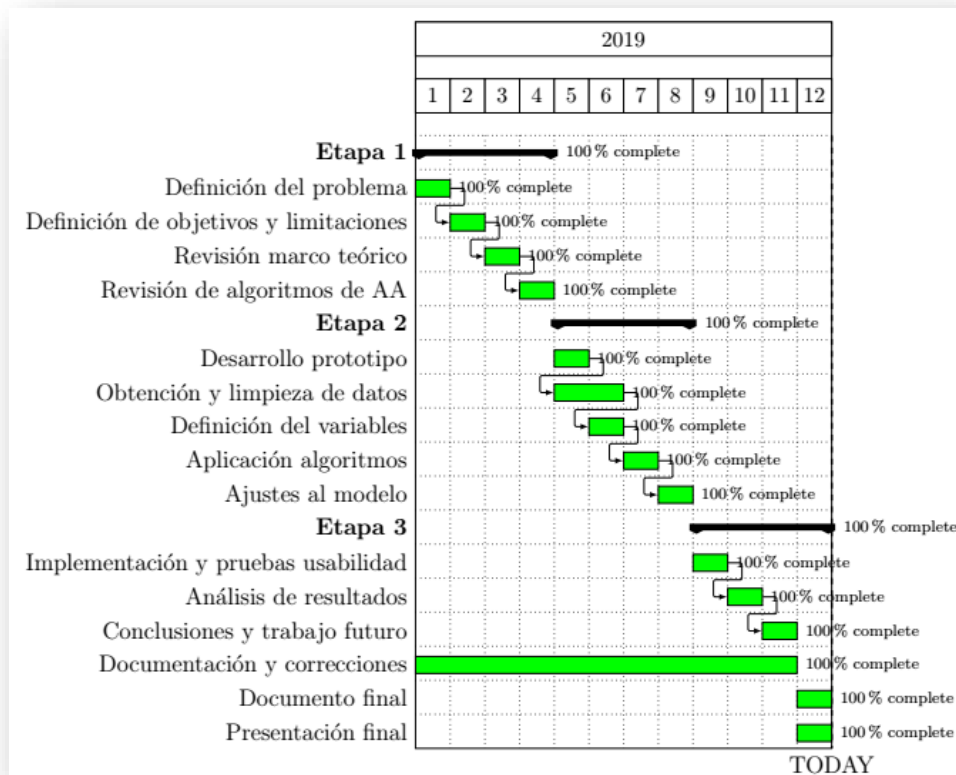


Figura 6. Cronograma de actividades.

2 Estado del arte y marco teórico

A continuación, se hace una revisión de los trabajos relacionados entorno a la deserción escolar para poder entender la problemática, sus causas, consecuencias y posibles acciones aplicando tecnología y metodologías en ciencia de datos, que puedan contribuir a reducir los índices en las Instituciones de Educación Superior (IES).

2.1 Deserción escolar

De acuerdo a Tinto (1989), la deserción en la educación superior es un tema difícil de explicar, por esta razón, propone analizar el fenómeno desde diferentes perspectivas: del estudiante, institucional y estatal o nacional. Desde la perspectiva de estudiante están sus expectativas, metas, capacidades intelectuales y origen socio-económico. En el aspecto institucional se

encuentran aquellas opciones ofertadas al estudiante para incluirlo en la vida universitaria, darle seguimiento y retenerlo. En el ámbito estatal o nacional, la deserción debe considerarse como la interrupción de estudios en cualquier modalidad y analizar las políticas que fomentan la retención, como becas; también es importante contemplar que algunos factores pueden cambiar con el tiempo. Debido a esta complejidad, el autor recomienda definir y delimitar correctamente el abandono al momento de realizar una investigación y tener en cuenta que pueden existir diferentes tipos y algunos pueden no ser objeto de análisis para determinar sus causas.

Rodríguez y Hernández (2008), mencionan que son diferentes las causas de la deserción escolar y que no pueden ser explicadas completamente, este análisis se realizó en la Universidad Autónoma Metropolitana Campus Iztapalapa en las divisiones académicas de Ciencias Básicas e Ingeniería, Ciencias Biológicas y de la Salud, así como de Ciencias Sociales y Humanidades, usaron la encuesta aplicada como metodología y análisis cualitativo de grupos focales. En un sentido regional, los autores comentan que la deserción en Latinoamérica se encuentra afectada por presiones económicas, integración familiar, violencia intrafamiliar y social, problemas de salud psicosomática, con la justicia, disciplina y llegada a la escuela a una edad tardía. Por tal motivo, plantean tres núcleos explicativos: a) Familia-Económica-Trabajo, b) Rendimiento y orientación y c) Estructura y organización académica.

Los resultados demuestran que una deserción temprana ocurre en el primer trimestre y una tardía en trimestres posteriores, siendo la tasa de deserción definitiva temprana más alta que la tardía 85.4% vs 46.8% respectivamente, también encontraron que el género no es relevante, la edad y el estatus civil son factores importantes. Las causas detectadas por orden de importancia fueron: actividad laboral, economía familiar, rendimiento escolar, cultura de estudio, orientación hacia la carrera y gestión académica. Los autores recomiendan acciones como reforzar la relación docente-alumno y observar el perfil del alumno que está en riesgo: asistencia a clases, irregularidad en sus estudios, retraso en estos, calificaciones, etc.

La Secretaría de Seguridad Pública (2011), presenta un análisis de la deserción escolar y sus posibles implicaciones en la sociedad en México. El estudio aporta un análisis de la problemática y la expone para ayudar a la implementación de mecanismos que apoyen a tomar acciones preventivas. El estudio se basó en un análisis estadístico en Guadalajara, Monterrey y Distrito Federal. Se expone, que una de las consecuencias de la deserción se refleja en fuerza laboral poco competente, con una capacitación más complicada y poco crecimiento económico. La investigación concluye que se debe buscar la permanencia escolar como un Factor de

Protección, así como el aprovechamiento del tiempo en jóvenes, para el bien de ellos mismos y de nuestra sociedad.

En 2016, el reporte de Panorama Educativo de México, especifica que en el ciclo 2014-2015 en educación media-superior existía una tasa de eficiencia terminal del 67.3% indicando que el 37% desertó de sus estudios, en este escenario, la investigación reveló que el impacto de la deserción no solo está en indicadores académicos, sino también en aspectos económicos, además de incrementar las desigualdades sociales, terminando en pobreza y en desintegración social. Por lo tanto, recomiendan que las instituciones desarrollen actividades para la detección temprana y también acciones preventivas, considerando que la problemática se presenta más en los primeros semestres de estudio del alumno. Como trabajo futuro, proponen realizar un análisis predictivo.

Ramírez, Espinosa y Millán (2016), elaboraron una estrategia con tecnologías de la información para abordar el tema de la deserción académica en los pre-gradados en universidades de Colombia. La metodología la abordaron desde la heurística y el ámbito proyectivo, con el objetivo principal de determinar si un estudiante se encuentra en riesgo de deserción y así cumplir con los estándares de calidad exigidos por los procesos de acreditación. Como principales causas del abandono señalan: bajo nivel de los estudiantes al ingresar a la institución, dificultades económicas, nulo interés vocacional, actitudes del estudiante y expectativas. De acuerdo a los autores, el retraso en el avance académico está altamente relacionado con la deserción, por lo tanto, la estrategia final propuesta con el apoyo de las TIC's se centró en estos puntos:

- E-learning para el refuerzo de áreas problema.
- Uso de tutorías especiales a estudiantes en factor de riesgo.
- Seguimiento de los apoyos o estímulos.
- Seguimiento de los factores de deserción y bajo rendimiento académico.
- Evaluación transversal y continua.
- Establecimiento de perfiles de riesgo a partir de las plataformas existentes.

La justificación para que la estrategia tenga éxito, radica en que las instituciones cuenten con sistemas de información y recursos digitales, lo que permite establecer perfiles de riesgo basado en datos existentes, por lo tanto, se puede integrar de manera causal para mitigar la deserción favorablemente.

López y Beltrán (2017) aplicaron un cuestionario para detectar las causas de deserción escolar en educación superior a 75 alumnos, 30 docentes y 30 padres de familia, la encuesta se realizó

a estudiantes de Contaduría del Centro Universitario UAEM Temascaltepec; cabe señalar que el estudio fue no probabilístico y propositivo. La investigación propone principalmente utilizar tres núcleos explicativos: económico, escolar e institucional. Dentro de sus hallazgos encontraron que las causas principales son la incorrecta elección de la carrera y la situación económica de los tutores.

Otro referente sobre el tema, es el trabajo realizado por Carvajal, González y Sarzoza (2017), en la Universidad de Playa Ancha en Chile. En su investigación analizaron los factores que influyen en la deserción en la educación superior con el objetivo de encontrar las variables más significativas. Como metodología utilizaron estadística descriptiva, correlacional e inferencial sobre estudiantes de Geografía, Matemáticas, Biología y Química con información del 2014 al 2015. En un primer paso se apoyaron de un análisis correlacional para encontrar significancia de variables y posteriormente implementaron Raíz Jerárquica de Tercer Orden, el modelo alcanzó un $r^2 = 0.568$ con 5% de nivel de significancia. Los resultados del modelo indican que son tres los factores que influyen en la deserción: calificaciones, puntaje de ingreso y nivel de estudios de los tutores. Adicionalmente, la utilización de un árbol jerárquico, permite analizar visualmente los factores de deserción y así contribuir a la toma de decisiones a nivel académico.

En Guaymas Sonora México, Zavala, Álvarez, Vázquez, Gonzales y Ramírez (2018), plantean que el problema de la deserción en estudiantes universitarios se puede explicar por tres factores: factores internos al alumno (FIA), factores externos al alumno (FEA) y factores bilaterales (FB). El trabajo se centró en realizar un análisis de factores de correlación con doce índices asociados a la deserción, incluyendo aquellos en donde la deserción fue voluntaria o forzada, el modelo resultante obtuvo un intervalo de confianza del 90% para los factores FIA, FEA y FB.

2.2 Minería de datos

En un aspecto metodológico, Moine (2013), realizó un análisis comparativo de metodologías para la aplicación de minería de datos en el campo médico en Argentina. En ellas expone que las fases identificadas en CRISP-DM son seis: I) entendimiento del negocio, II) entendimiento de los datos, III) preparación, IV) modelado, V) despliegue y VI) análisis de resultados. Por su parte en SEMMA son cinco: I) muestreo, II) exploración, III) modificación, IV) modelado y V) evaluación. Por otro lado, CATALYST se caracteriza por: I) Modelado de negocio, II) preparación, III) modelado y herramientas, IV) refinamiento y V) comunicación. Y adicionalmente en KDD

(Knowledge Discovery in Databases, por sus siglas en inglés), son nueve y se resumen en cinco pasos: I) selección, II) pre-procesamiento, III) transformación, IV) minería y V) evaluación de resultados. En su proceso de evaluación utilizó árboles de decisión, vecinos más cercanos (KNN), clasificador Naive Bayes y regresión logística; con R y Weka como herramientas de software. Como resultado, las metodologías con mejores resultados fueron Catalyst y CRISP-DM, además, recomienda utilizar su investigación para formular metodologías propias.

Por otro lado, Ahmed y Kumar (2015), discuten los beneficios de aplicar un modelo de inteligencia de negocios en una universidad de Quebec, Canadá. Puesto que en una institución se genera mucha información, y agrega complejidad a la administración académica, crearon un modelo para mejorar la toma de decisiones. Evaluaron herramientas de IN aplicando encuestas y análisis cuantitativo como metodología, después resolvieron el modelo con el uso de *Oracle Business Intelligence* para proporcionar tableros y reportes. Como resultado concluyeron que aplicar IN es una forma rápida de hacer a la universidad competitiva ya que la mantiene al día al utilizar los últimos avances en la tecnología e investigación, organiza el conocimiento de la institución y mejora la toma de decisiones y da un mejor seguimiento a los estudiantes. Proponen en trabajos posteriores tener un plan de capacitación para su utilización y evaluar el modelo después de un tiempo determinado.

Siguiendo la misma línea de investigaciones, se encuentra el trabajo realizado por Chinkes (2017), que propone la minería de datos para realizar un pronóstico de deserción en alumnos de educación superior. Los datos utilizados comprenden del 2007 al 2014, siendo un total de 421,282 registros, los cuales fueron proporcionados por el almacén de datos de la universidad. Su información fue protegida por temas de confidencialidad y contenían datos como edad, género, localidad, nivel de estudios de los tutores, año de ingreso a la carrera, materias reprobadas y aprobadas. Para generar el modelo, utilizó el servicio de la nube de Microsoft Azure Machine Learning con el algoritmo de Two-Class Bayes Point Much y Redes neuronales. Finalmente, el modelo tuvo una precisión del 66% y el autor concluye que los resultados de los pronósticos se deben tomar con cierta cautela ya que, aunque se puedan mejorar, existen varios factores que posiblemente no se estén considerando para poder asumir que se encontró una manera de pronosticar la deserción.

Cendejas, V., Acuña, L., Cortez, M. y Bolaños, J. (2017) mencionan a CRISP-DM y SEMMA como dos de las metodologías más importantes en la aplicación de minería de datos, ya que pueden garantizar una apropiada planificación y resultados efectivos. Realizaron una evaluación de metodologías con diferentes herramientas de software (Weka, SAS, Orange, entre otros).

Ellos propusieron su propia metodología para la toma de decisiones y recomiendan a SPSS como una herramienta que satisface satisfactoriamente sus fases metodológicas.

En Venezuela, Muñoz, S., Gallardo, T., Muñoz, M. y Muñoz, C. (2018), implementaron una metodología para calcular la probabilidad de deserción universitaria. En su trabajo, los autores analizaron factores que detectaron en la Universidad de los Andes-Venezuela del 2011 al 2016 en las facultades de Matemáticas, Arquitectura y Diseño y Ciencias Económicas. La metodología usada fue del tipo cuantitativo y estudio de caso con el apoyo de un modelo de Elección Discreta (Logit), utilizando información institucional. En su trabajo, definieron como variable dependiente a la deserción estudiantil y como variables explicativas el periodo académico, tipo de curso, género y número de estudiantes. Los resultados demostraron que el curso y tipo de periodo fueron las variables más significativas, y que los estudiantes que estuvieron en un periodo regular tienen mayor probabilidad de desertar que los que estuvieron inscritos en periodos cortos, además, se indicó que existe poca investigación sobre la deserción con un enfoque predictivo, lo cual puede ser un área de oportunidad para el objetivo de este documento.

Otro trabajo interesante es el realizado por Gallegos, Campos, Canales y González (2018), en el cual utilizaron un modelo probabilístico *Logit* para poder detectar los factores que motivan a una deserción en la Universidad Católica de la Santísima Concepción (Chile). En su trabajo se menciona que la deserción podría deberse a una ineficacia del estado en el sistema educativo, argumentando que según datos de la OCDE del 2013 el 30% de los estudiantes en educación superior desertan, en comparación con el 20% de otros países como Japón o Francia.

En su investigación, se apoyaron de modelos de probabilidad Logit con STATA y con datos de 1,500 estudiantes del 2004 al 2013, incluyendo variables demográficas, vocacionales y socioeconómicas como: edad, género, lugar de procedencia, escolaridad tutores, ingresos y becas. En sus descubrimientos, encontraron que las variables de región de procedencia, becas y promedio fueron las más significativas, mientras que los ingresos y escolaridad de los tutores no lo fue tanto, además se reduce la probabilidad de deserción con el seguimiento del rendimiento académico. Como trabajos futuros proponen la utilización de nuevas variables como satisfacción de la carrera o calidad de la docencia.

Ramírez y Grandón (2018), proponen el uso de árboles de decisión como herramienta para predecir la deserción académica de una universidad estatal en Chile, donde más del 50% de alumnos de educación superior no concluyen sus estudios. La investigación utilizó 5,288 registros de estudiantes de cuatro cohortes generacionales y un modelo de árbol de decisión

implementado en *RapidMiner Studio*, además reunieron variables demográficas, situación económica y algunos datos recolectados al momento de ingresar, como puntuación en examen de conocimientos.

El árbol utilizado tuvo una profundidad máxima de 20 y detectó tres factores que explican la deserción: promedio de calificaciones, periodo de avance y promedio de su examen de ingreso. Los resultados son prometedores, obtuvieron una precisión del 87.27%, lo cual alienta a usar algoritmos de árboles de decisión para poder contrarrestar la deserción estudiantil. Cabe señalar que sus resultados son incluso mejores que otras investigaciones que utilizan redes neuronales donde obtuvieron alrededor de 82% de precisión.

Esta revisión de la literatura, ofrece un panorama sobre los trabajos realizados entorno a deserción universitaria. En la Tabla 3, se puede observar que la mayor parte de esfuerzos está enfocado a detectar aquellos factores que inciden en una deserción y son pocos los que están aplicando métodos de aprendizaje computacional para poder hacer un pronóstico. En este escenario, se podría esperar que las variables como la del rendimiento escolar, algunos datos demográficos y de ingreso a la institución, se puedan utilizar para crear un modelo ajustado al comportamiento que se desea analizar. También se puede observar que los arboles de decisión y las redes neuronales se utilizan para casos de pronósticos, y como el proyecto busca anticipar una deserción, entonces son los algoritmos ideales para implementar el modelo.

Tabla 3. Datos relevantes sobre la deserción escolar

<i>Investigación</i>	<i>Año</i>	<i>Objetivo</i>	<i>Descripción</i>	<i>Método</i>
<i>Tinto</i>	1989	Análisis Factores	Perspectivas: alumno, institucional y estatal.	Teoría de retención
<i>Rodríguez y Hernández</i>	2008	Análisis Factores	Factores: laboral, economía familiar, rendimiento escolar, estudio y orientación.	Encuesta y análisis cualitativo
<i>SEP</i>	2011	Análisis Factores	Implicaciones sociales y acciones preventivas.	Análisis estadístico
<i>Moine</i>	2013	Metodologías minería	Evaluación de metodologías y software.	KDD, CRISP-DM, SEMMA y KATALYST
<i>Ahmed y Kumar</i>	2015	Ventajas IN	Ventajas de un modelo de IN en Universidad	Encuesta y análisis cuantitativo
<i>Reporte PEM</i>	2016	Análisis Factores	Recomendaciones acciones preventivas en primeros semestres.	Análisis estadístico
<i>Ramírez, Espinosa y Millán</i>	2016	Análisis Factores	Diferencias económicas, vocación, actitudes, expectativas	Heurística y ámbito proyectivo
<i>López y Beltrán</i>	2017	Análisis Factores	Núcleo económico, escolares e institucional.	No probabilístico y propositivo

Chinkes	2017	Pronóstico	Uso datos edad, género, localidad, nivel estudios, etc.	Machine Learning con el algoritmo de Two-Class Bayes Point Much y Redes neuronales
Carvajal, González y Sarzoza	2017	Análisis Factores	Calificaciones, puntaje de ingreso y nivel de estudios de los tutores.	Metodología utilizaron estadística descriptiva, correlacional e inferencial
Cendejas, Acuña, Cortez y Bolaños	2017	Metodologías minería	Evaluación de metodologías y software.	CRISP-DM y SEMMA
Zavala, Álvarez, Vázquez, Gonzales y Ramírez	2018	Análisis Factores	Factores internos alumno, externos y bilaterales.	Factores de correlación.
Muñoz, S., Gallardo, T. Muñoz, M. y Muñoz, C.	2018	Análisis Factores	Tipo de curso académico y periodo académico.	Modelo de Elección Discreta Logit
Gallegos, Campos, Canales y González	2018	Análisis Factores	Factores significativos: región de procedencia, becas y promedio.	Modelo probabilístico Logit
Ramírez y Grandón	2018	Pronóstico	Uso de variables demográficas, económicas y otros adicionales.	Árboles de decisión

2.2.1 Proceso KDD

De acuerdo a Moine (2013), el descubrimiento de conocimiento en datos KDD (por sus siglas en inglés, Knowledge Discovery in Databases) fue el primer modelo para la extracción de conocimiento de una manera metodológica y funciona como una herramienta para la toma de decisiones. En la Figura 7 se expone de manera resumida las siguientes fases:

- I. Selección: selecciona los datos.
- II. Pre-procesamiento: actividades como tratamiento de datos faltantes, reducción de ruido, etc.
- III. Transformación: selección de atributos.
- IV. Minería: aplicación de algoritmos.
- V. Evaluación de resultados: evaluación y explotación del conocimiento.



Figura 7. Etapas resumidas en KDD.

Actualmente el proceso KDD se refiere a todos los pasos desde la definición de objetivos, hasta la utilización del conocimiento. Y en este sentido es como se va a definir la metodología para el proyecto.

2.2.2 Selección de atributos

La selección de atributos por medio de algoritmos permite mejorar la calidad de los datos de entrada con la eliminación atributos que no son relevantes (Aguilar., y J. Díaz. N, 2005).

Algunos beneficios de aplicar esta reducción de dimensionalidad son:

- Reduce complejidad para obtener un modelo con mejor interpretación
- El rendimiento computacional aumenta.

De acuerdo a la técnica de selección de características se pueden clasificar en tres métodos:

- Completo: analizan todas las posibles combinaciones.
- Heurísticos: no recorre todas las opciones de características lo que se traduce una ejecución del método con mayor velocidad. Relief y DTM (Decisión Tree Model, por sus siglas en ingles).
- Aleatorios utilizan metodologías aleatorias para la selección. LVM y algoritmos genéticos.

También pueden ser clasificados por el algoritmo clasificador:

- Filter: el proceso de selección es independiente del algoritmo clasificador.
- Wrapper (empaquetador): usan el mismo algoritmo clasificador para seleccionar los atributos y la clasificación. Tiene un alto costo computacional. Evalúa subconjuntos de los atributos.
- Híbridos: toma las ventajas de ambos modelos *filter* y *wrapper*.

Para en análisis de atributos se va a usar *filter*, *cfs* y algún método *wrapper*.

2.2.3 Evaluación de desempeño

Las métricas para medir y comparar su desempeño se van a obtener de la *matriz de confusión*, Figura 8, la configuración de posición de columnas esta emparejada a la que trabaja la librería *caret* en R.

		REALES	
		TOTAL NEGATIVOS	TOTAL POSITIVOS
PREDICCIÓN	N	Verdaderos negativos (VN)	Falsos positivos (FP)
	S	Falsos negativos (FN)	Verdaderos positivos (VP)

Figura 8. Matriz de confusión (librería **caret** en **R**).

Las métricas de evaluación son:

- *Exactitud*: porcentaje de verdaderos positivos y negativos contra todos los datos.

$$exactitud = \frac{VP + VN}{Total}$$

- *Precisión*: porcentaje de verdaderos positivos del total de positivos clasificados.

$$precisión = \frac{VP}{Total\ clasificados\ positivos}$$

- *Especificidad*: de los verdaderos negativos, ¿Cuántos clasificó correctamente?

$$especificidad = \frac{VN}{Total\ de\ negativos\ reales}$$

- *Sensibilidad (exhaustividad, recall)*: de los verdaderos positivos, ¿Cuántos clasificó correctamente?

$$sensibilidad = \frac{VP}{Total\ de\ positivos\ reales}$$

Rodríguez A., (2011) recomienda utilizar el coeficiente de correlación de Mathew (MCC) como métrica para evaluar de manera global el rendimiento de modelos de clasificación binario. En el cual, se evalúa en un rango de -1 a 1, donde 1 representa la clasificación perfecta, 0 una clasificación aleatoria y -1 una clasificación inversa. Por lo tanto, se va a agregar a las evaluaciones de los rendimientos el MCC y la exactitud balanceada. Para el caso de MCC se va a representar en porcentaje para efectos prácticos.

2.2.4 Árboles de decisión

Los algoritmos de árbol de decisión son técnicas de aprendizaje supervisado, fácil de implementar y muy efectivos el cual se compone de un nodo inicial único y debajo se desprenden otros árboles independientes que señalan los atributos de predicción (Menes, C., Arcos G., Moreno, P., y Gallegos C., 2015).

Se elige árboles de decisión como algoritmo ya que es un excelente método de clasificación (Ramírez y Grandón, 2018). En la literatura, este algoritmo ya ha sido probado para identificar características en la población universitaria para explicar la deserción académica.

Específicamente se va a implementar el algoritmo C5.0 con RStudio. C5.0 es una extensión del algoritmo C4.5 y de su predecesor ID3. Algunas mejoras de C5 son en velocidad, rendimiento y árboles con menor profundidad.

2.2.5 Redes neuronales artificiales

Las redes neuronales artificiales son modelos computacionales que tratan de imitar a las neuronas del cerebro humano y resolver problemas complejos de aprendizaje, se compone de algoritmos que procesan un conjunto de datos para encontrar relaciones no lineales (Sánchez, 2015). Las neuronas se organizan en capas (monocapa y multicapa) y la salida de unas son las entradas de otras, entonces son *hacia delante (feedforward)* si tiene conexiones hacia atrás, entonces son *feedback*, algunas de ellas son: *backpropagation*, *Hopfield*, *Convolutional*, *LSTM (Long Short Term Memory, por sus siglas en inglés)*, *DBN (Deep Belief Networks, por sus siglas en inglés)* y *GAN (Generative Adversarial Networks, por sus siglas en inglés)*.

El trabajo realizado por Chinkes (2017) expone la utilización de las RNA en el sector académico para contrarrestar la deserción escolar, razón por la cual, es una opción para la creación de un modelo predictivo.

Las métricas para evaluar el desempeño serán las mismas de árboles de decisión, con el objetivo de poder comparar y seleccionar el modelo con los mejores resultados.

2.3 Modelo de prototipos

El modelo de prototipado, permite concentrar los esfuerzos en crear el software en vez de en la documentación, lo cual le da una gran ventaja ya que enfoca el tiempo en otros aspectos del desarrollo (Ganpatrao R., 2012). Este modelo también se caracteriza por involucrar al usuario en

el prototipo y mejorar las características para la implementación final, resultando en un cliente satisfecho.

Las ventajas del modelo son:

- Alinea al usuario, desarrollador y cliente a lo que realmente se necesita.
- Permite visualizar nuevos cambios al diseño original.
- Se tienen los requisitos generales, pero no el detalle.
- Minimiza los riesgos e incertidumbre en la fase de desarrollo.
- Permite definir el alcance del desarrollo.

Desventajas

- El usuario puede tener la falsa ilusión que la transición del prototipo al producto final es un proceso rápido.
- La rapidez en respuesta de funcionalidad en prototipo, puede ser mayor que en la del producto final, lo que podría ocasionar en el usuario una desilusión del producto.

De acuerdo a Granollers, T., Lóres, V., y Delgado, J. (2005), existen dos tipos de prototipado, los cuales son muy parecidos en efectividad de resultados:

- De baja fidelidad: muestran lo general sin detalle, rápidos y de fácil implementación, son baratos.
- Alta fidelidad: permite un analizar la interacción con el usuario. Es muy semejante al producto final.

Además, el prototipado puede ser para *presentación* o para *interacción*. Algunas técnicas de prototipado son: *bocetos*, *storyboard*, *en papel*, *de software*, entre otros.

Para el proyecto se va a utilizar el prototipado como metodología para diseñar el producto final y para involucrar al usuario. Se va a utilizar prototipos de alta fidelidad ya que existen los recursos para hacerlo: plataforma web de rápidos diseños y accesible a los evaluadores. Así, el usuario final puede tener una idea palpable del producto terminado y emitir comentarios de mejora. Se podrían crear tantos prototipos como ajustes se vayan encontrando en las *evaluaciones*, sin embargo, por cumplimiento de fechas, existe la limitante de tiempo, de tal forma que se podrían hacer hasta tres prototipos incluyendo evaluación de los mismos.

2.3.1 Evaluación de usabilidad

La *evaluación* es un proceso necesario para analizar la usabilidad de software con el usuario para encontrar mejoras en las diferentes fases del desarrollo de un proyecto (Granollers, T., Lóres, V., y Delgado, J., 2005). Esto también va a permitir medir el nivel de satisfacción de los usuarios. Por lo tanto, la literatura indica que los objetivos de la evaluación de la usabilidad son:

- Verificar la funcionalidad.
- Analizar la interacción usuario-interfaz.
- Detectar errores o problemas.

Las evaluaciones pueden realizarse en:

- Entornos controlados: permite centrarse en aspectos específicos y sin distractores.
- Entorno final: muestra mucha información sobre el uso real del software.

Los evaluadores pueden ser: a) diseñadores, b) programadores, c) expertos en usabilidad y d) usuario final.

En cuanto a los métodos para realizar las evaluaciones, se van a mencionar dos que son los que se van a utilizar y pertenecen al *tipo de evaluación de indagación*:

- Recorrido Cognitivo. Basado en la teoría de aprendizaje por exploración, permite observar y evaluar al usuario ejecutando ciertas tareas: a) se realiza un prototipo, b) reúnen expertos, c) definen tareas. c) definen acciones para cada tarea *happy path*. d) se obtienen las opiniones de acuerdo a cuatro preguntas. Por lo general se utiliza en fases iniciales de algún proyecto debido a los *descubrimientos* rápidos que se pueden obtener.
- Métodos heurísticos. Método más popular basado en heurísticas: a) el evaluador inspecciona el sistema y lo evalúa, según las tareas. b) Entre varios evaluadores, se genera una lista de evaluación, y se agrega una nota de severidad (0-4) y una de frecuencia (0-4) para determinar la criticidad de los ítems evaluados. C) Finalmente, se analizan los puntajes.

Para cumplir con los estándares de calidad en temas de desarrollo de software institucional, es necesario evaluar las interfaces web, razón por la cual, se van a utilizar las dos metodologías antes mencionadas: *métodos heurísticos* para detectar errores de diseño y *recorrido cognitivo* para involucrar al usuario en el proceso.

3 Sistema embebido

A continuación, se revisan la propuesta de proyecto con el fin de alcanzar los objetivos revisados en el apartado 1.3.

3.1 Descripción de la propuesta

El proyecto se centra en el desarrollo de un sistema embebido en la plataforma de Inteligencia de Negocio con el que cuenta UNIVR. Para entender su implementación, se hace una breve descripción de la arquitectura tecnológica actual, seguido del proceso de baja y de la propuesta.

La institución tiene un almacén de datos y sus propias aplicaciones de Inteligencia de Negocios, como tableros de control, reportes, KPI's, etc. En la Figura 9 se puede ver de manera general la interacción del sistema de IN con los usuarios en los diferentes departamentos de gestión académica y su interacción con otros sistemas.

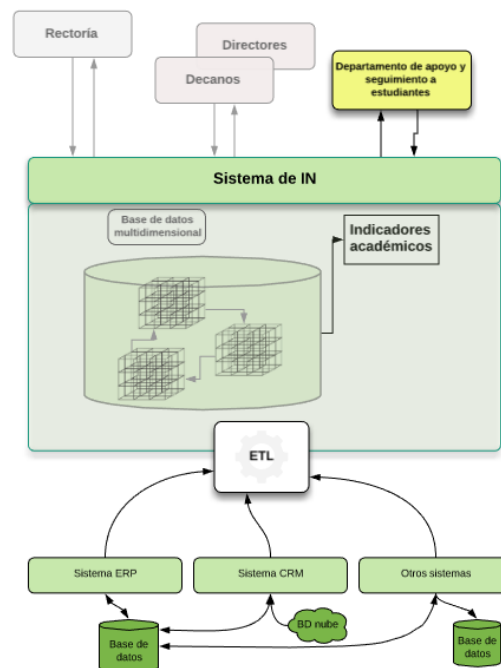


Figura 9. Arquitectura actual entre sistemas de inteligencia de negocios y otros sistemas transaccionales.

UNIVR cuenta con un Departamento para el Seguimiento Estudiantil (DSE), que tiene como objetivo apoyarlos con el fin de que continúen sus estudios. Una de sus tareas es la de realizar encuestas para detectar motivos de baja.

Este proceso comienza con la captura de información por medio del sistema ERP institucional y después es descargado manualmente en formato Excel cada cierto periodo de tiempo. Existen algunos indicadores al respecto, como bajas por programa, tendencias por periodo, % de bajas nuevo ingreso y otros. Para un nivel más departamental, se analiza la información de manera manual, es decir, del documento Excel se realizan gráficos, filtros y otros análisis muy particulares.

Para el caso de las deserciones, en el momento que se hace la consulta ya ocurrió la baja y lo que procedería es analizar los datos para encontrar factores en común o que estén incidiendo demasiado y aplicar las medidas necesarias a nivel directivo, ver Figura 10.

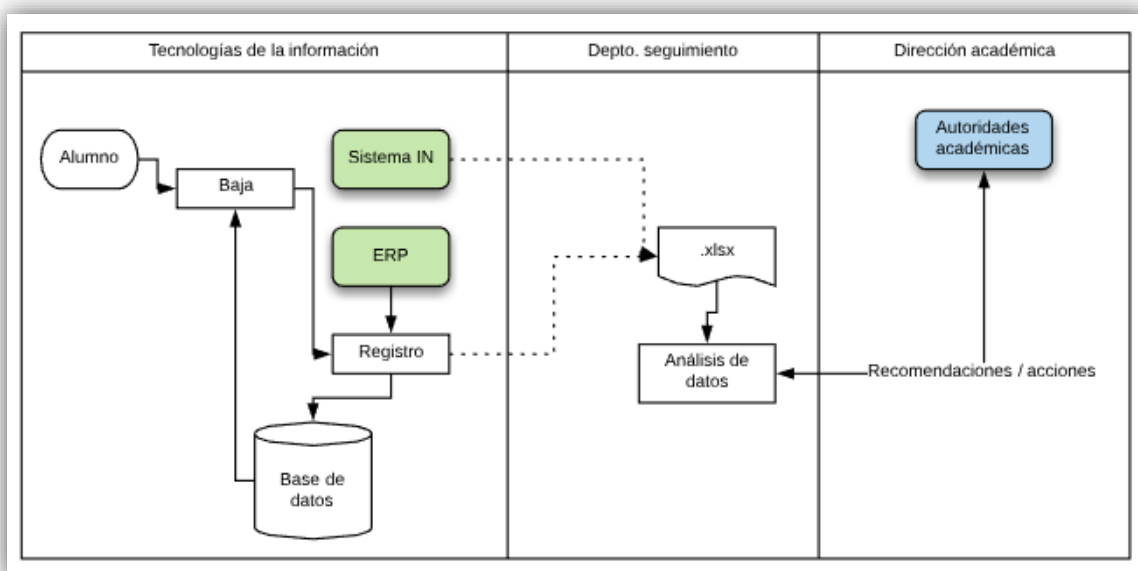


Figura 10. Proceso de baja.

El sistema embebido que se pretende realizar deberá trabajar junto al sistema de IN, como se puede observar en la Figura 11, esto quiere decir que va a utilizar las tablas de hechos y dimensiones existentes, a fin de explotar los datos recabados y procesados por el Departamento de Tecnologías de la Información (DTI). El modelo predictivo propuesto va a utilizar esta

información para su entrenamiento y se tendría que re-entrenar cada cierto tiempo con el objetivo de mejorar su exactitud y precisión, dicho de otra manera, entre más datos se utilicen, el modelo estará aprendiendo el comportamiento académico de la deserción. Finalmente, una interfaz gráfica (web) al usuario podrá mostrar los resultados del análisis predictivo por alumno o listado de alumnos (por programa).

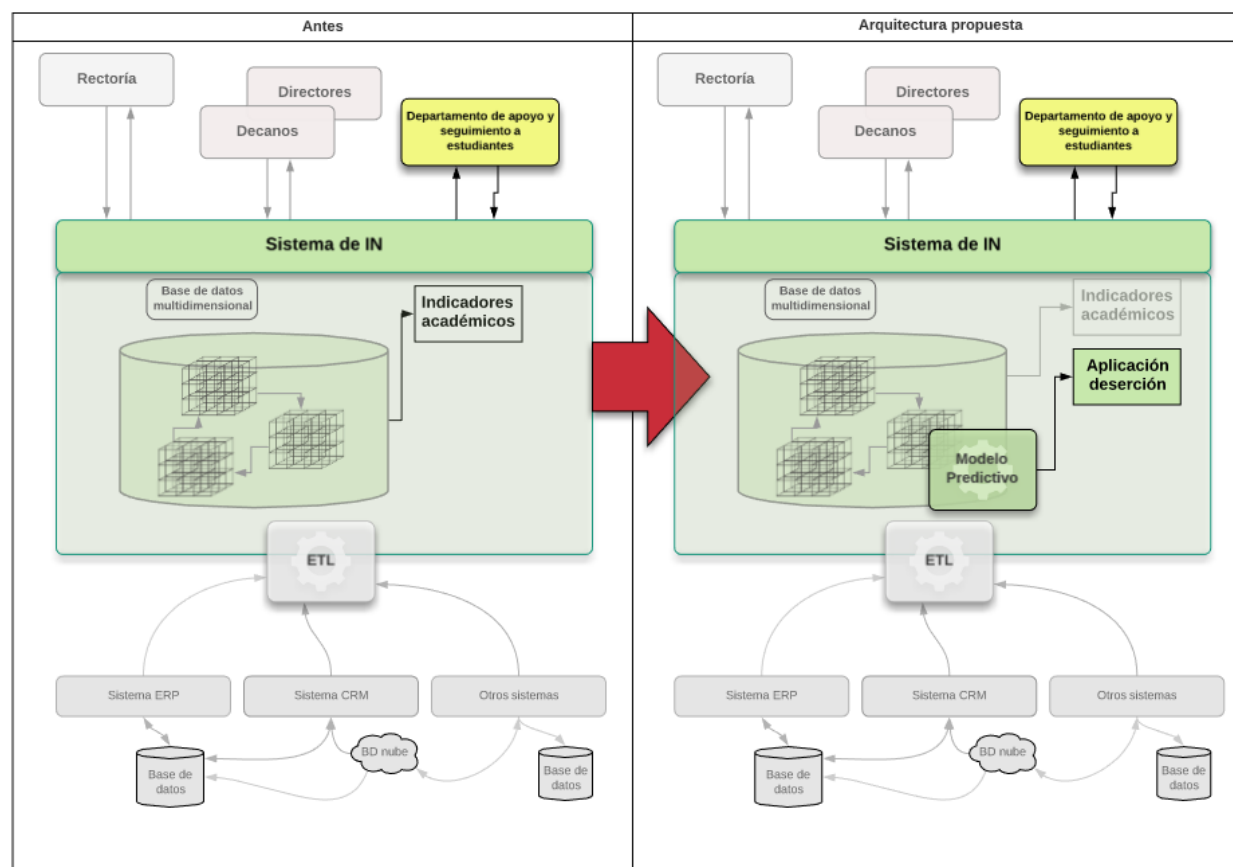


Figura 11. Aplicación embebida en el sistema de IN.

3.1.1 Requisitos funcionales y no funcionales

Para llegar al cumplimiento de los objetivos del proyecto, se lista a continuación los requerimientos funcionales y no funcionales que deberán de cumplirse con el fin de asegurar el éxito de la implementación y uso del sistema.

Requerimientos funcionales

De acceso

- El acceso a la aplicación embebida debe ser por medio del administrador de usuarios del sistema de IN.

De interfaz gráfica

- La aplicación web debe ser capaz de consultar la información procesada por el modelo de aprendizaje computacional, ya sea por alumno o por programa, mostrando información de riesgo de abandono.
- Filtro de datos por alumno, programa académico y periodo escolar (licenciatura).
- El tiempo de respuesta en el formulario por alumno no deberá de ser mayor a 3 segundos.
- El tiempo de respuesta en el formulario por programa no deberá ser mayor a 4 segundos.

Requerimientos no funcionales

De eficiencia

- El sistema debe ser capaz de procesar los datos históricos de los alumnos por periodo y por programa, para los entrenamientos de los modelos.
- Se seleccionará el modelo que mejor se ajuste a la solución de la problemática, dejando aquel que obtenga el mayor coeficiente de correlación Matthew y exactitud balanceada en las fases de “Test”.
- El entrenamiento de los modelos no debe tardar más de 20 minutos por ejecución, se espera ejecute un horario nocturno y una sola vez por periodo.
- La clasificación de las predicciones no debe superar los 5 minutos, ya que es un proceso que se ejecutará diario en el horario nocturno.

Seguridad y datos

- La comunicación de la interfaz web con el usuario deberá estar encriptada.
- Guardar un estadístico de los accesos de los usuarios.

Usabilidad

- El tiempo de capacitación del usuario debe ser menor a 3 sesiones de 1 hora.
- Se debe contar con un manual de usuario.

Desarrollo

- Los lenguajes para la implementación de los algoritmos son R y/o Python.
- Para el desarrollo web es *HTML5*, *JavaScript* y *Asp.NET* para las interfaces (según contingencias).
- Se dará preferencia a los conectores y/o API's de la plataforma de IN para facilitar la implementación de los modelos predictivos.

3.1.2 Producto esperado

El producto final será una aplicación web que utilizará un modelo predictivo que permita al usuario anticiparse y concentrarse en alumnos que puedan estar en riesgo de deserción y así, apoyararlos en la continuidad de sus estudios. La aplicación web contará con dos interfaces, la primera con información de riesgo por alumno y la segunda con información por programa académico, ver Figura 12 a) y b). En ambas interfaces deberá mostrar datos académicos del alumno y de contacto, información del programa y un porcentaje indicando la probabilidad de riesgo.

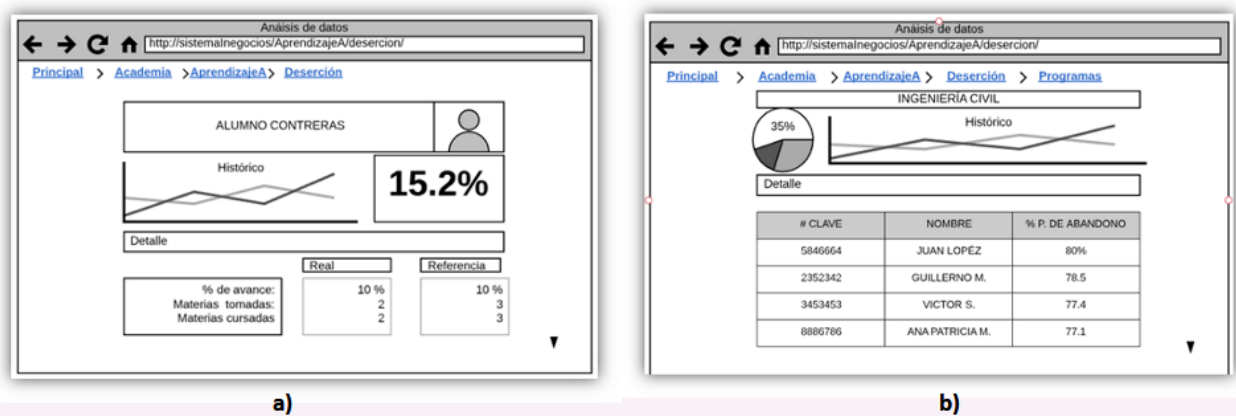


Figura 12. a) Aplicación web por programa académico b) Aplicación web por estudiante.

El modelo estará implementado en R embebido en la plataforma de IN por medio de API's y algunos conectores, ver Figura 13 .

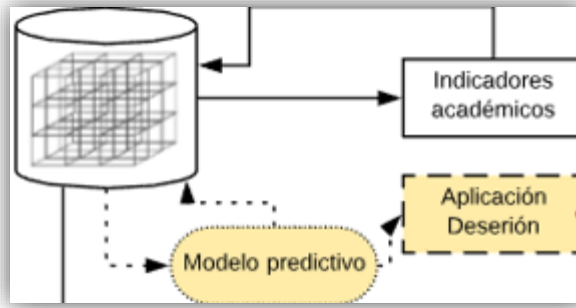


Figura 13. Sistema embebido en la arquitectura de IN institucional.

Adicionalmente, con la utilización de la herramienta se espera fomentar el uso de herramientas avanzadas de análisis para pronósticos, aprovechar los recursos tecnológicos institucionales y disminuir la utilización de Excel para los análisis.

4 Prototipo

El prototipo se desarrolló en la plataforma de IN de la institución con la herramienta Dossier 2019 y alineado a estos dos objetivos:

- Analizar el riesgo de deserción por alumno: debe mostrar métricas, como número de materias acreditadas, materias tomadas, porcentaje de avance según el plan de estudios del programa, etc.
- Análisis de riesgo de deserción desde la perspectiva de programa académico.

Características de los usuarios

Se dispone de dos tipos de usuarios teniendo así dos perfiles diferentes:

- Usuarios expertos: son cuatro personas y son desarrolladores de interfaces y también tienen a su cargo funciones de consultoría. Sus actividades principales están en el área de sistemas CRM, por lo tanto, interactúan con usuario finales, como alumnos, egresados, prospectos. Cumplen con los conocimientos en diseño de interfaces y usabilidad ya que todos son ingenieros de software.

- **Usuarios finales:** Es el usuario principal de esta aplicación y corresponde a los responsables del departamento de seguimiento a estudiantes. Tienen perfiles administrativos, como Lic. En administración de empresas. Son los principales responsables de recabar la información de los alumnos, analizarla y detectar aquellos alumnos que tienen necesidades de asesoramiento y seguimiento.

El prototipo será evaluado por cinco usuarios, los cuales serán los mismos para evaluar el producto final.

Escenarios de la evaluación

La evaluación con los expertos es en sus propias oficinas, ya que el prototipo fácilmente se puede visualizar en cualquier computadora que tenga un navegador y acceso a la red local. En cuanto al horario, se puede solicitar un espacio entre las 9 a.m. y las 6 p.m.

Con los usuarios finales, se dispone del mismo horario en sus oficinas, pero su tiempo es más demandado debido a sus actividades administrativas, por lo cual es más tardado concretar una cita.

Duración

Las sesiones están puestas de 20 – 25 minutos para en análisis de recorrido cognitivo y para el heurístico 40 mints.

Tareas

El primer paso que realiza el departamento de seguimiento de estudiantes es obtener un listado con los alumnos en algún tipo de riesgo académico en el periodo actual. A partir de ahí, los responsables toman decisiones para apoyar con tutorías a los alumnos. Entonces, la aplicación web debe realizar las tareas descritas en la Tabla 4.

Tabla 4. Tareas principales que realiza el usuario final en su análisis de riesgo académico.

#	Descripción de la tarea
1	Consultar aquellos alumnos en riesgo de deserción por programa.
2	Consultar aquellos alumnos en riesgo de deserción de todos los programas.
3	Consultar aquellos alumnos en riesgo de deserción por periodo de ingreso.
4	Consultar el expediente del alumno con información académica, datos de contacto e información del seguimiento personalizado.

4.1 Interfaz inicial

Con base a la definición de propuesta en la sección 3, se crea la primera versión del prototipo (1.1) con las siguientes características generales:

- Aplicación web con acceso al usuario desde su computadora conectada a la red local.
- Simulación de interacción con datos demostrativos, solamente para que el usuario pueda interactuar con los componentes de la interfaz y evaluar el prototipo.
- Información por periodos escolares de primavera y otoño. Existe diferencia en número de estudiantes por periodo, la cual es regular en sus históricos. El verano es un periodo corto y los ingresos son irregulares, razón por la cual se descarta. Esta política de análisis la establece el departamento de seguimiento estudiantil.

Interfaz web del análisis por alumno

En la Figura 14, se pueden apreciar los siguientes elementos:

- Filtro por alumno con la opción de buscar.
- Filtro por periodo actual: no modificable y oculto al usuario para mostrar información sobre el periodo corriente, permite mostrar la última información sobre él % de riesgo y algunas métricas, como materias tomadas, acreditadas, etc.
- Información histórica del % de riesgo por periodo.

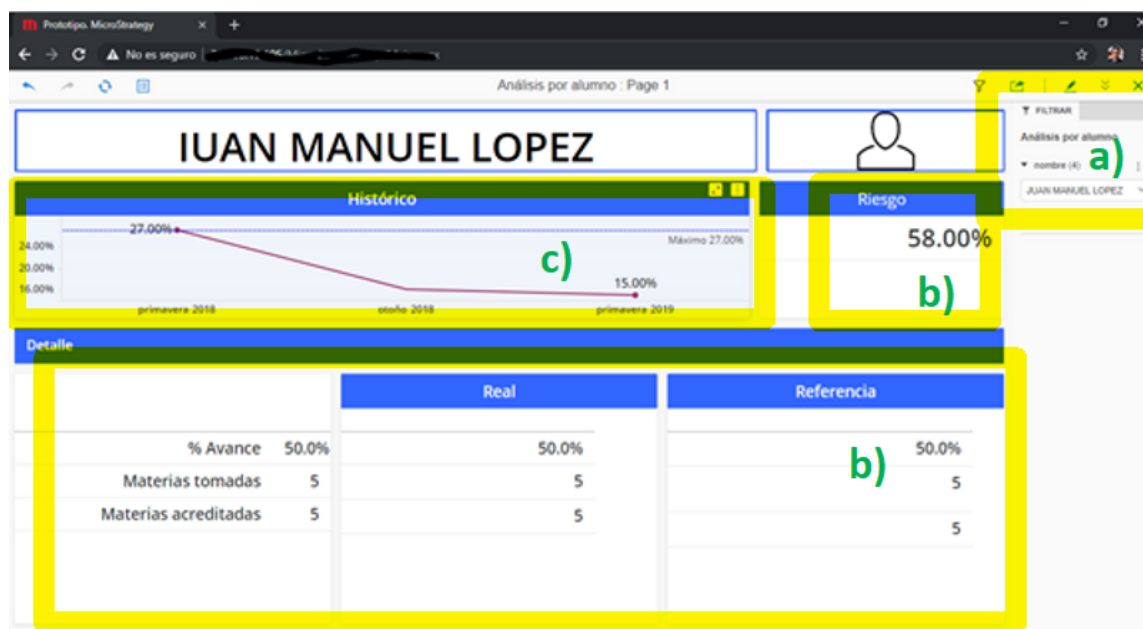


Figura 14. Primera versión de Interfaz web del análisis por alumno

Interfaz web del análisis por programa académico

En la Figura 15 se pueden visualizar los siguientes elementos:

- Selector para el programa académico.
- Sección de posibles bajas al siguiente periodo escolar.
- Histórico del comportamiento de la probabilidad de deserción por periodo académico.
- Listado de los alumnos que se encuentran en riesgo ordenados de mayor a menor.

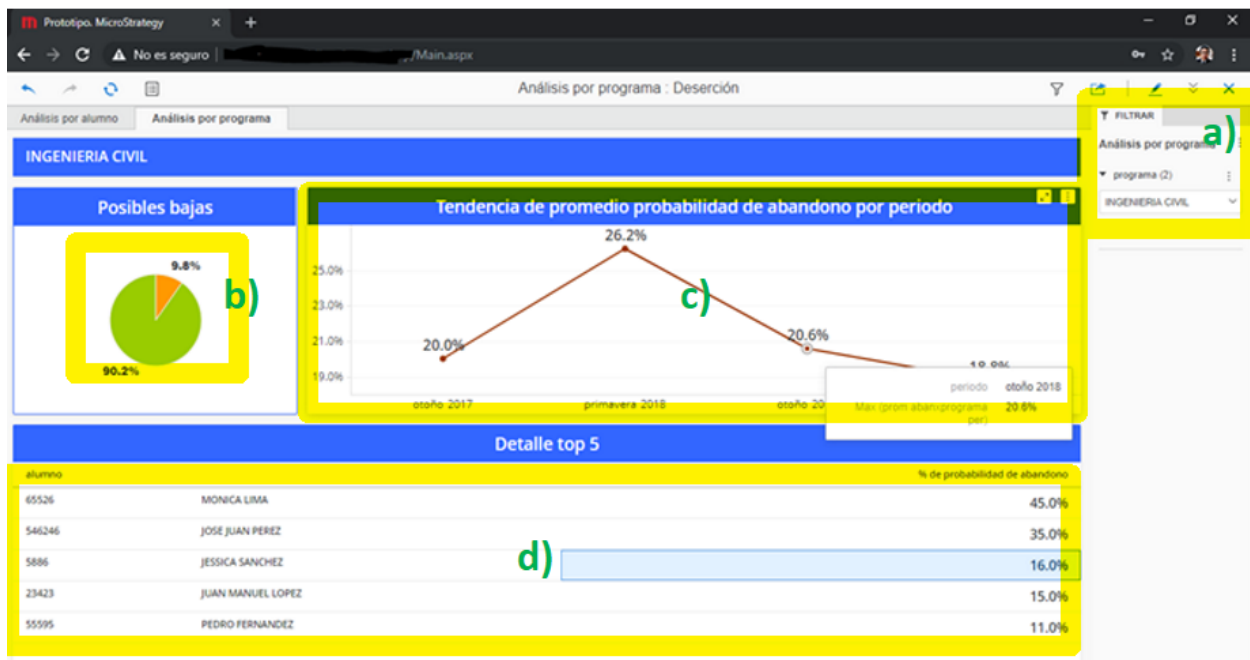


Figura 15. Primera versión de Interfaz web del análisis por programa académico.

4.2 Instrumento de evaluación

Para evaluar el prototipo se creó un instrumento basado en las 10 heurísticas de J. Nielsen (1994) y las 3 agregadas de Pierotti., D. (2004). Se obtuvo un total de 53 sub-heurísticas, distribuidas en las heurísticas de la

Tabla 5, solo para evaluar el prototipo.

Tabla 5. Heurísticas para la evaluación del prototipo.

Clave Heurística	Descripción
1	Visibilidad del estado del sistema
2	Empate entre el sistema y el mundo real(lenguaje de los usuarios)
4	Consistencia y estándares
6	Reconocimiento mejor que recuerdo.
8	Diseño estético
10	Ayuda y documentación
12	Iteración con el usuario de manera respetuosa y placentera

El criterio de evaluación fue:

- 0 = no es problema
- 1 = problema sin importancia
- 2 = problema de poca importancia
- 3 = problema grave
- 4 = catástrofe

Se decidió imprimir el instrumento para facilitar su llenado, ya que, en formato digital, se tendría que dividir la pantalla o tener otro dispositivo como Tablet o pantalla adicional: uno para visualizar la interfaz y otro para evaluarla, ver Figura 16 y Anexo 1. Para hacer referencia al instrumento, será **Instrumento tipo A**.

Evaluación de Prototipo

El objetivo de la encuesta dos interfaces web, las cuales van a mostrar información de riesgo académico para implementar tutorías efectivamente, los aspectos a evaluar son: visibilidad, lenguaje, estándares, facilidad de reconocimiento e interacción con el usuario.
Duración 20-30 minutos

Dirigido a personas con actividades en áreas tecnológicas con algún tipo de conocimiento en desarrollo de aplicaciones web: requerimientos de usuario, diseño web, programación, procesos o testing.

Criterio de evaluación
 0 = no es problema
 1 = problema sin importancia
 2 = problema de poca importancia
 3 = problema grave
 4 = catástrofe

Versión: _____
 Instrucciones: por cada ítem, evaluar según criterio del recuadro

	1.- Visibilidad del estado del sistema	Puntuación
1H01	¿Cada parte de la interfaz comienza con un título o encabezado que describa el contenido de la pantalla?	
1H02	¿El esquema de diseño de los íconos y su estética es consistente en todo el sistema?	
1H04	Los menús de instrucciones, puntos de entrada de datos y mensajes de error ¿aparecen en el mismo lugar de la pantalla o en el mismo menú?	
1H08	¿Hay un tipo de "respuesta" para cada acción u operación?	
1H09	¿El sistema provee algún tipo de "respuesta" visual en menús o cuadros de dialogo que indiquen las opciones que pueden seleccionarse?	
1h25	La terminología utilizada en los menús, ¿es consistente con el dominio del conocimiento del usuario en relación a la tarea a realizar?	
1h26	¿El sistema provee visibilidad?, es decir ¿puede expresar verbalmente el estado del sistema y que alternativas de acción posee en determinado momento?	
	2.- Empate entre el sistema y el mundo real(lenguaje de los usuarios)	
2h01	¿Los iconos son concretos y familiares para el usuario?	
2h02	Dada una determinada lista de nombres de ítems y variables para realizar tareas, ¿las opciones en los menús (nombres de los ítems) están ordenadas en la manera más lógica para el usuario?	

Figura 16. Recorte de instrumento (heurísticas) para evaluar prototipo.

Para alinearse a la funcionalidad que se espera, se creó otro instrumento basado en la metodología de recorrido cognitivo, para evaluar las tareas a realizar en la interfaz, ver Figura 17 y Anexo 2.

El objetivo de la encuesta es validar la funcionalidad según las tareas que el usuario puede realizar

Duración 20-25 minutos

Dirigido a: 1) personas con actividades en áreas tecnológicas con algún tipo de conocimiento en desarrollo de aplicaciones web: requerimientos de usuario, diseño web, programación, procesos o testing. 2) usuarios finales con algún perfil administrativo en el cual tengan a su cargo la explotación de la información académica para el seguimiento de alumnos.

PROT1.1

Instrucciones: evaluar cada acción contestando las 4 preguntas: S(si) N(no)

TAREAS / acciones	ACCIONES	¿El usuario intentará lograr el efecto correcto?	¿Percibirán los usuarios que está disponible la acción correcta?	¿Los usuarios asociarán la acción correcta con el resultado que espera lograr?	Si se realiza la acción correcta ¿Los usuarios verán que se está avanzando hacia el resultado?
1 Consultar expediente de alumno en prototipo	Seleccionar la pestaña de interfaz por alumno: "Análisis por alumno"				
	Clic en el buscador de alumnos (cuadro de texto, lado derecho)				
	Escribir el uid de un alumno: 12221				
	Seleccionar el iud que aparece y darle clic.				
	Consultar la información del alumno en la sección superior.				
	Consultar el histórico de riesgo académico del alumno.				
2 Consultar aquellos alumnos en riesgo de deserción por programa	Seleccionar la pestaña de interfaz "Análisis por programa"				
	En la sección de filtros por programa seleccionar la opción de "INGENIERIA CIVIL"				
	Consultar la tendencia de bajas del programa				
	En la sección "detalle de alumnos" mover el scroll inferior a la derecha				
	Verificar que las materias acreditadas corresponda con el % de riesgo (deben ser menores)				

Figura 17. Recorte de instrumento (recorrido cognitivo) para evaluar prototipo.

Para hacer referencia, el instrumento se llamará **Instrumento tipo B**.

4.2 Presentación de resultados

La evaluación con los instrumentos A y B se realizaron en las mismas oficinas de los usuarios expertos, ver Figura 18, el procedimiento fue el siguiente:

- Se dio una breve explicación del contexto del proyecto a cinco usuarios.
- Se les proporcionó una URL para visualizar el prototipo en línea con una credencial temporal.
- Se les facilitó los instrumentos impresos tipo A y B.
- Se respetó los tiempos asignados.

Algunas consideraciones observadas al aplicar las encuestas fueron:

- En ocasiones no se podía aplicar las dos encuestas, entonces hubo la necesidad de aplicarlo en dos momentos.
- La duración de las sesiones puede ser menor a lo planeado.
- Se puede apoyar de otras herramientas como Tablet, para no interrumpir a los usuarios de sus actividades normales.

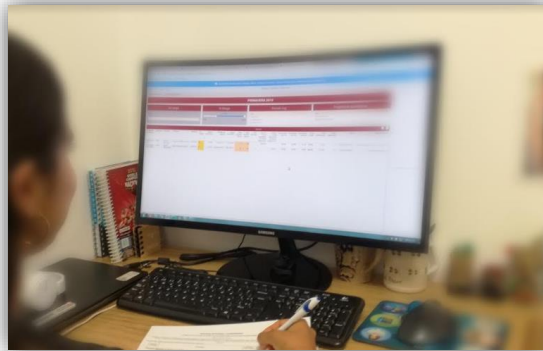


Figura 18. Aplicación de evaluación de usabilidad.

Después de aplicar las 5 encuestas tipo A, se obtuvo con los resultados la gráfica de la Figura 19.

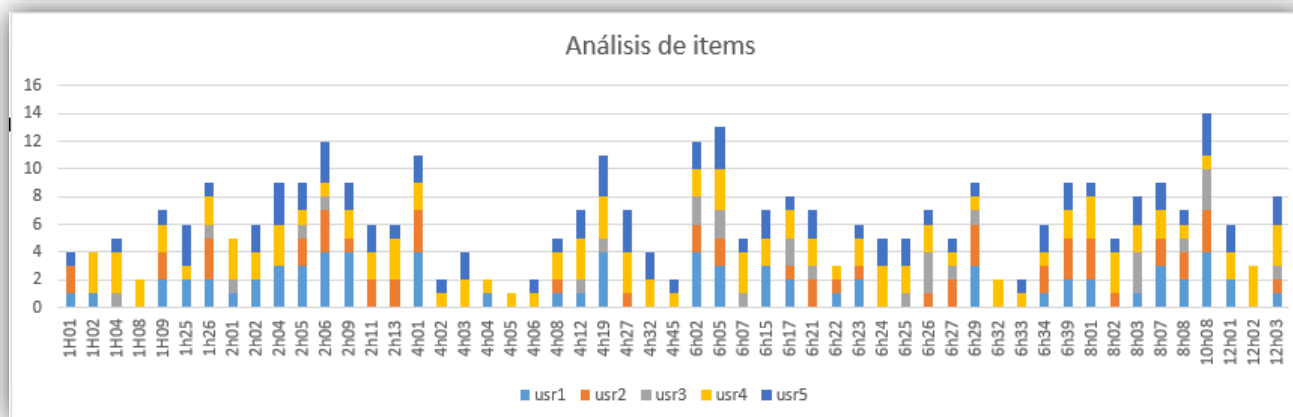


Figura 19. Aplicación de evaluación de usabilidad tipo A.

Se puede notar que sobresalen las siguientes claves:

- 2H06 Los colores seleccionados ¿Corresponden a valores esperados según los códigos de los usuarios?
- 4H01 ¿Los formatos de la institución han sido respetados de manera consistente a lo largo de las distintas pantallas del sistema?
- 4H19 ¿Las etiquetas de campos mantiene una forma consistente entre una pantalla y otra?
- 6H02 ¿El despliegue de datos comienza en la parte superior de la izquierda de la pantalla?
- 6H05 Los apuntadores, pistas visuales y mensajes ¿Están posicionados en lugares de la pantalla en donde es probable que el usuario dirija su mirada?
- 10H08 ¿La función de ayuda del menú es visible?

Verificamos obteniendo el valor de *criticidad*, Figura 20.



Figura 20. Valor de criticidad de las heurísticas.

Las heurísticas con valores críticos son:

- 2h04. Los campos relacionados e interdependientes, ¿Aparecen en la misma pantalla?
- 6h05 Los apuntadores, pistas visuales y mensajes están posicionados en lugares de la pantalla en donde es probable que el usuario dirija su mirada?
- 10h08 ¿La función de ayuda del menú es visible?

De este ejercicio, se concluye que las heurísticas 2h04, 6h05 y 10h08 se les puede considerar un problema mayor y se le debe dar la importancia correspondiente.

Se evaluaron dos tareas con el instrumento tipo B, los resultados se pueden visualizar en la Figura 21. Esta evaluación permitió observar si los elementos en la interfaz tienen sentido al momento de efectuar una tarea. Aunque fueron sencillas, permitió ver que, aun siguiendo los pasos que se les indican, puede ser posible que no sepan que esperar después de realizar una actividad. Esto indica que es necesario que la interfaz debería ayudar a aclarar los resultados esperados al usuario.

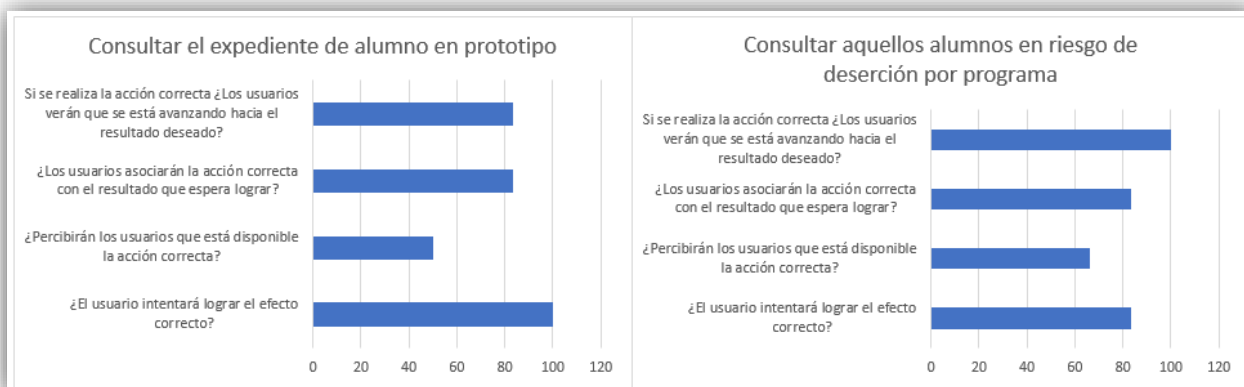


Figura 21. Resultados de la evaluación **tipo B**.

4.3 Ajustes al prototipo

Después de hacer las evaluaciones, y entrevistando nuevamente al usuario final, se aplicaron cambios al prototipo (versión 2.1), quedando las interfaces con algunas mejoras, Figura 22 y Figura 23.

Análisis General : Deserción

Análisis General | Análisis por alumno

PRIMAVERA 2019

En riesgo a) **% Riesgo** **Periodo Ing.** **Programas académicos** b)

(All) **TODOS** 0.0% 78.0%

(All) otoño 2017
primavera 2016
primavera 2017

(All) INGENIERIA CIVIL
INGENIERIA INDUSTRIAL

Detalle

Programa Académico	Matrícula	Nombre	Correos	Telefonos	En Riesgo	Avance académico	Periodo de ingreso	Apoyo Financiero	Baja materia 1 fecha	Baja materia 2 fecha	Mat. Inscritas	¿Dos (2) veces reprobada la misma asignatura?	¿Tres (3) veces reprobada la misma asignatura?	Reprobada 1 parcial	Reprobada 2 parcial	Reprobada 3 parcial	% de riesgo	% de reprobación
INGENIERIA CIVIL	5886	JESSICA SANCHEZ	jess.sanch23@univer.edu	22262343	S	50.0%	otoño 2017	CULTURAL	1	0	3	NO LERR		30.0%	20.0%	10.0%	78.0%	22.00%
	23423	JUAN MANUEL LOPEZ	joem.jlope2@univer.edu	24158847	S	50.0%	otoño 2017	BECA	0	0	4			0.0%	0.0%	0.0%	0.0%	0.00%
	65526	MONICA LIMA	monica.limans@univer.edu	22258781	S	45.0%	primavera 2017	CULTURAL	0	0	5			0.0%	0.0%	0.0%	0.0%	0.00%
	546246	JOSE JUAN PEREZ	jose.jm@univer.edu	22265784	S	80.0%	primavera 2017	BECA	0	0	3			0.0%	0.0%	0.0%	0.0%	0.00%
INGENIERIA INDUSTRIAL	3323	CARLA GOMEZ	carla.gomme2@univer.edu	22265554	S	85.0%	primavera 2016		0	1	3			0.0%	0.0%	0.0%	0.0%	0.00%
	3333	CARLOS FERNANDEZ	pedro.feers@univer.edu	22262557	N	86.0%	primavera 2016		0	1	2			0.0%	0.0%	0.0%	0.0%	0.00%
	12221	PATRICIA GONZALES	pat.gonzal@univer.edu	22286659	S	86.0%	primavera 2016		0	0	2			0.0%	0.0%	0.0%	0.0%	0.00%
	55434	FERMANDO	carla.gomme2@univer.edu	22286659	N	86.0%	primavera 2016		0	0	2			0.0%	0.0%	0.0%	0.0%	0.00%

Figura 22. Prototipo 2.1 con la interfaz "Análisis General".

En la interfaz de “Análisis general” se puede observar lo siguiente:

- Se agruparon y se hace visibles las opciones donde el modelo predictivo realiza su trabajo, etiqueta alumnos en riesgo y permite filtrar (S o N).
- Son otras opciones de filtrado, más usados por el usuario final.
- Existen varias métricas que el usuario final revisa, antes de planear contactar y dar seguimiento a un alumno.

Análisis por alumno : Individual

Análisis General | Análisis por alumno

INFORMACIÓN DEL ESTUDIANTE

INGENIERIA CIVIL

Datos del estudiante

Nombre	JESSICA SANCHEZ
Género	femenino
Estado	PUEBLA
Ciudad	PUEBLA
Correo	jess.sanch23@univer.edu
Teléfonos	22262343
Periodo de ingreso	otoño 2017
¿Alumno en riesgo?	☐

Desempeño académico

Reprobadas 1er parcial	30.0%
Reprobadas 2er parcial	20.0%
Reprobadas 3er parcial	10.0%

Riesgo académico

16.0%

Seguimiento tutoría

Tutor asignado	ALEJANDRO AYALA
Observaciones tutoría	En espera de entrevista con el alumno.

Figura 23. Prototipo 2.1 con la interfaz “Análisis por alumno”.

En la interfaz de análisis por alumno:

- a) En esta sección se encuentra toda la información de contacto del estudiante.
- b) En *desempeño académico* se encuentran las variables que el usuario final necesita conocer para determinar si se debe contactar al alumno.
- c) En *Seguimiento tutoría* muestra información de las tutorías del alumno y es parte de las acciones remediales. Esta sección dependerá si se encuentra disponible en el almacén de datos.

5 Resultados

En las siguientes secciones se trata la descripción del conjunto de datos utilizado, su preparación, como se hizo la selección de atributos, la experimentación con modelos de clasificación, la implementación de la solución web y la evaluación con el usuario.

5.1 Conjunto de datos utilizado

Como se especificó en la sección 1.5.2, se trabajó con un conjunto de datos de 51,497 registros correspondientes al decanato de Ingenierías, ciencias sociales y administrativas, con 26 atributos más la clase. El conjunto corresponde a 19 periodos escolares (Conjunto 1). Se agregó un segundo conjunto de datos con 2,244 registros (con la misma estructura), del último periodo (primavera 2020), dicha información se excluyó de todas las etapas por el hecho de que no está etiquetada la clase DESERTOR, por lo tanto se sumó al conjunto de test final (Conjunto 2), ver Figura 24. Los datos se obtuvieron en un archivo con formato csv.

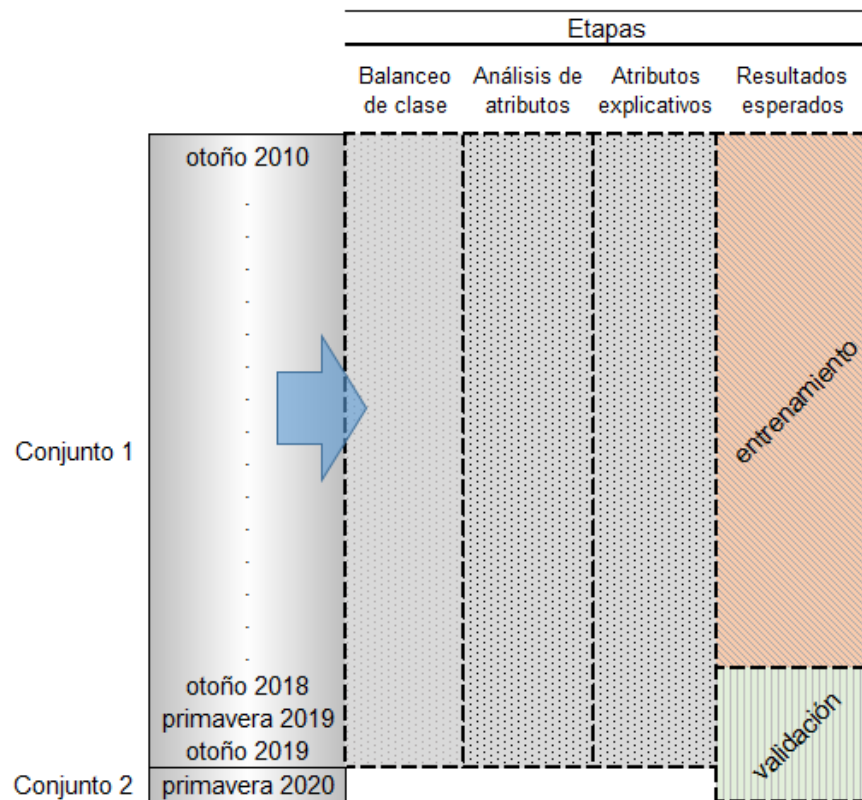


Figura 24. Conjunto de datos.

Es importante señalar que, para la fase de preparación de los datos, análisis descriptivo y entrenamiento de modelos, se trabajó con el archivo .csv, sin embargo, para su implementación en producción, estos registros se pueden obtener de manera automática cada vez que se ejecuta el proceso de entrenamiento del modelo, ver Figura 25. Este paso se hace a través de una consulta SQL que es invocada por una tarea programada que ejecuta un script en R, ver Sección 6.4.

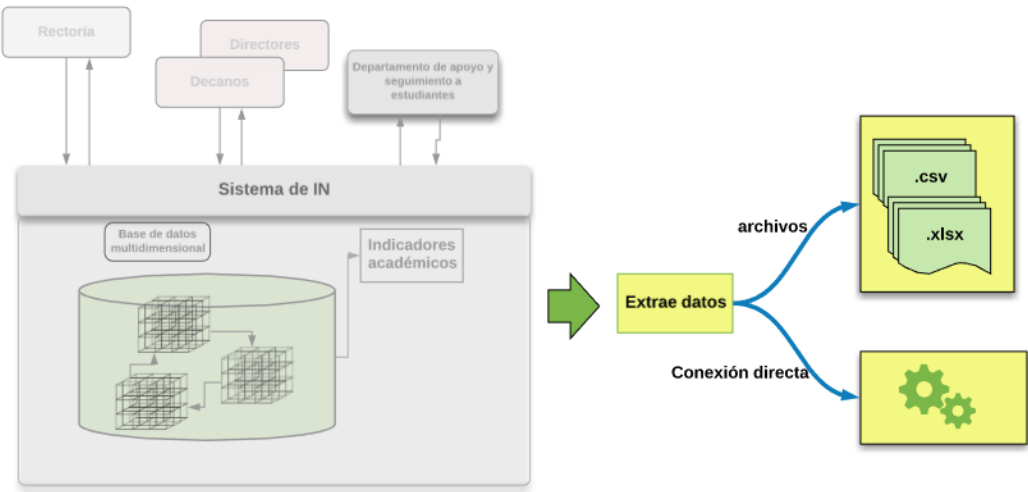


Figura 25. Extracción de datos de la plataforma de IN.

La Tabla 6 muestra el número de registros por programa académico del *Conjunto 1*. En total son 25 programas de licenciatura.

Tabla 6. Número de instancias por programa académico.

Programa	# de instancias
Adm. Empresas	3309
Adm. Educación	3801
Adm. Financiera	1467
Comercio	2718
Contaduría	2128
Adm. Restaurantes	3404
Negocios	833
Logística	1049

<i>Mercadotecnia</i>	2522
<i>C. políticas</i>	830
<i>P. Audiovisual</i>	2172
<i>Comunicación</i>	1563
<i>Derecho</i>	4035
<i>Economía</i>	851
<i>R. Internacionales</i>	1641
<i>Ing. civil</i>	2541
<i>Ing. Industrial</i>	2591
<i>Ing. Biónica</i>	1289
<i>Ing. Computación</i>	1063
<i>Ing. en Diseño</i>	4049
<i>Ing. Electrónica</i>	326
<i>Ing. en Manufactura</i>	901
<i>Ing. Química</i>	1482
<i>Mecatrónica</i>	4124
<i>Software</i>	808
TOTAL	51,497

En la Figura 26 se puede observar el número de instancias, según la clase *Desertor*. Se nota que existe una desproporción entre clases, lo cual puede llevar a obtener un modelo con resultados favoreciendo a la clase mayoritaria. (Arnejo, 2016). La proporción de la clase *Desertor(S)* representa el 12.19% del total de los datos.

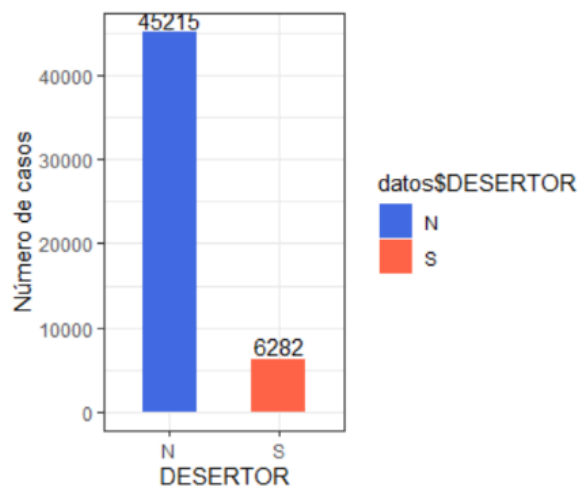


Figura 26. Número de casos por clase DESERTOR.

Balanceo de clase

Por lo tanto, el siguiente paso fue balancear la clase *Desertor*. Se utilizaron los datos del *Conjunto 1* y se probaron diferentes configuraciones para realizar entrenamientos con incremento de dos y tres veces los datos de la clase *Desertor(S)* y de muestras aleatorias al 50%, 60%, 70% y 80% de la clase *Desertor(N)*. Las pruebas se realizaron con todos los atributos disponibles. La Tabla 7 muestra las configuraciones y el resultado en cuanto número de registros y proporción de los nuevos datos.

Tabla 7. Configuraciones para balanceo de clase.

	REGISTROS		PROPORCIÓN		TOTAL
	S	N	S	N	
(1x, 100%)	6282	45215	12.20%	87.80%	51497
(2x, 50%)	12564	22607	35.72%	64.28%	35171
(2x, 60%)	12564	27129	31.65%	68.35%	39693
(2x, 70%)	12564	31650	28.42%	71.58%	44214
(2x, 80%)	12564	36172	25.78%	74.22%	48736
(3x, 50%)	18846	22607	45.46%	54.54%	41453
(3x, 60%)	18846	27129	40.99%	59.01%	45975
(3x, 70%)	18846	31650	37.32%	62.68%	50496
(3x, 80%)	18846	36172	34.25%	65.75%	55018

Finalmente, el resultado se puede apreciar en la Figura 27, en la cual la configuración con los máximos porcentajes en precisión, exactitud, sensibilidad y especificidad fue la de aplicar: 3x para la muestra de la clase *Desertor(S)* y el 80% de la clase *Desertor(N)*, Esto significa incrementar la proporción de la clase *Desertor(S)* hasta un 34.25% del total de los nuevos datos, ver datos resaltados de la Tabla 7. Visualmente en el gráfico se muestra con las líneas punteadas como se ajustan en las diferentes métricas de manera homogénea.

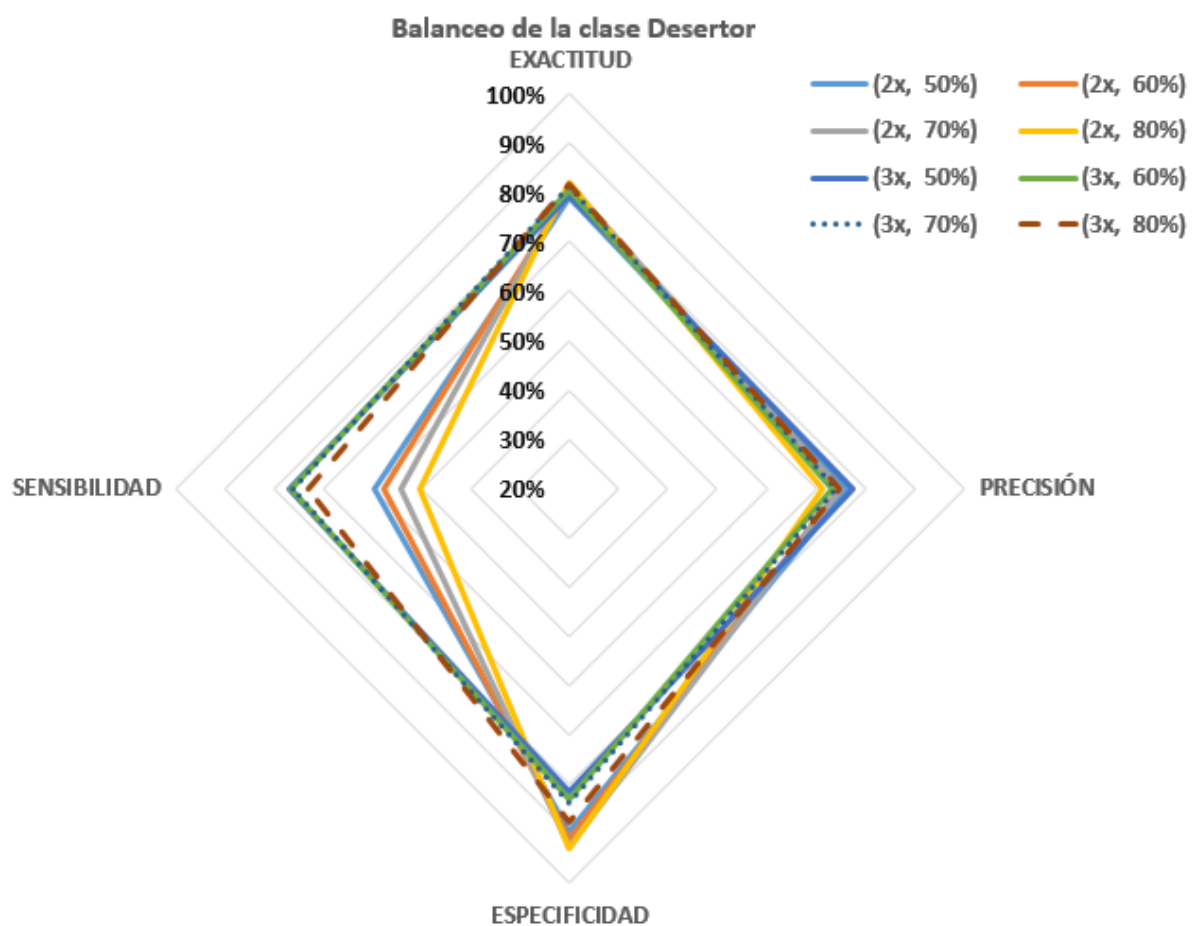


Figura 27. Balanceo de clase.

Los datos de la configuración (3x,80%), representados en la Figura 28, serán usados para los entrenamientos y validaciones de los modelos clasificadores.

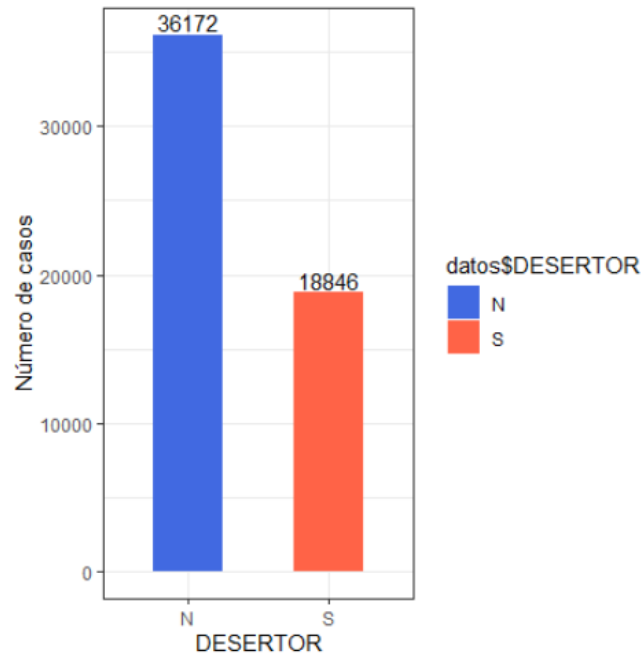


Figura 28. Conjunto 1 con la Clase DESERTOR balanceada.

Análisis de atributos

A continuación, la Tabla 8 muestra la descripción de los atributos facilitados por la institución, este análisis se realizó con el conjunto total de registros. Las variables obtenidas pueden ser redundantes, es por eso que, en una etapa posterior se obtendrán las más significativas. La mayoría de la información es del tipo académica ya que es consistente en los históricos, además que es la más usada por el departamento de seguimiento.

Tabla 8. Descripción de los atributos.

<i>Campo</i>	<i>Descripción</i>	<i>Tipo</i>	<i>Subtipo</i>	<i>Temporalidad del cálculo</i>	<i>Tipo de dato</i>
<i>PERIODO</i>	Periodo escolar	Académica	Periodo	Dato al periodo consultado	Alfanumérico
<i>PROM_PANTE</i>	Promedio al periodo anterior. Cuando es nuevo ingreso, no tiene promedio de periodo anterior.	Académica	Promedios	Dato al periodo anterior	Numérico
<i>PROM_INI</i>	Promedio al inicio de inscripción.	Académica	Promedios	Dato inicial.	Numérico
<i>EDAD</i>	Edad del alumno.	Demográfica	Edad	Dato al periodo consultado	Numérico
<i>FALTAS_PANTE</i>	Faltas hasta su último periodo.	Académica	Faltas	Dato al periodo anterior	Numérico
<i>ASIST_PANTE</i>	Asistencias hasta su último periodo.	Académica	Asistencias	Dato al periodo anterior	Numérico
<i>GÉNERO</i>	M masculino, F Femenino.	Demográfica	Género	Dato inicial.	Alfanumérico
<i>SEMESTRE_PANTE</i>	Semestre al periodo anterior.	Académica	Avance	Dato al periodo anterior	Numérico
<i>APOYO</i>	Indica si recibe algún tipo de apoyo como descuentos.	Económico	Económico	Dato al periodo consultado	Numérico

<i>REPRO_PANTE</i>	Asignaturas reprobadas al hasta su último periodo.	Académica	Reprobación	Dato al periodo anterior	Numérico
<i>REPRO_1X</i>	Asignaturas reprobadas una sola vez.	Académica	Reprobación	Dato al periodo anterior	Numérico
<i>REPRO_2X</i>	Asignaturas reprobadas dos veces	Académica	Reprobación	Dato al periodo anterior	Numérico
<i>REPRO_3X</i>	Asignaturas reprobadas tres o más veces.	Académica	Avance	Dato al periodo anterior	Numérico
<i>PERIODOS_INSCRIPTOS</i>	Todos los periodos en los que se ha inscrito.	Académica	Avance	Dato al periodo anterior	Numérico
<i>PROMEDIO_AC</i>	Promedio actual.	Académica	Avance	Dato al periodo anterior	Numérico
<i>CRED_PANTE</i>	Créditos cursados.	Académica	Avance	Dato al periodo anterior	Numérico
<i>AVANCE_XCRED</i>	Avance según créditos aprobados.	Académica	Avance	Dato al periodo anterior	Numérico
<i>ASIG_INSCRITASP</i>	Asignaturas inscritas en el periodo.	Académica	Asignaturas	Dato en el periodo consultado	Numérico
<i>FALTASP</i>	Faltas en el periodo.	Académica	Faltas	Dato en el periodo consultado	Numérico
<i>ESTADOCIVIL</i>	Estatus civil.	Demográfica	Estatus	Dato inicial	Numérico
<i>NINGRESO</i>	Indica que es alumno de nuevo ingreso.	Académica	Nuevo ingreso	Dato al periodo consultado.	Numérico
<i>FORANEO</i>	Indica si proviene fuera del estado.	Demográfica	Origen	Dato inicial.	Numérico
<i>REPROBADAS1P</i>	Proporción de asignaturas reprobadas en el primer parcial.	Académica	Reprobación	Dato en el periodo consultado.	Numérico
<i>REPROBADAS2P</i>	Proporción de asignaturas reprobadas en el segundo parcial.	Académica	Reprobación	Dato en el periodo consultado.	Numérico
<i>REPROBADAS3P</i>	Proporción de asignaturas reprobadas en el tercer parcial.	Académica	Reprobación	Dato en el periodo consultado.	Numérico
DESERTOR	S N (atributo clase)				Alfanumérico

La gráficas de la Figura 29, Figura 30, Figura 31 y Figura 32 muestran la densidad de cada atributo por la clase *Desertor*.

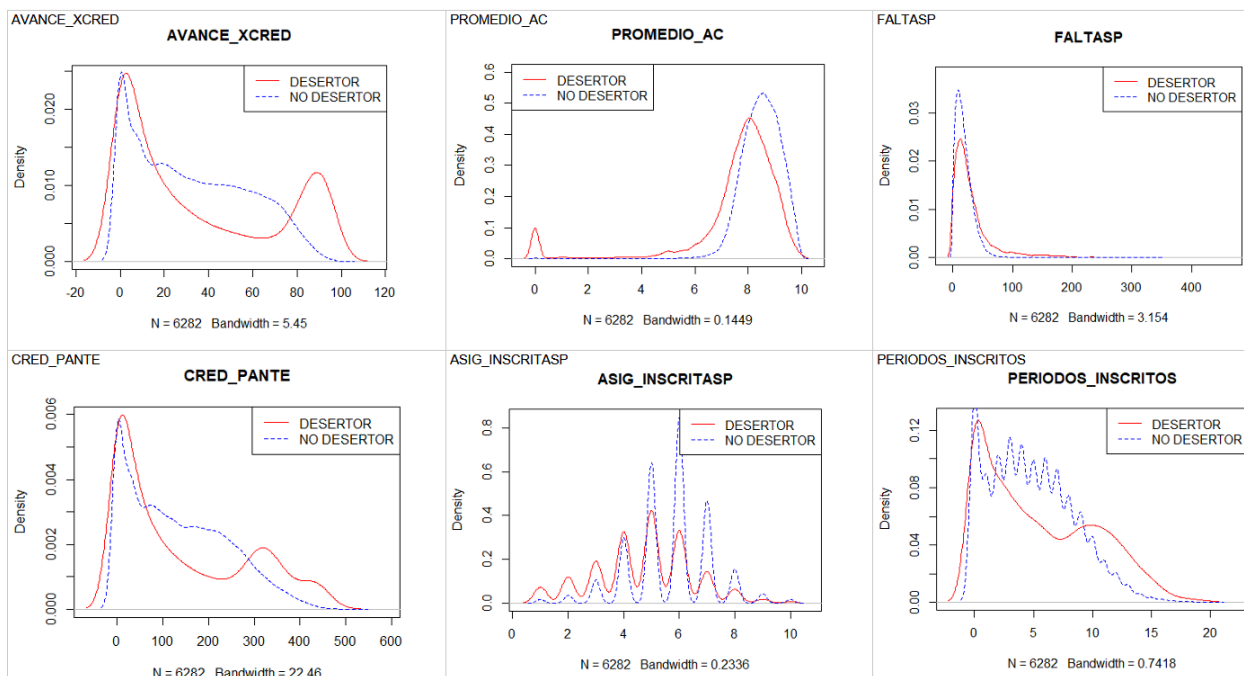


Figura 29. Diagrama de densidad de los atributos AVANCE_XCRED, PROMEDIO_AC, FALTASP, CRED_PANTE, ASIG_INSCRITAS Y PERIODOS_INSCRITOS.

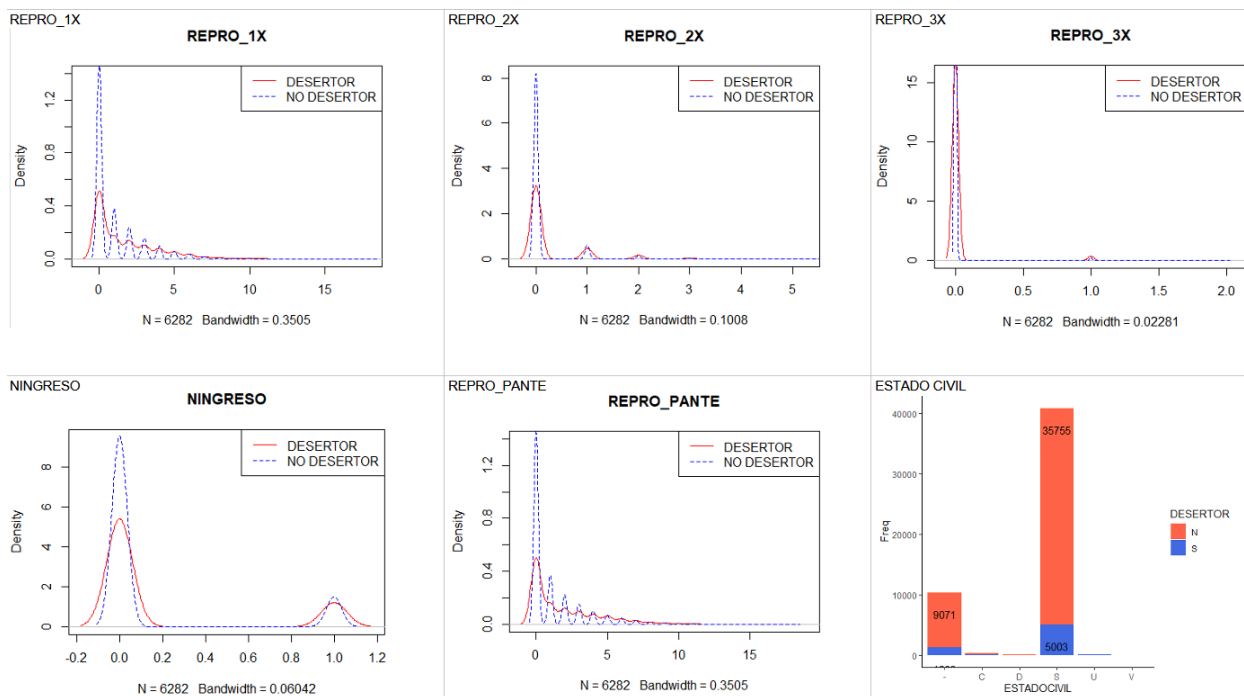


Figura 30. Diagrama de densidad de los atributos REPRO_1X, REPRO_2X, REPRO_3X, NINGRESO, REPRO_PANTE, ESTUS_CIVIL.

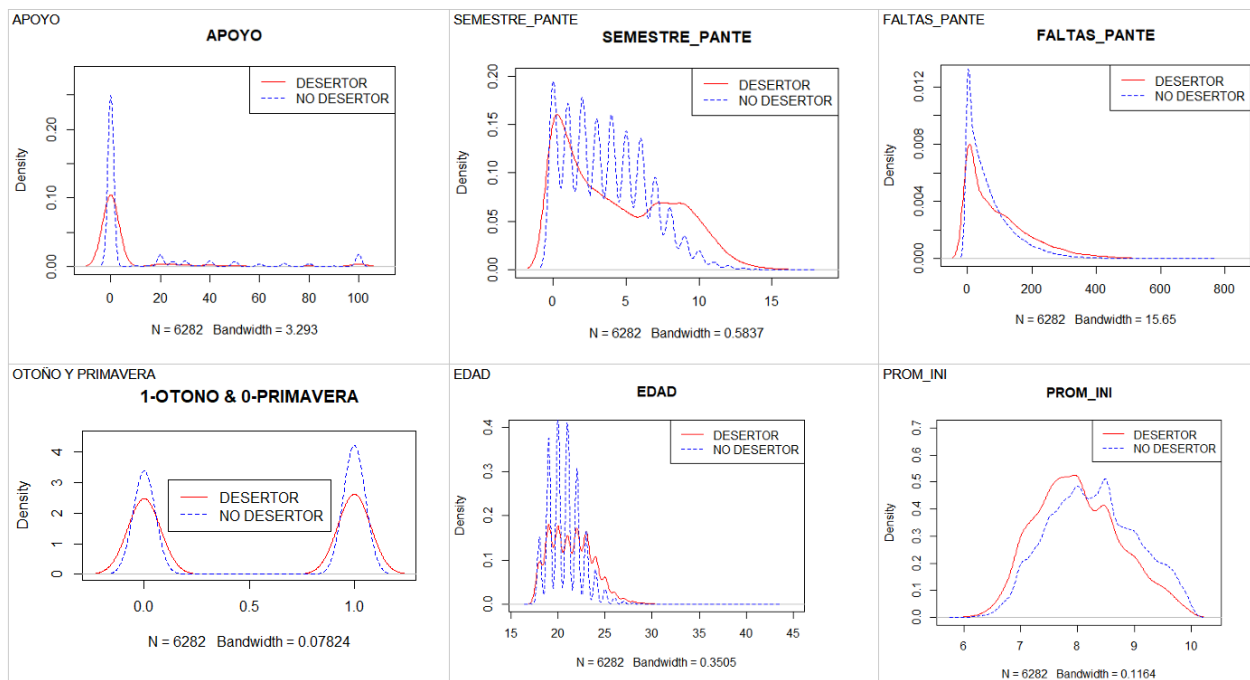


Figura 31. Diagrama de densidad de los atributos APOYO, SEMESTRE_PANTE, FALTAS_PANTE, OTOÑO Y PRIMAVERA, EDAD Y PROM_INI.

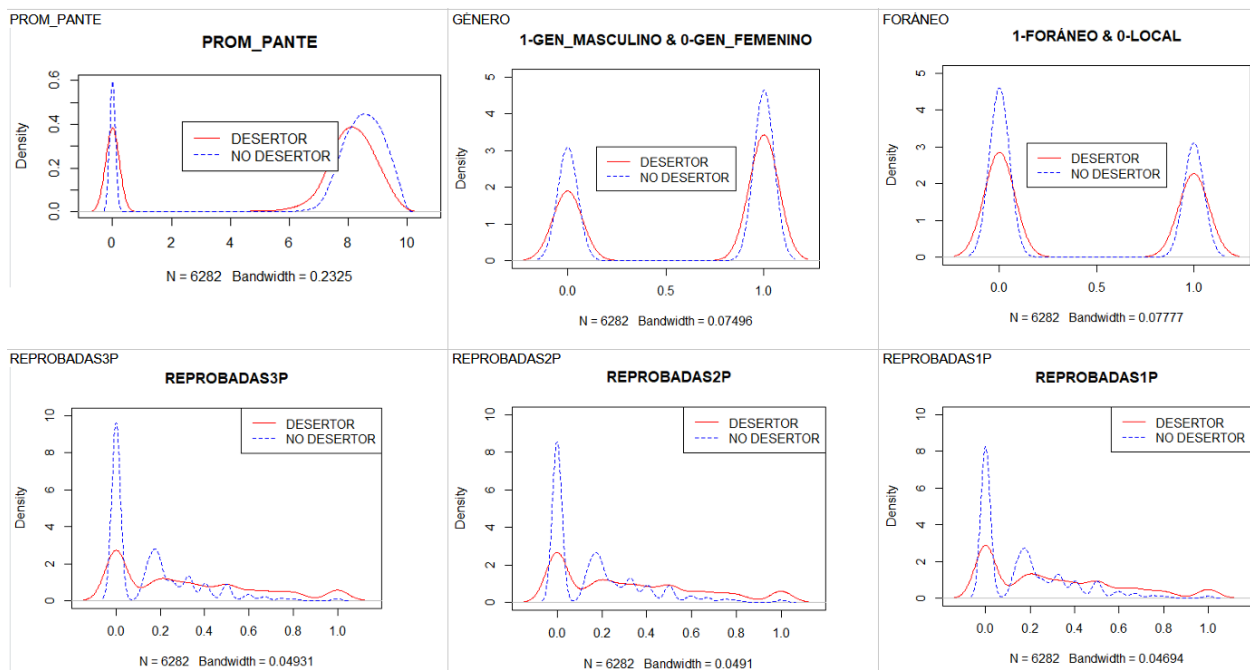


Figura 32. Diagrama de densidad de los atributos PROM_PANTE, GEN_MASCULINO, GEN_FEMENINO Y FORÁNEO.

Las gráficas muestran la dispersión de los datos y permite detectar a primera vista algunas características entre la clase *Desertor*:

- PROM_INI. Los desertores tienen una dispersión hacia promedios de ingreso más bajos.
- AVANCE_XCRED. Existen dos momentos en los que hay más desertores, al inicio de su vida académica, y al final de su carrera (en menor medida). Se puede notar que aquellos que están cerca del 60% de avance son los que menos desertan.
- REPROBADAS1P, 2P y 3P. Los desertores se diferencian más de la otra clase cuando comienzan a reprobar alrededor del 10% de sus asignaturas por parcial. Después, a partir del 50% de reprobación, nuevamente se nota una diferencia.
- PROMEDIO_AC. Promedio actual mantiene una dispersión similar al promedio de ingreso para los desertores.

En seguida, se analizaron los atributos con “algoritmos de selección” para poder obtener las variables más significativas, y eliminar redundancia.

5.2 Atributos explicativos

Se utilizaron algoritmos para *selección de atributos* con R y las librerías *FSelector* y *Boruta* (R, 2018). En este paso se usó en *Conjunto 1* de datos.

Primeramente, se realizó un análisis con métodos *rank* para determinar el orden de importancia de los atributos. Se utilizó *chi_cuadrada* y *gain information*, ver Figura 33. Se puede observar que el nivel de importancia es prácticamente el mismo en ambos casos.

CHI_CUADRADA		GAIN_INFORMATION	
	attr_importance		attr_importance
AVANCE_XCRED	0.34458698	AVANCE_XCRED	4.115061e-02
PROMEDIO_AC	0.31153500	PROMEDIO_AC	3.447550e-02
REPROBADAS3P	0.28156859	REPROBADAS3P	2.884411e-02
REPROBADAS2P	0.26307504	REPROBADAS2P	2.535619e-02
ASIG_INSCRITASP	0.24858566	ASIG_INSCRITASP	2.480687e-02
FALTASP	0.24185137	CRED_PANTE	2.187127e-02
CRED_PANTE	0.23146964	FALTASP	1.905666e-02
REPROBADAS1P	0.21974842	REPROBADAS1P	1.867782e-02
SEMESTRE_PANTE	0.20828786	SEMESTRE_PANTE	1.720971e-02
PERIODOS_INSCRITOS	0.18545544	PROM_PANTE	1.460921e-02
PROM_PANTE	0.17864263	PERIODOS_INSCRITOS	1.418701e-02
EDAD	0.17559131	EDAD	1.319310e-02
FALTAS_PANTE	0.15888917	FALTAS_PANTE	1.136049e-02
REPRO_PANTE	0.14272368	REPRO_PANTE	9.053311e-03
REPRO_1X	0.13235762	REPRO_1X	7.993178e-03
REPRO_2X	0.11832858	PROM_INI	5.854719e-03
PROM_INI	0.10770536	REPRO_2X	5.742552e-03
APOYO	0.09206147	APOYO	4.717201e-03
REPRO_3X	0.06210735	REPRO_3X	1.367918e-03
NINGRESO	0.03813314	NINGRESO	6.858257e-04
OTONO	0.03104093	OTONO	4.801689e-04
PRIMAVERA	0.03104093	PRIMAVERA	4.801689e-04
GEN_MASCULINO	0.02969886	GEN_MASCULINO	4.458202e-04
GEN_FEMENINO	0.02969886	GEN_FEMENINO	4.458202e-04
ESTADOCIVIL	0.01023613	ESTADOCIVIL	4.889438e-05
FORANEO	0.00000000	FORANEO	0.000000e+00

Figura 33. Métodos “rank” aplicados al conjunto de datos.

Un segundo análisis se realizó con métodos *filter*, para determinar un subconjunto con los atributos más significativos. Además, se utilizó la librería *Boruta*, para aplicar un algoritmo de *Envoltura* con el clasificador *random forest*, la ejecución del script tardó cerca de 2 horas en un servidor de la institución, ver Tabla 9 y Figura 34.

Tabla 9. Mejores atributos.

Atributos	WRAPPER- RANDOM FOREST	CONSISTENCY- BASED	CFS
ASIG_INSCRITASP	X	X	X
FALTASP	X	X	X
AVANCE_XCRED	X	X	X
PROMEDIO_AC	X	X	X
PROM_PANTE	X	X	
REPROBADAS3P	X	X	X
CRED_PANTE	X	X	
PROM_INI	X	X	
EDAD	X	X	
REPROBADAS2P	X	X	X
FALTAS_PANTE	X	X	
SEMESTRE_PANTE	X	X	X
PERIODOS_INSCRITOS	X	X	
REPROBADAS1P	X	X	X
APOYO	X	X	
REPRO_PANTE	X		
REPRO_1X	X	X	
OTONO	X	X	
PRIMAVERA	X		
REPRO_2X	X	X	X
GEN_FEMENINO	X		
GEN_MASCULINO	X	X	
NINGRESO	X	X	
REPRO_3X	X		
ESTADOCIVIL	X	X	
FORANEO			
Total de atributos	25	21	9
Tiempo ejecución	1.99 hrs	50.31 mints	3.33 seg

La Figura 34 muestra el orden de los atributos de menor a mayor de importancia con diagramas de caja, mostrando así, la dispersión de los datos de cada uno. Se puede notar que la variable FORANEO no es significativa, y a partir de la variable de REPROBADASP aumenta la significancia. Esto quiere decir que las variables más significativas están relacionadas a índice de reprobación, faltas, semestres cursados, edad, promedio de ingreso, promedio actual y avance según créditos. Este análisis también permitió detectar algunas inconsistencias en los atributos PROM_PANTE, AVANCE_XCRED y FALTASP:

- PROM_PANTE. En su mayoría representan valores en ceros, lo cual no necesariamente es incorrecto, ya que estas variables se encuentran afectadas por nuevo ingreso y reingreso, es decir, cuando son nuevo ingreso, es normal que no tengan promedio anterior porque no existe histórico, lo mismo sucede con su avance por créditos y faltas. Sin embargo, se encontraron datos atípicos, donde no es normal que existan valores mayores de 0 y menores a 6, ya que indica que alguien que reprobó el semestre anterior, y se inscribió al siguiente periodo sin acreditar. 90 registros, se descartaron por inconsistencia desde sistema.
- FALTASP. Se descartaron 5 casos mayores a 500 lo cual también indica una captura inadecuada en sistema.
- AVANCE_XCRED. Dos casos con datos incorrectamente calculados desde sistema arriba del 100%.

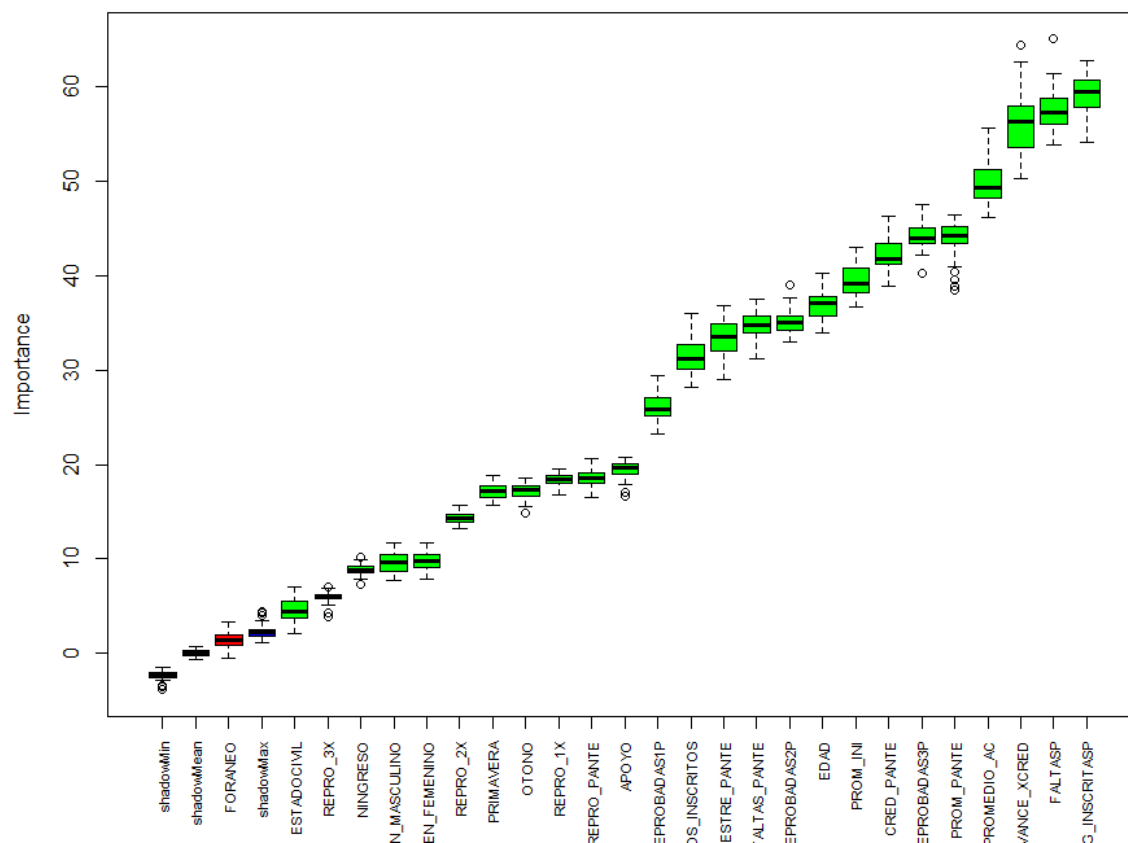


Figura 34. Análisis con método wrapper + random forest para la selección de atributos.

La siguiente fase fue determinar que atributos son los que se van a utilizar para los entrenamientos de los modelos. Para esta tarea, se utilizó como algoritmo clasificador árboles de decisión con validación cruzada base-10.

La Figura 35 muestra los resultados con tres proporciones de entrenamiento, 70%, 80% y 90%, y con atributos según los métodos de *Correlación* (CFS), *Consistencia* y *Envoltura* (*random-forest*). Se puede notar que los mejores desempeños fueron con una proporción de entrenamiento del 90% y 80% con métodos de selección por *Consistencia* y *Envoltura-RF*. La selección de atributos por *CFS* provocan que el modelo pierda sensibilidad con cualquier proporción de entrenamiento, siendo la opción menos deseada, sin embargo, destaca que reduce hasta nueve atributos significantes y esto pudiera ser útil cuando se deba de optimizar el tiempo de entrenamiento a costa de pérdida de ajuste.

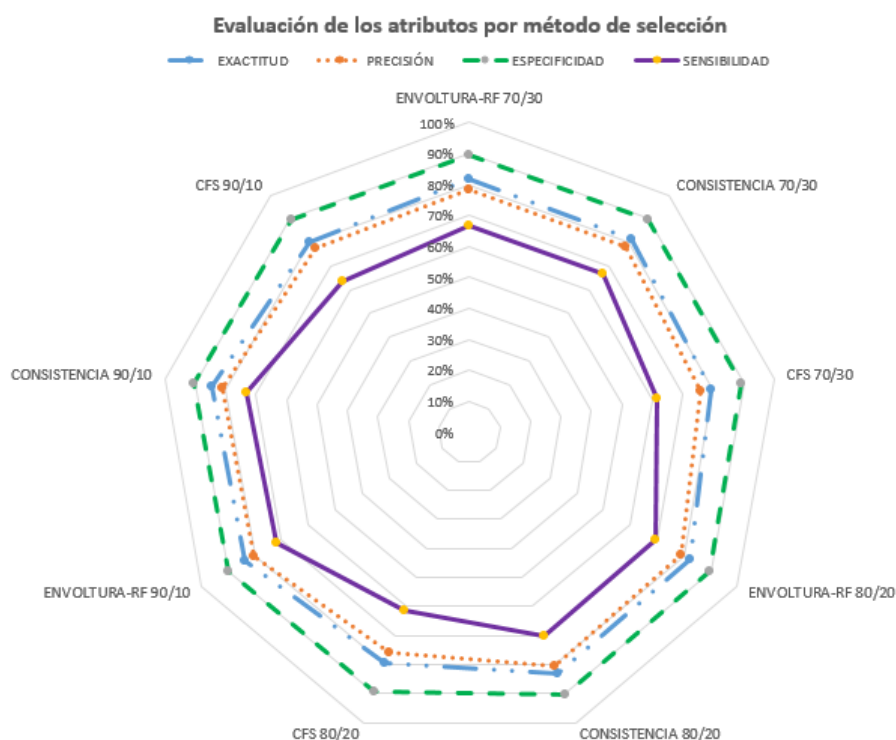


Figura 35. Métodos de selección de atributos y su desempeño con árboles de decisión.

A partir de estos resultados, se determina que los atributos a utilizar son los obtenidos por *Envoltura-RF*: ASIG_INSCRITASP, FALTASP, AVANCE_XCRED, PROMEDIO_AC, PROM_PANTE, REPROBADAS3P, CRED_PANTE, PROM_INI, EDAD, REPROBADAS2P, FALTAS_PANTE, SEMESTRE_PANTE, PERIODOS_INSCRITOS, REPROBADAS1P, APOYO, REPRO_PANTE,

REPRO_1X, OTONO, PRIMAVERA, REPRO_2X, GEN_FEMENINO, GEN_MASCULINO, NINGRESO, REPRO_3X y ESTADOCIVIL.

La integración y la transformación de los datos en el modelo de IN de la institución se realiza a partir de la media noche y en modo *off-line*, lo que permite agregar un proceso de entrenamiento para ejecutarse en ese horario, con un margen máximo de dos horas. El proceso de entrenamiento se puede ejecutar una vez al concluir el periodo escolar y el proceso de clasificación se puede ejecutar de manera diaria o semanal.

5.3 Modelos de clasificación

Los algoritmos a los que se hace referencia en la creación de los modelos, fueron desarrollados con las librerías *C50* para árboles de decisión y *neuralnet* para redes neuronales artificiales con R.

Para la evaluación de los rendimientos, se utilizaron las métricas de precisión, exactitud, especificidad y sensibilidad; así, como también, la exactitud balanceada y el coeficiente de correlación de Matthew (MCC), debido a la naturaleza de clasificación binaria del proyecto, ver sección 2.2.3.

Para la validación de los modelos, se usó *validación cruzada con base 10*. En la parte de los entrenamientos, se utilizó el *sub-conjunto balanceado* del *Conjunto 1* de datos. La proporción para la clase *Desertor(S)* corresponde al 34.5%. La selección aleatoria fue controlada por un valor de *semilla* en las sentencias R.

Arboles de decisión

Se utilizó el algoritmo C5.0 con datos balanceados de la Figura 28 y los atributos sugeridos por CFS, Consistencia y Envoltura-RF. Además, el algoritmo se utilizó con un valor de *Boosting* de 10 iteraciones para incrementar su desempeño.

En la Figura 36 se confirma que los atributos seleccionados por *Envoltura-RF* dan los mejores resultados, aunque son muy parecidos a los obtenidos por *Consistencia*. La evaluación más baja se obtiene usando solo los atributos por *CFS*. Se puede notar que sólo la métrica de *precisión*

es más baja (~ 88%), esta métrica nos indica cual es la proporción de verdaderos positivos del total de los *clasificados* como desertores.

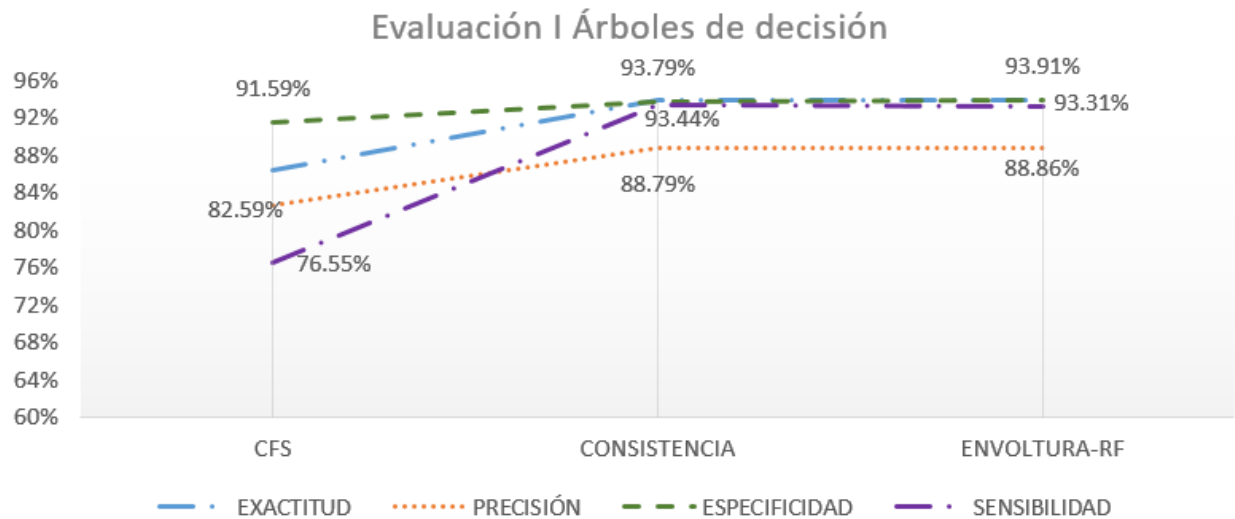


Figura 36. Árboles de decisión.

Para mejorar la evaluación del modelo, se obtuvo la *exactitud balanceada* y el *coeficiente de correlación de Matthew*, ver Figura 37. En este gráfico se puede observar el desempeño global de los modelos: la *exactitud balanceada* se mantiene para los atributos obtenidos por *consistencia* y *envoltura RS*; el coeficiente de correlación de Matthew nos indica que hay una relación fuerte positiva ya que se encuentra arriba de 70%.

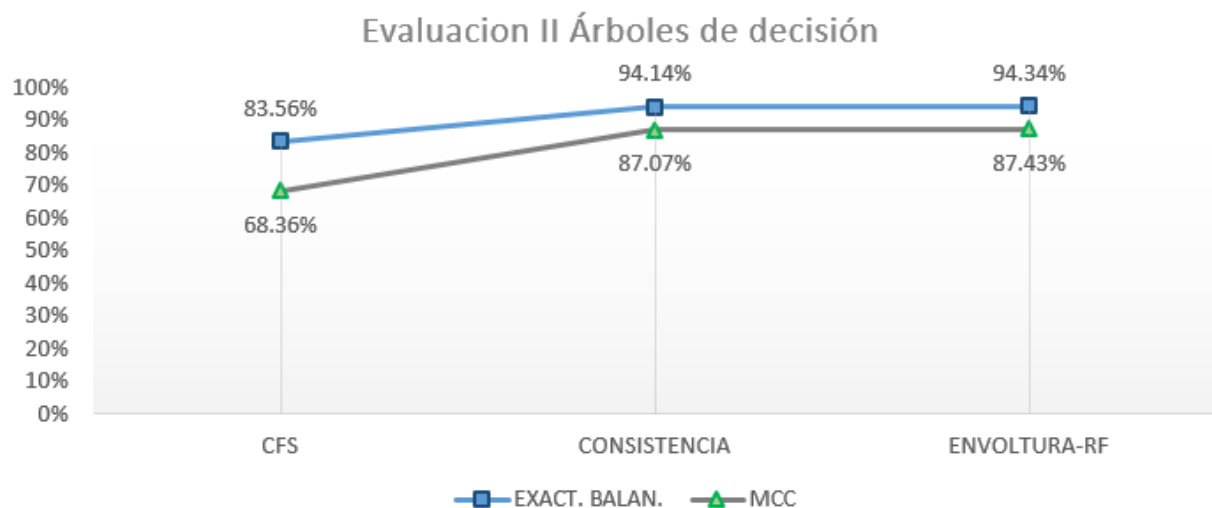


Figura 37. Exactitud balanceada y el coeficiente de correlación de Matthew.

En la Figura 38 se enlista el porcentaje de uso de cada atributo en la generación del árbol, indicando así, los criterios que se utilizaron para cualquier ramificación del árbol. Estas variables concuerdan con los criterios que el departamento de seguimiento revisa para analizar los reportes que les proporcionan. Se puede observar que se ocupan las *faltas*, el *avance del estudiante* y sus *calificaciones* y *reprobación de asignaturas* principalmente para dar seguimiento a cualquier estudiante.

```
Attribute usage:
100.00% AVANCE_XCRED
92.53% FALTAS_P
88.88% PROMEDIO_AC
83.00% ASIG_INSCRITAS_P
20.97% REPRO_3X
18.32% REPROBADAS1P
15.23% REPRO_2X
14.96% REPROBADAS3P
8.07% EDAD
7.19% PROM_INI
5.83% PROM_PANTE
4.95% FALTAS_PANTE
4.83% NINGRESO
4.02% REPRO_1X
2.64% OTONO
1.97% PERIODOS_INSCRITOS
1.96% REPROBADAS2P
1.61% CRED_PANTE
0.85% ESDOCIVIL
0.78% APOYO
0.48% SEMESTRE_PANTE
0.47% GEN_FEMENINO
0.35% REPRO_PANTE
```

Figura 38. Atributos usados por porcentaje en el modelo.

La Figura 39, muestra un listado de las primeras reglas que componen al árbol de decisión de tamaño 194, con atributos seleccionados por *Envoltura-RF*, ver Anexo 5. La Figura 40 muestra una visión general del árbol con todos sus nodos. Al momento de la redacción el paquete C50 no cuenta con un visualizador para desplegar el árbol correctamente. Y la Figura 41, muestra un subárbol a partir del nodo 12.

```

AVANCE_XCRED <= 83:
....FALTASP > 54:
.....REPROBADAS3P > 0.56: S (1609/131)
.....REPROBADAS3P <= 0.56:
.....PROMEDIO_AC <= 7.07: S (161/19)
.....PROMEDIO_AC > 7.07:
.....ASIG_INSCRITASP <= 4:
.....FALTASP > 58: S (91/16)
.....FALTASP <= 58:
.....PERIODOS_INSCRITOS > 6: N (8)
.....PERIODOS_INSCRITOS <= 6:
.....EDAD <= 20: N (4)
.....EDAD > 20: S (16/3)
.....ASIG_INSCRITASP > 4:
.....REPROBADAS1P > 0.75:
.....ESDOCIVIL <= 0: S (13)
.....ESDOCIVIL > 0:
.....REPRO_2X <= 0:
.....PROM_PANTE <= 3.6: S (3)
.....PROM_PANTE > 3.6: N (8)
.....REPRO_2X > 0:
.....REPRO_1X <= 3: N (3)
.....REPRO_1X > 3: S (14)
.....REPROBADAS1P <= 0.75:
.....PROM_PANTE <= 8.01:
.....ASIG_INSCRITASP > 6:
.....EDAD <= 20:
.....ASIG_INSCRITASP > 7: S (27/7)
.....ASIG_INSCRITASP <= 7:
.....ESDOCIVIL <= 0: S (13/3)
.....ESDOCIVIL > 0: N (11)
.....EDAD > 20:
.....REPROBADAS2P > 0.63: S (12/2)
.....REPROBADAS2P <= 0.63:
.....REPRO_1X <= 3: N (33)
.....REPRO_1X > 3:
.....GEN_FEMENINO <= 0: N (43/11)
.....GEN_FEMENINO > 0: S (20/6)
.....ASIG_INSCRITASP <= 6:
.....NINGRESO > 0:
.....REPROBADAS2P <= 0.57: N (14/2)
.....REPROBADAS2P > 0.57: S (6/1)
.....NINGRESO <= 0:
.....REPROBADAS3P > 0.33: S (101/30)
.....REPROBADAS3P <= 0.33:
.....PROMEDIO_AC <= 7.33: N (11)
.....PROMEDIO_AC > 7.33:
.....PROM_INI > 8.49: S (16/1)
.....PROM_INI <= 8.49:
.....FALTASP <= 77: S (88/37)
.....FALTASP > 77: N (16/1)
.....PROM_PANTE > 8.01:
.....PROM_PANTE > 9.13:
.....CRED_PANTE > 58: N (5)
.....CRED_PANTE <= 58:
.....PROMEDIO_AC <= 8.16: N (3)
.....PROMEDIO_AC > 8.16: S (15)
.....PROM_PANTE <= 9.13:
.....APOYO > 5: N (22)
.....APOYO <= 5:
.....FALTASP <= 55: N (24)
.....FALTASP > 55:
.....FALTAS_PANTE > 305: S (17/4)
.....FALTAS_PANTE <= 305:
.....PERIODOS_INSCRITOS <= 3:
.....CRED_PANTE <= 27: N (15)
.....CRED_PANTE > 27:
.....PROM_PANTE > 8.65: N (12/2)

```

Figura 39. Segmento del modelo generado por árboles de decisión.

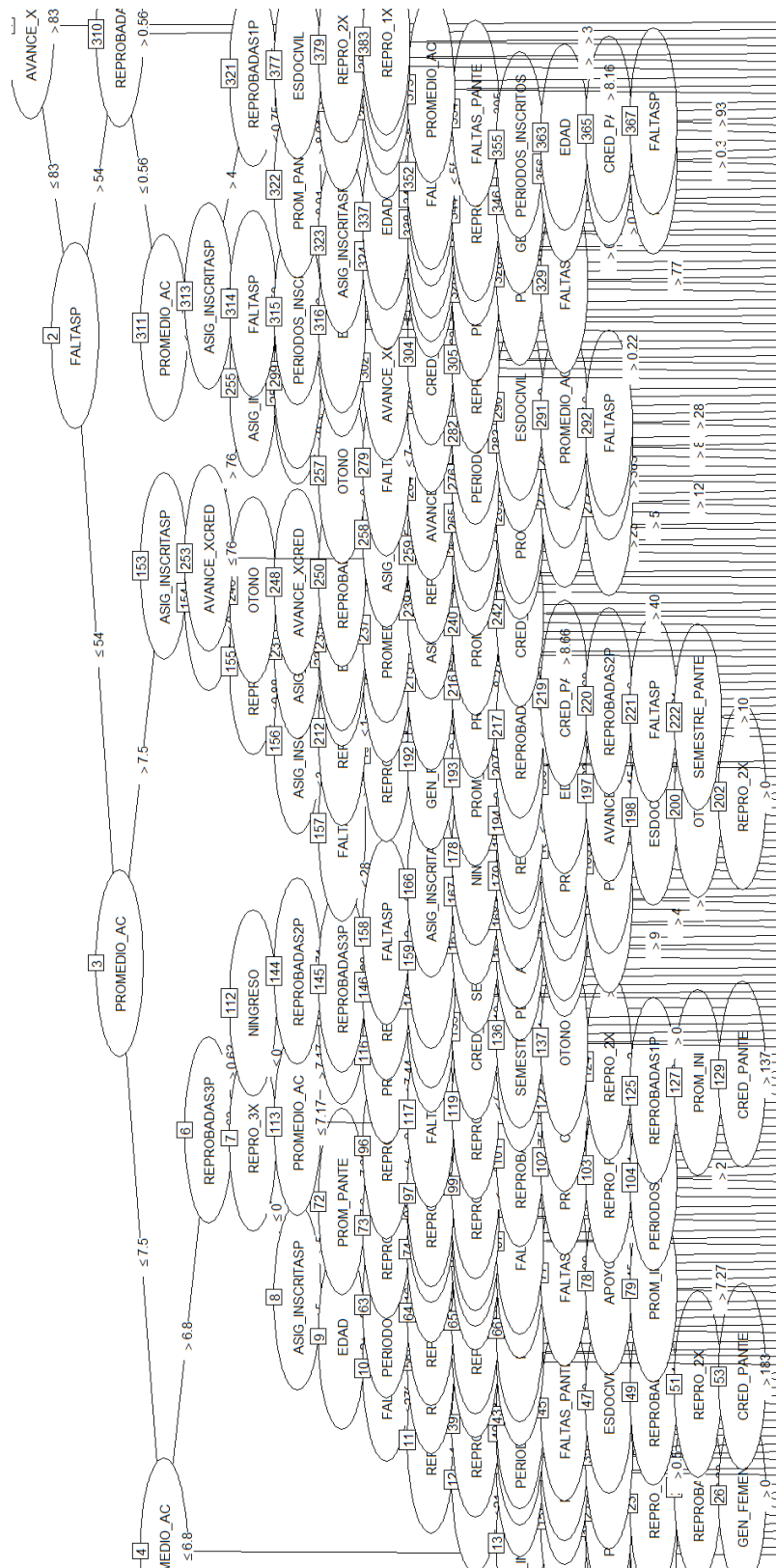


Figura 40. Árbol de decisión.

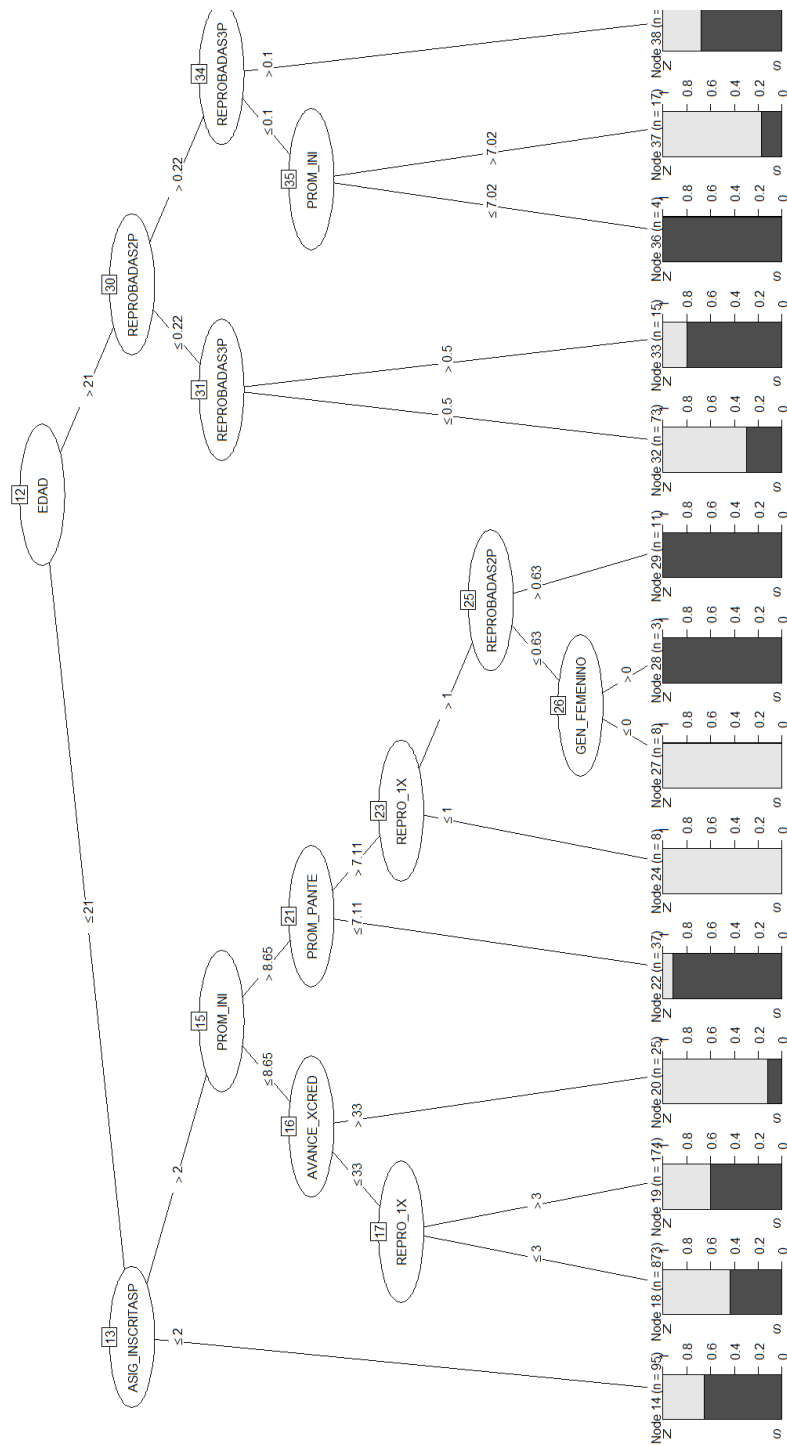


Figura 41. Sub-árbol a partir del nodo 12.

Al final de las hojas muestra gráficamente la probabilidad de clasificación (S o N), este dato puede ser usado para clasificar a los posibles desertores en la aplicación web o mostrar alguna probabilidad de clasificación.

Redes neuronales artificiales

Se utilizó el paquete *neuralnet* de R con el algoritmo *Resilient Backpropagation* y el conjunto de datos balanceado mencionado en árboles de decisión, los atributos también fueron los mismos. Se usaron dos configuraciones de capas ocultas, una de (2,2) y otra de (12) neuronas para contrastar resultados.

Se aplicó validación cruzada base 10 para la evaluación de las redes neuronales con sus configuraciones. La Figura 42 muestra los resultados de los modelos creados medidos por *exactitud*, *precisión*, *especificidad* y *sensibilidad*.

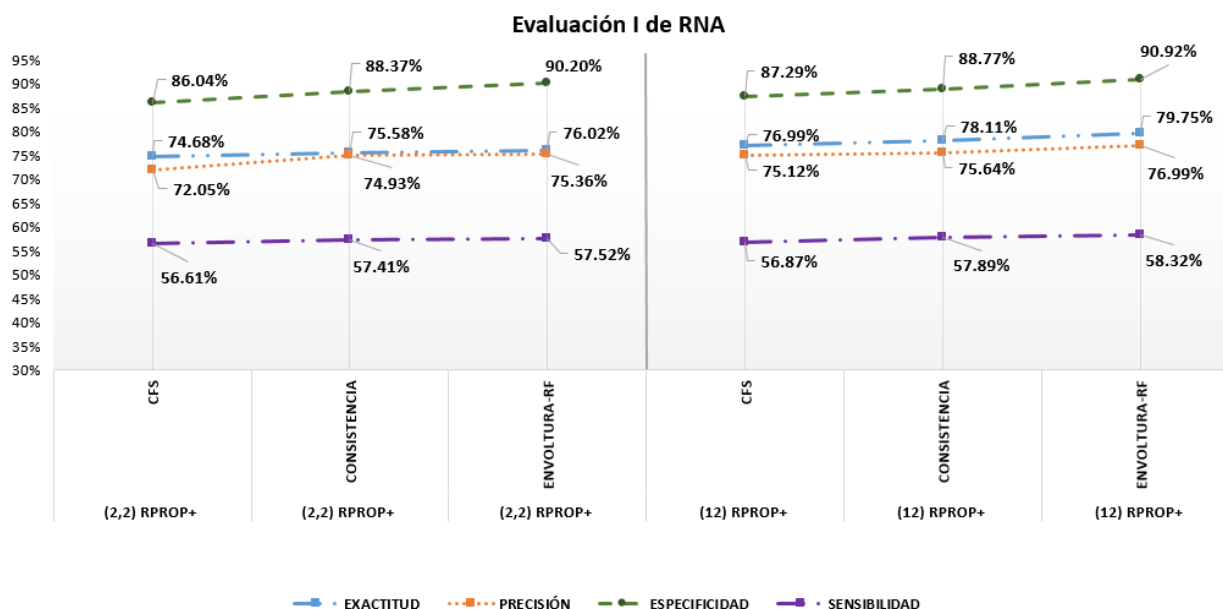


Figura 42. Configuraciones para redes neuronales artificiales con evaluación por exactitud, precisión, especificidad y sensibilidad.

Se puede observar que los atributos por CFS da los valores más bajos, sin embargo, conforme se van agregando atributos, existe una mejora en los resultados. Aunque la precisión es alta, la sensibilidad es baja, esto quiere decir que puede detectar alrededor de 57% de desertores (según configuración red y número de atributos).

En la Figura 43 se analizan los resultados con el *coeficiente de correlación Matthews* y *exactitud balanceada*. Este gráfico proporciona un mejor entendimiento de los resultados, donde se observa que el MMC está entre el 49% y 54%, mientras que la exactitud balanceada está en 71% y 75%.

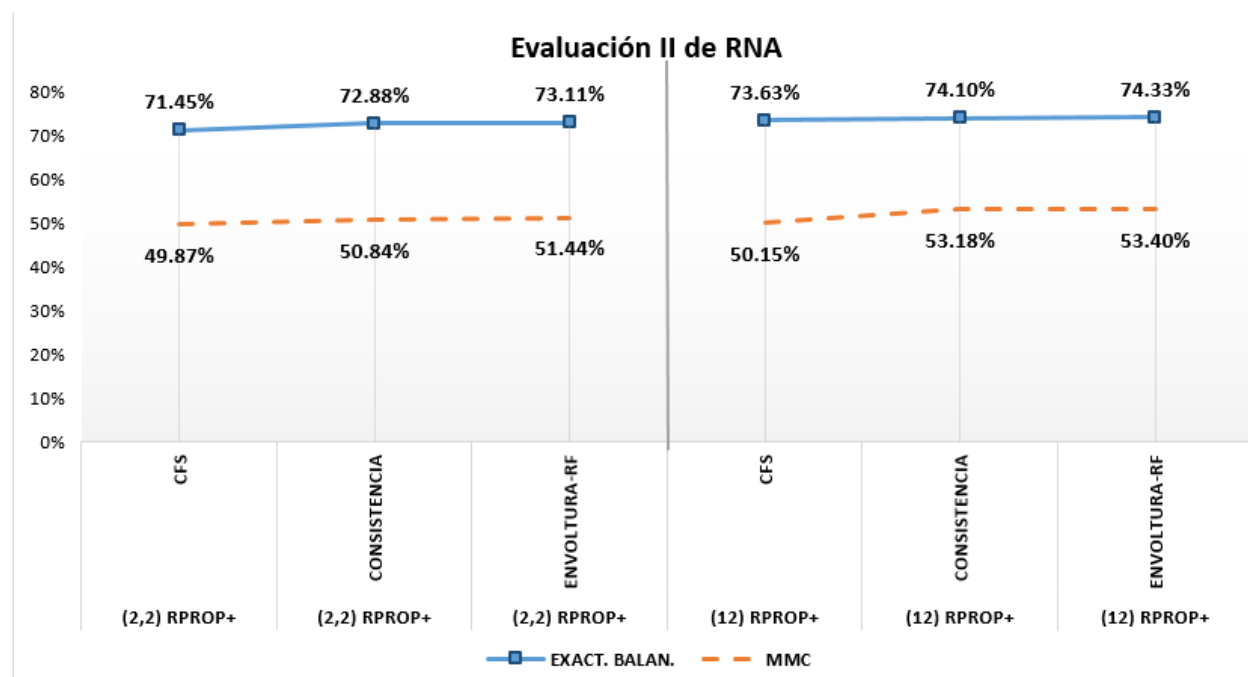


Figura 43. Configuraciones para redes neuronales con evaluación por exactitud balanceada, coeficiente de correlación Matthews y la proporción de la clase DESERTOR (S)

El coeficiente de Matthew indica que las evaluaciones se encuentran entre el 49.66% y 53.69 lo que indica que existe una fuerte relación positiva. Cabe mencionar que existe una ligera mejora con el aumento de neuronas en la RNA. Además, se observa que en ambas métricas los valores mínimos son por CFS y los máximos por *Envoltura-RF*.

Para efectos de contrastar los resultados con otros modelos, se tomó la configuración usada con atributos *Envoltura-RF* y una capa oculta de 12 neuronas. Esta configuración se encuentra limitada por los recursos del servidor en productivo: servidor Xeón Platinum y memoria RAM 16 gb. Se realizaron pruebas con otras configuraciones, sin embargo, algunos se descartaron por la

limitante del tiempo, ya que exigen en ocasiones tiempos de entrenamiento de más de dos horas. El modelo seleccionado, se puede apreciar en la Figura 44.

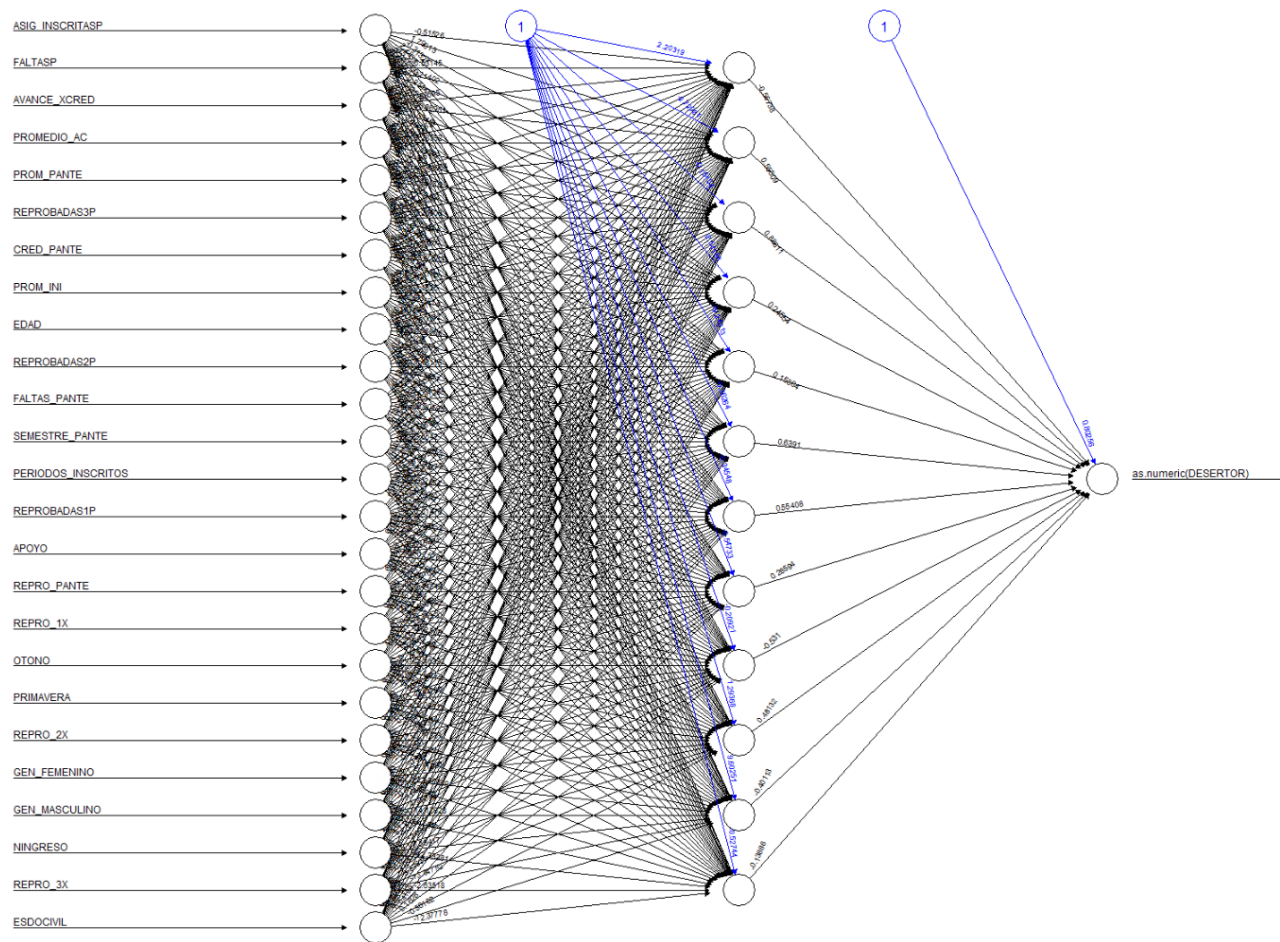


Figura 44. Red neuronal entrenada con 12 neuronas en capa oculta.

Selección del modelo

La Figura 45 contrasta los mejores resultados de redes neuronales artificiales y árboles de decisión, según clasificador y configuración utilizada. Para el caso de árboles de decisión se tomó los mejores resultados dados por atributos por *Envoltura-RF*. Para Redes Neuronales, la configuración más apropiada fue la de una capa oculta con 12 neuronas. Los resultados demuestran con creces que árboles de decisión es el mejor algoritmo clasificador para la detección de *Desertores*.

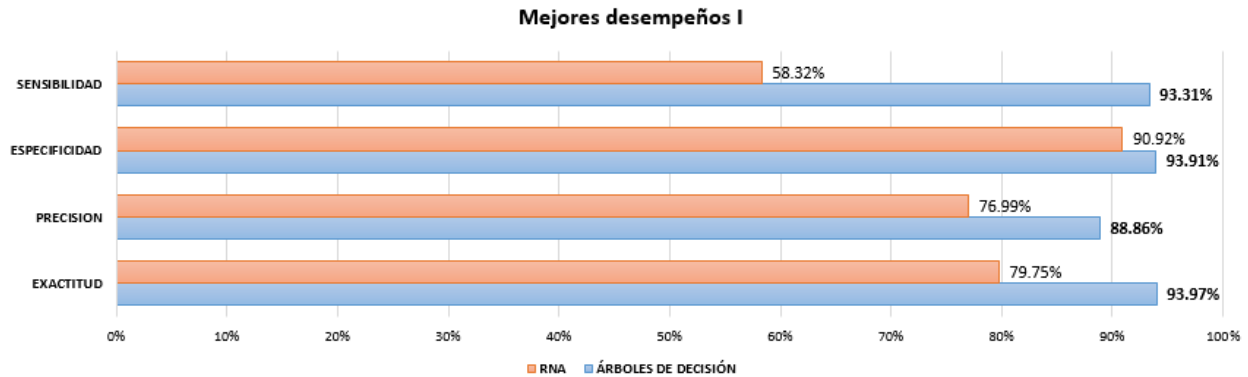


Figura 45. Comparativa entre redes neuronales y árboles de decisión: exactitud, precisión, especificidad y sensibilidad.

De manera global, en la Figura 46 se compara la exactitud balanceada y el coeficiente de correlación de Matthew para la evaluación del desempeño de ambos modelos clasificadores. Al igual que el gráfico anterior, se puede observar que árboles de decisión tiene un mejor desempeño que redes neuronales ya que alcanza un 94.34% de exactitud balanceada y un coeficiente de correlación de Matthew de 87.43%, los cuales son superiores a los de RNA: 74.33% y 53.40% respectivamente.

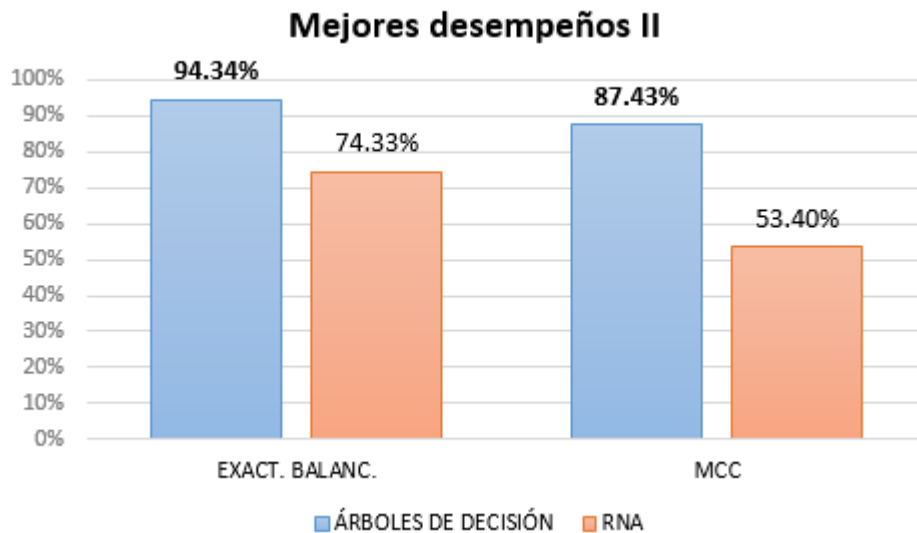


Figura 46. Comparativa entre redes neuronales y árboles de decisión: exactitud balanceada y coeficiente de correlación Matthew (MCC).

De esta forma, árboles de decisión resultó ser la mejor opción como algoritmo clasificador para la realización de este proyecto.

Resultados esperados

Puesto que el objetivo es poder detectar los posibles desertores, se hizo un análisis por periodo con deserciones de tres periodos anteriores(otoño 2018, primavera 2019 y otoño 2019) + el actual (primavera 2020), ver Figura 24. Se obtuvieron dos modelos por clasificador según algoritmo y con clase balanceada (3x,80%) y datos del otoño 2010 a la primavera 2018.

En la Figura 47 se graficaron los desertores reales a partir del otoño 2018 hasta el otoño 2019 y se proyectó la predicción para primavera 2020.

La predicción realizada por árboles de decisión resulta más próxima a los valores reales desde primavera 2019, pero conserva una uniformidad en las predicciones entre periodos, mientras que redes neuronales artificiales se separa más de los valores reales que árboles de decisión a partir de primavera 2019, este comportamiento se ve representado por la línea punteada en color gris. Por otro lado, en el caso de redes neuronales, los resultados en el entrenamiento, demuestran que aunque tiene mejor desempeño en las validaciones, su resultado es aceptable, según coeficiente de correlación de Mathew ya que se encuentra por encima del 50% lo que indica que existe una correlación fuerte positiva.

Según la predicción por árboles de decisión, se puede esperar que para la primavera 2020 los posibles desertores sean alrededor de 436.

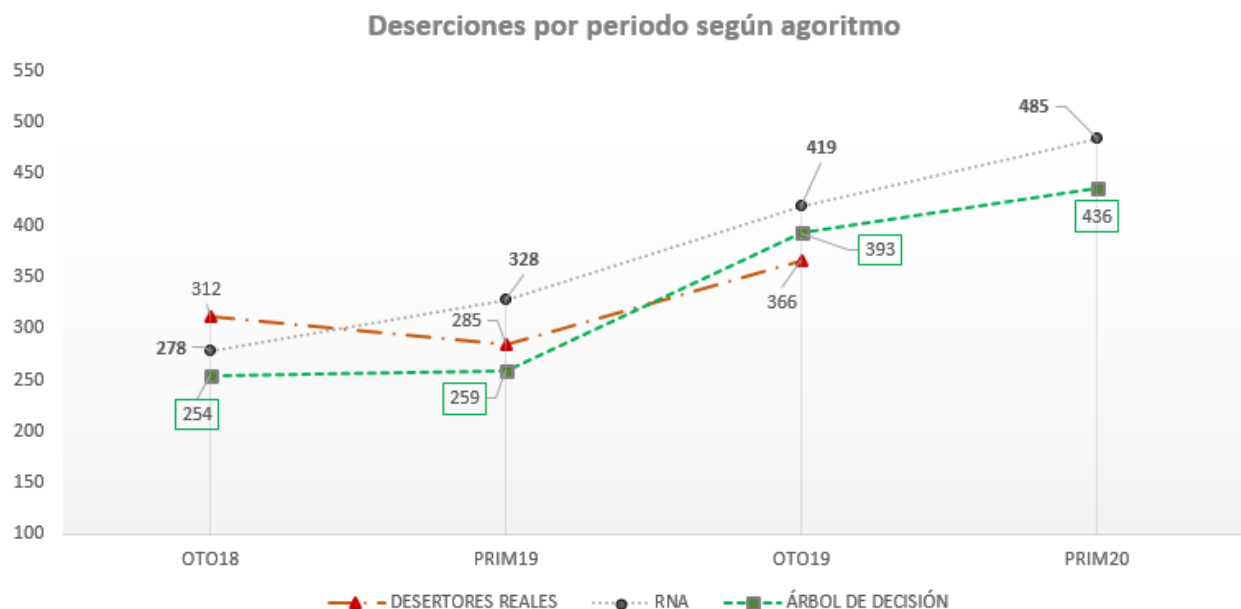


Figura 47. Desertores por periodo con datos históricos.

Es importante destacar que se predicen más desertores de los que realmente resultan por periodo, lo cual no necesariamente es incorrecto, ya que, según el área de seguimiento estudiantil, el objetivo es apoyar a los alumnos antes de que se encuentren con algún problema, tantos como sea posible, por lo tanto, el sistema cumpliría con su función de concentrar la atención sobre aquellos alumnos que den señales de riesgo académico, los cuales son más al comienzo de los periodos.

5.4 Implementación

El proceso de integración de datos en la institución se ejecuta en las noches y su objetivo es consolidar la información del día en modo *off-line* para que esté disponible a la mañana siguiente. Se integró los entrenamientos del modelo y las predicciones por medio de *scripts* en R con conexiones a la base de datos multidimensional. El modelo generado se guarda en formato binario *.rds* (en el servidor) para que posteriormente sea procesado por R. De esta manera, se carga en memoria solo cuando va a procesar los registros de los alumnos, los resultados de la clasificación se guardan en el almacén de datos, ver Figura 48.

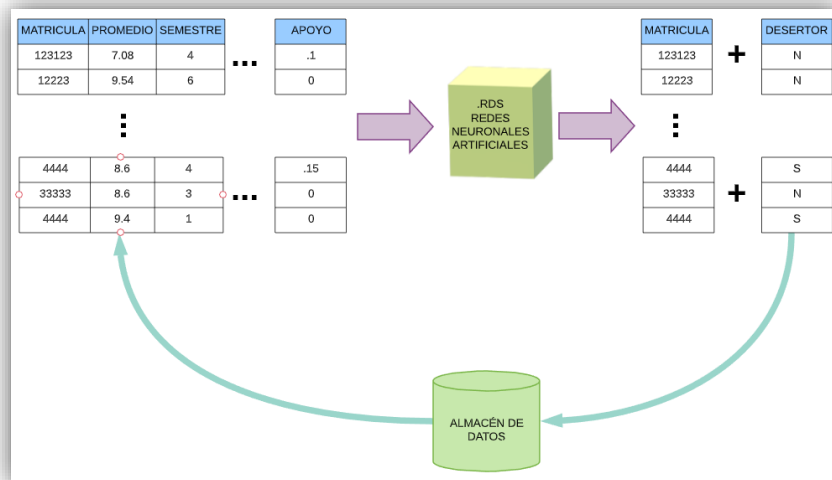


Figura 48. Funcionamiento del modelo predictivo.

El proceso general se puede ver en la Figura 49 y es el siguiente:

1. Se invocan dos *scripts* en R, uno para ejecutar el entrenamiento y otro para la predicción. El primero inicia a las 3:45 y el siguiente a las 4:30, dejando tiempo suficiente para concluir su ejecución. Se cuidaron los márgenes de tiempo, puesto que existen otros procesos que se ejecutan en este horario. En promedio se ejecutan tareas de menos de 1 minuto.
 - a. El *script* de entrenamiento genera un archivo *.rds* de 12 a 13 *megabytes* en el servidor. Se entrena con un histórico de alumnos por periodo desde el otoño 2010 hasta el otoño 2019.
 - b. El *script* de clasificación lee el archivo y pasa el listado de los alumnos inscritos en el periodo actual. Dicho listado es regresado en una tabla temporal con las predicciones hechas por los modelos. Enseguida, se ejecutan procedimientos almacenados en la base de datos para realizar las transformaciones necesarias y alojar la información en una tabla de hechos que está a nivel de alumno y periodo escolar.
2. Por medio de otros procesos del modelo de IN, se propagan las predicciones hacia otras tablas de hechos. En este momento queda lista la plataforma de IN para poder ser consultada por medio de una aplicación web.

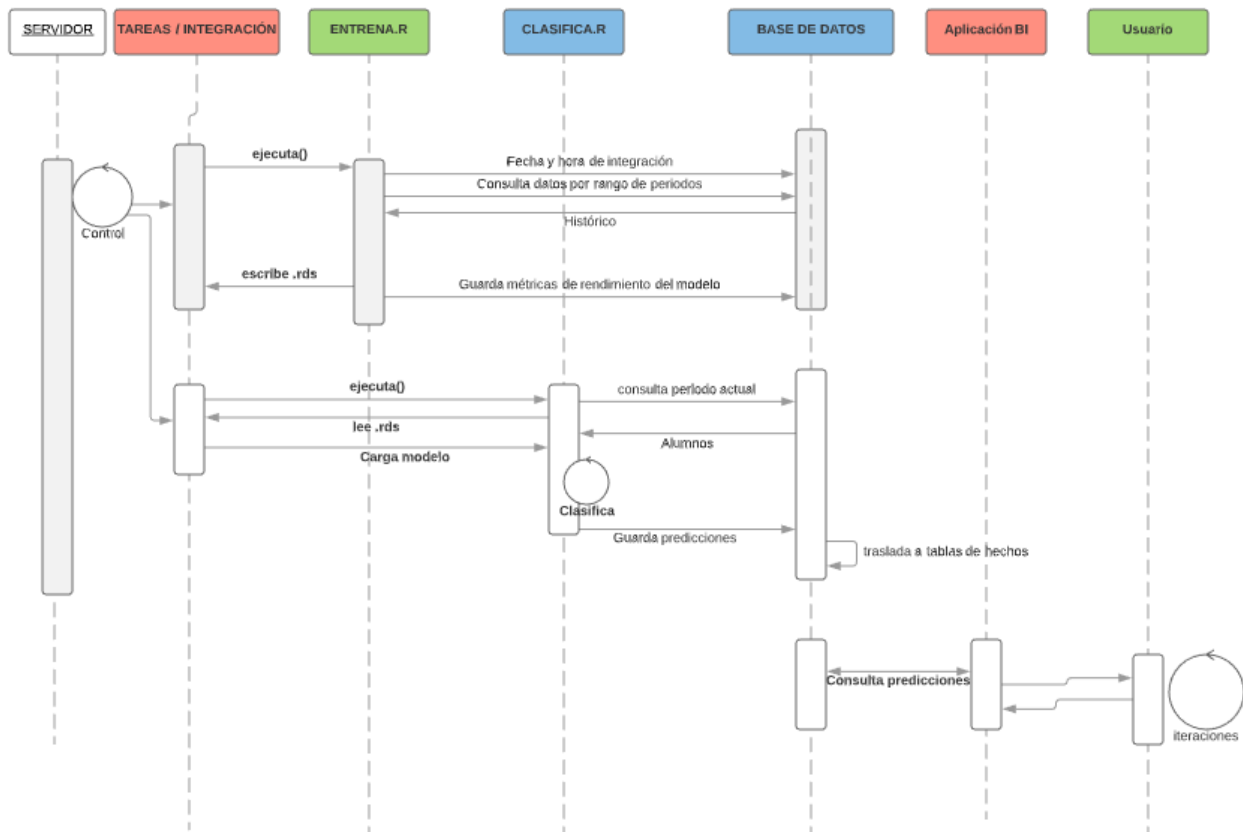


Figura 49. Despliegue del modelo.

Para el despliegue en web, se utilizó el prototipo validado para realizar la versión en producción. Por cuestiones de políticas del Área de Tecnologías de la Información de la institución, se alojó en la plataforma de IN la cual usa Dossier 2019, y utiliza *Asp.Net*, *Ajax* y *Html*. La aplicación tiene 5 interfaces, de las cuales 3 están orientadas al usuario final y 2 para visualizar el desempeño del modelo predictivo, Figura 50. El modelo predictivo seleccionado es el que mejor desempeño tiene en *precisión*, *exactitud*, *sensibilidad* y *especificidad*, exactitud balanceada y coeficiente de correlación Matthew mostrado en la Figura 45 y en la Figura 46.

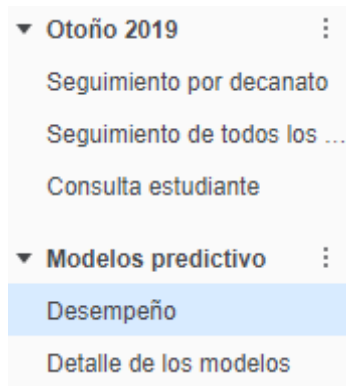


Figura 50. Menú principal de la aplicación web.

La Figura 51 muestra un resumen por decanato de las predicciones para el periodo actual:

- Muestra el porcentaje de alumnos en riesgo de deserción por decanato académico según modelo predictivo.
- Muestra el total de alumnos en posible riesgo.
- Es una tabla con detalle por programa académico de posibles desertores, ordenados de menor a mayor, según proporción de riesgo por programa. Adicionalmente en la parte inferior se muestra un [link](#) hacia la documentación del usuario, Anexo 5.

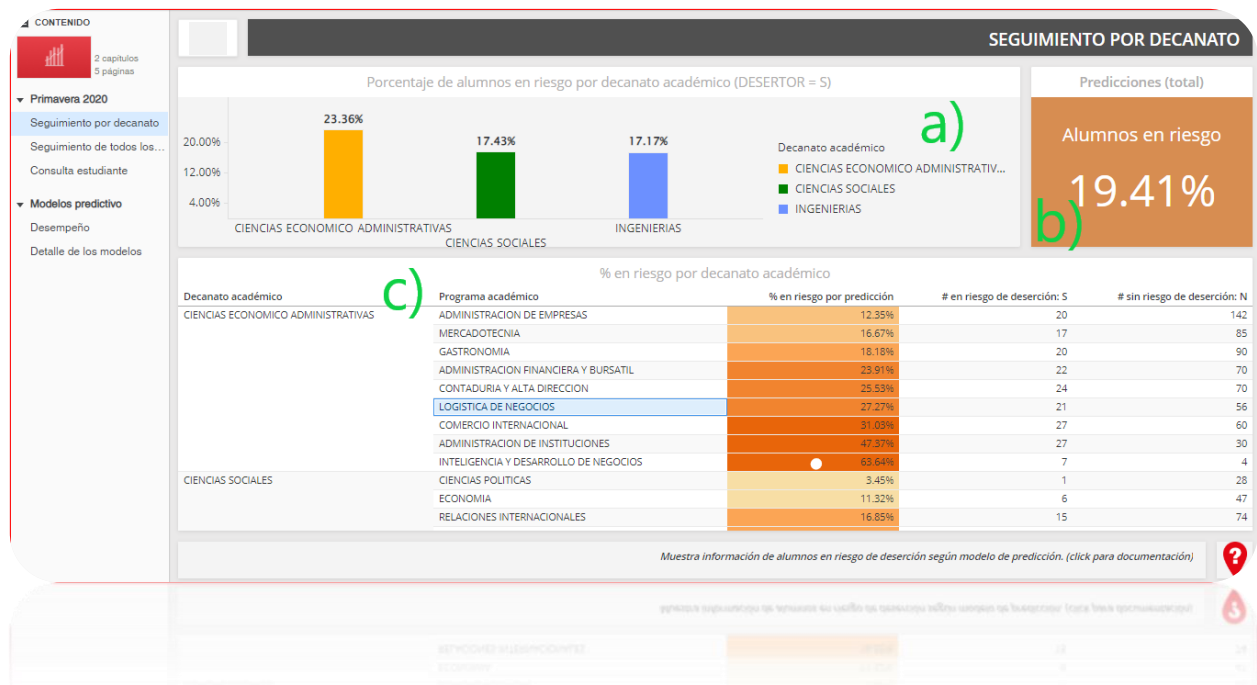


Figura 51. Seguimiento por decanato.

La Figura 52 muestra un listado de todas las predicciones por alumno:

- Filtros por red neuronal o árbol de decisión: los resultados se visualizan en la tabla de la sección b).
- Detalle a nivel de alumno con datos académicos, de contacto por programa y decanato.
- Número de inscritos por decanato y programa.
- Proporción en cuanto a número de alumnos con riesgo de deserción.

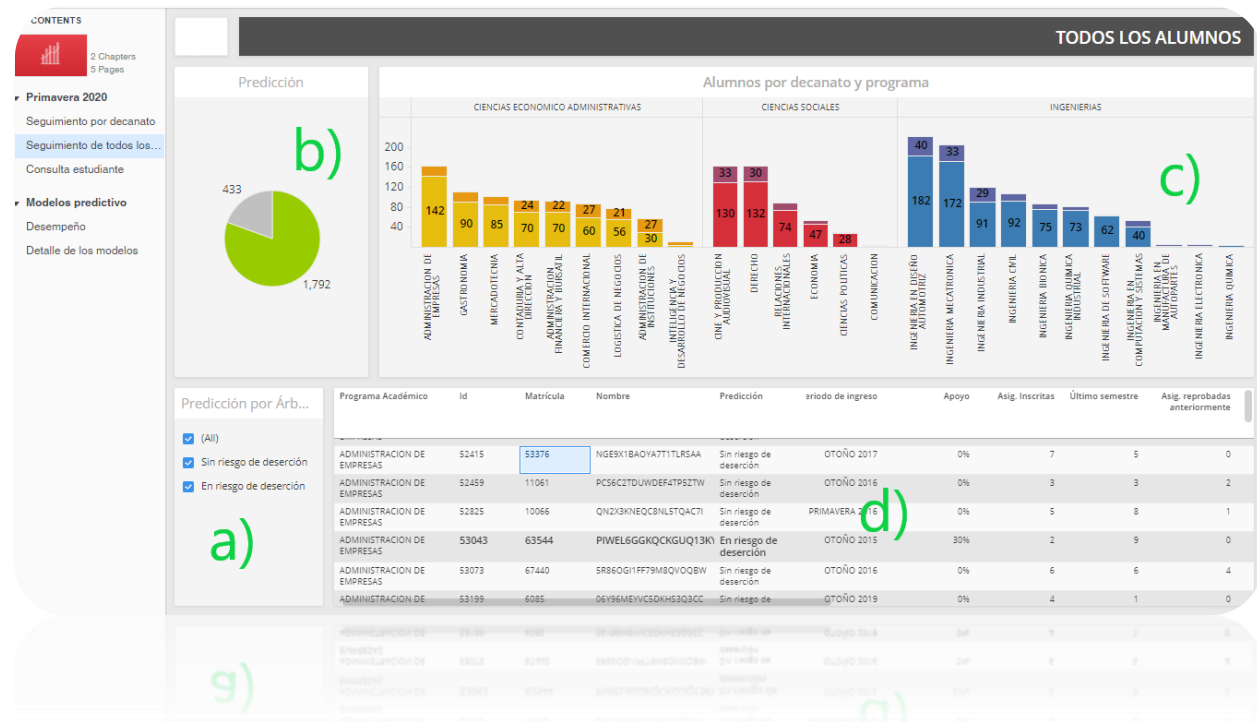


Figura 52. Seguimiento de los alumnos.

La Figura 53 muestra la interfaz por alumno, permitiendo al usuario hacer búsquedas y revisar su avance escolar individual:

- Filtro por nombre o por clave ID.
- Muestra información general del estudiante como datos de contacto, dirección, género, etc.
- Contiene información académica como periodo de ingreso, estatus, programa escolar adscrito, etc.
- Muestra datos de su avance académico a partir de su histórico.

- e) Resultado de la predicción, S indica que es un posible desertor.
- f) En la parte superior muestra el promedio actual y su avance por créditos. En la parte inferior se encuentra la sección con datos de reprobación.

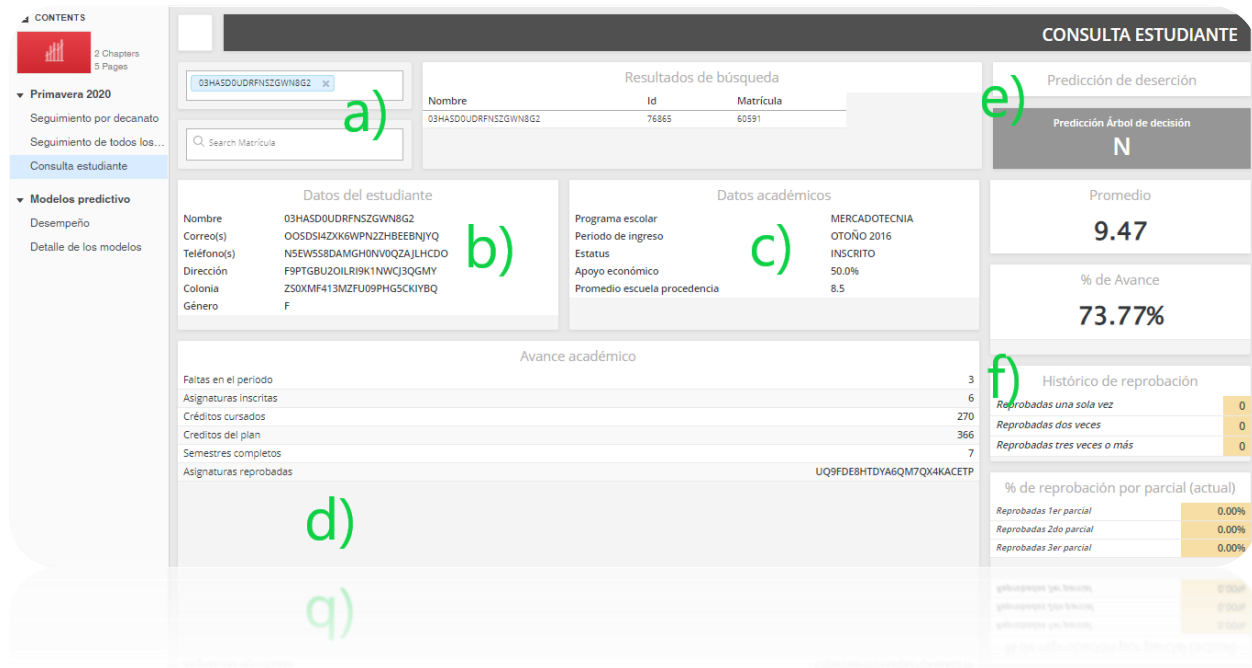


Figura 53. Consulta por estudiante.

Para el monitoreo de modelo, se creó la interfaz de la Figura 54 que contiene:

- a) Exactitud, precisión, especificidad y sensibilidad del modelo predictivo, los cuales calcularon la predicción mostrada en las interfaces.
- b) Muestra un comparativo de promedios de los desempeños de los modelos.
- c) Contiene información de los archivos físicos que contienen los modelos.

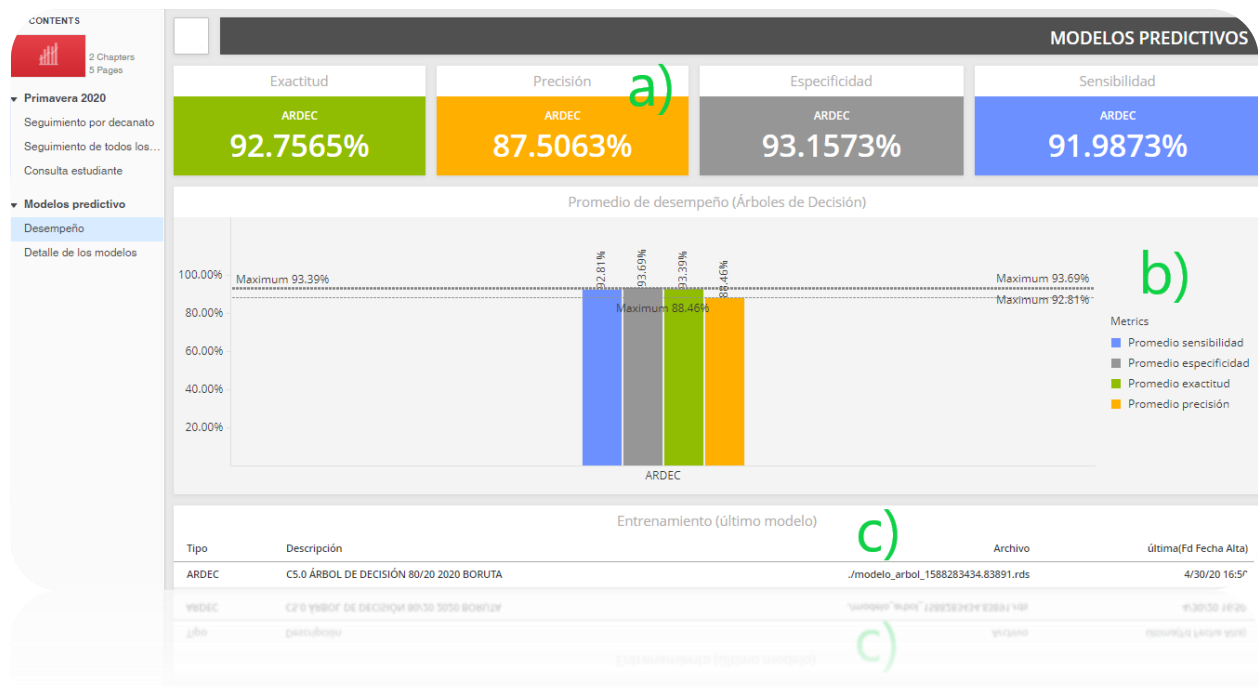


Figura 54. Desempeño de los modelos.

La Figura 55 contiene un detalle de los entrenamientos, muestra la fecha de creación y que atributos se usaron para su entrenamiento. También muestra cual es el modelo actual.

DETALLE					
Modelos entrenados					
Tipo	Nombre archivo	Estatus activo	Descripción	Fecha Alta	Atributos usados
RNA	./modelo_rna_1586940880.20765.rds	S	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/15/2020 3:54:40 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586855327.74236.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/14/2020 4:08:48 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586768248.45303.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/13/2020 3:57:28 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586681812.82998.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/12/2020 3:56:52 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586595041.06862.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/11/2020 3:50:43 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586509251.17916.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/10/2020 4:00:52 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586422348.17222.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/9/2020 3:52:28 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586335942.06444.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/8/2020 3:52:22 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586249752.85923.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/7/2020 3:55:53 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1586076938.83545.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/5/2020 3:55:38 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"
	./modelo_rna_1588283434.83891.rds	N	RED NEURONAL ARTIFICIAL 80/20 (4.3) (0.1) (rprop=10000) (WRAPPER: R.FOREST)	4/30/2020 16:50:00 AM	"ASIG_INSCRITASPI","FALTASPI","AVANCE_XCRED","PROMEDIO_AC","PROM_PANTE","REPROBADAS3PI","CRED_PANTE","PROM_PANTE"

Figura 55. Detalle de entrenamiento de modelos.

5.5 Evaluación

Se aplicó la **encuesta tipo C** con la ayuda de 5 usuarios para evaluar la interfaz web por medio de *google forms*, Figura 56 y Anexo 3. La implementación de la encuesta fue flexible para el usuario ya que le permitió tomarse su tiempo para contestar 107 *ítems* según su disponibilidad. El instrumento estuvo en línea, al igual que los accesos web, lo que favoreció para contestar en cualquier parte fuera de sus lugares de trabajo. El tiempo proporcionado para contestar la encuesta fue de 5 días.

docs.google.com/forms/d/e/1FAIpQLScx_Q5FSm7dvhai5cWwJ22

Evaluación de implementación

El objetivo de la encuesta es evaluar las interfaces web, las cuales van a mostrar información de riesgo académico para implementar tutorías efectivamente, los aspectos a evaluar son: visibilidad, lenguaje, estándares, facilidad de reconocimiento e interacción con el usuario.

* Required

Tomar en cuenta los siguientes criterios:

Criterio de evaluación
0 = no es problema
1 = problema sin importancia
2 = problema de poca importancia
3 = problema grave
4 = catástrofe

Nombre o alias *

Your answer

NEXT

Page 1 of 14

Never submit passwords through Google Forms.

Figura 56. Evaluación con google forms, **Instrumento tipo C**.

La Figura 57 muestra la puntuación total por cada ítem de la encuesta, si se compara con la evaluación del prototipo de la Figura 19, se pudo notar que el máximo total de calificación llegó a los 14 puntos, mientras que en la evaluación de la interfaz final no supera los 8 puntos lo cual indica que ha alcanzado un nivel de usabilidad aceptable según heurísticas.

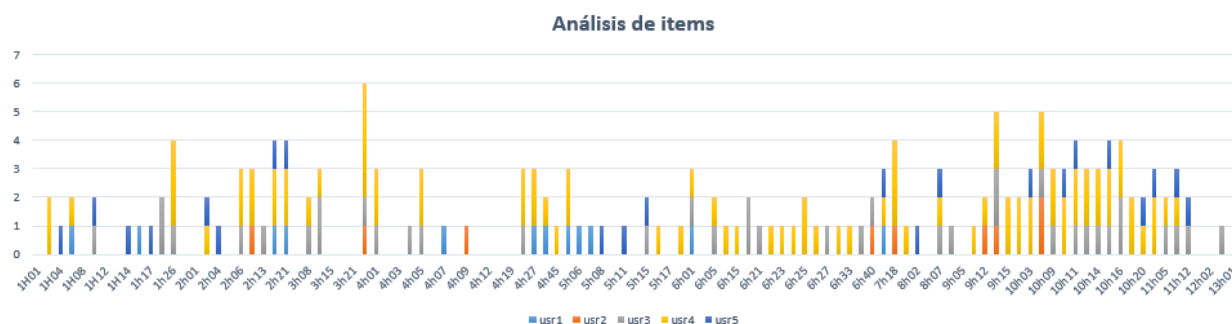


Figura 57. Resultados de evaluación final.

El ítem que tiene mayor puntaje es el 3h23 que corresponde a: *¿Los usuarios pueden configurar la apariencia de su propio sistema, sesión y valores por defecto para la pantalla?* Esta característica no está incorporada y es del tipo estético por lo tanto no es prioritario agregarlo.

La Figura 58 muestra el resultado de análisis de criticidad de los ítems, en el cual se puede observar que ninguno superó el grado 2, el cual corresponde a *Problema de usabilidad menor*.



Figura 58. Análisis de criticidad de versión final.

Los resultados demuestran que la aplicación web satisface con los estándares en temas de usabilidad.

Por otro lado, para asegurar la funcionalidad, se evaluaron dos tareas principales:

- Obtener el listado de alumnos en riesgo de deserción.
- Consultar alumnos en riesgo académico por nombre de alumno.

La Figura 59 muestra los resultados realizados por 5 usuarios con el instrumento **tipo D**, que es una variante del tipo B, orientado a la aplicación final. De manera general son buenos resultados ya que superan el 95% de aceptación, mientras que el ítem de menor calificación es del 75%: *Si se realiza la acción correcta ¿Los usuarios verán que están avanzando hacia el resultado*

esperado?, es decir, el usuario duda sobre los resultados esperados. Sin embargo, esto puede ser un tema de adopción el cual se puede resolver con el uso y tiempo.

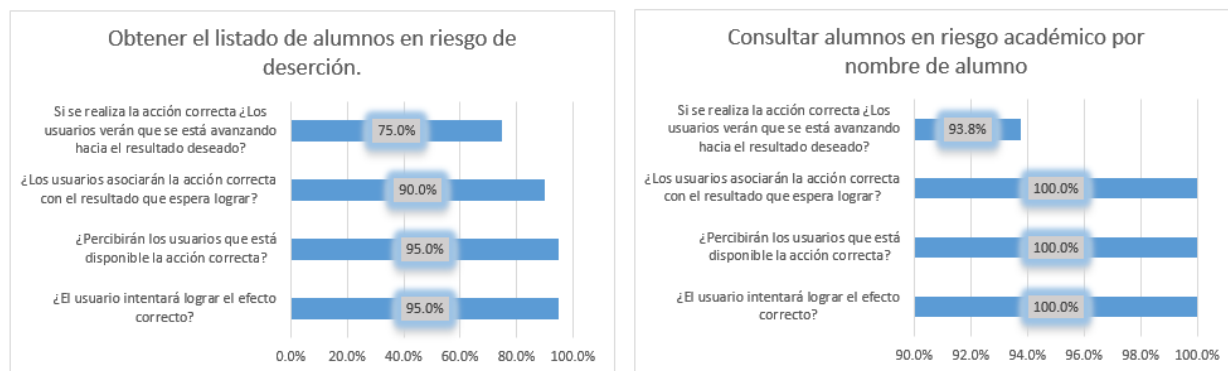


Figura 59. Evaluación de tareas sobre la aplicación web.

6 Conclusiones y trabajo futuro

En un primer análisis de atributos, pareciera que se pueden descartar algunos que no parecen tan relevantes, como es el caso de asignaturas reprobadas, periodo (otoño o primavera) y género. Sin embargo, un análisis más exhaustivo con selección de atributos por *envoltura – random forest*, demuestran que al utilizar la mayoría de los atributos se obtiene el máximo desempeño en los modelos.

El balance de clase, permitió mejorar las métricas de desempeño, como se demostró en la Figura 27 y Tabla 7.

Los atributos usados por el método de *CFS* proporcionaron un menor tiempo de ejecución en los entrenamientos, pero con un menor desempeño en cuanto a la sensibilidad; esta métrica de rendimiento es de importancia para el proyecto, ya que indica realmente cuantos desertores se aciertan de todos los que deberían de ser.

El coeficiente de correlación de Mathew y la exactitud balanceada proporcionaron una mejor evaluación de los modelos, permitiendo que los resultados no se vean afectados por las variaciones de las métricas de exactitud, precisión, especificidad y sensibilidad.

Árboles de decisión fue el clasificador con mejor coeficiente de correlación Matthew de 87.43% y exactitud balanceada de 94.34%

Redes neuronales, aunque quedó por debajo de árboles de decisión en su evaluación, las pruebas indican que aumentando las neuronas en la capa oculta podría mejorar su desempeño, sin embargo, requiere más poder de procesamiento ya que los entrenamientos en servidor llegan a superar la hora.

De este análisis surge una contraparte, y es que también se puede ver como un pronóstico para la reinscripción al siguiente periodo. El cual puede servir para cálculos presupuestales, ya que se hacen con base a los posibles reinscritos.

El beneficiario final es el alumno, por que recibe el acompañamiento que requiere para no interrumpir sus estudios; este esfuerzo se traduce en monitorear cerca de 450 alumnos por periodo pertenecientes a los decanatos de *Ciencias sociales, administrativas e Ingenierías* de UNIVR.

Naturalmente el área de seguimiento estudiantil mejora sus tareas administrativas con el uso del sistema embebido por que obtiene un listado en cuestión de segundos y evita consolidar un reporte de varios documentos *Excel*, incluso de diferentes áreas (admisiones y control escolar).

La aplicación web permite al usuario enfocarse en los alumnos en riesgo y en su avance académico..

El sistema embebido permite aprovechar los recursos con los que cuenta la institución como bases de datos, servidores, plataformas web, administración de usuarios, seguridad y acceso a datos transformados y estandarizados. Este aprovechamiento se tradujo en reducción de tiempos en el desarrollo e implementación de la aplicación web.

R resolvió todas las necesidades de aprendizaje computacional en la implementación y no fue necesario interactuar con alguna otra herramienta semejante para este fin. Aunque existen alternativas dentro de la plataforma de IN en la institución se optó por dejar en *R* el análisis de datos para proporcionar modularidad al proyecto.

El modelo de prototipos resultó ser una técnica efectiva para evaluar la usabilidad y funcionalidad, ya que antes de implementar las interfaces, se pudieron hacer cambios en el diseño lo cual agilizó los tiempos de desarrollo evitando hacer cambios de mayor complejidad.

Un entorno *off-line* favorece la arquitectura propuesta por qué los procesos de entrenamiento no son tan recurrentes (una vez al periodo) y la clasificación una vez al día. Permitiendo así ejecutar

los análisis en horarios de bajo consumo de recursos sin afectar los procesos principales de la institución.

Los resultados de las predicciones se pueden agregar donde se necesiten en el modelo multidimensional, sin afectar ni modificar información consolidada de la institución. Básicamente se puede guardar en alguna tabla de hechos a nivel de alumno y periodo.

La deserción es un reto para cualquier institución educativa y con la ayuda de algoritmos y plataformas tecnológicas es posible colaborar a la toma de decisiones para evitar bajas o abandonos: gestionar mejores tutorías y apoyos a estudiantes. A nivel institucional esto permite una mejor gestión de los recursos académicos y económicos de la institución optimizando así sus procesos y reduciendo tiempos de respuesta.

Trabajo futuro

- Usar atributos adicionales como como historial de pago, adeudos, acceso a campus, etc.
- Implementar muestreo estratificado por decanato para el balanceo de clase en el entrenamiento del modelo.
- Podría extenderse el proyecto con la implementación de alertas tempranas de deserción por medio de notificaciones a usuarios encargados de seguimiento estudiantil.
- Dividir datos en nuevo ingreso y reingreso, ya que tienen características diferentes, es decir, en nuevo ingreso no hay historial en la institución, mientras que los de reingreso, por lo general ya cursaron un semestre: faltas, promedio en periodo, asignaturas reprobadas, etc.
- Utilizar otros algoritmos de clasificación como *vecinos más cercanos*, *máquinas de soporte vectorial*, *regresión logística*, etc., los cuales están disponibles para R.
- Usar una combinación de clasificadores para generar una solución más robusta: bagging.
- Utilizar otras opciones en el campo del aprendizaje computacional como Python, Weka, Kmine, RapidMiner, etc.
- Usar servicios en la nube, tales como *Machine Learning en AWS*, *Azure Machine Learning*, *BigML*, *IBM Watson*, *TensorFlow* o alguna otra solución en aprendizaje computacional que mejore los tiempos de procesamiento.
- Para despliegue de modelos para clasificar en modo online, analizar la viabilidad de usar PMML (Predictive Model Markup Language por sus siglas en inglés) o la utilización de herramientas como OpenCPU.

Referencias

- Ahmed, B.A.F. y Kumar, D, S. (2015). Business Intelligence Design Model (BIDM) for University. International Journal of Computer Applications (0975-8887). Volume 111 - No. 14. febrero 2015.
- Alonso, C., Blanco, E., Fernández, T., y Solís, P. (2014). Caminos desiguales: trayectorias educativas y laborales de los jóvenes en la ciudad de México. México: INEE-El Colegio de México. Primera edición, 2014, INEE: 978-608-7675-61-7. https://www.inee.edu.mx/images/stories/2015/mesapublica/Caminos_desiguales.pdf, 01 de enero 2019.
- Aguilar., y J. Díaz. N (2005). Selección de atributos relevantes basada en bootstrapping. Departamento de lenguajes y sistemas informáticos. Universidad de Sevilla. ETS ingeniería informática. Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, PP. 21-30 ISBN: 84-9732-449-8.
- Arnejo, H. A. (2016). Métodos para la mejora de predicciones en clases desbalanceadas en el estudio de bajas de clientes (CHURN). Universidad de Santiago de Compostela. Noviembre de 2016.
- Banco Mundial (2018). Informe sobre el desarrollo mundial 2018: Aprender para hacer realidad la promesa de la educación, cuadernillo del “Panorama general”, Banco Mundial, Washington DC. Licencia: Creative Commons de Reconocimiento CC BY 3.0 IGO. doi: 10.1596/978-1-4648-1096-1. <https://openknowledge.worldbank.org>, 18 de julio 2019.
- Carvajal, C., González, J., y Sarzoza, S. (2017). variables sociodemográficas y académicas explicativas de la deserción de estudiantes en la facultad de ciencias naturales de la universidad de playa ancha (chile). Formación Universitaria, Vol. 11(2), 3-12 (2018), publicado Abril 2017. https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-50062018000200003&lng=en&nrm=iso&tlng=en, 10 de marzo 2019.
- Cendejas, V., Acuña, L., Cortez, M. y Bolaños, J. (2017). El uso de modelo y metodologías de minería de datos para la inteligencia de negocios. Revista de sistemas computacionales y

TIC'S. Junio 2017 Vol.3 No. 8 54-63.
http://www.ecorfan.org/spain/researchjournals/Sistemas_Computacionales_y_TICs/vol3num8/Revista_de_Sistemas_Computacionales_y_TIC%60S_V3_N8_7.pdf

Chinkes, E. (2017). Pronósticos y data mining para la toma de decisiones, pronóstico sobre la deserción de alumnos de una facultad. Cuadernos del CIMBAGE No.20 107-132, aceptado el 12 de diciembre 2017. http://bibliotecadigital.econ.uba.ar/download/cuadcimbage/cuadcimbage_n20_05.pdf, 10 de marzo 2019.

Díaz, U. (2012). Encabezan mexicanos deserción universitaria. Reforma - México DF. https://www.inee.edu.mx/images/stories/2015/mesapublica/Caminos_desiguales.pdf, 01 de enero 2019.

Gallegos, J., Campos, N., Canales, K., y González, E. (2018). Factores determinantes en la deserción universitaria. caso facultad de ciencias económicas y administrativas de la universidad católica de la santísima concepción. Formación Universitaria, Vol. 11(3), 11-18 (2018), Publicado Jun. 2018. <http://ezproxy.upaep.mx:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edssci&AN=edssci.S0718.50062018000300011&lang=es&site=eds-live>, 10 de marzo 2019.

Ganpatrao R. (2012). Comparative Study of Prototype Model For Software with Engineering With System Development Life Cicle. International Institute of Information Technology, Pune (Maharashtra), India. IOSR Journal of Engineering (IOSRJEN). ISSN: 2250-3021 Volume 2, Issue 7(July 2012), PP 21-24. www.iosrjen.org

Granollers, Perdrix y Lorés (2004). Incorporación de usuarios en la evaluación de la usabilidad por recorrido cognitivo. Interacción '04, Lleida, España. Mayo 2004. <https://aipo.es/articulos/3/42.pdf>, 11 julio 2019.

Granollers, T., Lóres, V., y Delgado, J. (2005). Diseño de sistemas interactivos centrados en el usuario. Universitat Oberta de Calalunya. Colección Manueales. ISBN 8497883209, ISBN pdf 9788490293041. Barcelona España.

Han, J., Kamber, M. y Pei, J (2012). Data Mining, Concepts and Techniques. Third Edition.

Elsevier, Morgan Kaufmann. USA. ISBN 978-0-12-381479-1.
<http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>

Kimball, Ralph (2008). The data warehouse lifecycle toolkit. Practical Techniques for bulding data warehouse and business intelligence systemes. Second Edition, chapter 9. Pag. 371

López Villafaña, L. y Beltrán Solache, A. (2017). La deserción en estudiantes de educación superior: tres percepciones en estudio, alumnos, docentes y padres de familia. Escuela Normal de Tejupilco y Centro Universitario UAEM Temascaltepec. Consultado de Pistas Educativas, No. 126, diciembre 2017. México, Tecnológico Nacional de México en Celaya. <http://www.itcelaya.edu.mx/ojs/index.php/pistas/article/view/776/849>, 30 de enero 2019.

Menes, C., Arcos G., Moreno, P., y Gallegos C., (2015). Desempeño de algoritmos de minería en indicadores académicos: Árbol de decisión y Regresión Logística. Revista Cubana de Ciencias Informáticas. Vol. 9, No. 4, Octubre-Diciembre, 2015. ISSN: 2227-1899 | RNPS: 2301. <https://rcii.uci.cu> Pág. 104-117.

Moine, Juan., M. (2013). Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo. Universidad Nacional de la Plata. Facultad de Informática. Tesis presentada en la Facultad de Informática de la Universidad Nacional de la Plata. Abril de 2013. http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1&isAllowed=y

Muñoz, S., Gallardo, T., Muñoz, M., y Muñoz, C. (2018). Probabilidad de deserción estudiantil en cursos de matemáticas básicas en programas profesionales de la Universidad de los Andes Venezuela. Formación Universitaria, Vol. 11(4), 33-42 (2018) Publicado Ago. 2018. https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-50062018000400033&lng=en&nrm=iso&tlng=en, 10 de marzo 2019.

Nielsen, J. y Mack, R. (1994). Usability Inspection Methods. Jhon Wiley & Sons, Inc. New York. SunSoft yIBM T. J. Watson Research Center.

- Organización para la Cooperación y el Desarrollo Económicos (2019). Higher Education in Mexico: Labour Market Relevance and Outcomes / Educación Superior en México, resultados y relevancia para el mercado laboral, series Higher Education, OECD Publishing, Paris, ISBN 978-92-64-30940-1, ISBN 978-92-64-30943-2 <https://doi.org/10.1787/9789264309432-en>.
- R Core Team (2018). R: A Language and Environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Ramírez, E., Espinosa, D., y Millán, E. (2016). Estrategia para afrontar la deserción universitaria desde las tecnologías de la información y las comunicaciones. Revista Científica, 24, 52-62. Doi: 10.14483/udistrital.jour.RC.2016.24.a5. <https://revistas.udistrital.edu.co/ojs/index.php/revcie/article/download/10344/11430>, 10 de marzo 2019.
- Ramírez, P. y Grandón, E. (2018). Predicción de la deserción académica en una universidad pública chilena a través de la clasificación basada en árboles de decisión con parámetros optimizados. Formación Universitaria, Vol. 11(3), 3-10 (2018), Publicado Jun. 2018. <http://ezproxy.upaep.mx:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edssci&AN=edssci.S0718.500620180003000003&lang=es&site=eds-live>, 10 de marzo 2019.
- Rodríguez, A. (2011). Parte diagnóstica: Aplicación de la Web semántica y técnicas de Inteligencia Artificial para la creación de sistemas inteligentes de diagnóstico Diferencial en Medicina. Universidad Carlos III de Madrid. Madrid, España. Junio del 2011. <http://nadir.uc3m.es/alejandro/phd/thesisEvFinal.pdf>
- Rodríguez, J. y Hernández, J. (2008). La deserción escolar universitaria en México, la experiencia de la universidad autónoma metropolitana campus Iztapalapa. Revista Electrónica publicada por el Instituto de Investigación en Educación Universidad de Costa Rica. ISSN 1409-4703. <http://revista.inie.ucr.ac.cr>, 01 de enero 2019.
- Ruiz., R, Aguilar, J. y Riquelme, J. (2002). Evaluación de Rankings de Atributos para Clasificación. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Sevilla, Sevilla, España. <http://www.lsi.us.es/redmidas/Capitulos/LMD28.pdf>, 10 de julio 2019.

Saavedra, E. R. R., Sarmiento, D. M. E., y Rojas, E. E. M. (2016). Estrategias para afrontar la deserción universitaria desde las tecnologías de la información y las comunicaciones. *Revista Científica*, 24, 52-62. Doi: 10.14483/udistrital.jour.RC.2016.24.a5. <https://revistas.udistrital.edu.co/ojs/index.php/revcie/article/download/10344/11430>, 11 de febrero 2019.

Sánchez, Nicolás (2015). Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. Magíster en Finanzas (Enfoque Mercado de Capitales), Universidad Externado de Colombia. *DEON* , 9, pp. 113-172. DOI: <http://dx.doi.org/10.18601/17941113.n9.04>

Secretaría de Educación Pública (2016). Principales cifras del sistema educativo nacional 2015-2016, cifras preliminares. Biblioteca de publicaciones oficiales del gobierno de México. https://backend.aprende.sep.gob.mx/media/uploads/proedit/resources/principales_cifras_d_8d13ebbb.pdf, 09 de marzo 2019.

Secretaría de Educación Pública de Educación Media Superior (2012). Reporte de la encuesta nacional de deserción. Secretaría de Educación Pública, Consejo para la Evaluación de la Educación del Tipo Medio Superior, Indicadores educativos del Instituto Nacional para la Evaluación de la Educación. http://www.sniesep.gob.mx/descargas/estadistica_e_indicadores/estadistica_e_indicadores_educativos_21PUE.pdf, 11 de febrero 2019.

Secretaría de Seguridad Pública (2011). Deserción escolar y conductas de riesgo en adolescentes. Subsecretaría de prevención y participación ciudadana. dirección general de prevención del delito y participación ciudadana. gobierno federal SSP. Dirección General de Prevención del Delito y Participación Ciudadana. Gobierno Federal SSP. <https://www.uv.mx/dgdaie/files/2013/09/Embarazo-en-Adolescentes-del-Sureste-de-Mexico-2.pdf>, 10 de marzo 2019.

Solano, A. F., y Cardona, J. D. (2015). Evaluación Colaborativa de la Usabilidad en el Desarrollo de Sistemas Software Interactivos. Universidad Autónoma de Occidente. Santiago de Cali, Colombia. ISBN: 978-958-8713-87-8

Tinto, V. (1989). Definir la deserción: una cuestión de perspectiva. *Revista de educación superior*

71.18 (1989): 1-9. https://www.researchgate.net/profile/Vincent_Tinto2/publication/252868573_DEFINIR_LA_DESERCION_UNA_CUESTION_DE_PERSPECTIVA/links/571d596008aee3ddc56ac879/DEFINIR-LA-DESERCION-UNA-CUESTION-DE-PERSPECTIVA.pdf, 09 de marzo 2019.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., y Steinberg, D. (2006). Top 10 algorithms in data minning. Springer-verlag Londong Limited 2007. Knowl Inf Syst (2008) 14:1–37. DOI 10.1007/s10115-007-0114-2.

XindongWu,Vipin Kumar,J. Ross Quinlan,Joydeep Ghosh,Qiang Yang, Hiroshi Motoda,Geoffrey J. McLachlan,Angus Ng,Bing Liu,Philip S. Yu, Zhi-Hua Zhou,Michael Steinbach,David J. Hand,Dan Steinberg 2007, Top 10 algorithms in data mining, 2006. Springer-verlag Londong Limited 2007. Knowl Inf Syst (2008) 14:1–37. DOI 10.1007/s10115-007-0114-2.

Zavala, M. Álvarez, M., Vázquez, M., González, I., y Bazán, A. (2018). Factores internos, externos y bilaterales asociados con la deserción en estudiantes universitarios. Interacciones Revista de Avances en Psicología. 2018 vol. 4. No.1, Enero-Abril, 59-69. ISSN 2411-5940 (impreso) / e-ISSN 2413-4465, (digital). https://www.researchgate.net/publication/322748710_Factores_internos_externos_y_bilaterales_asociados_con_la_desercion_en_estudiantes_universitarios, 09 de marzo 2019.

Anexo 1 Ejemplo de instrumento tipo A

1.1

Evaluación de Prototipo

El objetivo de la encuesta dos interfaces web, las cuales van a mostrar información de riesgo académico para implementar tutorías efectivamente, los aspectos a evaluar son: visibilidad, lenguaje, estándares, facilidad de reconocimiento e interacción con el usuario.

Duración 20-30 minutos

Dirigido a personas con actividades en áreas tecnológicas con algún tipo de conocimiento en desarrollo de aplicaciones web: requerimientos de usuario, diseño web, programación, procesos o testing.

Criterio de evaluación
 0 = no es problema
 1 = problema sin importancia
 2 = problema de poca importancia
 3 = problema grave

Versión: _____

Instrucciones: por cada ítem, evaluar según criterio del recuadro

1.- Visibilidad del estado del sistema

Puntuación

1H01	¿Cada parte de la interfaz comienza con un título o encabezado que describa el contenido de la pantalla?	
1H02	¿El esquema de diseño de los iconos y su estética es consistente en todo el sistema?	
1H04	Los menús de instrucciones, puntos de entrada de datos y mensajes de error ¿aparecen en el mismo lugar de la pantalla o en el mismo menú?	
1H08	¿Hay un tipo de "respuesta" para cada acción u operación?	
1H09	¿El sistema provee algún tipo de "respuesta" visual en menús o cuadros de dialogo que indiquen las opciones que pueden seleccionarse?	
1h25	La terminología utilizada en los menús, ¿es consistente con el dominio del conocimiento del usuario en relación a la tarea a realizar?	
1h26	¿El sistema provee visibilidad?, es decir ¿puede expresar verbalmente el estado del sistema y que alternativas de acción posee en determinado momento?	
2.- Empate entre el sistema y el mundo real(lenguaje de los usuarios)		
2h01	¿Los iconos son concretos y familiares para el usuario?	
2h02	Dada una determinada lista de nombres de ítems y variables para realizar tareas, ¿las opciones en los menús (nombres de los ítems) están ordenadas en la manera más lógica para el usuario?	
2h04	Los campos relacionados e interdependientes, ¿aparecen en la misma pantalla?	
2h05	Si las formas de los objetos de la interfaz son utilizados como pistas visuales, ¿conducen con las convenciones culturales de los usuarios?	
2h06	Los colores seleccionados, ¿Corresponden a valores esperados según los códigos de los usuarios?	
2h09	Cuando se ingresan datos en la pantalla, ¿la terminología utilizada para describir la tarea es familiar para los usuarios?	
2h11	Cuando la pantalla incluye preguntas que deben ser respondidas, ¿el lenguaje de esas preguntas es simple y claro?	
2h13	Los títulos de los menús, ¿siguen el mismo estilo gramatical?	
4.- Consistencia y estándares		
4h01	¿Los formatos de la institución han sido respetados de manera consistente a lo largo de las distintas pantallas del sistema?	
4h02	¿El abuso de las letras en mayúsculas en la pantalla ha sido evitado?	
4h03	¿Las abreviaturas no incluyen punto?	
4h04	¿Los números enteros están justificados a la derecha y los números reales alineados con respecto al punto decimal?	
4h05	¿Los iconos poseen etiqueta?	
4h06	¿No hay más de 12/20 tipos de iconos?	
4h08	¿Cada ventana posee un título?	
4h12	¿Los menús son presentados de manera vertical?	
4h19	¿Las etiquetas de campos mantiene una forma consistente entre una pantalla y otra?	
4h27	Color: hasta cuatro colores diferentes (adicionales utilizados ocasionalmente.	
4h32	Se evitan los pares de colores espectralmente extremos y altamente cromáticos?	
4h45	¿Todas las abreviaturas poseen la misma longitud?	

6.- Reconocimiento mejor que recuerdo.

- 6h02 ¿El despliegue de datos comienza en la parte superior de la izquierda de la pantalla?
- 6h05 Los apuntadores, pistas visuales y mensajes están posicionados en lugares de la pantalla en donde es probable que el usuario dirija su mirada?
- 6h07 Las áreas de texto tienen márgenes (espacios de respiración) que los rodean?
- 6h15 Las etiquetas de los campos están cercanas a los mismos, pero separadas de estos por al menos un espacio en blanco?
- 6h17 Los campos de entrada de datos que son opcionales ¿están claramente marcados?
- 6h21 Se utiliza tamaño de letra, realce de fuente, subrayado, color, sombreado o tipografía especial para mostrar la cantidad relativa o importancia de los diferentes ítems en pantalla?
- 6h22 Se utilizan los bordes para identificar grupos significativos
- 6h23 Se utiliza el mismo color para agrupar elementos relacionados
- 6h24 La codificación de color es consistente dentro de todo el sistema?
- 6h25 El color se usa en conjunción con algún otro elemento redundante?
- 6h26 ¿Existe buen contraste de brillo y de color entre los colores usados para imágenes y fondo?
- 6h27 Los colores suaves, brillantes y saturados se han utilizado para enfatizar datos, mientras que los colores oscuros, opacos y no saturados han sido usados para des-enfatizar datos?
- 6h29 ¿El sistema provee un mapeo que hace que el usuario perciba que existen relaciones entre los controles y las acciones asociadas?
- 6h32 Aquellas secuencias de números o letras que tiene gran longitud, ¿se han descompuesto en "bloques"?
- 6h33 Los ítems inactivos en un menú aparecen en gris o están omitidos?
- 6h34 ¿Existen elecciones por defecto dentro del menú?
- 6h39 ¿Las pantallas de entrada de datos y los cuadros de diálogo indican donde los campos son opcionales?

8.- Diseño estético

- 8h01 La información esencial para tomar decisiones (y solo esta información) ¿es mostrada en la pantalla?
- 8h02 Los iconos son visualmente distinguibles de acuerdo a su significado conceptual
- 8h03 Los objetos extensos, las líneas resaltadas y las áreas simples de la pantalla, ¿se distinguen de los iconos?
- 8h07 ¿Cada pantalla de entrada de datos incluye un título simple, corto, claro y suficientemente distintivo?
- 8h08 ¿Las etiquetas de los campos son familiares y descriptivas?

10.- Ayuda y documentación

- 10h08 ¿La función de ayuda del menú es visible?

12.- Iteración con el usuario de manera respetuosa y placentera

- 12h01 ¿Es cada icono individual un miembro armonioso dentro de una familia de iconos?
- 12h02 ¿Se ha evitado el detalle excesivo en el diseño de iconos?
- 12h03 ¿Se ha usado el color con discreción?

1.1

Evaluación de Prototipo / implementación

El objetivo de la encuesta es validar la funcionalidad según las tareas que el usuario puede realizar

Duración 20-25 minutos

Dirigido a: 1) personas con actividades en áreas tecnológicas con algún tipo de conocimiento en desarrollo de aplicaciones web; requerimientos de usuario, diseño web, programación, procesos o testing. 2) usuarios finales con algún perfil administrativo en el cual tengan a su cargo la explotación de la información académica para el seguimiento de alumnos.

PROT1.1

Instrucciones: evaluar cada acción contestando las 4 preguntas; S(si) N(no)

TAREAS / acciones	ACCIONES	¿El usuario buenista correcto?	¿Permitirán los usuarios que está disponible la acción correcta?	¿Los usuarios asociarán la acción correcta con el resultado que se espera según?	Si se realiza la acción correcta ¿Los usuarios verán que se está avanzando en el resultado?	¿Se realiza la acción correcta ¿Los usuarios verán que se está avanzando en el resultado?
1 Consultar expediente de alumno en prototipo	Seleccionar la pestaña de interfaz por alumno: "Análisis por alumno". Clic en el buscador de alumnos (cuadro de texto, lado derecho). Escribir el uid de un alumno: "12221". Seleccionar al uid que aparece y darle clic. Consultar la información del alumno en la sección superior. Consultar el historico de riesgo académico del alumno.					
2 Consultar aquellos alumnos con riesgo de deserción por programa	Seleccionar la pestaña de interfaz "Análisis por programa". En la sección de filtros por programa seleccionar la opción de "INGENIERIA CIVIL". Consultar la tendencia de bajas del programa. En la sección "Detalle de alumnos" mover el scroll inferior a la derecha. Verificar que las materias acreditadas corresponda con el % de riesgo (deben ser menores)					

Observaciones:

Anexo 3 Ejemplo de instrumento tipo C

1.1

Evaluación de Implementación		
El objetivo de la encuesta es que por medio de dos interfaces web, se van a mostrar información de riesgo académico para implementar tutorías efectivamente, los aspectos a evaluar son: visibilidad, lenguaje, estándares, facilidad de reconocimiento e interacción con el usuario.		
Duración 30-40 minutos		
Dirigido a personas con actividades en áreas tecnológicas con algún tipo de conocimiento en desarrollo de aplicaciones web: requerimientos de usuario, diseño web, programación, procesos o testing.		
Instrucciones: por cada ítem, evaluar según criterio del recuadro 1.- Visibilidad del estado del sistema		Criterio de evaluación 0 = no es problema 1 = problema sin importancia 2 = problema de poca importancia 3 = problema grave 4 = catástrofe
		Puntuación
1H01	¿Cada parte de la interfaz comienza con un título o encabezado que describa el contenido de la pantalla?	
1H02	¿El esquema de diseño de los íconos y su estética es consistente en todo el sistema?	
1H04	Los menús de instrucciones, puntos de entrada de datos y mensajes de error ¿aparecen en el mismo lugar de la pantalla o en el mismo menú?	
1H07	Si se utilizan ventanas emergentes (pop-up) para mostrar mensajes de error, ¿permiten estas ventanas que el usuario visualice el error en la interfaz cuando se despliegan?	
1H08	¿Hay un tipo de "respuesta" para cada acción u operación?	
1H09	¿El sistema provee algún tipo de "respuesta" visual en menús o cuadros de dialogo que indiquen las opciones que pueden seleccionarse?	
1H12	¿El sistema provee algún tipo de "respuesta" visual en menús o cuadros de dialogo que indiquen en cual de las posibles opciones se encuentra posicionado el cursor o el puntero del ratón?	
1H13	¿Hay un tipo de "respuesta" cuando los objetos de la interfaz son seleccionados o movidos?	
1H14	¿El estado actual de cada ícono es claramente indicado?	
1h16	Si existen demoras mayores a 15 segundos en las respuestas del sistema, ¿El usuario es informado del progreso en la culminación de la respuesta?	
1h17	¿Los tiempos de respuestas son apropiadas para cada tarea?	
1h25	La terminología utilizada en los menús, ¿es consistente con el dominio del conocimiento del usuario en relación a la tarea a realizar?	
1h26	¿El sistema provee visibilidad?, es decir ¿puede expresar verbalmente el estado del sistema y que alternativas de acción posee en determinado momento?	
1h29	¿Si los usuarios navegan entre diferentes pantallas del sistema, ¿el sistema utiliza etiquetas conceptuales, mapa de menús o marcas de navegación a modo de ayudas para esa navegación?	
2.- Empate entre el sistema y el mundo real (lenguaje de los usuarios)		
2h01	¿Los iconos son concretos y familiares para el usuario?	
2h02	Dada una determinada lista de nombres de ítems y variables para realizar tareas, ¿las opciones en los menús (nombres de los ítems) están ordenadas en la manera más lógica para el usuario?	
2h04	Los campos relacionados e interdependientes, ¿aparecen en la misma pantalla?	
2h05	Si las formas de los objetos de la interfaz son utilizados como pistas visuales, ¿conducen con las convenciones culturales de los usuarios?	
2h06	Los colores seleccionados, ¿Corresponden a valores esperados según los códigos de los usuarios?	

2h09	Cuando se ingresan datos en la pantalla, ¿la terminología utilizada para describir la tarea es familiar para los usuarios?	
2h11	Cuando la pantalla incluye preguntas que deben ser respondidas, ¿el lenguaje de esas preguntas es simple y claro?	
2h13	Los títulos de los menús, ¿siguen el mismo estilo gramatical?	
2h19	¿El sistema ingresa/elimina automáticamente espacios en blanco?	
2h21	¿El sistema ingresa de manera automática comas en valores superiores a los tres dígitos?	
3.- Control y libertad del usuario		
3h04	Cuando una tarea efectuada por el usuario se completa, ¿El sistema espera alguna señal del usuario antes de procesar la tarea?	
3h06	Se pregunta al usuario que confirme acciones que tendrán consecuencias drásticas, negativas o destructivas?	
3h07	¿Existe una función para "deshacer" al nivel de cada acción simple, cada entrada de datos y cada grupo de acciones completadas?	
3h08	¿Los usuarios pueden cancelar operaciones en progreso?	
3h10	¿Los usuarios pueden reducir el tiempo de entrada de datos copiando y modificando datos existentes?	
3h15	Si el sistema posee menús de niveles múltiples, ¿existe algún mecanismo que permita a los usuarios regresar al menú previo?	
3h17	¿Los usuarios pueden moverse hacia delante o hacia atrás entre las opciones de campos o cuadros de diálogo?	
3h21	¿Los usuarios pueden revertir sus acciones de manera sencilla?	
3h23	¿Los usuarios pueden configurar la apariencia de su propio sistema, sesión y valores por defecto para la pantalla?	
4.- Consistencia y estándares		
4h01	¿Los formatos de la institución han sido respetados de manera consistente a lo largo de las distintas pantallas del sistema?	
4h02	¿El abuso de las letras en mayúsculas en la pantalla ha sido evitado?	
4h03	¿Las abreviaturas no incluyen punto?	
4h04	¿Los números enteros están justificados a la derecha y los números reales alineados con respecto al punto decimal?	
4h05	¿Los iconos poseen etiqueta?	
4h06	¿No hay más de 12 o 20 tipos de iconos?	
4h07	¿Existe algún elemento visual que identifique la ventana activa?	
4h08	¿Cada ventana posee un título?	
4h09	¿Es posible utilizar las barras de desplazamiento horizontal y vertical en cada ventana?	
4h10	¿La estructura de los menús corresponde con la de las tareas?	
4h11	¿Han sido establecidos los estándares de la institución para el diseño de los menús, están aplicados de manera consistente en todas las pantallas?	

	4h12	¿Los menús son presentados de manera vertical?	
	4h13	¿Si la opción de un menú es la de "salir", esta opción aparece como último ítem del menú?	
	4h19	¿Las etiquetas de campos mantiene una forma consistente entre una pantalla y otra?	
	4h22	¿Las técnicas para atraer la atención del usuario están utilizadas de manera cuidadosa?	
	4h27	Color: hasta cuatro colores diferentes (adicionales utilizados ocasionalmente).	
	4h32	Se evitan los pares de colores espectralmente extremos y altamente cromáticos?	
	4h45	¿Todas las abreviaturas poseen la misma longitud?	
	5.- Prevención de errores.		
	5h05	¿Los mensajes de error están expresando de manera tal que es el sistema y no el usuario, quien se hace cargo de los errores?	
	5h06	Si se usan mensajes de error con humor, ¿son apropiados y respetuosos para la comunidad de usuarios?	
	5h07	¿Los mensajes de error son gramaticalmente correctos?	
	5h08	¿Los mensajes de error evitan el uso de signos de admiración?	
	5h09	¿Los mensajes de error evitan el uso de palabras violentas u hostiles?	
	5h11	¿Todos los mensajes de error del sistema utilizan un estilo gramatical, una terminología, una forma y abreviaturas consistentes?	
	5h12	¿Los mensajes colocan al sistema bajo el control del usuario?	
	5h15	Si se detecta un error en un campo de entrada de datos, ¿el sistema posiciona el cursos en ese campo o lo resalta de alguna manera?	
	5h16	¿Los mensajes de error informan al usuario sobre la severidad del error cometido?	
	5h17	¿Los mensajes de error sugieren la causa del problema que los ha ocasionado?	
	5h20	¿los mensajes de error indican que acción debe realizar el usuario para corregir el peor correspondiente?	
	6.- Reconocimiento mejor que recuerdo.		
1	6h01	¿Existen pistas visuales y espacios en blanco para distinguir preguntas, apuntadores, puntos de inserción de respuestas e instrucciones en las interfaces de preguntas y respuestas?	
2	6h02	¿El despliegue de datos comienza en la parte superior de la izquierda de la pantalla?	
3	6h05	Los apuntadores, pistas visuales y mensajes están posicionados en lugares de la pantalla en donde es probable que el usuario dirija su mirada?	
4	6h07	Las áreas de texto tienen márgenes (espacios de respiración) que los rodean?	
5	6h15	Las etiquetas de los campos están cercanas a los mismos, pero separadas de estos por al menos un espacio en blanco?	
6	6h17	Los campos de entrada de datos que son opcionales ¿están claramente marcados?	
7	6h21	Se utiliza tamaño de letra, realce de fuente, subrayado, color , sombreado o tipografía especial para mostrar la cantidad relativa o importancia de los diferentes ítems en pantalla?	
8	6h22	Se utilizan los bordes para identificar grupos significativos	
9	6h23	Se utiliza el mismo color para agrupar elementos relacionados	
10	6h24	La codificación de color es consistente entro de todo el sistema?	
11	6h25	El color se usa en conjunción con algún otro elemento redundante?	
12	6h26	¿Existe buen contraste de brillo y de color entre los colores usados para imágenes y fondo?	
13	6h27	Los colores suaves, brillantes y saturados se han utilizado para enfatizar datos, mientras que los colores oscuros, opacos y no saturados han sido usados para des-enfatizar datos?	

6h29	¿El sistema provee un mapeo que hace que el usuario perciba que existen relaciones entre los controles y las acciones asociadas?	
6h32	Aquellas secuencias de números o letras que tienen gran longitud, ¿se han descompuesto en "bloques"?	
6h33	¿Los ítems inactivos en un menú aparecen en gris o están omitidos?	
6h34	¿Existen elecciones por defecto dentro del menú?	
6h39	¿Las pantallas de entrada de datos y los cuadros de dialogo indican donde los campos son opcionales?	
6h40	En las pantallas de entrada de datos y en los cuadros de dialogo, ¿los campos dependientes se muestran solo cuando es necesario?	
7.- Flexibilidad y eficiencia de uso		
7h01	Si el sistema soporta tanto usuarios novatos y expertos, ¿se encuentran disponibles múltiples niveles de mensaje de error?	
7h18	En los cuadros de dialogo, ¿los usuarios tienen la opción de hacer "clic" directamente en la opción del cuadro de dialogo o de utilizar un atajo del teclado?	
8.- Diseño estético		
8h01	La información esencial para tomar decisiones (y solo esta información) ¿es mostrada en la pantalla?	
8h02	Los iconos son visualmente distinguibles de acuerdo a su significado conceptual	
8h03	Los objetos extensos, las líneas resaltadas y las áreas simples de la pantalla, ¿se distinguen de los iconos?	
8h07	¿Cada pantalla de entrada de datos incluye un título simple, corto, claro y suficientemente distintivo?	
8h08	¿Las etiquetas de los campos son familiares y descriptivas?	
9.- Ayuda a reconocer y recuperarse de errores		
9h05	¿Las entradas de datos son no sensibles a mayúsculas siempre que sea posible?	
9h11	¿El sistema previene a los usuarios de cometer errores siempre que esto es posible?	
9h12	¿El sistema alerta a los usuarios si están a punto de cometer un error potencialmente serio?	
9h14	¿Las pantallas para entrada de datos y cuadros de dialogo indican el número de espacios en caracteres que están disponibles para un campo?	
9h15	Los campos en las pantallas de entradas de datos y los cuadros de diálogo, ¿contienen valores por defecto cuando corresponde?	
10.- Ayuda y documentación		
10h02	¿Las instrucciones en línea se distinguen visualmente?	
10h03	¿Las instrucciones siguen la secuencia de las acciones del usuario?	
10h08	¿La función de ayuda del menú es visible?	
10h09	¿La interfaz de ayuda del sistema es consistente con las interfaces de la aplicación que soporta?	
10h10	Navegación, ¿La información es fácil de encontrar?	
10h11	Presentación, ¿La disposición visual esta bien diseñada?	
10h13	Orientación a la meta ¿Qué puedo hacer yo con este programa?	
10h14	Descriptivo ¿Para qué es esta cosa?	
10h15	Procedimental: ¿Cómo hago yo para hacer esta tarea?	
10h16	Interpretativo ¿Por qué sucedió eso?	
10h17	Navegacional ¿Dónde estoy?	
10h19	¿Puede el usuario cambiar el nivel de detalle disponible?	
10h20	¿Pueden los usuarios cambiar fácilmente entre la ayuda y su trabajo?	
10h21	Tras haber accedido a la ayuda ¿pueden los usuarios continuar con su trabajo desde donde lo dejaron interrumpido?	
11.- Habilidades		
11h05	El sistema codifica automáticamente los ítems con color, con esfuerzo escaso o nulo por parte del	
11h07	¿Son los usuarios los iniciadores de las acciones antes de ser quienes deben responder ante ellas?	
11h11	Cuando el usuario accede a una pantalla o una caja de dialogo, ¿el cursor ya esta posicionado en el campo que mas probablemente el usuario vaya a necesitar?	
11h12	¿Pueden los usuarios moverse hacia adelante y hacia atrás dentro de un campo?	
12.- Iteración con el usuario de manera respetuosa y placentera		
12h01	¿Es cada icono individual un miembro armonioso dentro de una familia de iconos?	
12h02	¿Se ha evitado el detalle excesivo en el diseño de iconos?	
12h03	¿Se ha usado el color con discreción?	
13.- Privacidad		
13h01	¿Las áreas protegidas son completamente inaccesibles?	

Anexo 4 Ejemplo de instrumento tipo D

1.1

Evaluación de tareas

El objetivo de la encuesta es validar la funcionalidad según las tareas que el usuario puede realizar

Duración 20-25 minutos

Dirigido a: 1) personas con actividades en áreas tecnológicas con algún tipo de conocimiento en desarrollo de aplicaciones web; requerimientos de usuario, diseño web, programación, procesos o testing. 2) usuarios finales con algún perfil administrativo en el cual tengan a su cargo la explotación de la información académica para el seguimiento de alumnos.

Instrucciones: evaluar cada acción contestando las 4 preguntas: S(s) N(no)

C 1.1

TAREAS / acciones	ACCIONES	¿El usuario intentará lograr el efecto correcto?	¿Peribirán los usuarios que está disponible la acción correcta?	¿Los usuarios asociarán la acción correcta con el resultado que espera lograr?	Si se realiza la acción correcta ¿Los usuarios verán que se está avanzando hacia el resultado deseado?
1 Consultar que programa académico tiene el mayor número de alumnos en riesgo académico.					
Clic en la opción del menu izquierdo "Seguimiento por decanato".					
En la sección "Porcentaje de alumnos en riesgo por decanato académico (DESERTOR = S)" click sobre el gráfico de barra que tenga el mayor porcentaje de riesgo.					
En la sección "%" en riesgo por decanato académico" ubicar con un clic el programa con mayor % de riesgo.					
En esa misma sección visualizar el número de alumnos en riesgo que se encuentra en la columna "n" en riesgo de deserción.S"					
2 Obtener el listado de alumnos en riesgo de deserción.					
Clic en la opción del menu izquierdo "Seguimiento de todos los alumnos".					
Clic en la sección "Predicción por RNA" y seleccionar solo "En riesgo de deserción" El listado se deberá filtrar.					
Clic en el listado, en la parte izquierda derecha superior, pasar el mouse, aparecerá un cuadro pequeño con puntos.					
Clic en el cuadro y deberá aparecer una opción "Export", clic en la opción "Excel"					
Guardar el archivo en una ubicación segura y abrir.					
3 Consultar alumnos en riesgo académico por nombre de alumno					
Clic en la opción del menu izquierdo "Consulta estudiante".					
En el cuadro de la esquina superior izquierda clic en el que aparece como "Search Nombre"					
Ingresar: REWVF080GJ5C63RFXM, deberá aparecer una opción con ese mismo dato, dar clic.					
Consultar datos de correo o teléfono para contactarlo.					
Observaciones:					

Anexo 5 Reglas generadas por árbol de decisión

```
UBICATION_TREE:
AVANCE_XCRED > 83: S (3287/471)
AVANCE_XCRED <= 83:
... FALTASP > 54:
... REPROBADAS3P > 0.56: S (1609/131)
... REPROBADAS3P <= 0.56:
... PROMEDIO_AC <= 7.07: S (161/19)
... PROMEDIO_AC > 7.07:
... ASIG_INSCRITASP <= 4:
... FALTASP > 58: S (91/16)
... FALTASP <= 58:
... PERIODOS_INSCRITOS > 6: N (8)
... PERIODOS_INSCRITOS <= 6:
... EDAD <= 20: N (4)
... EDAD > 20: S (16/3)
ASIG_INSCRITASP > 4:
... REPROBADAS1P > 0.75:
... ESDOCIVIL <= 0: S (13)
... ESDOCIVIL > 0:
... REPRO_2X <= 0:
... PROM_PANTE <= 3.6: S (3)
... PROM_PANTE > 3.6: N (8)
... REPRO_2X > 0:
... REPRO_1X <= 3: N (3)
... REPRO_1X > 3: S (14)
REPROBADAS1P <= 0.75:
... PROM_PANTE <= 8.01:
... ASIG_INSCRITASP > 6:
... EDAD <= 20:
... ASIG_INSCRITASP > 7: S (27/7)
... ASIG_INSCRITASP <= 7:
... ESDOCIVIL <= 0: S (13/3)
... ESDOCIVIL > 0: N (11)
... EDAD > 20:
... REPROBADAS2P > 0.63: S (12/2)
... REPROBADAS2P <= 0.63:
... REPRO_1X <= 3: N (33)
... REPRO_1X > 3:
... GEN_FEMENINO <= 0: N (43/11)
... GEN_FEMENINO > 0: S (20/6)
ASIG_INSCRITASP <= 6:
... NINGRESO > 0:
... REPROBADAS2P <= 0.57: N (14/2)
... REPROBADAS2P > 0.57: S (6/1)
... NINGRESO <= 0:
... REPROBADAS3P > 0.33: S (101/30)
... REPROBADAS3P <= 0.33:
... PROMEDIO_AC <= 7.33: N (11)
... PROMEDIO_AC > 7.33:
... PROM_INI <= 8.49: S (16/1)
... PROM_INI > 8.49:
... FALTASP <= 77: S (38/37)
... FALTASP > 77: N (16/1)
PROM_PANTE > 8.01:
... PROM_PANTE > 9.13:
... CRED_PANTE > 58: N (5)
... CRED_PANTE <= 58:
... PROMEDIO_AC <= 8.16: N (3)
... PROMEDIO_AC > 8.16: S (15)
PROM_PANTE <= 9.13:
... APOVO > 5: N (22)
... APOVO <= 5:
... FALTASP <= 55: N (24)
... FALTASP > 55:
... FALTAS_PANTE > 305: S (17/4)
... FALTAS_PANTE <= 305:
... PERIODOS_INSCRITOS <= 3:
... CRED_PANTE <= 27: N (15)
... CRED_PANTE > 27:
... PROM_PANTE > 8.65: N (12/2)
... PROM_PANTE <= 8.65: [S1]
PERIODOS_INSCRITOS > 3:
... EDAD <= 20: N (34/2)
... EDAD > 20:
... CRED_PANTE <= 73: S (5)
... CRED_PANTE > 73:
... FALTASP <= 93: N (86/16)
... FALTASP > 93: S (6/1)

FALTASP <= 54:
... PROMEDIO_AC <= 7.5:
... PROMEDIO_AC > 7.5:
... REPROBADAS3P <= 0.63:
... REPROBADAS3P > 0.63:
... REPRO_3X <= 0:
... ASIG_INSCRITASP > 5:
... PROM_PANTE <= 7.69:
... REPROBADAS1P > 0.67:
... REPROBADAS3P > 0.38: S (12)
... REPROBADAS3P <= 0.38:
... REPROBADAS3P <= 0.25: S (9/2)
... REPROBADAS3P > 0.25: N (7)
... REPROBADAS1P <= 0.67:
... REPRO_2X > 2:
... ASIG_INSCRITASP > 6: N (12)
... ASIG_INSCRITASP <= 6:
... FALTASP <= 12: N (3)
... FALTASP > 12: S (23/1)
... REPRO_2X <= 2:
... EDAD > 18: N (527/144)
... EDAD <= 18:
... ASIG_INSCRITASP > 7: N (8)
... ASIG_INSCRITASP <= 7:
... FALTASP > 39: N (7)
... FALTASP <= 39:
... APOVO > 45: S (6)
... APOVO <= 45:
... PROM_INI <= 7.27: N (8)
... PROM_INI > 7.27: S (72/28)
PROM_PANTE > 7.69:
... REPRO_2X > 0: S (6)
... REPRO_2X <= 0:
... REPROBADAS2P <= 0.3: S (28/3)
... REPROBADAS2P > 0.3:
... REPROBADAS3P <= 0.18: N (8)
... REPROBADAS3P > 0.18:
... REPROBADAS2P > 0.75: N (6)
... REPROBADAS2P <= 0.75:
... PROMEDIO_AC > 7.48: N (7)
... PROMEDIO_AC <= 7.48:
... REPRO_PANTE > 1: S (13/1)
... REPRO_PANTE <= 1: [S2]
ASIG_INSCRITASP <= 5:
... EDAD > 24:
... PERIODOS_INSCRITOS > 16: N (6)
... PERIODOS_INSCRITOS <= 16:
... REPROBADAS1P > 0.44: S (44/3)
... REPROBADAS1P <= 0.44:
... REPROBADAS1P > 0.33: N (10/1)
... REPROBADAS1P <= 0.33:
... PROM_INI <= 6.9: N (4)
... PROM_INI > 6.9: S (38/10)
EDAD <= 24:
... FALTAS_PANTE > 270:
... REPROBADAS3P <= 0.33: N (29)
... REPROBADAS3P > 0.33:
... EDAD > 23: S (6)
... EDAD <= 23:
... FALTAS_PANTE <= 434: N (10)
... FALTAS_PANTE > 434: S (5)
FALTAS_PANTE <= 270:
... REPRO_2X <= 1:
... EDAD > 21:
... REPROBADAS2P <= 0.22:
... REPROBADAS3P <= 0.5: N (73/22)
... REPROBADAS3P > 0.5: S (15/3)
... REPROBADAS2P > 0.22:
... REPROBADAS3P > 0.1: S (238/76)
... REPROBADAS3P <= 0.1:
... PROM_INI <= 7.02: S (4)
... PROM_INI > 7.02: N (17/3)
... EDAD <= 21:
... ASIG_INSCRITASP <= 2: S (95/33)
... ASIG_INSCRITASP > 2:
... PROM_INI <= 8.65:
... AVANCE_XCRED > 33: N (25/3)
... AVANCE_XCRED <= 33:
... REPRO_1X <= 3: N (873/383)
... REPRO_1X > 3: S (174/69)
... PROM_INI > 8.65:
... PROM_PANTE <= 7.11: S (37/3)
... PROM_PANTE > 7.11:
... REPRO_1X <= 1: N (8)
... REPRO_1X > 1: [S3]
REPRO_2X > 1:
... REPROBADAS3P <= 0.1:
... PROMEDIO_AC <= 7.14: S (12/3)
... PROMEDIO_AC > 7.14: N (13)
... REPROBADAS3P > 0.1:
... PERIODOS_INSCRITOS <= 9: S (155/37)
... PERIODOS_INSCRITOS > 9:
... FALTAS_PANTE <= 170: N (9)
... FALTAS_PANTE > 170:
... ESDOCIVIL <= 0: S (5)
... ESDOCIVIL > 0: [S4]

REPROBADAS3P > 0.63:
... NINGRESO > 0:
... REPROBADAS2P > 0.71: S (17/1)
... REPROBADAS2P <= 0.71:
... REPROBADAS3P > 0.88: N (7)
... REPROBADAS3P <= 0.88:
... REPROBADAS3P > 0.75: S (32/9)
... REPROBADAS3P <= 0.75:
... FALTASP <= 7: S (6)
... FALTASP > 7: N (35/7)
NINGRESO <= 0:
... PROMEDIO_AC <= 7.17: S (234/49)
... PROMEDIO_AC > 7.17:
... PROM_PANTE > 8.05: S (38/2)
... PROM_PANTE <= 8.05:
... PROM_PANTE > 7.44:
... EDAD > 21: S (84/13)
... EDAD <= 21:
... CRED_PANTE > 119: N (10)
... CRED_PANTE <= 119:
... SEMESTRE_PANTE > 1: S (80/24)
... SEMESTRE_PANTE <= 1:
... OTONO <= 0: N (12/1)
... OTONO > 0: S (3)
... PROM_PANTE <= 7.44:
... FALTASP <= 9: N (13)
... FALTASP > 9:
... REPROBADAS2P <= 0: S (7)
... REPROBADAS2P > 0:
... PROM_INI > 8.55: S (10)
... PROM_INI <= 8.55:
... CRED_PANTE <= 19: S (22/3)
... CRED_PANTE > 19:
... REPRO_2X > 2: S (7/1)
... REPRO_2X <= 2: [S5]
PROMEDIO_AC > 7.5:
... ASIG_INSCRITASP > 4:
... AVANCE_XCRED <= 76: N (25379/4015)
... AVANCE_XCRED > 76:
... ASIG_INSCRITASP > 7:
... OTONO <= 0: S (157/22)
... OTONO > 0:
... EDAD > 25: S (9)
... EDAD <= 25:
... AVANCE_XCRED <= 79: N (15)
... AVANCE_XCRED > 79:
... CRED_PANTE > 383: N (4)
... CRED_PANTE <= 383:
... REPROBADAS3P <= 0.22: S (23/3)
... REPROBADAS3P > 0.22: N (3)
ASIG_INSCRITASP <= 7:
... PROM_INI > 9.38: N (114/12)
... PROM_INI <= 9.38:
... OTONO <= 0:
... ASIG_INSCRITASP <= 5:
... REPROBADAS1P <= 0.5: N (118/32)
... REPROBADAS1P > 0.5:
... EDAD <= 23: N (3)
... EDAD > 23: S (8)
... ASIG_INSCRITASP > 5:
... AVANCE_XCRED > 79:
... CRED_PANTE <= 385: S (161/41)
... CRED_PANTE > 385: S (19/2)
... AVANCE_XCRED <= 79:
... FALTASP <= 11:
... SEMESTRE_PANTE <= 5: S (3)
... SEMESTRE_PANTE > 5: N (43/2)
... FALTASP > 11:
... PROMEDIO_AC <= 7.94: N (12)
... PROMEDIO_AC > 7.94:
... FALTAS_PANTE > 133: S (61/19)
... FALTAS_PANTE <= 133:
... FALTASP <= 12: S (5/1)
... FALTASP > 12: N (27/4)
```

Anexo 5 Manual de usuario

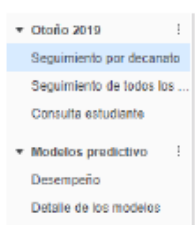
Ayuda

Muestra información de alumnos en riesgo de deserción según modelos de predicción: redes neuronales artificiales y árboles de decisión. La información mostrada corresponde al periodo actual.

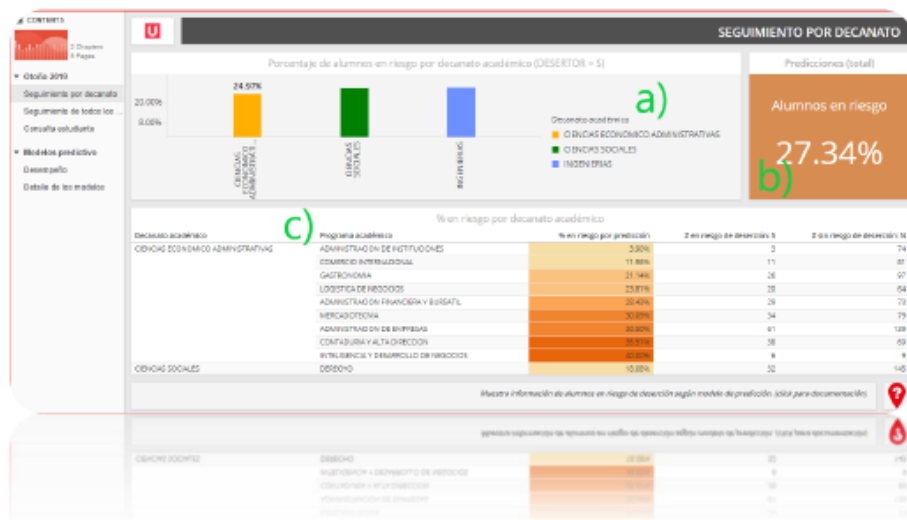
Seguimiento por decanato

Muestra un resumen por decanato de los alumnos en riesgo de deserción según modelos de predicción: redes neuronales artificiales y árboles de decisión. La información del periodo mostrado corresponde al periodo actual.

1. Seleccione la opción "Seguimiento por decanato" en el menu de la parte izquierda de la pantalla.



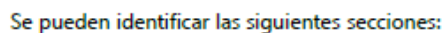
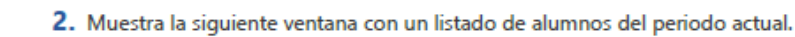
2. Enseguida aparecera la pantalla



Se puede encontrar la siguiente información:

- Todos los alumnos

1. Selecciona la opción del menú principal.

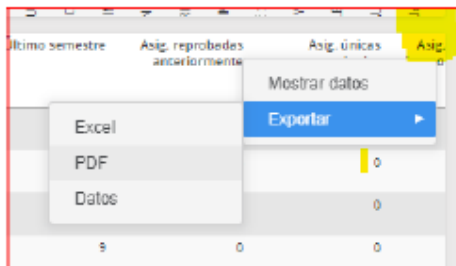


- Página | 2

-
- d) Muestra información del alumno como periodo de ingreso, asignaturas inscritas, avance académico, reprobación, promedio, etc.

3. Exportar los datos

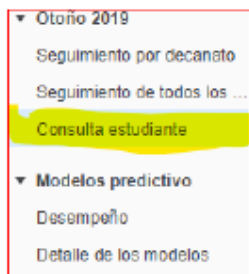
En el lado derecho del listado se puede dar click y aparecerá la opción de exportar:



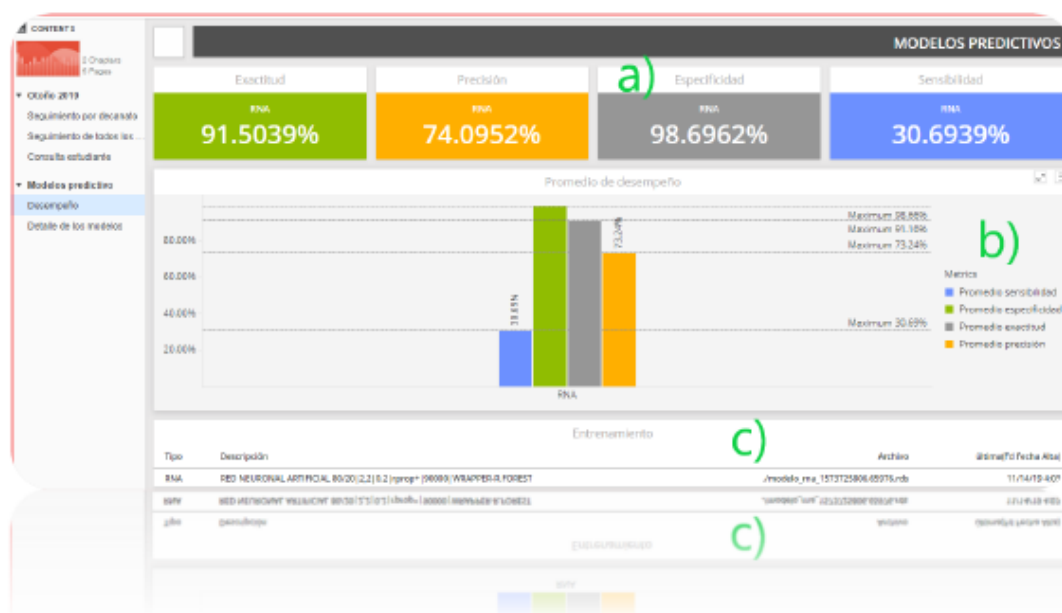
Consulta por estudiante

Muestra información detallada del estudiante filtrada por nombre o por matrícula.

1. Selecciona la opción del menu principal.



2. Aparecerá la siguiente interfaz:



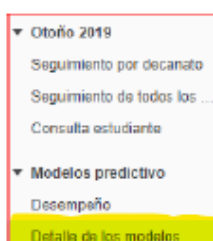
Se puede observar las siguientes secciones:

- Principales métricas de los modelos.
- Muestra un comparativo de promedios sobre las principales métricas de los modelos entrenados.
- Detalle de archivos físicos de los modelos.

Detalle de entrenamientos de los modelos

Muestra información detallada de los entrenamientos, como nombres de archivos, fecha de creación, atributos usados, etc.

1. Selecciona la opción del menú.



Acontinuación, aparece un listado con el tipo de modelo, archivo generado, hora de creación, atributos usados y metricas de desempeño: