Predictor de deserción universitaria

Giselle V. Romero[1], Joaquín S. Toranzo Calderón[1], Sesbastián E. Jaremczuk[1], Juan C. Gómez[1,2], Claudio Verrastro[1,3]

[1] Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Grupo de Inteligencia Artificial y Robótica (GIAR), Av. Medrano 951 (C1179AAQ) Ciudad Autónoma de Buenos Aires, Argentina
[2] Instituto Nacional de Tecnología Industrial (INTI) Instrumentación y Control, Electrónica e Informática, Avenida General Paz 5445 Edificio 38, (B1650WAB) San Martín, Provincia de Buenos Aires, Argentina
[3] Comisión Nacional de Energía Atómica (CNEA) Centro Atómico Ezeiza, CNEA. Instrumentación y Control, AYA, Camino Real Presbítero González y Aragón 15, (B1802) Ezeiza, Provincia de Buenos Aires, Argentina

juanca@inti.gob.ar

Recibido el 24 de diciembre de 2020, aprobado el 2 de febrero de 2021

RESUMEN
La deserción estudiantil siempre ha sido un tema de preocupación debido a sus múltiples implicancias. En este trabajo se propone la aplicación de técnicas de reconocimiento de patrones para exponer información útil y formular reglas de inferencia en sistemas de diagnóstico automático. De esta manera se generan modelos predictivos de deserción universitaria en la UTN.BA, a partir de bases de datos de estudiantes de la carrera de Ingeniería en Sistemas de la Información del plan K08. Se construyeron dos modelos, uno basado sobre Máquinas de Vectores de Soporte y otro sobre Redes neuronales. Ambos presentan resultados muy similares reconociendo a estudiantes en situación de deserción con una exactitud de 79%.

PALABRAS CLAVE: MINERÍA DE DATOS - PREDICTOR DE DESERCIÓN UNIVERSITARIA - SVM - REDES NEURONALES - APRENDIZAJE AUTOMÁTICO

ABSTRACT
Dropping out has been a cause of concern due to its multiple implications. In this work, the application of pattern recognition techniques is proposed to make explicit meaningful information to be used later in expert systems. The application of these techniques was aimed at generating predictive models of university dropout at the UTN.BA, from databases of students of the Information Systems Engineering career of the K08 plan. Two models were built, one based on Support Vector Machines and the other on Neural Networks. Both present very similar results, recognizing dropout students with an accuracy of 79%.

KEYWORDS: KNOWLEDGE DATA DISCOVERY - DATA MINING - UNIVERSITY DROPOUT PREDICTOR - SVM - NEURAL NETWORKS - MACHINE LEARNING

Introducción

La deserción de estudiantes siempre ha sido un fenómeno de interés en la investigación educativa, considerado un problema en cada uno de sus niveles por sus distintas consecuencias. En particular, la deserción universitaria tiene un impacto en las vacancias en las áreas vinculadas a las TICs, y por consiguiente con el desarrollo de éstas de manera local (INET, 2016) (SPU, 2018). Esto muestra que no es una excepción para el sector universitario, por lo que las Universidades han volcado esfuerzos para abordar la deserción, esfuerzos que se reflejan en la búsqueda de causas para el abandono de los estudios y en el diseño de dispositivos institucionales que se dirijan a estas causas (Kuna, *et al*., 2010).

El presente trabajo propone la aplicación de una metodología de análisis de información en pos de ofrecer herramientas para un acercamiento a la comprensión del fenómeno de la deserción universitaria, y así dirigir acciones de las políticas institucionales que busquen modificar la naturaleza y condiciones que lo posibilitan. De esta forma se mejorará la experiencia educativa estudiantil y el uso de los recursos institucionales. Este trabajo se concentra en la generación de modelos que permitan identificar posibles causas y/o contar con indicadores de alerta temprana de deserción para la carrera de *Ingeniería en Sistemas de la Información* (Plan K08), UTN.BA.

La metodología que se expone consiste en una serie de etapas, o procedimientos, orientadas a reconocer patrones "ocultos" en la información registrada en un conjunto de datos. El reconocimiento de estos patrones en la información ya disponible y su posterior análisis puede entenderse como un descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés: *Knowledge Data Discovery*).

El resultado final de los procedimientos aquí desarrollados ha sido la generación de modelos que predigan si ciertos elementos del conjunto de datos seguirán o no un comportamiento específico (en este caso, si un estudiante representado en este conjunto de datos abandonará o no sus estudios).

A continuación, se explica en qué consiste la metodología KDD y de qué manera puede abordarse el problema de la deserción universitaria con ella. Se introduce brevemente el fenómeno de deserción y debido a su naturaleza elusiva, se propone una definición en términos cuantitativos, necesaria para la aplicación del procedimiento mencionado. Luego se describe la base de datos utilizada, sus limitaciones y una exploración estadística de sus datos. En ese punto se realizan una serie de consideraciones sobre las transformaciones necesarias y el alcance de los resultados obtenidos así como los modelos que se construyen posteriormente. Se describen los pasos del proceso KDD realizado y finalmente se exponen los resultados y conclusiones.

Descripción de la metodología KDD

El descubrimiento de conocimiento en bases de datos (KDD) (Fayyad, *et al.* 1996) es una metodología consolidada en la investigación de problemas de diversas ciencias sociales debido a su capacidad para tratar con sistemas de mucha complejidad, no sólo por manejar grandes volúmenes de datos, sino por evaluar e interpretar numerosas variables que describen o relacionen a estos datos. Esta capacidad para predecir eventos de un determinado fenómeno tiene valor en sí misma, pero además goza de la posibilidad de generar modelos adecuados a la interpretación humana, lo que ayuda a explicar determinados eventos sobre la base de posibles causas.

En tanto metodología, KDD consiste en una serie de procedimientos que se suceden unos a otros para la obtención de conocimiento útil a partir de un conjunto de datos inicial. A este proceso de descubrimiento de conocimientos subyacentes se lo denomina *elicitación*, y estos conocimientos (típicamente nuevos) pueden ser la clave para entender el fenómeno estudiado, o por lo menos identificar variables potencialmente relevantes para tener una mejor comprensión.

En el centro de la metodología KDD están las técnicas de *Minería de Datos* (DM, por sus siglas en inglés: *Data Mining*) y los algoritmos de *Reconocimiento de Patrones* (PR, por sus siglas en inglés: *Pattern Recognition*) (Bishop, 2006). Estas técnicas y algoritmos (típicamente estadísticas) resultan en modelos descriptivos (modelos que destacan o identifican las relaciones entre las variables vinculadas al evento a explicar) y predictivos (modelos que permiten indicar para nuevos datos, su eventual resultado). Dada su importancia, DM designa a toda una etapa de la metodología, la de la explotación de los datos para el reconocimiento de los patrones ocultos; aún más, DM suele designar incluso en los ámbitos más formales a la serie completa de las etapas, a la metodología KDD[1].

En la Figura 1 se observa una representación de la secuencia que siguen las etapas que conforman a la metodología. Nótese que la secuencia no es necesariamente lineal, pues luego de cada etapa debe existir una instancia de supervisión en la que se decida si debe retrocederse a una etapa previa o no, en virtud de los resultados ofrecidos por la etapa precedente.

[1] Nótese que esto puede generar en ciertos casos equívocos provocados por el uso de la metonimia. En lo que sigue del artículo, el contexto es suficiente para dejar en claro a qué designa el nombre "DM", y se desambiguará cuando no sea suficiente para resolver el equívoco.

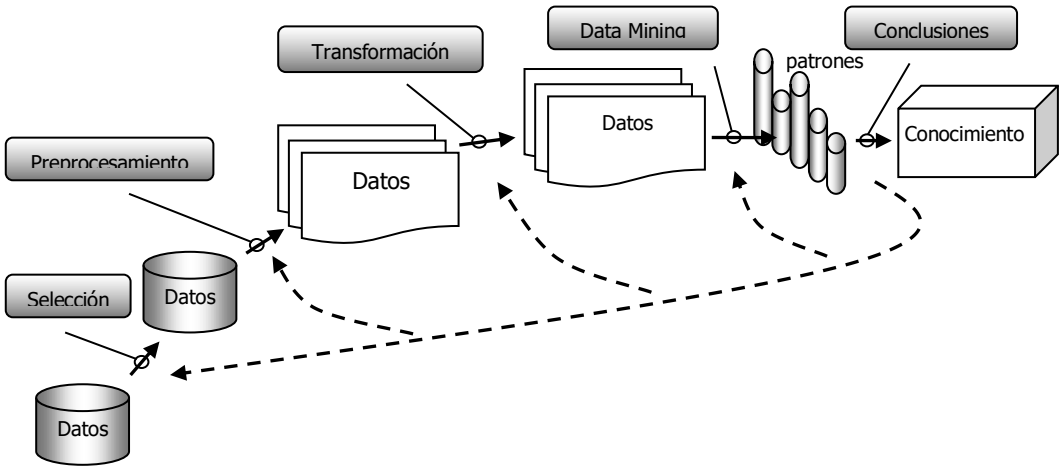

Fig. 1 Proceso KDD

Al comienzo del proceso, se necesita seleccionar qué datos se usarán, escogiendo tanto las fuentes pertinentes, como los registros específicos de entre los disponibles. Para esta *selección* se utiliza conocimiento previo[2], típicamente el consejo de expertos de la disciplina involucrada, en este caso las relativas a las Ciencias de la Educación. Además, se recurre a un análisis estadístico preliminar iterativo de los datos disponibles evaluando la correlación entre los mismos.

La segunda etapa, el *preprocesamiento* de los datos, tiene por objeto la modificación de la base de datos para su "limpieza", i.e. detectar y corregir errores en los valores que toman los campos, eliminar información evidentemente inválida, y completar campos cuando sea posible o necesario.

Luego se realiza una *transformación* de los datos, construyendo una nueva base de datos, en función de relaciones que se pueden establecer de manera directa con los datos ya disponibles. La caracterización de esta base de datos dependerá del tipo de comportamiento que se desee descubrir en ella, pero también del tipo de técnicas y algoritmos que se aplicarán en la etapa siguiente, dado que debe haber una compatibilidad formal entre ellos para que funcionen, i.e. tipo de datos de un campo y el tipo consumido por el algoritmo.

La siguiente etapa, la que recibe la denominación propia de "DM", realiza la búsqueda de patrones que subyacen en los datos, para lo cual explota las relaciones que pudiera haber entre las variables. De esa manera, al

[2] Vale destacar que la *selección* de los datos es posible sólo tras su recolección. La recolección no es parte de KDD, y tienen criterios que pueden ser independientes a su posterior análisis, como información pertinente para determinar la regularidad de un estudiante para determinar su derecho a acceder a becas o el cálculo de sus notas para determinar su prioridad en la inscripción a materias.

descubrir esos patrones se puede construir un modelo y dependiendo de qué técnica se use, este puede ser descriptivo o predictivo.

La última etapa, la interpretación del modelo, es una actividad humana y es el último paso del proceso, donde posiblemente se genera nuevo conocimiento, el conocimiento *elicitado*. A partir de éste, muchas veces surge la necesidad de hacer iteraciones del proceso, parcial o totalmente; por ejemplo, iteraciones parciales consistirán en regresar a la etapa de preprocesamiento para modificar el modo en el que se corrigen errores encontrados o regresar a la etapa de transformación para construir más variables que mostraron ser valiosas, mientras que iteraciones totales podrían consistir en repetir todas las etapas al tener una noción poco precisa pero ya informada del patrón oculto en la base de datos.

Deserción Universitaria

La deserción universitaria es un fenómeno que se manifiesta de maneras muy distintas según la institución educativa, por lo que su definición es problemática. En primer lugar, las instituciones de educación privada apelan a la interrupción en la matriculación de sus estudiantes para determinar su continuidad académica, mientras que las instituciones de educación pública no siempre determinan la continuidad de sus estudiantes en términos de su matriculación. Dado que este trabajo se dirige a una carrera dictada en la UTN.BA, perteneciente al sector público, y que la institución no cuenta con ninguna definición explícita (o al menos no se halla disponible), es preciso referirse a las definiciones propuestas en la literatura y enunciar una definición clara y precisa con la cual trabajar.

Las diferencias entre instituciones conllevan a una serie de nociones difusas y hasta informales de la deserción. Viale Tudela (2014) recoge algunos intentos de desarrollar y precisar estas nociones. En ellas, la idea generalizada apunta hacia el cese de actividades deliberado o forzado del alumno en la institución educadora, pero a la vez remarca que no todo cese de actividades representa una "verdadera" deserción. Barrera Rebellón (2008) propone un enfoque similar, pero da un paso más al recomendar la construcción de una definición y metodología de cálculo, apropiada para diferenciar a desertores "verdaderos" de estudiantes que no han desertado a pesar de haber interrumpido sus actividades académicas. Fernández-Hileman, *et al.* (2014) realizan una reconstrucción bastante exhaustiva del estado de la cuestión, en la que se distinguen nociones emparentadas con la deserción, tales como la "retención" y la "persistencia", o el "fracaso" y el "abandono", así como distinguen entre nociones de deserción vinculadas al momento de la deserción, a la duración, al alcance o al mecanismo de la deserción; la reconstrucción muestra la falta de acuerdo entre los investigadores para fijar un criterio, pero también la laxitud con la que se habla de deserción.

Siguiendo las distinciones recogidas por Fernández-Hileman *et al.* (2014), y bajo la recomendación del trabajo de Barrera Rebellón (2008), la presente propuesta se dirige a una noción de la deserción en tanto un posible abandono definitivo de las actividades académicas, representado

por una discontinuidad prolongada en las actividades del estudiante. Para establecer una definición que permita una metodología de clasificación, se categoriza a los estudiantes como potenciales desertores si no registran actividades por dos o más ciclos lectivos consecutivos (hasta el momento del análisis). Por "actividades" se entienden los exámenes finales que rinde un estudiante y las materias que cursa, de manera que un estudiante que no se haya presentado a rendir exámenes o no se haya inscrito en materias, será considerado como un potencial desertor (por haber abandonado sus actividades). Esta definición para la deserción no es dependiente del momento en el que se produce el abandono, pudiendo abarcar tanto la "deserción precoz", como la "deserción temprana" y la "deserción tardía". Asimismo, no distingue el mecanismo que produce el abandono ni el alcance de la deserción. Con esta definición es posible categorizar a los estudiantes descriptos por la base de datos disponible, recategorizar a los estudiantes con el paso del tiempo, y categorizar a nuevos estudiantes.

Base de Datos disponible

El proyecto prevé el uso de las bases de datos de estudiantes de UTN.BA Desafortunadamente, estas bases son limitadas en cuanto a su contenido. Han sido desarrolladas sólo para llevar el registro académico de estudiantes y su identificación. Esto es, las cursadas, exámenes parciales, firma de materias y aprobación de finales con sus respectivas fechas. Hay algunos datos extras como género y tipo de educación secundaria, pero no hay datos relacionados a otras actividades y características, tales como la dedicación laboral, datos históricos de la misma, domicilio actualizado, etc. Se cuenta casi exclusivamente con datos indirectos de desempeño universitario.

Esta limitación impide, por ejemplo, realizar un análisis de condiciones socioeconómicas y laborales, que a priori parecerían de interés a los fines de enfocar los esfuerzos de ayuda desde la Universidad en pos de bajar la deserción no deseada.

La base de datos utilizada fue entregada en tres archivos: el documento Datos-Estudiantes-SIGA[3] contiene un registro por estudiante, con datos personales; el documento Cursadas-Estudiantes-SIGA contiene un registro por cada materia en la que se inscribe cada estudiante; el documento Finales-Estudiantes-SIGA contiene un registro por cada examen final rendido por cada estudiante. Los registros de las tres tablas se encuentran relacionados a través de un ID anonimizado que corresponde a cada estudiante.

A continuación, se describe cada tabla y se realiza una primera exploración estadística de sus contenidos.

[3] SIGA es el Sistema Integrado de Gestión Académica de la UTN-FRBA

Datos Estudiantes SIGA:

Contiene un registro por estudiante con los siguientes datos personales: *ID*, *nacionalidad*, *localidad* (del domicilio), *provincia, estudios secundarios*, *estado civil*, *sexo*, *año de nacimiento* y *año de ingreso* (a la Universidad).

En la Tabla 1 se muestran parcialmente los datos. Se destaca que la gran mayoría son de nacionalidad argentina, viven cerca a la facultad, son solteros y que un 87% son de sexo masculino. Por otra parte, la educación secundaria se encuentra más equilibrada. Sobre el dato del estado civil, que suena razonable que tenga algún impacto en la continuidad de los estudios, la Secretaría de Planeamiento y Gestión de Procesos indicó que es un dato que seguramente no se encuentre actualizado, y si estuviera actualizado, no se registra el momento de la actualización.

Tabla 1. Análisis de datos de la BD de estudiantes (muestra parcial)

Variable	Característica	Frecuencia	Frecuencia %	Ranking
Nacionalidad	Argentina	8242	98,88	1
Nacionalidad	Perú	33	0,4	2
Nacionalidad	Bolivia	22	0,26	3
Sexo	M	7260	87,1	1
Sexo	F	1075	12,9	2
….	….	….	….	….

Cursadas Estudiantes SIGA:

La Tabla 2 muestra parcialmente los datos de cursadas. Tienen un registro por cada cursada realizada por una determinada persona y se relaciona con las otras tablas mediante un *ID* anonimizado con la persona que la cursó. Es de destacar que puede haber registros de la misma materia para la misma persona, pero realizado en momentos diferentes.

Tabla 2. Análisis de datos de la BD de cursadas (muestra parcial)

Variable	Característica	Frecuencia	Frecuencia %	Ranking
Modalidad	Anual	90208	45,15	1
Modalidad	Cuat ½	58496	29,28	2
Modalidad	Cuat 2/2	50233	25,14	3
Descripción.de.recursada.regular	No Recursó	152426	76,28	1
Descripción.de.recursada.regular	Recursó 1 vez	30665	15,35	2
Descripción.de.recursada.regular	Recursó 2 v.	10057	5,03	3
Descripción.de.recursada.libre	No Recursó	173493	86,83	1
Descripción.de.recursada.libre	Recursó 1 v.	19345	9,68	2
Descripción.de.recursada.libre	Recursó 2 v.	4603	2,3	3

Esto se debe a que se está registrando cada evento relacionado a las cursadas. Las variables que describen a cada registro son las siguientes (y entre paréntesis se indican los valores que puede tomar cada variable, o una breve descripción de la variable): *curso* (*código*); *materia* (*nombre*); *departamento (nombre del Dto.)*; *modalidad* (*anual*, *cuat 1/2* o *cuat 2/2*); *turno* (*M*,*T*,*N*); *ciclo lectivo de cursada (año)*; *tipo de aprobación (firmó, no firmó, promocionó, libre, cambio de curso*, y *cursando)*; *descripción de recursada regular* (*recursó 1 vez, recursó 2 veces, ...3, ...4, ...5, más de 5*); *cantidad de veces de recursada regular* (ídem anterior pero con números: *0, 1, 2, 3, 4, 5, 99*); *descripción de recursada libre* (ídem a recursada regular); *cantidad de veces recursada libre* (*0, 1, 2, 3, 4, 5, 99*).

El contenido de esta tabla corresponde al desempeño académico de estudiantes durante las cursadas de cada materia, así como a la forma y oportunidad de la cursada. Explorando la tabla se observa que no hay una materia que se destaque por haber sido cursada mucho más que otras y, las que se encuentran dentro de las primeras 3 más frecuentes son aquellas de niveles iniciales. Una distribución así es esperable. A su vez, se distingue una cierta predilección por turnos mañana y noche.

Finales Estudiantes Siga

La Tabla 3, muestra parcialmente los datos de los finales. Esta tabla tiene un registro por cada examen final donde se indica: *materia* (*nombre*); *nota* (0...10); *aprobó* (*0, 1*); *promocionó* (0, 1), y por supuesto el *ID* para relacionar con las otras tablas. Se destaca que, de la misma forma que la tabla de cursadas, se registra cada evento de final pudiendo repetirse tanto los registros de estudiantes como de materias. La información está referida al desempeño académico dado únicamente por la nota obtenida y si aprobó por promoción o no.

Tabla 3. Análisis de datos de la BD de cursadas (muestra parcial)

Variable	Característica	Frecuencia	Frecuencia %	Ranking
Materia	Química	5325	6,39	1
Materia	Ingeniería y Sociedad	4935	5,93	2
Materia	Sistemas y Organizaciones	4867	5,84	3
Aprobado	1	66694	80,07	1
Aprobado	0	16597	19,93	2
….	….	….	….	….

De los estadísticos extraídos de la tabla con información de finales, se puede observar que existen errores en los datos como la nota máxima y que existen registros del año 2017 cuando este estudio se hace hasta el 2016.

Una vez terminada la exploración estadística preliminar de los datos en estas tres tablas provistas, se concluye que los errores encontrados son los típicos producidos al realizar la entrada manual de datos. Por ejemplo,

errores numéricos, confusión con datos de un campo en otro, diferentes nombres o abreviaturas para la misma materia, etcétera.

Selección del conjunto inicial de datos

Esta operación es la primera etapa de la metodología KDD. Para este trabajo se decidió excluir directamente a aquellos campos de datos personales que, debido a su muy baja varianza, no aportan mucha información, sumado al hecho de que, además, algunos son datos que no fueron actualizados sino que se completaron en una primera y única vez. Un ejemplo del primer caso es la nacionalidad donde el 99% de los registros pertenecen a la clase *argentina*. Un ejemplo del segundo es el *estado civil*.

Se excluyen entonces: *Nacionalidad*, *Localidad*, *Provincia*, *Estado civil*.

Limpieza de datos y preprocesamiento

Se procedió a una primera etapa de limpieza donde se corrigieron todos aquellos errores que fueran fácilmente identificables y donde no hubiese ambigüedad en la interpretación. Las acciones tomadas fueron:

• La unificación de los nombres de las materias ya que figuran con distintas denominaciones y que a su vez son diferentes de los del plan de estudios.
• La restricción de los datos a los alumnos que tienen cursadas y finales en el periodo analizado (2008 - 2016)
• Corrección de registros con diferentes errores de formato.
• Eliminación de registros con errores groseros que no se pudieron corregir. Se destaca que, al eliminar un registro, se borran todos los que tienen el mismo ID.

Luego de la limpieza de datos, y contando con las tres tablas, se aplica el criterio de clasificación previamente definido y se genera una tabla nueva con un registro por cada *ID* y un nuevo campo con la etiqueta que indica si el estudiante es desertor o no: *desertó* (1, 0).

A esta altura del proceso de KDD pudo decirse que la calidad de los datos luego de enmendar estas particularidades, es aceptable para realizar el estudio. No sucede lo mismo con la cantidad de datos, ya que, según la experiencia, para la cantidad de variables en juego y de los hiperparámetros de los modelos resultan muy escasos.

Se decidió entonces, reducir la gran cantidad de variables en juego agrupando las características de datos personales, cursada y finales en una única tabla general con un registro por estudiante. A la vez, para evaluar si en estos datos existe o no la información necesaria para detectar los casos de deserción con anticipación, se decidió también comenzar con la construcción de dos predictores, Máquinas de Vectores de Soporte (SVM, *Support Vector Machine*) y Redes Neuronales (NN, *Neural Networks*).

Transformación y agrupamiento

Se unieron los datos presentados y analizados en una sola tabla general. Se construyó así una única tabla con un registro por estudiante que contiene: los campos de datos personales; los campos de las características agrupadas de cursadas; los campos de las características agrupadas de finales. Además, hubo que considerar que los modelos seleccionados para esta etapa, SVM y NN (Bishop, 2006) (Tsoukalas & Uhrig, 1997) requieren los datos en forma numérica exclusivamente.

De esta forma se redujo en promedio la cantidad de campos de cada registro, a la vez que se dejó esta cantidad en un número fijo. Por otro lado, con el objetivo doble de generar campos numéricos y de agrupar características fue necesaria la creación de variables auxiliares que acumularan, por ejemplo, la cantidad de veces que en diferentes cursadas se repetía una misma característica. Para el caso del campo *turno* que cada cursada tiene, con tres valores posibles, *M*, *T* y *N*, se generaron tres variables cuantitativas que acumulan la cantidad de cursadas que hizo cada estudiante en los turnos correspondientes. Cada estudiante cuenta ahora en su registro con los campos *TurnoM*, *TurnoT* y *TurnoN*.

De manera similar se procedió con los tipos de aprobación de cursadas, generando variables auxiliares acumuladoras, una por cada alternativa: *tipo de aprobación firmó*, *tipo de aprobación libre*, *tipo de aprobación cambio de curso*, *tipo de aprobación promocionó* y *tipo de aprobación no firmó*. Cada una acumula la cantidad de materias cursadas con ese tipo de aprobación.

Se generó el campo *cantidad de veces recursada regular* cuyo contenido es la suma de la cantidad de veces que recursó materias en general.

El campo original *descripción de cursada regular* se desdobló en un campo por alternativa donde se acumula el número de recursadas correspondiente. Se generaron así: *no recursó*; *recursó 1 vez*; y así para *2*, *3*, *4*, *5*, *nVeces*, donde este último se reservó para cantidad de veces que recursó más de cinco oportunidades.

Para el campo de estudios secundarios se generó la variable *es técnico* con dos valores posibles 1: si viene de un colegio técnico, 0: si no. Se eliminaron los registros de los que no se tenía información.

Se generó la variable cuantitativa *edad al ingreso*, calculada como la diferencia entre el *año de ingreso* a la universidad y el *año de nacimiento*. De manera similar se construyó *cantidad de años* que lleva la cantidad de años desde que se inscribió hasta la última actividad registrada.

Para la tabla de finales se procedió de manera similar. Así, estudiante por estudiante, se tomó la nota más alta de los finales rendidos para una misma materia y se construyó un promedio en el campo *promedio de notas máximas*, excluyendo de esta forma los desaprobados o notas aprobadas de la misma materia si resultan más bajas.

Asimismo, se generó un campo con el promedio sobre todas las notas obtenidas en finales: *promedio de todas las notas*.

Finalmente, se agregaron los campos: *aprobó*, con la cantidad finales aprobados; *no aprobó*; y el *índice de aprobación*, que está calculado como la cantidad de finales aprobados sobre el total de finales rendidos.

En esta instancia del proceso KDD, luego de haber realizado los procesos de: limpieza, preprocesamiento, transformación, agrupamiento más el agregado de un nuevo índice, quedan 3814 registros de 26 campos numéricos cada uno en un único archivo denominado "base de datos estudiantes.csv".

Como preparación para el siguiente proceso, el de DM, se separaron los datos disponibles en los conjuntos de *entrenamiento* y *prueba* (*test*), quedando el de *prueba* sólo reservado para evaluar el modelo. En la Tabla 4 se muestran las cantidades de cada conjunto.

Luego, se revisó la relación de los registros de estudiantes contra el total y se observó que quienes desertaron representaban un 40% aproximadamente, lo que implicaba un pequeño desbalance en el *dataset*. Si bien esto no es en rigor una base de datos desbalanceada, se realizaron algunas pruebas preliminares y se verificó que balanceando los conjuntos mejoraba el resultado de la clasificación en aproximadamente un 2%. Por este motivo se decidió utilizar la técnica de sobremuestreo, o muestreo con repetición, (en inglés '*Oversampling*') (Chawla, 2005) para balancear los datos de entrenamiento y validación. Esta técnica consiste en repetir registros de manera que, al momento de realizar nuevamente la proporción, la cantidad de datos de cada clase (en este caso sólo dos) se encuentren balanceadas. Por tal motivo, sólo en el conjunto de entrenamiento, se duplicaron 589 registros de la clase desertor, elegidos al azar, quedando un total de 3579 registros balanceados como se aprecia en la Tabla 5. Dejando los de prueba en la proporción original.

Tabla 4. Cantidad de muestras de cada conjunto

Conjuntos	Cant. total de muestras	Cant. de desertores	Proporción de desertores
Entrenamiento	2990	1217	40,7 %
Prueba	824	323	39,2 %
Totales	3814	1540	40,3%

Tabla 5. Cantidad de muestras de cada conjunto (entrenamiento aumentado)

Conjuntos	Cant. total de muestras	Cant. de desertores	Proporción de desertores
Entrenamiento	3579	1806	50,4 %
Prueba	824	323	39,2 %

Por último, se procedió a estandarizar los datos del conjunto de entrenamiento. Una vez realizado se almacenaron los valores medios y desvíos estándar para utilizarlos en el tratamiento del conjunto de prueba al momento de necesitarlos. Los datos del conjunto de prueba no formaron parte de la normalización.

Finalmente, se realizó un nuevo análisis exploratorio de los datos mediante el armado de una matriz de correlación, ver Figura 2. La matriz de correlación es una herramienta que muestra cuán relacionadas están entre sí las características de un conjunto de datos (Shetye, 2019). De esta se obtuvieron algunas conclusiones preliminares tales como que la deserción no está relacionada con el *tipo de aprobación* (*firmó*, *libre*, *cambio de curso*, *promocionó*, *no firmó*). También se mostraron algunas relaciones lógicas entre algunas variables. Se pueden mencionar algunos casos: *edad al último final* y *edad*; *cantidad de años* (de cursada) y *turnoN*; *promedio de notas máximas* y *promedio de todas las notas*.

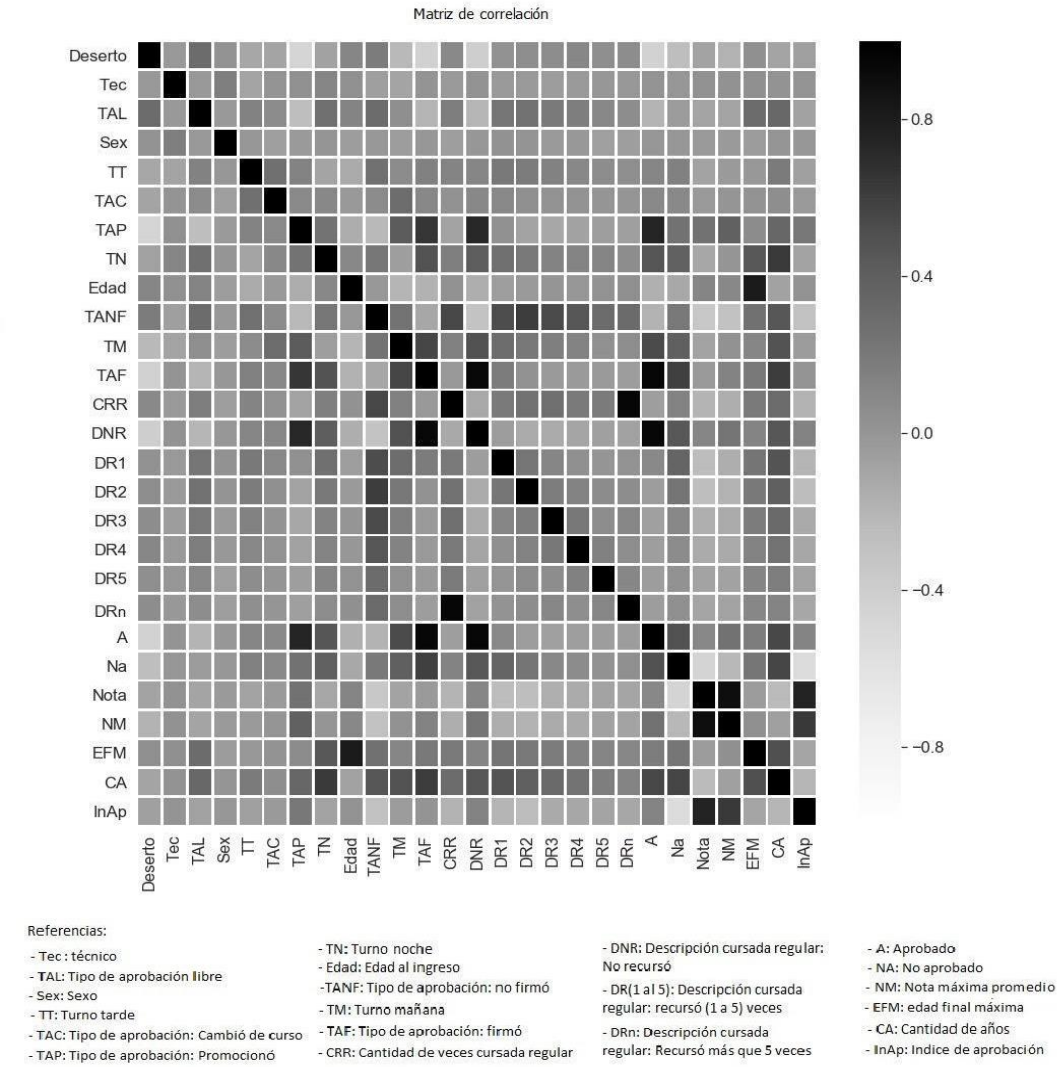

Fig. 2 Matriz de correlación

Minería de Datos

En esta etapa del proceso KDD se procede a la búsqueda de patrones, es el proceso de DM propiamente dicho. En esta instancia se construyeron dos clasificadores, se sintonizaron sus hiperparámetros y se evaluaron con los datos disponibles.

Estamos ante un problema de clasificación binario ya que la salida esperada es un 1 si el alumno es considerado desertor y un 0 si no lo es. Los métodos seleccionados para llevar a cabo esta tarea fueron SVM y NN. Ambos modelos requieren datos de tipo numéricos.

En primer lugar, se utilizó SVM, que es un algoritmo de aprendizaje automático supervisado utilizado principalmente con propósitos de clasificación binaria, aunque posee extensiones para clasificación multiclase y regresión. La idea principal es encontrar un hiperplano capaz de separar ambas clases, pero maximizando la distancia entre la frontera de decisión y los elementos de los datos más cercanos de cada clase. Estos últimos son los vectores de soporte, es decir son los puntos críticos de datos en donde si alguno es eliminado, la posición del hiperplano cambiaría. Dichos puntos son los más difíciles de clasificar y son los únicos relacionados con la función de decisión.

Si las clases no son linealmente separables, los datos originales pueden ser mapeados usando funciones núcleo (*kernel*) a un espacio de dimensionalidad mayor donde las clases sean linealmente separables. Como el modelo de determinación de parámetros es un problema convexo de optimización, cualquier solución local es un óptimo global (Bishop, 2006) (Numerentur.org, 2020)

Durante la construcción del clasificador SVM, se evaluaron varios hiperparámetros y *kernels* utilizando validación cruzada (*cross validation*) con el conjunto de entrenamiento, y *grid search* (Pedregosa, *et al.*, 2011) para encontrar la configuración más adecuada entre ellos. *Grid search* realiza una búsqueda exhaustiva de los parámetros de una lista, y utiliza validación cruzada para evitar el sobreentrenamiento (*overfitting*).

En la Tabla 6 se muestran los valores obtenidos durante el proceso de ajuste. De todos los conjuntos evaluados de *kernels* e hiperparámetros, el mejor fue el *kernel* Gaussiano con C igual a 1000 y gamma con valor 0,001 (Pedregosa, *et al.*, 2011). Donde *C*, es un parámetro de regularización común a todos los *kernels* (Witten, Frank, & Hall, 2011) y gamma define cuánta influencia tiene cada ejemplo de entrenamiento (Witten, *et al.*, 2011).

Tabla 6. Búsqueda de hiperparámetros SVM

Kernel	Hiperparámetros				Exactitud media	
Poly	Degree 2				0,734	(+/-0,036)
Poly	Degree 3				0,781	(+/-0,035)
Rbf(gaussian)	C	1	Gamma	0,001	0,747	(+/-0,027)
		1		0,0001	0,686	(+/-0,034)
		10		0,001	0,771	(+/-0,016)
		10		0,0001	0,744	(+/-0,026)
		100		0,001	0,795	(+/-0,020)
		100		0,0001	0,764	(+/-0,018)
		1000		0,001	0,803	(+/-0,014)
		1000		0,0001	0,786	(+/-0,019)
Linear	C	1			0,786	(+/-0,026)
		10			0,787	(+/-0,022)
		100			0,788	(+/-0,025)
		1000			0,789	(+/-0,020)

En segundo lugar, se utilizaron las NN, que son un modelo de algoritmo de aprendizaje automático supervisado inspirado en las neuronas biológicas y sus funciones. Este tipo de redes son capaces de aprender de la experiencia. Para ello es necesario un conjunto de muestras compuesto por características que sirven de entrada al sistema de clasificación y por sus correspondientes etiquetas que son los valores deseados de la salida (Bishop, 2006)

Las NN también adolecen del fenómeno conocido como sobreaprendizaje (*overfitting*). Este resulta de ajustar excesivamente el modelo a los datos de entrenamiento. De esta forma pierden poder de generalización, esto es clasificar correctamente ejemplos no utilizados durante el proceso de entrenamiento. Para solucionarlo se dividen los datos de entrenamiento en dos partes, uno para el entrenamiento propiamente dicho y otro, de validación para medir su poder de generalización y cortar el proceso de entrenamiento si este tiende a sobreaprender. La proporción empleada para esta subdivisión es de 70/30.

En la Figura 3 se ve la estructura de la red perceptrón multicapa utilizada. Se diseñó una red perceptrón con 26 entradas, una salida y una capa oculta de 16 neuronas. Para la determinación del número de neuronas de la capa oculta se realizaron pruebas con el conjunto de entrenamiento hasta lograr el mejor resultado.

Debido a que se pretende un clasificador binario, en ambas capas (oculta y de salida) se utilizó la función de activación sigmoidea (Ec. 1), y que da valores de salida entre 0 y 1.

$$fact = \frac{1}{(1 + e^{-x})}$$

Se utilizó el optimizador Adam (Keras, 2019) y el error cuadrático medio como función de pérdida (*loss function*). Para la evaluación se utilizó principalmente la exactitud (*accuracy*).

En el entrenamiento se fijaron 500 épocas y luego de varias pruebas empíricas se determinaron los parámetros *paciencia* y *delta mínimo* en 100 y 0,005 respectivamente (Keras, 2019).

Donde:
paciencia es la cantidad de épocas que se espera antes de cortar el entrenamiento cuando la validación no mejora, y
el *delta mínimo* es el mínimo cambio que califica como una mejora.

Estos parámetros se utilizan para activar la detención temprana (*early stopping*), y así evitar el sobreentrenamiento. En este punto se realizaron 10 pruebas para verificar si había una diferencia significativa entre entrenamientos en el rango de épocas en el que se activa la detección temprana y los resultados finales.

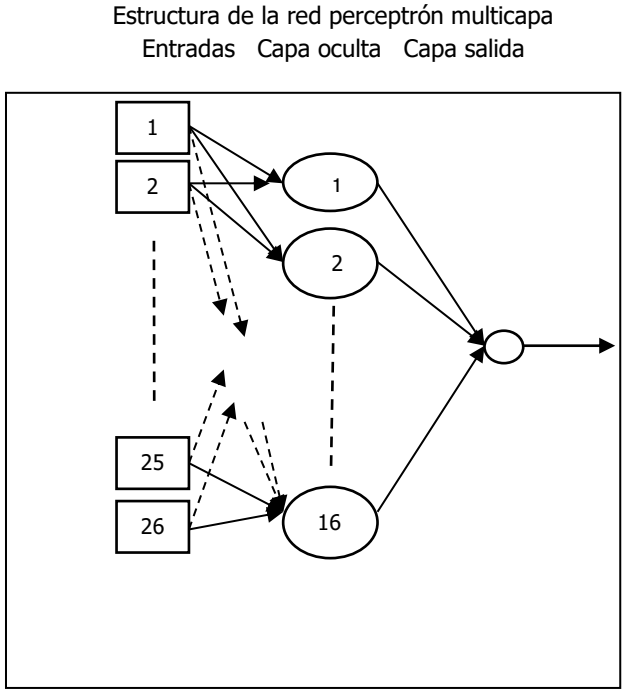

Fig. 3. Perceptrón multicapa

Dichos resultados mostraron que en promedio se necesitan 230 épocas de entrenamiento con test de validación para alcanzar un valor de 0,799 de exactitud. Ver Tabla 7.

Tabla 7. Estadísticas de 10 pruebas con NN

10 pruebas	Épocas	Pérdida	Exactitud	Pérdida para validación	Exact. para Validación
Promedio	230	0,128	0,834	0,140	0,799
Desvío Est.	32,1	0,0043	0,0069	0,0016	0,0061

En la Figura 4 puede apreciarse cómo deja de mejorar la generalización activándose la detección temprana, a pesar de que el error para el conjunto de entrenamiento continúa bajando.

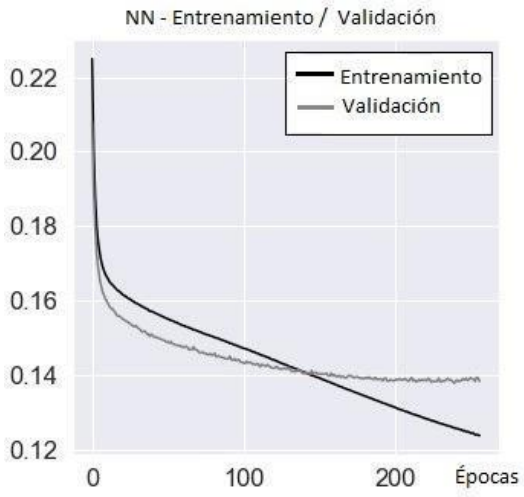

Fig. 4. Error de entrenamiento y validación por época

Resultados y discusión

La evaluación final de los clasificadores construidos se hizo sobre el total de los 824 datos del conjunto de prueba separado inicialmente. Para ello se utilizaron varias métricas: Matriz de confusión, el área bajo la curva ROC, *Precision-Recall*, y *F1-Score*.

Matriz de confusión

La matriz de confusión ofrece una medida de rendimiento y muestra qué tan bien el clasificador categoriza las distintas clases.

En las Tabla 8 y Tabla 9 se muestran los resultados en la matriz de confusión, donde en las filas se muestran los valores verdaderos de los ejemplos y en columnas los indicados por el clasificador. Donde *VN* es verdadero negativo, *FP* falso positivo, *FN* falso negativo, *VP* verdadero positivo, *Pred-* y *Pred+* cantidad de predicciones negativas y positivas, *N* y *P* la cantidad de ejemplos negativos y positivos y *Exac* es la exactitud. Se utilizó el orden de matriz de confusión de *Scikit-learn* (Pedregosa, *et al.*, 2011).

La exactitud obtenida por los clasificadores es muy similar, fue 79,2% para SVM y 78,4% para NN. Para NN se utilizó la mejor red hallada de las 10 mencionadas anteriormente durante el entrenamiento.

Tabla 8. Matriz de confusión para el clasificador SVM

SVM	**Predicciones**		
Observaciones (actual classes)	VN=394	FP=107	N = 501
	FN=64	VP=259	P = 323
	Pred-=458	Pred+ =366	Exac= 79,2%

Tabla 9. Matriz de confusión para el clasificador NN

NN	**Predicciones**		
Observaciones (actual classes)	VN=380	FP=121	N = 501
	FN=57	VP=266	P = 323
	Pred-=437	Pred+ =387	Exac=78,4%

El reporte completo para cada clasificador se muestra en la Tabla 10. Se presentan considerando a cada clase como positiva. Allí se muestran los resultados para las métricas empleadas: *precision*, *recall* y *F1-Score*. (Witten, *et al.*, 2011) (Russell, 2004).

Donde:
precision es la proporción de identificaciones positivas que resultó correcta.
Recall (también *sensibilidad*) es la proporción de positivos reales que se identificó correctamente.
F1-Score es una relación de compromiso (media armónica) entre *precision* y *recall* y genera un único valor que refleja la bondad del clasificador en presencia de clases desbalanceadas (Saito, *et al.*, 2015).

Se observa la similitud entre los valores obtenidos por uno y otro clasificador.

Tabla 10. Resultado de la clasificación sobre el conjunto de test para SVM y NN

	Precision		***Recall***		***F1- score***		**Soporte**
	SVM	**NN**	**SVM**	**NN**	**SVM**	**NN**	
No desertor	0,86	0,87	0,79	0,76	0,82	0,81	501
Desertor	0,71	0,69	0,80	0,82	0,75	0,75	323

ROC - AUC

Se comparan ambos clasificadores utilizando la curva ROC (Witten, *et al.*, 2011) Figura 5. Los resultados indican que son buenos clasificadores y que no hay diferencias significativas entre ambos. El área bajo la curva (AUC) es de 0,86 y 0,85 para SVM y NN respectivamente. El punto donde la especificidad iguala a la sensibilidad es (0,2; 0,8) con un valor de 0,8.

Precision-Recall

Se completa la presentación de los resultados con la curva *Precision-recall* Figura 6, donde se observa nuevamente que los modelos hallados de SVM y NN dan resultados muy similares. Se muestra, por ejemplo, que para un valor de *recall* (sensibilidad) de 0,8 corresponde uno de *precision* de 0,7 aproximadamente, lo que significa que estos clasificadores pueden identificar a un 80 % de entre todos los potenciales desertores, con un nivel de falsa alarma de aproximadamente el 30 %.

El valor de precisión promedio (*Average Precision*, medida parecida al AUC)) es de 0,76. (Pedregosa, *et al.*, 2011)

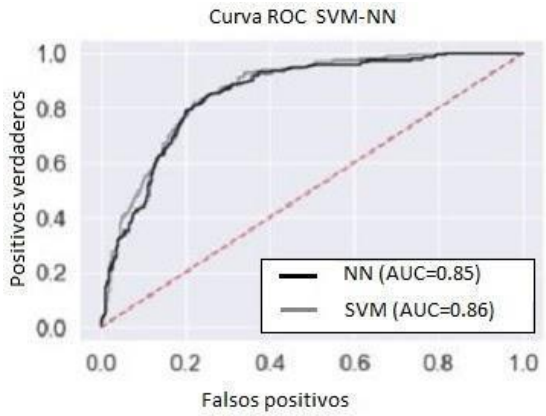

Fig. 5. Curva ROC para SVM y NN

Nótese que en el punto de *recall* mínimo (0,4 aproximadamente) se muestra el valor de la proporción de la clase de representación minoritaria ya que el conjunto de prueba conserva las proporciones originales de desertores/no desertores.

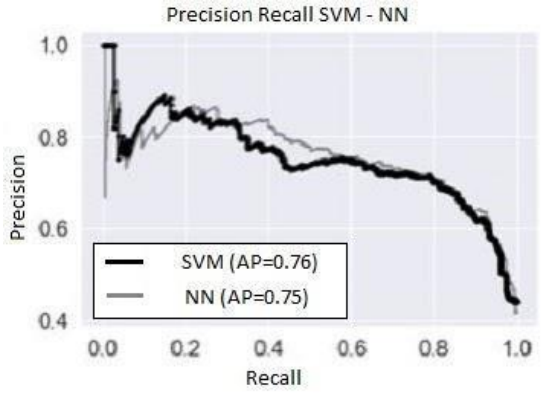

Fig. 6 Curvas Precision-*Recall*

Conclusiones

Los resultados de exactitud obtenidos para ambos clasificadores, 79,2% y 78,4% para SVM y NN respectivamente son considerados buenos. Esto indica que en la base de datos hay información suficiente para predecir el fenómeno de deserción.

El preprocesamiento realizado, agrupando los datos en una única tabla no impidió un buen desempeño de los clasificadores. Asimismo, permite un tratamiento transversal a todas las carreras y a todos los planes.

La herramienta desarrollada mostró ser útil para detectar casos incipientes de deserción y se puede sintonizar para que sea más sensible. En rigor, la sintonía debe realizarse considerando los costos asociados a los errores de los clasificadores, buscando minimizarlos.

Hay que remarcar, sin embargo, que para mejorar la comprensión del fenómeno y por ende ayudar a paliar algunos de sus efectos negativos, es necesario que las bases de datos incluyan otro tipo de información (laboral, demográfica, socioeconómica, etcétera.). Este podría ser objeto de un trabajo multidisciplinario a futuro, en el que también se pueda aumentar la cantidad de registros de la base de datos. Además, como producto de ese mismo trabajo a futuro, se deben establecer los costos asociados para optimizar su desempeño

Cabe mencionar que para realizar la clasificación de estudiantes como desertores o no desertores, fue menester generar una definición medible del fenómeno, dado que no existe una en la UTN FRBA.

Agradecimientos

Se agradece muy especialmente a la Ingeniera Vanina de los Ángeles Bottini y a la Licenciada Patricia Cibeira así como al personal de sus respectivas Secretarías por la gestión para la obtención de los datos necesarios para la realización de este trabajo.

Referencias

BARRERA REBELLÓN, M., (2008). *Análisis de Supervivencia Aplicado al Problema de la Deserción Estudiantil en la Universidad Tecnológica de Pereira.* Universidad Tecnológica de Pereira - Colombia.

BISHOP, C. M., (2006). *Pattern Recognition and Machine Learning.* Springer.

CHAWLA, N. V., (2005). *Data Mining for Imbalanced Datasets: An Overview.* IN; USA: Department of Computer Science and Engineering; University of Notre Dame.

FAYYAD, U. ; PIATETSKY-SHAPIRO, G. & SMYTH, P., (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framwork. *KDD-96 AAAI.*

FERNÁNDEZ-HILEMAN, M.; CORENGIA, A. & DURAND, J. C., (2014). Deserción y retención universitaria: una discusión bibliográfica. (W. T. Agudelo, Ed.) *Pensando Psicología, 10*(17), 85-96.
INET, (2016). *Demanda de capacidades 2020. Análisis de la demanda de capacidades laborales en la Argentina.* CABA: Instituto Nacional de Educación Tecnológica. Ministerio de educación.
KERAS, (19 de 12 de 2019). *Keras API reference / Callbacks API / EarlyStopping*. Obtenido de Keras API reference / Callbacks API / EarlyStopping
KERAS, (20 de 12 de 2019). *Keras API reference/Optimizers/ADAM*. Obtenido de https://keras.io/api/optimizers/adam/
KUNA, H.; GARCÍA MARTÍNEZ, R. & VILLATORO, F., (2010). Identificación de causales de abandono de estudios universitarios. Uso de procesos de explotación de la información. *Revista iberoamericana de tecnología en educación y educación en tecnología, 5*, 39-44.
Numerentur.org. (08 de 01 de 2020). *Numerentur.org Máquina de vectores DE SOPORTE*. OBTENIDO DE HTTP://NUMERENTUR.ORG/SVM/
PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B.; GRISEL, O. & BLONDEL, M., (20 de 12 de 2011). Scikit-learn: Machine Learning in Python.org - Grid search. *Journal of Machine Learning Research*, 2825-2830. Obtenido de https://scikit-learn.org/stable/modules/grid_search.html
RUSSELL, N., (2004). *Inteligencia Artificial, un enfoque moderno.* Pearson.
SAITO, T. y REHMSMEIER, M., (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE 10(3): e0118432*, 1/21.
SHETYE, A., (28 de 12 de 2019). *Towards data science*. Obtenido de https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b
SPU, (2018). *Áreas de vacancia, vinculación, pertinencia y planificación del sistema universitario: una herramienta para abordar la expansión de la educación superior en territorio.* CABA: Secretaría de Políticas Universitarias -Ministerio de Educación – Presidencia de la Nación, Argentina.
TSOUKALAS, L. H., & UHRIG, R., (1997). *Fuzzy and Neural Aproaches in Engineering.* John Wiley & Sons Inc.
VIALE TUDELA, H., (2014). Una aproximación teórica a la deserción estudiantil universitaria. *Revista digital de investigación en docencia universitaria, 8*(1), 59-75.
WITTEN, I. H., FRANK, E., & HALL, M. A., (2011). *Data Mining. Practical Machine Learning Tools and Techniques.* Amsterdam: Elsevier.