## Proceso:

## Objetivo: Se quiere averiguar cuál es el porcentaje de estudiantes que abandonan su educación y cuáles no

## Dataset: Personalized Learning Dataset

## Link: <https://www.kaggle.com/datasets/adilshamim8/personalized-learning-and-adaptive-education-dataset/data>

## Descripción: Este dataset está diseñado para respaldar la investigación sobre sistemas de aprendizaje adaptativo, educación personalizada y modelos predictivos del éxito estudiantil

## Cantidad de ejemplos y de atributos:

Total de 10000 ejemplos con total de 15 atributos

|  |  |
| --- | --- |
| **Nombre** | **Tipo** |
| Student\_ID | Nominal |
| Age | Numérico |
| Gender | Nominal |
| Education\_Level | Nominal |
| Course\_Name | Nominal |
| Time\_Spent\_on\_Videos | Numérico |
| Quiz\_Attemps | Numérico |
| Quiz\_Scores | Numérico |
| Forum\_Participation | Numérico |
| Assingment\_Completion\_Rate | Numérico |
| Engagement\_Level | Nominal |
| Final\_Exam\_Score | Numérico |
| Learning\_Style | Nominal |
| Feedback\_Score | Numérico |
| Dropout\_Likelihood | Nominal |

## Tareas:

## Limpieza: La columna eliminada fue Student\_ID ya que no aporta información importante al igual que otras columnas que no son relevantes para lo que se quiere predecir.

## Selección: Para un primer vistazo al dataset se utilizaron todas las columnas incluyendo Student\_ID, así luego en base a los resultados ir sacando columnas las cuales sean irrelevantes para la predicción, el atributo seleccionado será Dropout\_likelihood

## Integración: No se integraron datos ya que el dataset proporcionó todo lo que se necesitaba para este proceso de minería de datos

## Transformación: No se transformaron datos ya que no fue necesario realizarlo

## Minería:

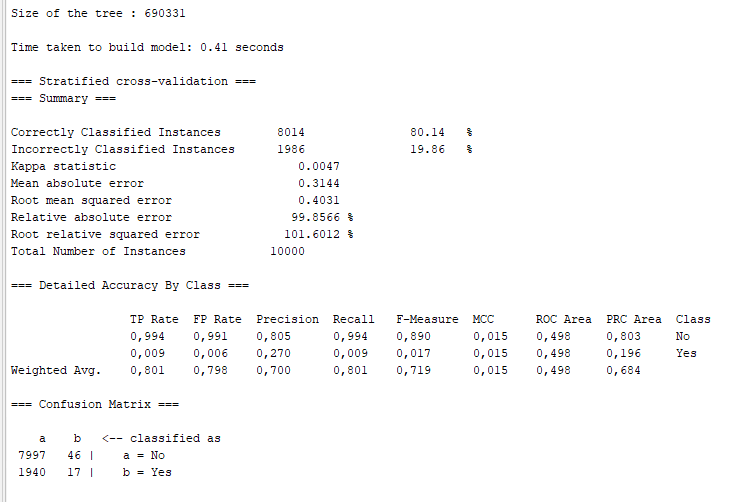
Se aplicaron los siguientes algoritmos:

Para el primer vistazo al dataset se utilizó randomForest, IBl, IBk para clasificación.

Utilizando el algoritmo RandomForest seleccionando el atributo Dropout\_Likelihood con todas las columnas del dataset sin limpieza, no pudo crear un resultado, se quedo “trabado” esto se da a que el volumen de datos es muy grande o tiene varios datos como para que el algoritmo pueda realizar una predicción

Después de limpiar y seleccionar los datos

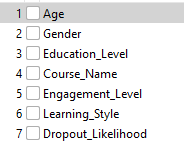
random tree



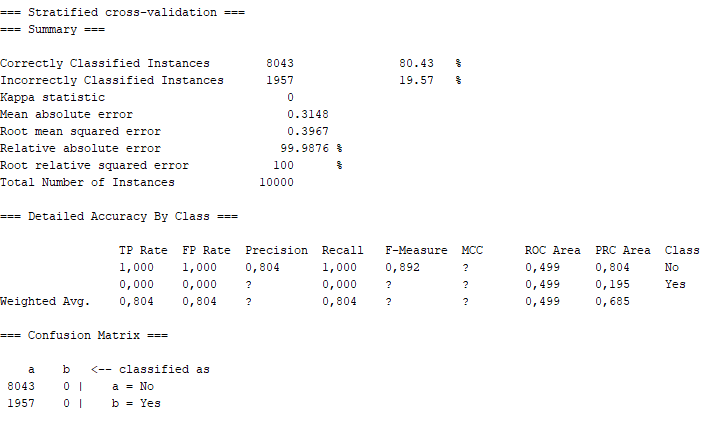
relación positiva del 80% para conocer quienes continuarán con sus estudios

la siguiente prueba del dataset se dará sacando algunas columnas como Student\_ID, Time\_Spent\_on\_Videos, Quiz\_Attempts y Scores, Forum\_Participation, Assingment\_Completition\_Rate, Final\_Exam\_Escore y Feedback\_Score

Las columnas restantes son:



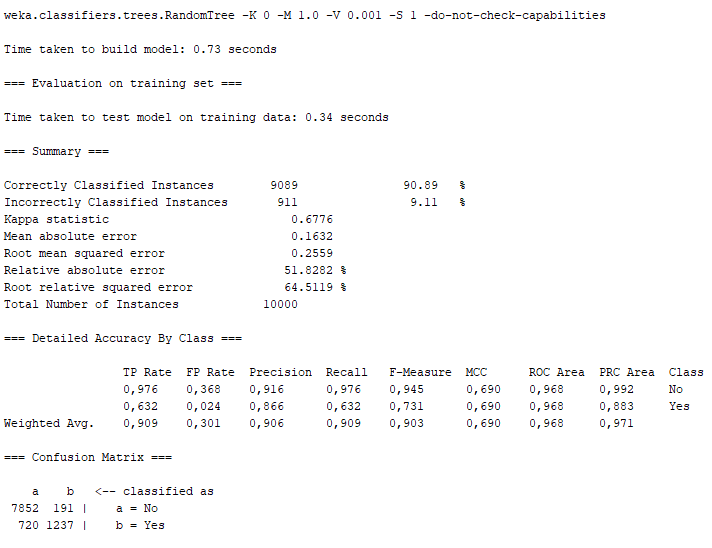
Utilizando J48 con 20 folds



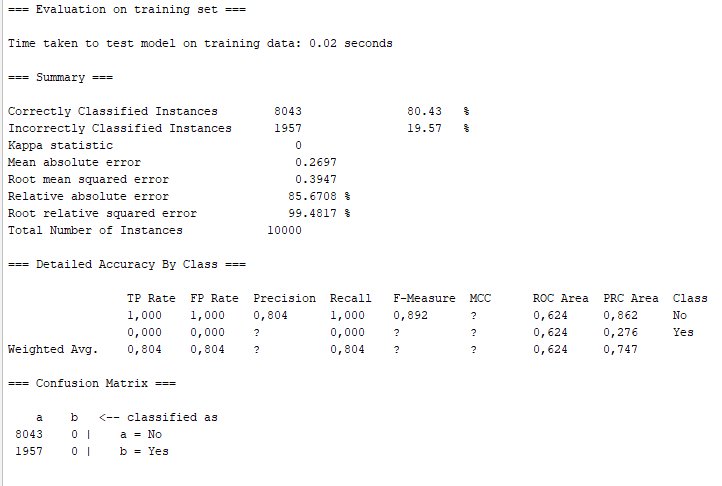
muestra el mismo resultado que en el primer resultado obtenido al principio

Cambiando el algoritmo por RandomForest y utilizando training set

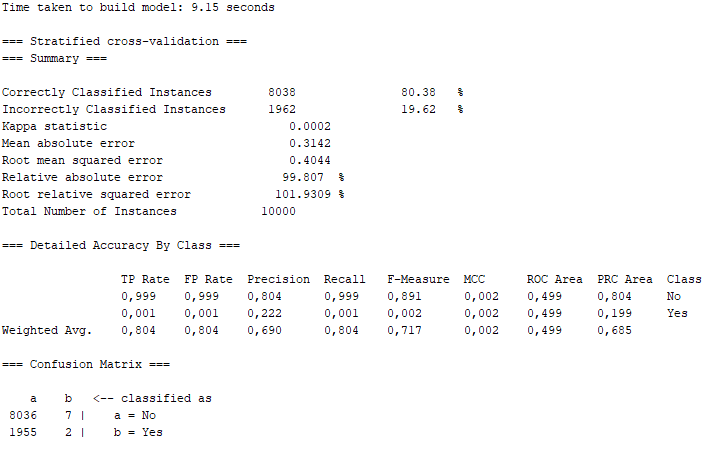
da como resultado lo siguiente (aplicando la limpieza realizada anteriormente):



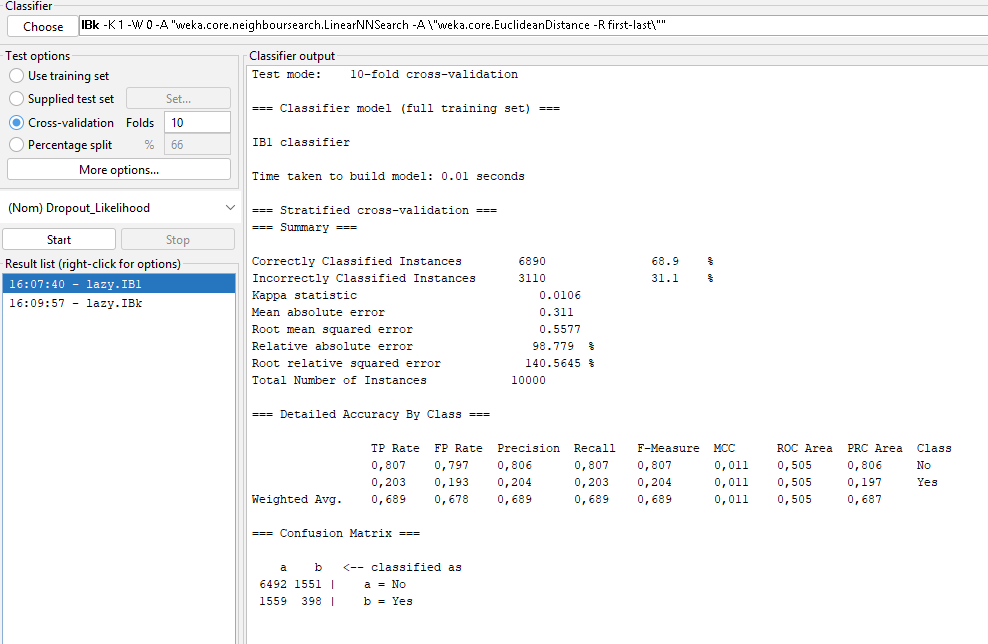
Cambiando al algoritmo multilayer perceptron da el siguiente resultado, siendo este idéntico en la matriz de confusión y en las instancias clasificadas correctamente y incorrectamente, cuenta con una pequeña diferencia en ROC y PRC



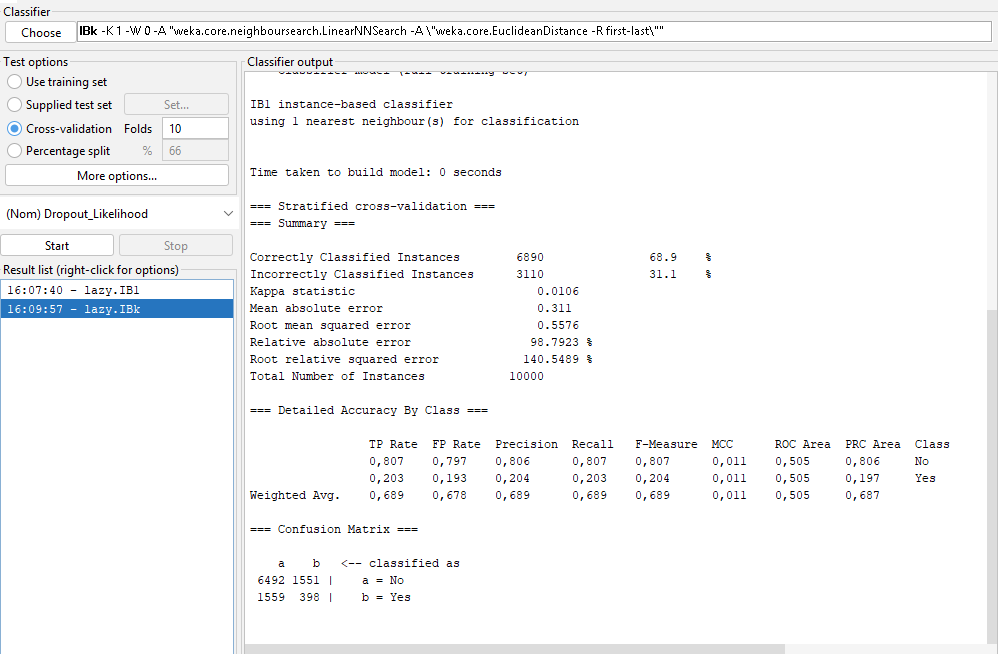
Mismo algoritmo pero usando cross validation de 20 folds



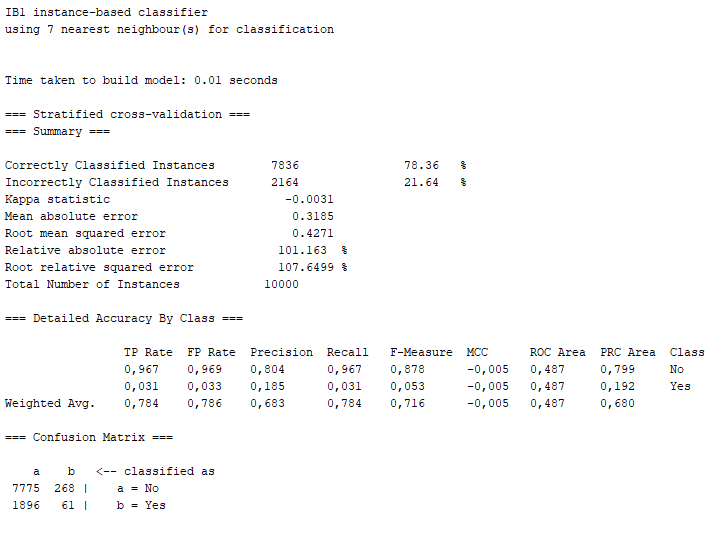
utilizando IB1 con CV de 10



utilizando IBk K = 1 y CV de 10



IBk con un total de K = 7 y CV de 10



## Evaluación:

Son las matrices de confusión presentes en las imágenes del apartado de **Minería.**

## Conclusiones:

Entre todos estos resultados la clasificación correcta llegó alrededor de 78.00-80.00 a 90.00 en los algoritmos seleccionados y junto a la selección y limpieza de los datos, se puede determinar que para esta tarea de clasificación para conocer el Dropout\_likelihod es satisfactorio, interpretando estos resultados se llega a la conclusión de que el 80 % de los estudiantes continuarán con sus estudios mientras que el 20 % aproximadamente abandonaran.