Nodes: Plataforma para la predicción de deserción escolar utilizando técnicas de inteligencia artificial

D. González Díaz[1], J. I. Picie Alcaraz[2], M. D. González Martínez [3], C. H. Hernández Jacome [4]*, E. Onofre Ruiz[5]
[1]Universidad Tecnológica del Centro de Veracruz,
Av. Universidad, No. 350, Dos caminos, 94910, Cuitláhuac, Veracruz, México.
daniel.gonzalez@utcv.edu.mx
***Área de participación**: Sistemas Computacionales*

Resumen

Uno de los principales problemas en los que se encuentra el sistema educativo en México es la deserción escolar, en el estado de Veracruz existe un índice del 14% de cinco millones de estudiantes que se matriculan en el nivel medio superior. Ante este fenómeno que incrementa año con año, la Secretaría de Educación Pública en el año 2018 se propuso disminuir dicho porcentaje, sin embargo, no se cuenta con alguna herramienta de apoyo para lograr ese objetivo. En este contexto, el presente proyecto hace énfasis en el estudio y análisis de la deserción escolar en la zona centro del Estado de Veracruz utilizando los servicios que proporciona Azure Machine Learning Studio para pruebas iniciales del comportamiento de los datos y elección del algoritmo para la implementación de una plataforma tecnológica que permita predecir el porcentaje de deserción que posee un estudiante al ser matriculado en el nivel medio superior.

***Palabras clave**: machine learning, deserción, media superior, árbol de decisión.*

Abstract

One of the main problems in the educational system in Mexico is the School Drop-out. In the State of Veracruz, there are 14% of five million students who are enrolled in High School, and they do not end up in their studies. This phenomenon is increasing every year. In 2018, the Ministry Public Education proposed to reduce this percentage, however, there were not sufficient tools to achieve the goal. In this context, this project makes emphasis on the study and the analysis of school drop-outs in the Central Zones of the State of Veracruz by using the services provided by Azure Machine Learning Studio for behaviors initial tests in order to implement a technological platform that helps to predict the percentage of school drop-outs of students enrolled in High School.

***Key words**: machine learning, school drop-out, high school, decision tree.*

Introducción

La Inteligencia Artificial permite aprender, tomar decisiones y formarse una idea determinada de la realidad, actualmente la inteligencia artificial es un campo extraordinariamente amplio y multidisciplinario, pero sobre todo complejo, pues supone un severo esfuerzo por entender la complejidad de los diversos comportamientos en los seres vivos en términos de procesos de información, mediante un extenso conjunto de técnicas, métodos y algoritmos que con base en investigaciones abren un mundo de posibilidades para poder implementarla en distintas áreas; una de las técnicas más empleadas de la inteligencia artificial es el aprendizaje automático (Machine Learning), pues se encarga de que las computadoras realicen acciones sin la necesidad de ser programadas explícitamente, es decir, a través de proporcionar datos a los algoritmos correspondientes, se es posible guiar decisiones u obtener predicciones; sin duda, la inteligencia artificial es un área en la que hay muchos hechos pero hay mucho más por hacer (Alfonso Galipienso, Cazorla Quevedo, Colomina Pardo, Escolano Ruiz, & Lozano Ortega, 2003).

El INEE (Instituto Nacional para la Evaluación de la Educación) en México, reporta que tan solo en el ciclo 2014-2015, cerca de 700.000 alumnos dejaron las escuelas. En primaria supone apenas un 0,6% y en secundaria representa un 4,4%, según los datos federales. Sin embargo, la tasa de deserción escolar representa un 14,4% de un total de cinco millones de estudiantes que se matriculan cada año en el nivel media superior en México. Este es el porcentaje más alto de abandono respecto a los otros grados de estudio, éste no es un acontecimiento espontáneo, es más bien el resultado de un proceso complejo que en muchas ocasiones tiene antecedentes en

etapas tempranas de la trayectoria educativa y deriva en una decisión (no estrictamente voluntaria) influenciada por preferencias, expectativas y restricciones que enfrentan los jóvenes, como la falta de apoyos familiares, escolares y comunitarios. En dicho proceso se entretejen factores de índole individual, social, económica y cultural que se refuerzan simultáneamente y se agravan con el tiempo.
Veracruz es uno de los estados del país de mayor inasistencia entre los alumnos de entre 15 y 17 años, donde las autoridades realizan un esfuerzo para solventar este fenómeno asignando metas por parte de la subsecretaría de EMS (Educación Media Superior); una de las cuales es reducir a 9% la tasa de deserción escolar en el bachillerato en el 2018. El resultado esperado de la ejecución de este proyecto es que las Instituciones Educativas de nivel Medio Superior (IEMS) puedan acceder a la plataforma tecnológica mediante una capa de servicios, que puede ser utilizada e integrada a las aplicaciones de control escolar utilizadas por las Instituciones Educativas de nivel Medio Superior (IEMS), permitiendo interactuar con los modelos de predicción y determinar si un candidato es propenso a deserción escolar a fin de brindar el apoyo necesario a tiempo.

Metodología

La implementación de la metodología de trabajo SCRUM en el proyecto NODES fue utilizada para la gestión y desarrollo del proyecto, controlando las iteraciones y flujos de trabajo en cuanto a tiempo y entregables (Rubin, 2015). Incluye junto con la descripción de este ciclo de vida iterativo e incremental para el proyecto, los artefactos o documentos con los que se gestionan las tareas de adquisición y suministro: requisitos, monitorización y seguimiento del avance, así como las responsabilidades y compromisos de los participantes en el proyecto.
El propósito del uso de SCRUM es la distribución de trabajo y control de entregables e iteraciones en el desarrollo del sistema ya que el desarrollo ágil es más útil para el desarrollo de productos personalizados (A. Srivastava, 2017). El alcance de SCRUM para NODES es la descripción del proceso de desarrollo del sistema en cuanto al tiempo establecido, además la generación de entregables que contribuyen a la integración y funcionalidad del sistema respecto al análisis de historias de usuario denotadas en la pila de artefactos.

Incepción

Se ha identificado la necesidad de contar con una herramienta tecnológica que implemente técnicas de inteligencia artificial para identificar estudiantes con mayor riesgo de abandonar sus estudios, a fin de brindarles el apoyo necesario a tiempo y coadyuvar en la meta de reducción del porcentaje de deserción escolar. Tras evaluar las diferentes variables que envuelven a los estudiantes desertores que los identifican entre sí, algunas de estas variables son: el género, el código postal, la edad, si cuenta con una capacidad diferente, si pertenece a un grupo étnico, nivel de estudios del padre/tutor y número de personas que viven con él. En la imagen 1, se observa la distribución de los datos en los cuales el color rojo representa quienes desertaron de la escuela en contra los que concluyeron sus estudios en color azul, se observa que las dos clases están perfectamente definidas.

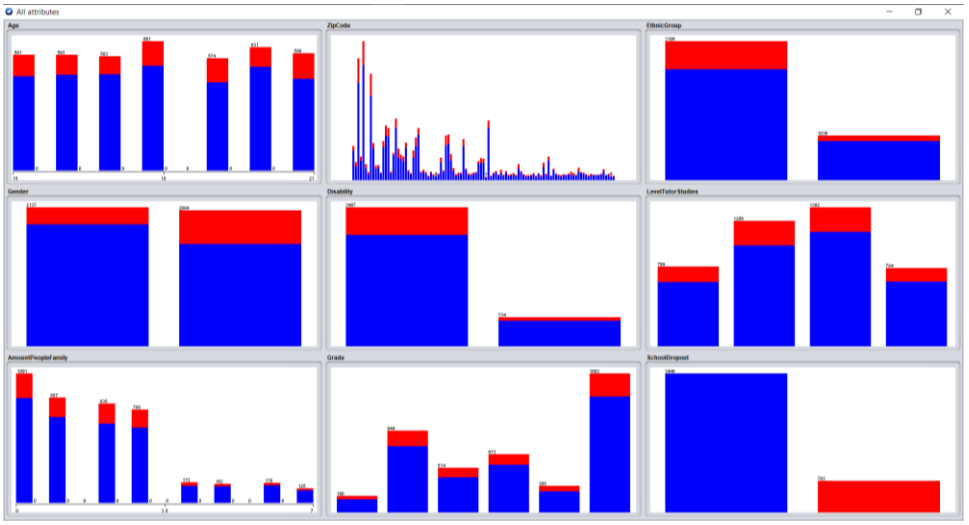

Imagen 1 Distribución de datos en WEKA
Fuente: WEKA 3.8

Elaboración

Con respecto a la selección del modelo de predicción de inteligencia artificial "*Two Class Boosted Decision Tree"* (Árbol de decisión impulsado de dos clases) es un algoritmo de clasificación que se asemeja a la estructura de un árbol, donde tiene comienzo en una raíz y se despliega en ramas, también llamados nodos, que se interrelacionan formando patrones (Microsoft, 2019). Este algoritmo solo tiene dos resultados por ello es llamado de dos clases, la respuesta esperada en un SI o NO tomando como media el valor de 0.55 entre 0 y 1. La elección de este algoritmo es porque muestra una precisión aceptable, necesaria para el funcionamiento que realiza NODES, tiempos de entrenamiento moderados y el uso de linealidad permitiendo, a través del estudio de diversas condiciones de las variables, la optimización de la función objetivo. Actualmente existe una variedad de algoritmos de clasificación que emergen cada año, la mayoría de ellos descuidan el hecho de que el rendimiento de los clasificadores inducidos depende en gran medida de los datos como indica (Rodrigo C. Barros, Márcio P. Basgalupp, & André CPLF, 2012), los datos con los que cuenta NODES son cuantitativos permitiendo la clasificación en diferentes patrones.

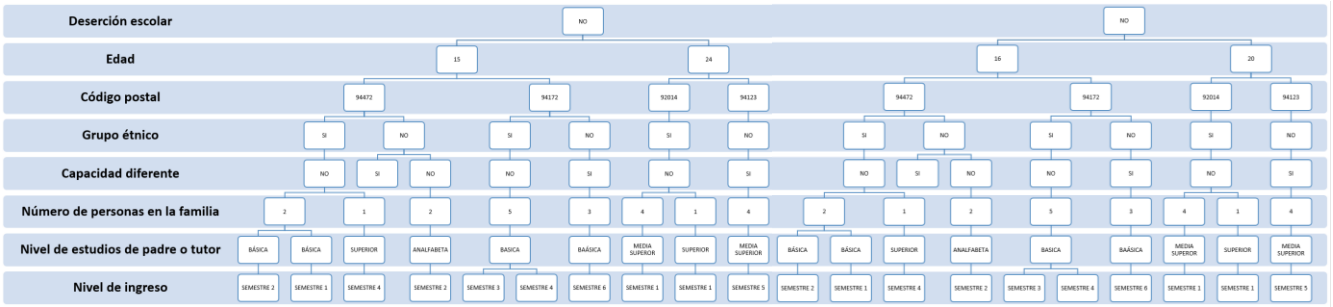

Imagen 2 Diagrama de árbol de decisión de dos clases de NODES
Fuente: Diagrama realizado para NODES

La imagen 2, se interpreta que el **NIVEL 1** está el campo de "deserción escolar" y en ella está el resultado del alumno, es por ello que algoritmo "*Two Class Boosted Decision Tree"* (Árbol de decisión impulsado de dos clases) al ser los únicos posibles valores SI y NO, el valor que el modelo debe predecir. En el **NIVEL 2** se encuentra la edad, donde es una variante numérica entre 15 y los 24 años, esta variable nos permite prescindir si el alumno se encuentra en rezago educativo, este sucede cuando el alumno repite algún año escolar o toma años sabáticos. El **NIVEL 3** es el código postal que identifica la zona donde vive el estudiante, de esta manera agrupamos a los estudiantes según su posición geográfica. El **NIVEL 4** corresponde si el estudiante pertenece a un grupo étnico, actualmente, en los municipios indígenas, el número de alumnos en Educación Media Superior (EMS) representa la menor proporción (12.4%) considerando, además, que el número de planteles ubicados en esos territorios es bajo cuenta con una capacidad diferente (Voces del compromiso, 2017). La capacidad diferente se encuentra en el **Nivel 5**, se evalúa si el alumno tiene una capacidad diferente. El **NIVEL 6** se considera al número de personas que comparten la dependencia económica del padre de familia o tutor, la intención es saber a gran escala la disposición económica en conjunto con el **NIVEL 7** que es el nivel de estudios del tutor considerando si es analfabeta, básica, media superior o superior.
(Rouse, 2017) Señala que, el aprendizaje automático permite que las computadoras aprendan de datos y experiencias y actúen sin ser programadas explícitamente. Desarrolle aplicaciones de Inteligencia Artificial (**IA**) que detectan, procesan y actúan sobre la información, aumentando las capacidades humanas, incrementando la velocidad y la eficiencia, y ayudándole a lograr más. El uso del *Servicio de Azure Machine Learning* ofrece diferentes servicios para experimentar e implementar diferentes algoritmos que abarca al área de Inteligencia Artificial, eso enfocado al estudio de los datos especialmente, debido a que Nodes busca brindar opciones que, mediante métodos estadísticos y probabilísticos.

Construcción

La arquitectura de NODES funciona con base en la tecnología de la Nube. En esta desprenden diferentes componentes tales como base de datos, servicios como API (Application Programming Interface, Interfaz de programación de aplicaciones) y servicios externos mediante la conexión a internet. En la imagen 3 se describe la arquitectura dentro de Nodes la cual se compone de:

- **Base de datos**: se almacena la información de los estudiantes, es llamada también base de datos experta, ya que el algoritmo se alimenta de ella para calcular las probabilidades de cada patrón que detecta y, de esta forma, estimar en que patrón se colocan los nuevos datos agregados.
- **Algoritmo de Inteligencia Artificial:** existen diferentes programas que apoyan en la utilización de algoritmos sin necesidad de programarlos de nuevo. Estas herramientas simplifican el trabajo de desarrollo, por ejemplo, los utilizados en este proyecto como Azure Machine Learning Studio y WEKA (*Waikato Environment for Knowledge Analysis*, Entorno para Análisis del Conocimiento de la Universidad de Waikato).
- **Componentes de integración de servicios externos**: Para complementar la información e identificar al individuo en su macro y micro entorno utilizando dos servicios:
    - **Banco de indicadores de INEGI (Instituto Nacional de Estadística y Geografía)**: en México se realiza un censo a la población en la cual el cumulo de información realizan estudios de probabilidad y estadística la cual, según los indicadores por zona geográfica, edad, sexo, entre otros nos dan una proyección del estudiante.
    - **SIID (Servicios de información para interoperabilidad de Datos) Coneval**: El Consejo Nacional de Evaluación de la Política de Desarrollo Social mide la pobreza en México y evalúa programas y políticas sociales del Gobierno Federal. De esta manera, se sabe el acceso a la educación, salud y servicios de primera necesidad en su comunidad.
- **Capa de servicios API REST (*Representational State Transfer*, Transferencia de Estado Representacional)**: es una interfaz entre sistemas que usa HTTP (*Hypertext Transfer Protocol*, Protocolo de transferencia de hipertexto) para obtener datos o generar operaciones sobre esos datos en el formato JSON (*JavaScript Object Notation*, notación de objeto de JavaScript). De esta manera, es posible utilizar este servicio en diferentes plataformas.

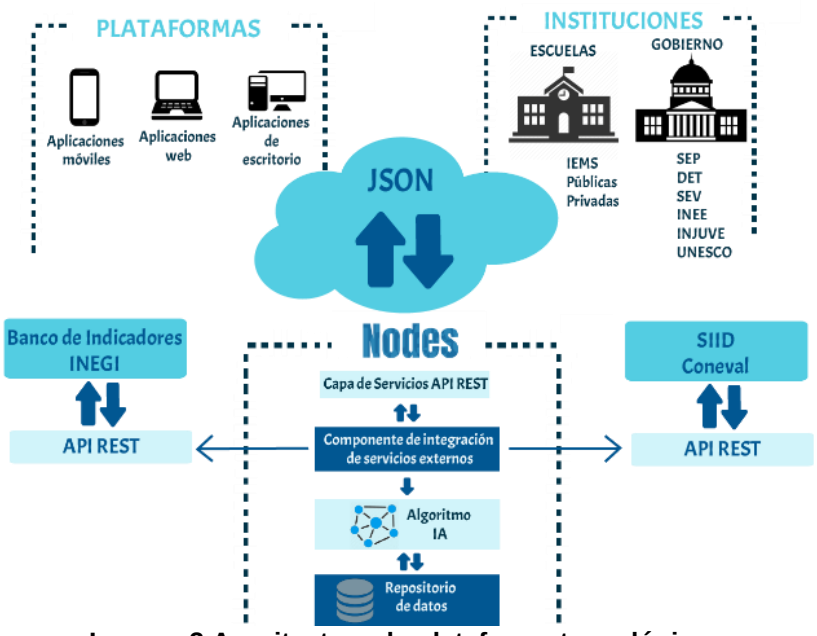

Imagen 3 Arquitectura de plataforma tecnológica.
Fuente: Arquitectura generada para Nodes

A continuación en la imagen 4, se muestra como cada cliente (las Instituciones de Educación Media Superior) consultan la información del estudiante para obtener el porcentaje y el dictamen, si es que el estudiante está en riesgo de deserción escolar, las consultas podrán de manera individual o cargando un archivo CSV (*comma-separated values*, Valores separados por comas) para consultar un número mayor de alumnos.

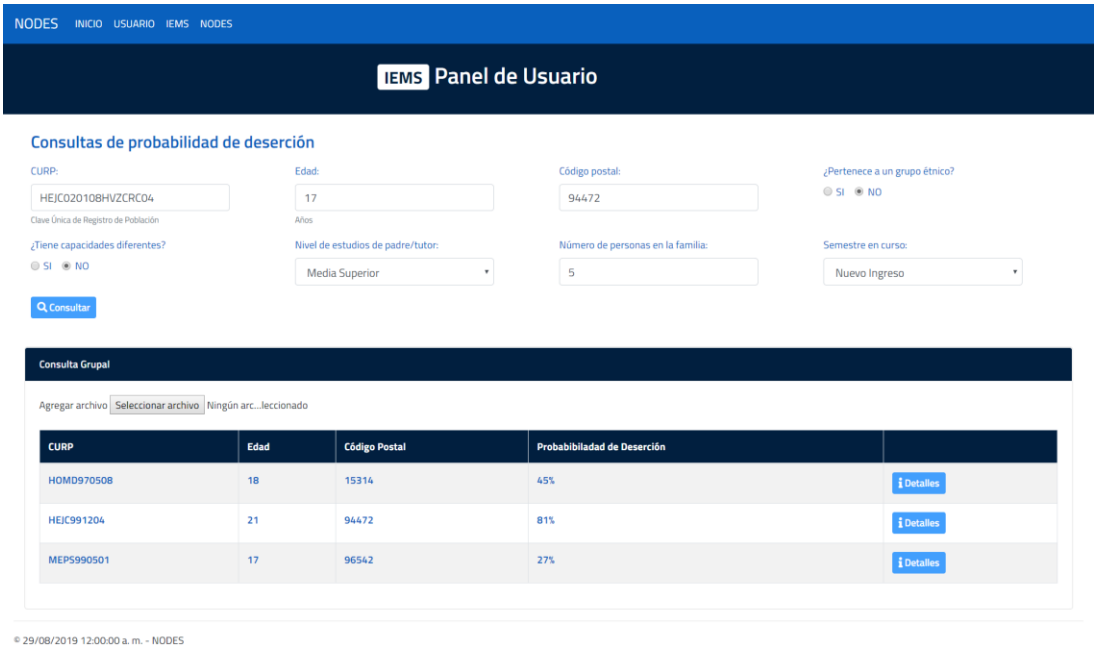

Imagen 4 Vista para el cliente para realizar las predicciones a Nodes
Fuente: Test generado para Nodes

Resultados y discusión

Resultados del modelo

El árbol generado del algoritmo se visualiza a continuación en la Imagen 5. En este se detalla el uso de *Two Class Boosted Decision Tree* (Árbol de decisión impulsado de dos clases) en el comienzo del diagrama, del mismo modo nuestra base de datos (Data Set) se encuentra en la parte superior del diagrama con el nombre "nodes_csv.csv", a continuación se realiza una distribución de Pareto (Split Data) y evaluación del modelo probabilístico generado (Train Model).

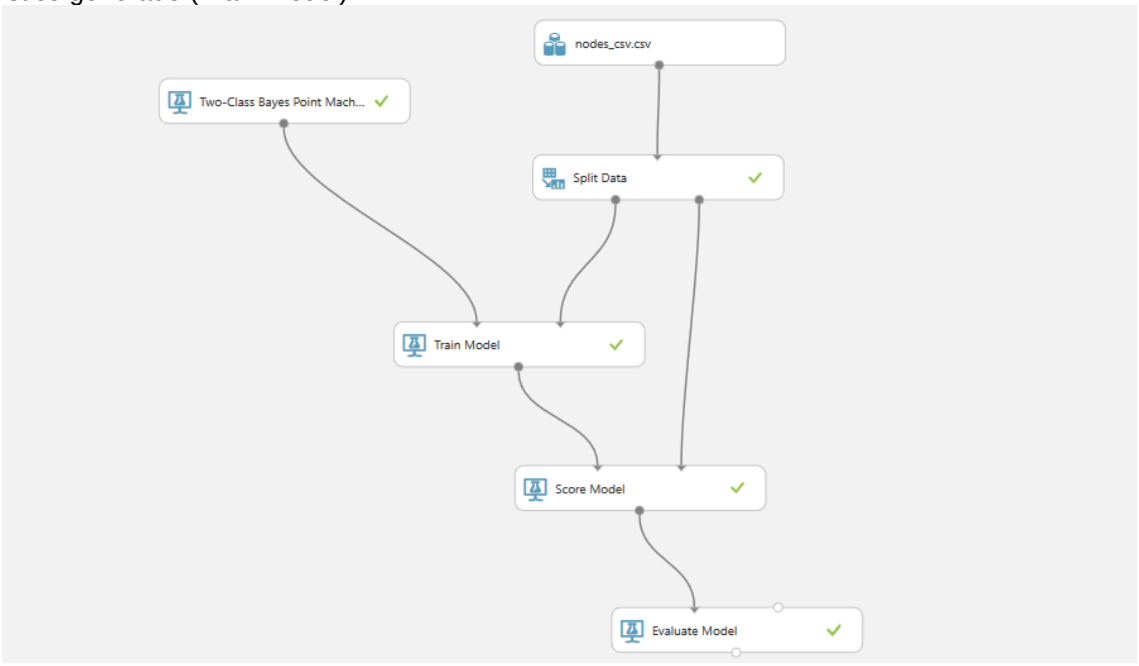

Imagen 5 Modelo probabilístico con base de datos, algoritmo árbol binario y evaluación de modelo
Fuente: Azure Machine Learning Studio

Implementación y administración

La implementación del modelo para producir predicciones se efectuará en la nube, y usar los modelos acelerados por hardware en los FPGA (Field-Programmable Gate Array, matriz de puertas programables), para obtener inferencias de pronta respuesta. Cuando el modelo está en producción, es necesario monitorear y determinar el rendimiento y la derivación de los datos, y volver a entrenar el modelo según sea necesario.

En la Imagen 6 se encuentra la arquitectura que se genera en *Azure Machine Learning* y en la cual consumimos desde servicio de NODES, de esta manera se divide el trabajo en partes, es decir infraestructura en la nube.

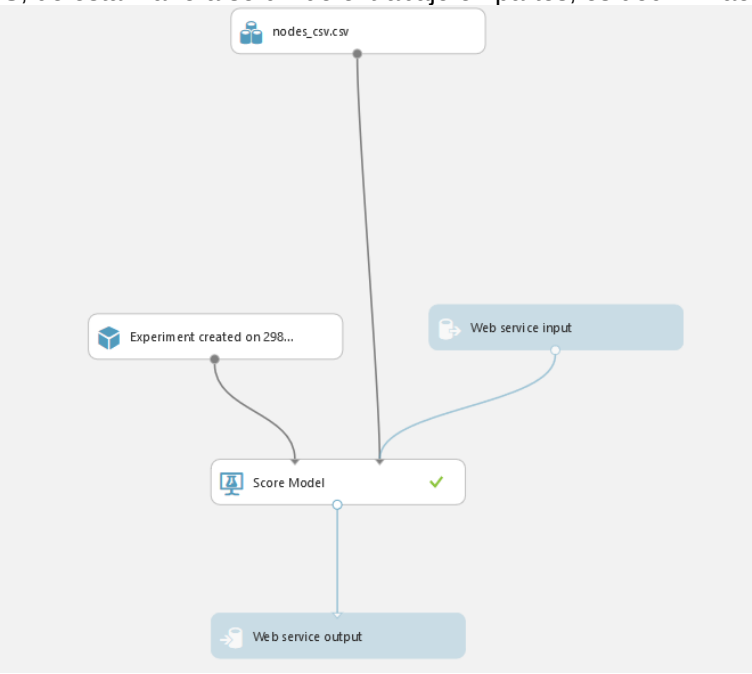

Imagen 6 Diagrama de interacción Web Service y modelo probabilístico

Fuente: Azure Machine Learning Studio

Descripción de resultados

La información obtenida marca una tendencia parabólica, en la siguiente gráfica (Imagen 7), se observa que tanto verdaderos positivos y falsos negativos van en tendencia proporcional, lo cual es bueno para el algoritmo.

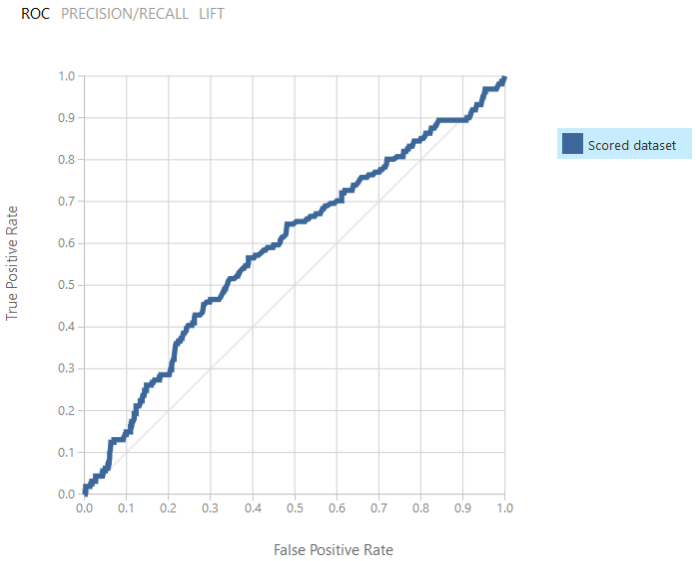

Imagen 7 Gráfica de tendencia de los datos respecto a verdaderos positivos y falsos negativos

Fuente: Azure Machine Learning Studio

En la siguiente tabla (Imagen 8) se aprecia la evaluación de resultado del algoritmo, en la primer columna se encuentran los valores ciertos, estos valores son los que predijo correctamente, es decir que aserto, en la siguiente columna se encuentran aquellos que predijo de manera incorrecta, es decir a los que no atino. A pesar de ser un gran número de valores verdaderos positivos y negativos el algoritmo cuenta con una presicion muy baja de 0.259.

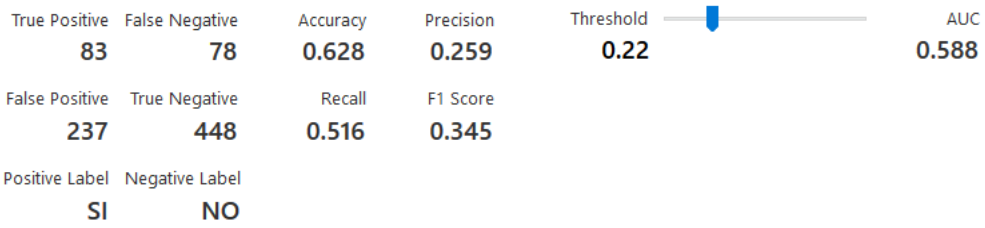

Imagen 8 Reporte de evaluación del modelo
Fuente: Azure Machine Learning Studio

Trabajo a futuro

A partir de las pruebas se identificaron áreas de oportunidad para el proyecto NODES el cual tiene ventajas de generación de ideas de una forma confiable y con resultados factibles al pertenecer al área de calidad educativa, atacando el fenómeno de deserción escolar que afecta a la sociedad en México y en el mundo. Es por ello por lo que el proyecto se puede adaptar en diferentes escenarios según su ámbito social, político y económico con la información necesaria gracias a la Inteligencia Artificial.
México ha tenido un magro desarrollo en materia de Inteligencia Artificial, por lo que tendría que implementar una estrategia híbrida a la brevedad (Arreola, 2018), por lo que es atractivo y funcional la dirección dada por Nodes, ofreciendo una ventaja competitiva al resolver las necesidades de la Instituciones de Educación Media Superior (IEMS), agregar otras variables para evaluar la condición de los estudiantes, crear modelos de mercado con base en el negocio de consultoría y ofrecer soluciones según el caso del estudiante (por ejemplo; ofrecer becas al cual postularse). Estas mejoras contribuyen en un marco de continuidad al aumento de la eficiencia y de la satisfacción del usuario y/o cliente del producto. La educación es una etapa elemental en la sociedad, es por ello que NODES se enfocó en un principio a la EMS (Educación Media Superior) por su gran porcentaje de deserción en México; la posibilidad de emerger en otros niveles educativos es factible, principalmente en el nivel universitario que se encuentra en segundo lugar de porcentaje de deserción escolar.

Conclusiones

Al concluir el desarrollo del proyecto, se realizó un análisis de los resultados expuestos por el algoritmo donde pueden surgir diferentes opciones para mejorar el proyecto realizado. Asimismo, se propone una línea de investigación de patrones detectados en el banco de información y realizar una investigación a fondo sobre el fenómeno de deserción por generación.
El desarrollo de esta plataforma para consulta de la información, ayuda a las EMS a conocer que probabilidad de deserción que tiene cada estudiante al ingresar al nuevo nivel de estudio, esto con la finalidad crear un plan de acción de cada uno de ellos y dar un seguimiento a cada uno de ellos, además aporta valor a la sociedad y ofrece una solución al creciente porcentaje de abandono escolar que ocurre en la EMS, problema que no solo ocurre en el nivel media superior sino en todos los niveles sufren de este fenómeno y que está acrecentando y amenaza a los estudiantes.

Agradecimientos

Estas líneas son dedicadas a familiares y amigos que de manera indirecta apoyan a cada uno de los autores de este proyecto, Carlos Heriberto Hernández Jacome quien colaboró con el desarrollo e implementación del modelo de inteligencia artificial, Josué Iván Picie Alcaraz encargado del desarrollo de la plataforma tecnológica. De igual forma al I.S.C. Daniel González Díaz quien brindo su conocimiento, asesoramiento y dirección del proyecto y al M.E.I. Eric Onofre Ruiz quién aporto sus conocimientos y brindó entrenamiento en el idioma inglés.

Agradecimiento al Centro de Investigación de Inteligencia Artificial por la iniciación con el Programa de Fomento a las Vocaciones en el Área de Inteligencia sobre esta rama de investigación.
En memoria de Dolores González Acuca, padre y apoyo de María Dolores González Martínez quien desarrolló la imagen y diseño del proyecto. Descanse en paz.

Referencias

A. Srivastava, S. B. (2017). SCRUM model for agile methodology,. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 864-869.

Alfonso Galipienso, M. I., Cazorla Quevedo, M. Á., Colomina Pardo, O., Escolano Ruiz, F., & Lozano Ortega, M. Á. (2003). *Inteligencia Artificial.* Spain: Thomson Editores.

Arreola, J. (20 de 07 de 2018). *www.forbes.com.mx.* Obtenido de https://www.forbes.com.mx/mexico-puede-triunfar-en-inteligencia-artificial/

Baza, F. G. (2005). *Deserción escolar y desigualdd económica en México: Un análisis empírico para los niveles medio superior y superior.* D.F., Mexico. Obtenido de https://tesis.ipn.mx/bitstream/handle/123456789/752/1121_2005_ESE_DOCTORADO_garcia_baza_florberto.pdf?sequence=1&isAllowed=y

Gladys, M. S. (2013). *Revista de Actualización Clínica Investiga.* Obtenido de Estudios de Correlacion: http://www.revistasbolivianas.org.bo/scielo.php?pid=S2304-37682013000600006&script=sci_arttext

Hernández Sampieri, R. (2000). Diseños Exploravitas.

Hernández, E. A.-E. (1995). *lifeder*.

Instituto Nacional de Estadística y Geografía (INEGI). (2010). *Principales resultados del Censo de Población y Vivienda 2010.* Ciudad de México: Gobierno Federal.

Instituto Nacional para la Evaluación de la Educación. (2018). *LA EDUCACIÓN OBLIGATORIA EN MEXICO: Informe 2018.* INEE, Ciudad de México. Obtenido de https://www.inee.edu.mx/portalweb/informe2018/04_informe/capitulo_020204.html

Microsoft. (16 de septiembre de 2018). *https://docs.microsoft.com*. Obtenido de https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/ef/overview

Microsoft. (2019). *Docs.microsoft.com*. Recuperado el 29 de septiembre de 2019, de https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/boosted-decision-tree-regression

Microsoft. (7 de marzo de 2019). *https://code.visualstudio.com*. Obtenido de https://code.visualstudio.com/docs/editor/whyvscode

Rodrigo C. Barros, Márcio P. Basgalupp, & André CPLF. (2012). Clus-DTI: improving decision-tree classification with a clustering-based decision-tree induction algorithm. *Journal of the Brazilian Computer Society*, 351–362.

Rouse, M. (2017). *Inteligencia Artificial* . Obtenido de https://searchdatacenter.techtarget.com/es/definicion/Inteligencia-artificial-o-AI

Rubin, K. S. (2015). *Essential Scrum.* Arbon, Michigan: Addison Wesley.

StarUML. (2017). *http://staruml.sourceforge.net/.* Obtenido de http://staruml.sourceforge.net/docs/developer-guide(en)/ch01.html

Universidad de Wakaito. (22 de enero de 2019). *www.cs.waikato.ac.nz*. Obtenido de https://www.cs.waikato.ac.nz/ml/weka/

Voces del compromiso. (14 de 08 de 2017). *compromisoporlaeducacion.mx.* Obtenido de http://compromisoporlaeducacion.mx/la-situacion-educativa-actual-de-la-poblacion-indigena-en-mexico/