

**Instituto Tecnológico y de Estudios Superiores de Monterrey**

**Maestría en Inteligencia Artificial Aplicada**

Proyecto Integrador

Aplicación de modelos de aprendizaje supervisado para el cálculo de probabilidad de pérdida de empleo en clientes de nómina

Avance 3. Baseline

Grupo 10:

Abraham Cabanzo Jiménez - A01795355

Kevin Alejandro Ayón Payán - A01740679

Pedro Ulises Meléndez Ortega - A00716301

Doctora: Grettel Barceló Alonso

9 de febrero de 2025

**Introducción**

A lo largo del proyecto se han realizado distintas estrategias de limpiado y mejoramiento de los datos para maximizar la eficiencia en el momento de implementar análisis que se lleven a cabo en un entorno de la vida real. En este avance, se realizará el primer desarrollo de modelos de aprendizaje automático, el cual será el punto de partida del equipo para evaluar la calidad, viabilidad del problema y solvencia de los datos.

En este nuevo avance, se construirán distintas métricas de evaluación de modelos, esto nos ayudará a proporcionar análisis y comparaciones de modelos avanzados, el cual nos permitirá determinar si los datos implementados en el análisis contienen la información suficiente (además de gran calidad) para hacer predicciones significativas que ayuden a resolver el problema.

A lo largo de todo este avance se abordarán preguntas claves para seleccionar el mejor algoritmo/modelo, dependiendo de las características, métricas, detección de ajuste y establecimiento de desempeños minimos esperados por el equipo para el proyecto.

**Preguntas a contestar**

**¿Qué algoritmo se puede utilizar como *baseline* para predecir las variables objetivo?**

Como punto de partida y de referencia que utilizará el equipo en el modelado del proyecto, principalmente observaremos el comportamiento de modelos aplicando regresión logística, siendo este el modelo “sencillo” de punto de referencia; además de también aplicar redes neuronales (también se tiene pensado bosque aleatorio) como modelos más “complejos” de punto de referencia, esto nos permitirá observar el comportamiento inicial de ambos tipos de modelos (sencillos y complejos) para poder tener referencia de cómo están los datos y cómo vamos a dirigir el análisis para desarrollar la mejor solución a la problemática empleada por el equipo.

Al tener resultados de modelos sencillos (como la regresión logística) y resultados de modelos complejos (como redes neuronales) podemos empezar las comparaciones del tipo de modelo que deberíamos emplear acorde a los datos y la estrategia que se espera implementar a lo largo del proyecto, lo cual serían puntos de partida para el equipo para brindar un enfoque más apegado al resultado que se espera tener.

**¿Se puede determinar la importancia de las características para el modelo generado? Recuerden que incluir características irrelevantes puede afectar negativamente el rendimiento del modelo y aumentar la complejidad sin beneficios sustanciales.**

Sí es posible determinar la importancia de las características, pero dependiendo del modelo que estamos implementando los datos. Si bien existen métricas que nos ayudan a observar el desempeño de los modelos como accuracy, precisión, f1-score, etc, se puede aplicar distintas herramientas y códigos de python para determinar la importancia de características de los modelos.

Un ejemplo de esto es aplicar el coeficiente de regresión, la cual nos proporciona el impacto de cada variable en la predicción que llevamos a cabo con el modelo. Para modelos más complejos como redes neuronales o bosque aleatorio, se pueden aplicar otras técnicas y herramientas como LIME o la impureza de Gini (esta técnica se emplea en python por medio de feature\_importances)

**¿El modelo está sub/sobreajustando los datos de entrenamiento?**

Para el caso de la red neuronal tradicional obtuvimos un desempeño relativamente bajo. De las 81 variables disponibles se utilizó PCA y se ocuparon 27 variables lo que nos dió el 95% de relevancia con el análisis realizado. Aún así creemos que en este modelo inicial tenemos un sobreentrenamiento porque la exactitud obtenida fue del 72%.

Ya en estas alturas nos estamos encontrando con retos importantes debido al tamaño del dataset y las limitaciones de hardware disponibles (principalmente en memoria RAM). Por lo que se ha tenido que realizar un proceso iterativo de refinamiento de código para tratar de ser lo mas eficiente posible en el uso de memoria.

**¿Cuál es la métrica adecuada para este problema de negocio?**

Para este problema, la métrica principal debe ser Recall, ya que queremos detectar la mayor cantidad posible de clientes en riesgo de perder su empleo. Con solo el 4% de la población en la clase positiva, un modelo trivial que siempre prediga 0 tendría un 96% de precisión, pero no detectaría ningún caso real. Por ello, el Accuracy no es una métrica útil en este contexto.

**¿Cuál debería ser el desempeño mínimo a obtener?**

Partiremos de un modelo de regresión logística que nos está arrojando un accuracy de 86% y un recall de 76%