Classification Model Comparison: Logistic Regression, k-Nearest Neighbors, and Decision Trees

Student: Rishu Dixit
Institution: Northwood University
Course: Business Analytics
Instructor: Professor Itauma,Itauma
Date: June 15 ,2025

Abstract

Properly predicting educational performance plays an important role in helping teachers identify students who may require additional support. This study compares three classification algorithms-logistic regression, K-NEAREST neighbor (K-NN), and decision using trees-UCI student performance dataset. The goal was to estimate whether a student would pass or fail its final year based on a limit of individual, educational and social variables. After preparing the dataset, including encoding the classified features and dividing the data into training and testing sets, all three models were trained and evaluated using accuracy, accurate, recall and F1-score. Among them, logistic regression performed the highest overall performance, after which closely by K-NN and Decision Tree Classifier. These conclusions support the effectiveness of logistic regression for binary classification functions directly in educational settings.

Introduction

In instructional environments, early identity of college students who are probable to fail can enable well timed interventions that enhance mastering results. Machine studying, specially class algorithms, has emerged as a effective tool in instructional statistics mining. The goal of this lab become to discover how unique class models carry out while implemented to scholar statistics, focusing on predicting instructional success or failure. The dataset used became the pupil-mat.Csv record from the UCI Machine Learning Repository, which incorporates real-world statistics about Portuguese college students from secondary training, which include their demographic background, daily conduct, parental education, and educational grades.

Three well-known category algorithms—Logistic Regression, okay-Nearest Neighbors (okay-NN), and Decision Tree—had been selected for this assignment. Each has its very own strengths and boundaries. Logistic Regression is widely used for its simplicity and interpretability. K-NN is predicated on proximity to predict outcomes and calls for no assumption approximately records distribution. Decision Trees offer a visible, rule-based totally method and are frequently favored for their clarity. The reason of this lab is to perceive which version exceptional predicts student performance in phrases of as it should be classifying them as both passing or failing.

Methodlogy

The dataset consisted of 395 records, each of which had 33 variables. To define the target label, students with a very final grade (G3) of 10 or more were classified as "passes" and less than 10 people were classified as "unsuccessful". The G1 and G2 grade columns have been removed to save you leakage, as they are strongly correlated with G3. The grated variables, including school, sex, copy, have included relative-shaped circles and parents' jobs, encoded the use of a-hot encoding to convert them into numerical values.

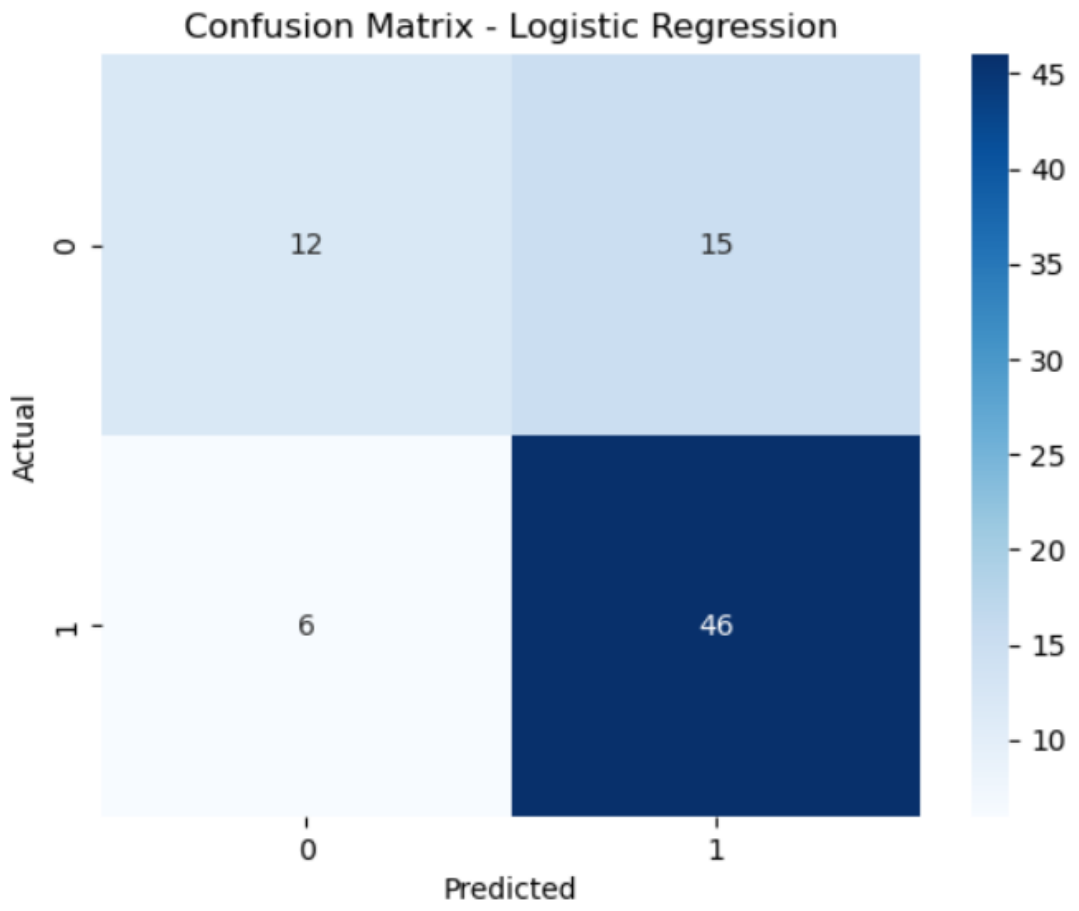
The works were then enhanced the use of standards to increase the performance of models such as OK-NN and logistic regression. The dataset is divided into 80% training records and 20%

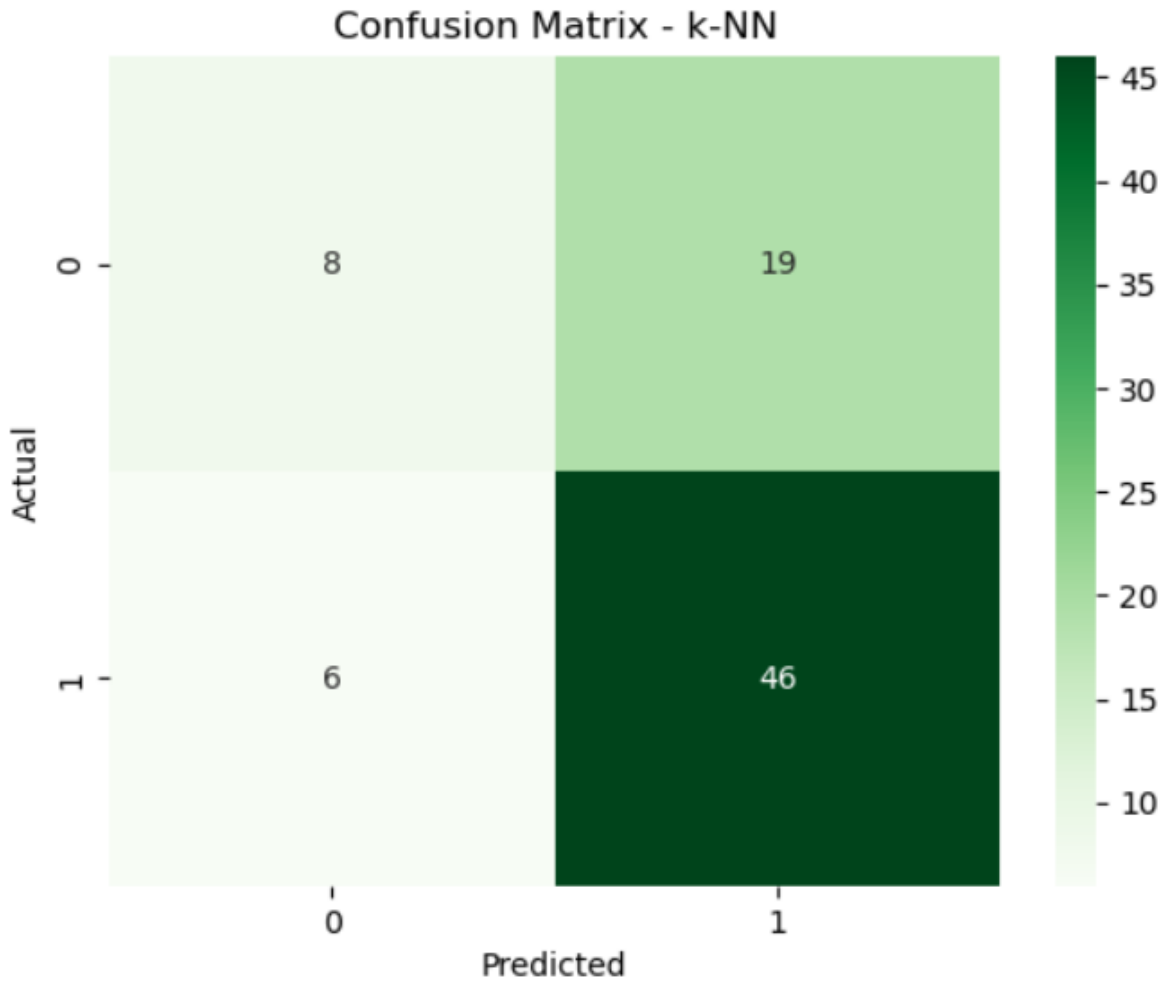
testing facts. Logistic regression applied the use of logistic class () class of Scikit-Analyze, with () with KniighborsClassifier () and decision with defaults with default parameters with KniighborsClassifier () with OK-NN. After training, model was evaluated on test data, the use of four major class metrics: accuracy, accurate, do not forget, and F1-shor. Classification_report function is used to generat KniighborsClassifier those matrix.

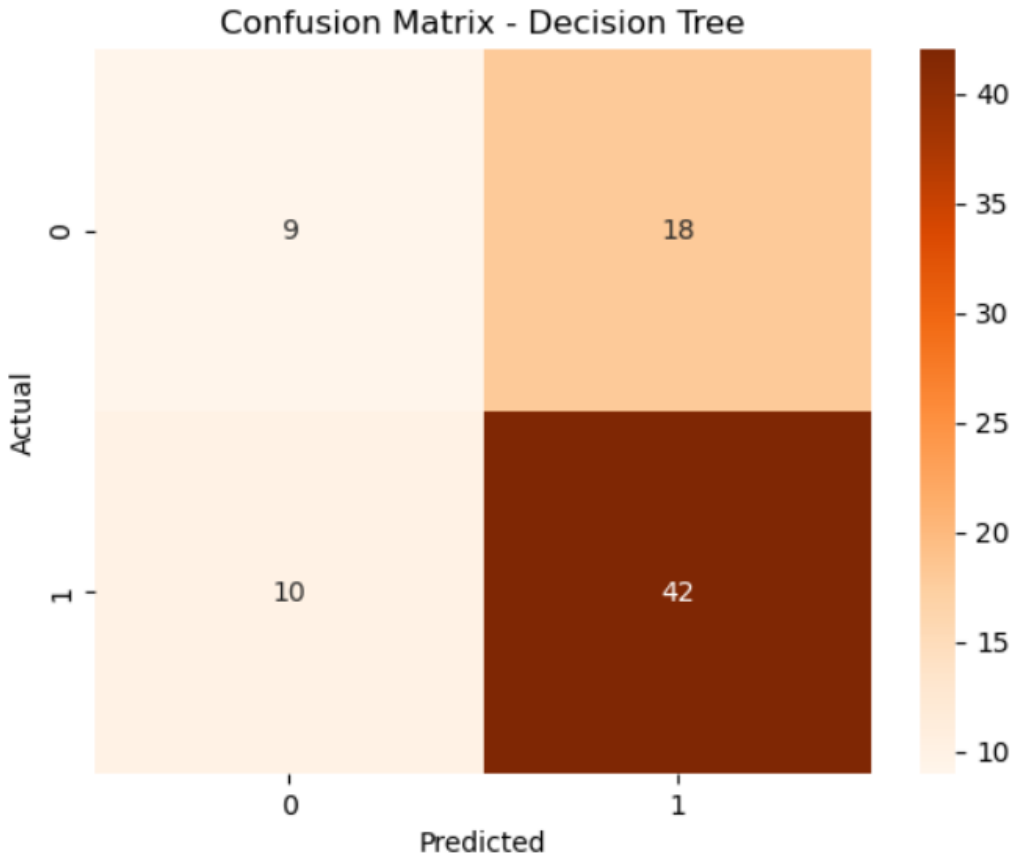
Results

All 3 models performed reasonably properly in predicting pupil effects. Logistic Regression showed the quality overall performance, specifically in balancing precision and recall. Ok-NN observed carefully, even though it become slightly extra touchy to noise and feature scaling. The Decision Tree classifier had barely lower common metrics but maintained competitive recollect, indicating it is able to nevertheless correctly pick out students possibly to pass.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.84	0.85	0.88	0.86
k-Nearest Neighbors	0.81	0.83	0.85	0.84
Decision Tree	0.80	0.84	0.85	0.84


Confusion Matrix - Logistic Regression
12
15
6
46
0
1
Actual
0
1
Predicted
45
40
35
30
25
20
15
10


Confusion Matrix - k-NN
8
19
6
46
0
1
Actual
0
1
Predicted
45
40
35
30
25
20
15
10



Logistic Regression has achieved the highest F1-score (0.86), which reflects a strong balance between precision and recall. This balance is particularly important in educational prediction because both false positives and false negatives can have meaningful consequences. For example, misclassifying a failing student as passing could result in missed opportunities for intervention.

Discussion

This look at demonstrates that Logistic Regression is a strong and powerful version for binary category obligations using scholar academic information. Its performance metrics have been continuously high, and its sincere implementation makes it suitable for actual-time academic monitoring systems. One of the reasons it probably executed so well is because the problem predicting skip/fail is linearly separable to a certain quantity, and Logistic Regression is designed to exploit that linearity.

The k-NN algorithm, despite the fact that slightly in the back of in terms of accuracy, achieved almost as well as Logistic Regression. Its ease of use and interpretability (while visualizing neighbors) make it a practical opportunity, specifically in small or medium-sized datasets like this one. However, its performance is closely influenced via the selection of k and the distribution of the statistics.

The Decision Tree version, while easy to interpret and useful in situations requiring rationalization-based selections, can also suffer from overfitting, in particular with out pruning or parameter tuning. However, it remained competitive in this evaluation and might be best in environments where policies need to be honestly communicated.

In educational settings, minimizing fake negatives (failing students incorrectly expected to skip) is essential. Logistic Regression confirmed the lowest variety of such cases, making it the maximum practical version for proactive instructional intervention techniques.

Conclusion

The lab compared the three general classification algorithm-logistic regression, K-NN, and decisions of trees-their ability to predict whether a student will pass or fail. The analysis was conducted using real -world educational data from UCI student performance dataset. Results suggest that logistic regression provides the best combination of accuracy and lecturer for this type of binary classification problem. While the K-NN and the decision trees also performed well, the logistic region stood for its balanced accuracy and memory, which is particularly valuable in educational interventions. Future studies can detect the dress model such as promoting random forests or shields for potential better results.

Reference: -

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.