

**DIPLOMADO EN IA PARA EDUCACIÓN**

**MÓDULO 5 “Proyecto integración de Machine Learning en Educación”**

**TÍTULO:** “Modelo Predictivo para la Detección Temprana de Riesgo de Reprobación en Biología en Estudiantes de 1º Medio”

|  |  |
| --- | --- |
| **Docente** | Leonardo Hernández Vera |
| **Alumno** | Carolina Quiroz Concha |

**INTRODUCCION**

En el contexto educativo actual, la integración de tecnologías basadas en inteligencia artificial y análisis de datos ofrece oportunidades concretas para mejorar el acompañamiento y el aprendizaje de los estudiantes. Particularmente, los modelos predictivos permiten anticiparse a situaciones de riesgo académico, generando alertas tempranas que habilitan intervenciones pedagógicas más eficaces y personalizadas. Esta perspectiva está alineada con enfoques de mejora continua y gestión por resultados en el ámbito educativo, los cuales destacan la importancia de utilizar herramientas tecnológicas como apoyo para la toma de decisiones informadas (Fundación Chile; G&P Consultores, 2006).

Este informe presenta el diseño de una arquitectura de solución en la nube, basada en servicios de Google Cloud Platform (GCP), cuyo propósito es predecir si un estudiante reprobará la asignatura de Biología al finalizar 1º medio. El modelo se alimenta de datos recolectados en etapas tempranas del año escolar —dos primeras notas parciales, asistencia a clases y asistencia a laboratorios— lo que permite actuar con anticipación frente a posibles dificultades.

La finalidad de este trabajo no es solo técnica, sino educativa: generar una herramienta práctica que facilite la toma de decisiones informadas por parte de profesores, apoderados y los propios estudiantes. Por ejemplo, los docentes podrían aplicar estrategias de reforzamiento focalizadas, ajustar metodologías o derivar a tutorías; los apoderados podrían recibir orientación temprana sobre cómo apoyar a sus hijos en casa; y los estudiantes, al conocer su situación de riesgo, podrían asumir compromisos concretos para mejorar su desempeño y asistencia. De esta forma, el modelo predictivo no solo entrega información, sino que promueve acciones concretas para prevenir la reprobación y favorecer trayectorias escolares exitosas.

**OBJETIVOS**

**Objetivo general.**

Desarrollar un modelo de aprendizaje automático que permita predecir, de forma temprana, si un estudiante de 1º medio reprobará la asignatura de Biología al finalizar el año escolar, utilizando datos académicos y de asistencia procesados en una arquitectura basada en servicios de Google Cloud.

**Objetivos específicos.**

* **Recopilar y preparar datos educativos relevantes,** incluyendo las dos primeras notas parciales, el porcentaje de asistencia a clases y la asistencia mensual a laboratorios de estudiantes de 1º medio, para su posterior análisis y uso en un modelo predictivo.
* **Entrenar y evaluar un modelo de clasificación binaria** que utilice dichos datos como variables predictoras, con el fin de identificar a estudiantes con alto riesgo de reprobación en Biología y apoyar así decisiones pedagógicas oportunas y fundamentadas.

**DESCRIPCION DE DATOS**

El conjunto de datos analizado proviene de un archivo llamado 1DATABAS1.csv, el cual recopila información académica de estudiantes de 1º medio correspondiente a la asignatura de Biología. Las fuentes de datos incluyen las notas parciales obtenidas en dos evaluaciones escritas, el porcentaje de asistencia a clases teóricas y el porcentaje de asistencia a sesiones prácticas en laboratorio. Cada registro representa un estudiante, identificado únicamente por un código.

Respecto a la distribución de los datos, se realizó una visualización de las variables numéricas mediante histogramas con curvas de densidad (KDE). Esta representación permitió observar cómo se comportan las calificaciones y asistencias dentro del grupo de estudiantes. Por ejemplo, se puede identificar si los datos están concentrados en torno a ciertos valores (como asistencias altas) o si existe variabilidad significativa.

En cuanto al preprocesamiento, se aplicaron varios pasos esenciales para asegurar la correcta lectura y tratamiento de los datos. En primer lugar, las columnas de notas se encontraban en formato texto con comas como separadores decimales, por lo que fueron transformadas a números flotantes utilizando puntos decimales. Este paso fue clave para habilitar análisis estadísticos posteriores. Posteriormente, se utilizaron funciones estadísticas para obtener medidas de tendencia central y dispersión (media, mínimo, máximo, desviación estándar), lo que permitió caracterizar los valores típicos de cada variable. Finalmente, las variables numéricas fueron preparadas para su uso futuro en tareas de visualización y análisis predictivo. La correcta gestión, organización y procesamiento de los datos es fundamental para construir modelos de aprendizaje automático útiles, especialmente en contextos educativos donde la toma de decisiones debe estar sustentada en evidencia (Hernández Vera, 2023).

**MÉTODOS Y MODELOS**

Para abordar el problema de predicción del rendimiento académico en la asignatura de Biología, se utilizó un modelo de **Regresión Logística**, el cual es apropiado para problemas de clasificación binaria, como en este caso donde se busca determinar si un estudiante aprobará (1) o reprobará (0).

La Regresión Logística pertenece a la categoría de modelos de aprendizaje supervisado y paramétrico, caracterizado por su capacidad para estimar la probabilidad de pertenencia a una de dos clases, en función de un conjunto de variables independientes. En este caso, las variables predictoras utilizadas fueron:

* Nota 1
* Nota 2
* Asistencia a clases (%)
* Asistencia a laboratorio (%)

1.- Partición de datos

Los datos fueron divididos en:

* 80% para entrenamiento (X\_train, y\_train)
* 20% para prueba (X\_test, y\_test)

La partición se realizó utilizando la función train\_test\_split de scikit-learn, fijando random\_state=42 para asegurar la reproducibilidad de los resultados.

2.- Hiperparámetros del modelo

El modelo fue instanciado utilizando los siguientes hiperparámetros:

* solver='liblinear': Este solucionador es ideal para conjuntos de datos pequeños o medianos y es compatible con problemas de clasificación binaria.
* C=1.0: Este parámetro controla la regularización inversa. Un valor de C más alto implica menos regularización, mientras que valores bajos implican más penalización. En este caso, se utilizó el valor por defecto (1.0), que ofrece un buen equilibrio entre ajuste y generalización.
* random\_state=42: Para asegurar resultados consistentes.

Este modelo permite una interpretación clara de los coeficientes asociados a cada variable, facilitando la identificación de los factores más influyentes en el riesgo de reprobación. Además, es computacionalmente eficiente, lo que lo hace apropiado para proyectos educativos que manejan volúmenes de datos moderados.

**EVALUACION DEL MODELO**

### **1.- Precisión general (accuracy)**

* **Entrenamiento: 87.5%**
* **Prueba: 83.3%**

Esto indica que el modelo mantiene un rendimiento consistente y no muestra señales claras de sobreajuste.

### **AUC (Área bajo la curva ROC)**

* **Entrenamiento: 0.98**
* **Prueba: 1.00**

**2.- Desempeño por clase**

#### ****Entrenamiento:****

* Clase 0 (Reprobado):
  + Precisión: 0.00
  + Recall: 0.00
  + F1-score:0.00  
    (El modelo no predijo correctamente ningún caso de reprobación en entrenamiento)
* Clase 1 (Aprobado):
  + Precisión: 87.5%
  + Recall: 100%
  + F1-score: 93.3%

#### ****Prueba:****

* Clase 0 (Reprobado):
  + Precisión: 100%
  + Recall: 50%
  + F1-score: 66.7%
* Clase 1 (Aprobado):
  + Precisión: 80%
  + Recall: 100%
  + Gráfico

    Descripción generada automáticamenteF1-score: 88.9%

El modelo de regresión logística presenta un **buen rendimiento en general,** especialmente en la **identificación de estudiantes que aprueban** la asignatura de Biología. Esto se refleja tanto en la **alta precisión y recall** de la clase "Aprobado", como en el valor del **Área Bajo la Curva (AUC),** que alcanza el máximo posible (1.00) en el conjunto de prueba.

Sin embargo, el modelo muestra **limitaciones al clasificar correctamente a los estudiantes que reprueban**. Esto se debe en parte a que la clase "Reprobado" está **subrepresentada** en el conjunto de datos, lo que genera un **desequilibrio de clases** que puede sesgar el aprendizaje del modelo hacia la clase mayoritaria. Esta situación es común en contextos educativos, donde los casos de reprobación suelen ser estadísticamente menos frecuentes.

Además del desbalance, creo podrían influir otros factores:

* **Cantidad reducida de datos:** el conjunto de prueba es pequeño (solo 6 casos), lo cual limita la capacidad de generalización y análisis robusto.
* **Poca complejidad del modelo:** si bien la regresión logística es efectiva e interpretable, puede no ser suficiente para captar patrones complejos entre las variables predictoras.
* **Ruido en los datos originales:** errores de medición o registros poco precisos pueden disminuir la sensibilidad del modelo frente a los casos minoritarios.

Aun así, el modelo es una base sólida para implementar un sistema de **alerta temprana,** capaz de detectar con buena precisión a estudiantes en riesgo de bajo rendimiento, siempre y cuando se complemente con estrategias para mejorar, como or ejemplo la recolección de más datos.

**INTERPRETACIÓN DE LOS RESULTADOS**

Los resultados del modelo de regresión logística muestran un desempeño satisfactorio en términos generales, con una **precisión del 87,5% en entrenamiento y 83,3% en el conjunto de prueba.** Esto indica que el modelo logra aprender patrones relevantes en los datos sin sobreajustarse, lo cual se confirma también por el alto valor del **AUC (Área Bajo la Curva ROC),** que fue de **0.98 en entrenamiento y 1.00 en prueba**. Un AUC cercano a 1 sugiere que el modelo tiene una excelente capacidad de discriminación entre las clases "aprobado" y "reprobado".

No obstante, el análisis más detallado revela que este buen rendimiento se concentra principalmente en la **clase mayoritaria (estudiantes aprobados).** En el conjunto de entrenamiento, el modelo predijo correctamente todos los casos de estudiantes aprobados, pero **falló en identificar correctamente a quienes reprobaron**, con una precisión y recall de 0 para esta clase. En el conjunto de prueba, el desempeño mejora parcialmente para la clase "reprobado", pero el número de casos sigue siendo muy bajo, lo que limita la robustez de las conclusiones.

Esto refleja un **problema clásico de desbalance de clases**, donde la clase minoritaria (reprobado) está insuficientemente representada y, por tanto, el modelo tiende a ignorarla para minimizar el error general. Aunque esto no afecta significativamente la precisión global, sí compromete la **utilidad pedagógica del modelo**, ya que su propósito principal es **anticipar casos de riesgo académico para intervenir a tiempo.**

Otro punto importante es que los resultados obtenidos dependen de variables cuantificables como las **notas parciales y porcentajes de asistencia**, que si bien son buenos predictores, no explican por sí solos el rendimiento de un estudiante. Factores como motivación, dificultades personales, métodos de estudio o apoyo familiar no están reflejados en los datos, lo que podría limitar la capacidad predictiva del modelo en contextos reales.

**CONCLUSIONES**

El presente proyecto tuvo como objetivo desarrollar un modelo predictivo capaz de anticipar la reprobación de estudiantes en la asignatura de Biología durante 1º medio, utilizando únicamente datos disponibles en etapas tempranas del año escolar: notas parciales y porcentajes de asistencia a clases y laboratorios. Para ello, se seleccionó la regresión logística como modelo de clasificación binaria, por su simplicidad, interpretabilidad y eficacia en problemas con variables independientes numéricas.

Los resultados obtenidos muestran un buen nivel de precisión general, tanto en el conjunto de entrenamiento como en el de prueba, y una excelente capacidad de discriminación entre clases según el área bajo la curva ROC. Sin embargo, el modelo presenta limitaciones al identificar correctamente a los estudiantes en riesgo de reprobación, lo que se explica principalmente por el desbalance en el conjunto de datos. Esto refuerza la necesidad de seguir perfeccionando el modelo, incorporando nuevas estrategias como recolección de más datos, técnicas de balanceo y eventualmente el uso de modelos más complejos.

Desde una perspectiva práctica, este modelo puede ser implementado como una herramienta de **alerta temprana,** que permita a docentes, orientadores y equipos directivos tomar decisiones informadas y oportunas respecto al acompañamiento académico de los estudiantes. Su valor no reside solo en el diagnóstico, sino en su potencial para **activar estrategias de intervención pedagógica personalizada,** tales como tutorías focalizadas, planes de apoyo y comunicación proactiva con apoderados.

En términos más amplios, este proyecto demuestra cómo el uso responsable y ético de la inteligencia artificial en educación puede contribuir a promover trayectorias escolares más justas, al identificar oportunidades de mejora antes de que se traduzcan en fracaso académico. La posibilidad de anticipar dificultades mediante datos objetivos ofrece un enfoque más preventivo y centrado en el estudiante, alineado con una visión de educación inclusiva y orientada al éxito formativo de todos los alumnos.

**REFERENCIAS**

* Fundación Chile; G&P Consultores. (2006). Cuaderno 1: Diseño del plan de mejoramiento de la gestión y adecuación del Proyecto Educativo Institucional. Programa de Dirección y Gestión Escolar, Fundación Chile.
* Hernández Vera, L. (2023). Importancia de los datos en el desarrollo de la inteligencia artificial y el machine learning.