

Modelo Predictivo de Aprobación Temprana en Espacios de Educación Superior

Gustavo Illescas¹, Maylen Dell' Oso², Florencia Paglione³, María Rosa Dos Reis¹,
José Arturo Mora Soto⁴

¹Instituto de Investigación en Tecnología Informática Avanzada, Facultad de Ciencias Exactas. Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA). Centro Asociado CIC. ²Becario EVC-CIN, Facultad de Ciencias Exactas, UNCPBA. ³Facultad de Ciencias Exactas UNCPBA. ⁴Universidad de Celaya, México.

*illescas@exa.unicen.edu.ar, {mpaglione, mdelloso}@alumnos.exa.unicen.edu.ar,
rosa.dos.reis@econ.unicen.edu.ar, jmora@udec.edu.mx*

RESUMEN

El presente trabajo trata sobre la creación de un clasificador de datos para estimar el rendimiento que tendrá un alumno en una asignatura puntual en función a las materias correlativas. Como datos de entrada se obtiene la información de los alumnos del Sistema de Información Universitaria (SIU Guarani) de la Facultad de Cs. Exactas de la UNICEN, en particular de la carrera de Ing. de Sistemas. Se utilizan un conjunto de técnicas y herramientas que permiten analizar estos datos con el fin de ayudar tanto a los alumnos de la facultad como a los miembros de los equipos de gestión, de cátedra y quienes tienen la oportunidad de utilizar los resultados en los procesos de tomas de decisiones con el fin de generar mejoras en la institución. El trabajo se enmarca en el proyecto de investigación "Gestión Informática del Conocimiento como soporte para la toma de decisiones Organizacionales" (03/C314) actualmente en desarrollo.

Palabras clave: *Predicción, Educación Superior, Analítica de datos.*

CONTEXTO

El trabajo se encuentra enmarcado dentro del proyecto de incentivos 03/C314 específicamente dentro de la línea "Analítica de Datos para la gestión del conocimiento orientado a la toma de decisiones en espacios de Educación Superior" desarrollado en el Instituto de Tecnología Informática Avanzada (INTIA) de la Facultad de Ciencias Exactas

(EXA), UNCPBA. Es importante señalar que el tema que aborda este trabajo fue presentado y finalmente aprobado bajo el marco de la beca de incentivo a la investigación EVC-CIN 2023, a manos de la becaria Maylen Dell'Oso, (director Dr. Gustavo Illescas y co-directora Mg. María Rosa Dos Reis). La presentación obtuvo un puntaje de 74,94 sobre 100, habiendo además obtenido los siguientes distintivos dentro del contexto UNCPBA:

- Tercer puesto para Ing. de Sistemas
- Quinto puesto dentro del área de Ingenierías y Tecnologías.
- Tercer puesto dentro de la Facultad de Cs. Exactas.

También se realizó la presentación de un trabajo en la 10^o edición del Congreso Nacional de Ingeniería Informática / Sistemas de Información (CONAIISI 2022) donde se mostraron avances utilizando el Clasificador Naive Bayes (Illescas, et. al. 2022a) para lo cual utilizó el entorno de desarrollo integrado (IDE) RStudio (<https://www.r-studio.com/es/>).

Seguido a esto se suma la presentación como plan de tesis tanto de Maylén Dell'Oso como de Florencia Paglione para la obtención del título de grado de Ingeniería de Sistemas de la EXA, UNCPBA.

1. INTRODUCCIÓN

En la actualidad la gran mayoría de las organizaciones gestionan la información digitalmente, lo que resulta en una masiva cantidad de datos para analizar y usar en beneficio propio. Poseer datos almacenados es una gran pérdida de tiempo y recursos si no se utilizan para tomar decisiones basadas en datos. Cada dato es una potencial herramienta para analizar el pasado, entender el presente y cambiar el futuro.

En este contexto la Facultad de Ciencias Exactas utiliza, entre otros sistemas (ver figura 1), el SIU Guarani (<https://www.siu.edu.ar/>) para gestionar la información de sus alumnos lo cual conlleva la responsabilidad de utilizarla, así sea para análisis estadísticos como para construir la visión del futuro de la facultad (Illescas, et. al. 2022b).

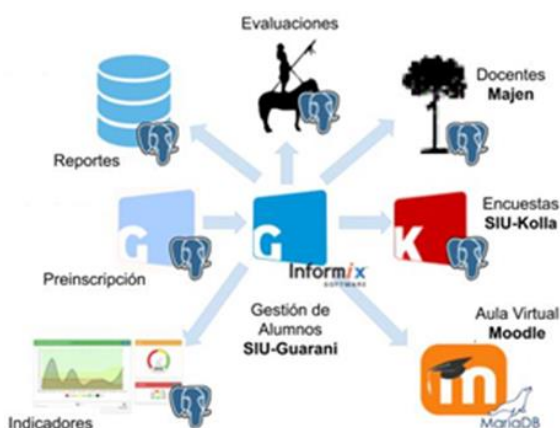


Figura 1. Esquema de Sistemas y Bases de datos disponibles en EXA. Fuente: área TICs EXA

Una gran parte de los datos almacenados en la herramienta son las notas de los alumnos, tanto de la cursada como la de los finales, lo que representa una gran oportunidad para realizar un análisis de los resultados de las evaluaciones y obtener predicciones sobre el futuro rendimiento académico de los alumnos basados en su historia académica.

Este proyecto supone calcular la probabilidad de aprobación de una materia en relación a sus materias correlativas lo que brindaría beneficios tanto para los alumnos como para la unidad académica.

Con respecto a la unidad académica, son los miembros de equipos de gestión y de cada cátedra, quienes van a poder utilizar los resultados para analizar el rendimiento de sus futuros estudiantes. Saber cuáles son los puntos fuertes y débiles de cada cuatrimestre para evaluar las posibles mejoras con el objetivo de crear un mejor entorno educativo para los estudiantes.

Además, y no menos importante, el alumno no solo se podrá ver beneficiado por las mejoras realizadas por la institución, sino también por tener el conocimiento de cuáles son las materias que probablemente le sean más desafiantes y por consiguiente, requieran más dedicación.

Estos objetivos tienen una relación estrecha con la búsqueda de la permanencia del alumno dentro de la carrera (Parra 2019a, Harrison, 2020). Tanto para brindarles una guía para sus estudios como así también un mayor acompañamiento en su experiencia académica.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Metodología

Siguiendo bajo la línea de trabajo que el proyecto acreditado se utilizará la metodología de desarrollo *Design Thinking* (Brown T., 2009). Este “es un método de trabajo en equipo que propicia la colaboración y la entrega frecuente de resultados a través de varias iteraciones. Si bien el método fue diseñado en el ámbito del diseño industrial, en la actualidad, es un marco de trabajo que también se está empleando en el desarrollo de proyectos de investigación y en el desarrollo de programas académicos debido a los resultados demostrados en la generación de nuevas ideas e innovaciones” (Illescas, et. al. 2022b) y se compone de cinco etapas como se muestra en la figura 2: empatía, definición, ideación, prototipado y testeado (Dinngo, 2020).

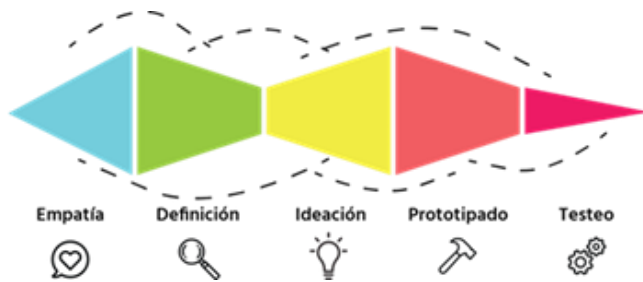


Figura 2. El proceso del Design Thinking (Dinngo, 2020)

3. RESULTADOS OBTENIDOS/ESPERADOS

La finalidad de este proyecto es la de utilizar un conjunto de técnicas y herramientas que permitan explorar los datos y detectar patrones en ellos que asistan a los tomadores de decisiones pertenecientes a la institución académica como a los alumnos de la misma.

En relación con el plan de trabajo, se verá generada una nueva base de conocimientos directamente asociadas a la probabilidad que tienen los alumnos de aprobar una materia en función a las notas obtenidas en las materias correlativas.

Se entrenarán distintos modelos de clasificadores de datos, como Naive Bayes (Parra, 2019b), Random Forest (Iarussi, 2020), Support Vector Machines (Parra, 2019c) y K Nearest Neighbors (Parra, 2019a; Harrison, 2018), con el fin de no limitarse al rendimiento de una sola técnica sino obtener los mejores resultados dentro de las posibilidades consideradas. Para el análisis de los resultados de cada herramienta se utilizarán distintas métricas intentando abarcar la mayor cantidad de enfoques posibles para la comparación de las mismas.

```
nb = naiveBayes(IO ~ ., data =
train)
# Se realiza la predicción con el
dataset de test
Predict <- predict(nb, test)
# Se muestran las métricas
confusionMatrix(Predict,
as.factor(test$IO), positive = "1")
F1_Score(test$IO, Predict,
positive = "1")
predNB <-
prediction(as.numeric(Predict),
test$IO)
perfNB <-
performance(predNB,measure="tpr",x.m
easure="fpr")
```

Figura 2. Extracto de código en R para aplicar las herramientas y sus métricas. (Illescas, et. al 2022a)

Una consideración crucial al momento de comparar modelos de clasificación es asegurarse de que el modelo no solo funcione bien con los datos que ya conoce, sino que también pueda hacer predicciones precisas sobre datos nuevos que nunca había visto. Para lograr esto, una práctica común es particionar los datos disponibles en tres conjuntos: entrenamiento, validación y prueba (Or B, 2023, Lee, 2019). La partición de datos en estos tres conjuntos es una práctica estándar en el aprendizaje automático que mejora la capacidad del modelo para aprender de manera efectiva, se ajusta sin sesgos y, finalmente, se evalúa de manera justa, asegurando que esté listo para enfrentarse a situaciones reales (Singh, 2021).

Por último, se podrían utilizar modelos de detección de anomalías ya que los datos están muy desbalanceados y estos modelos funcionan muy bien a la hora de clasificar muestras con mucho desbalance. Otra perspectiva, donde se podrían balancear los datasets, es el uso de otras técnicas como Both Samplin.

4. FORMACIÓN DE RECURSOS HUMANOS

La estructura del equipo de trabajo se muestra en la siguiente tabla:

Apellido y nombre	Título	Cargo	Funciones
Illescas, Gustavo	Dr.	Prof. UNCPBA	Director
Dos Reis, Rosa	Mg.	Prof. UNCPBA	Co-Dir.
Dell'Oso Maylen	Pre-grado	Alumno	Becaria/ Tesista
Paglione, Florencia	Pre-grado	Alumno	Becaria/ Tesista
Mora-Soto, Arturo	Dr.	Prof. Celeya	Integrante

Becario y tesis de grado

- Evaluación de un Modelo de Detección Temprana de Aprobación en Espacios de Educación Superior. Dirección: Illescas G., Dos Reis R. Becario: Dell'Oso Maylen (EVCIN 2023)

5. BIBLIOGRAFÍA

Illescas G., Mora Soto A., Dell'Oso M., Paglione F., Bessonart V. (2022a): *Evaluación de un Modelo de Detección Temprana de Aprobación en Espacios de Educación Superior*. 10º edición del Congreso Nacional de Ingeniería Informática / Sistemas de Información (CoNaIISI). Noviembre de 2022. Facultad Regional Concepción del Uruguay de la UTN.

Illescas G., Todorovich E., Aciti C., Rodriguez G., Silvestrini P. (2022b): *Aplicación de Analítica de Datos en espacios de Educación Superior*. XXIV Workshop de Investigadores en Ciencias de la Computación, Universidad Champagnat. Mendoza, Argentina, Abril 2022.

Dinngo (2020): *¿Cómo es el proceso de Design Thinking?*. Especialistas en formación y consultoría en Design Thinking. Sitio Web: <https://dinngo.es/como-es-el-proceso-de-design-thinking/>

Parra, F. (2019a): *Estadística y Machine Learning con R 6.4*. <https://bookdown.org/content/2274/metodos-de-clasificacion.html#algoritmo-k-vecinos-mas-cercanos>

Harrison, O. (2018): *Conceptos básicos de aprendizaje automático con el algoritmo de vecinos más cercanos K*. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

Lee, S.B., Gui, X., Manquen, M., Hamilton, E.R. (2019). *Use of Training, Validation, and Test Sets for Developing Automated Classifiers in Quantitative Ethnography*. In: Eagan, B., Misfeldt, M., Siebert-Evenstone, A. (eds) *Advances in Quantitative Ethnography*. ICQE 2019. Communications in Computer and Information Science, vol 1112. Springer, Cham. https://doi.org/10.1007/978-3-030-33232-7_10

Or, B. (2023). *Breaking the Mold: Challenging the Common Split for Training, Validation, and Test Sets in Machine Learning*. <https://pub.towardsai.net/breaking-the-mold-challenging-the-common-split-for-training-validation-and-test-sets-in-machine-271fd405493d>

Singh, V., Pencina, M., Einstein, A.J. (2021). *Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging*. *Sci Rep* 11, 14490 <https://doi.org/10.1038/s41598-021-93651-5>

Brown, Tim (2009). *Change by Design, How Design Thinking Transforms Organizations and Inspires Innovation*. Ed. Harper Collins New York ISBN 978-0-06-193774-3

Parra, F. (2019b): *Estadística y Machine Learning con R 6.9*. <https://bookdown.org/content/2274/metodos-de-clasificacion.html#clasificador-bayesiano>

Iarussi, F. (2020): *Caracterización de Asimetrías en Hipocampos Usando Técnicas de Inteligencia Artificial*. Tesis de grado. Facultad de Ciencias Exactas, UNCPBA.

Parra, F. (2019c): *Estadística y Machine Learning con R 6.7*. <https://bookdown.org/content/2274/metodos-de-clasificacion.html#maquina-soporte-vector>