Abstract:
Universities in the current era have primarily turned into electronic systems. As a result, universities are producing student data daily for official purposes. However, the vast data produced daily can be used far beyond formal purposes, such as predicting whether a student will struggle or fail in their studies from an early age so that these educational institutions can help them. This study aims to predict students' academic performance in advance, using pre-admission information and first-year subject marks. The study was applied to the data of 824 students from King Khalid University in the Kingdom of Saudi Arabia. In order to achieve our goal, we start by applying data processing and pre-analysis techniques to the data and then apply machine learning algorithms to the data. The experimental results showed that the Random Forest algorithm outperformed the decision tree in predicting students' academic performance. Moreover, it turned out that studying the pre-admission data with the first year marks gave the best prediction result than studying the data before admission only.

Design:

This project is provided to predict students' academic performance in advance. This study helps universities reduce student dropouts and help identify students who are likely to fail or stumble early. Moreover, this study assists universities to take advantage of their resources and use them for the benefit of the student.

Data:

In this project, we benefit from open data at King Khalid University.

https://data.kku.edu.sa/ar/open-data

The data contains 28 attributes and 824 records, we chose this data because it fits with the desired goal of this project, as it contains pre-admission data for university students while maintaining the privacy of these students.

Algorithm:

To achieve the purpose of this study, the data was modeled using two algorithms, namely Random Forest and Decision Tree for the purpose of predicting students' academic performance. This is done using (high school average, Standard Achievement Admission Test, General Aptitude Test) as attributes in addition to the grades of English, Computer and Mathematics for the first school year.

Tools:

All data analysis and modeling operations were done using the Python language. The following libraries were used to perform this task:

1. Pandas- a data processing library.
2. Matplotlib - library for creating statistical graphics.
3. Matplotlib.pyplot- A library for adding more details to graphs.
4. Seaborn - An advanced graphic library.
6. Sklearn- A library used to run different algorithms.

Communication:

By using two types of algorithms, the random forest and the decision tree to measure students' academic performance, the random forest gave a better result in predicting students' performance than the decision tree of our sample data and gave best result of F1-score that create a balance between precision and recall

Type of measure	RF	DT
Precision	0.79	0.62
Recall	0.70	0.72
F1-score	0.74	0.67
Accuracy	0.78	0.68

Moreover, a comparison was made between attributes and their impact on the model result.. In the first stage, we tested the work of pre-admission data only. Then we tested the effect of the pre-admission attributes in addition to the grades of the first semester.. Finally, we compared them with the pre-admission attributes with the grades of the first and second semesters. We concluded that by using all the attributes , the model gave a better result in predicting students' academic performance, and therefore we recommend that universities focus on trying to predict students' academic performance after the first year.

Depending on RF Classifier - At risk students				
Features	Precision	Recall	F1-score	Accuracy
Pre-admission	0.66	0.55	0.60	0.67
Pre-admission & First semester	0.71	0.59	0.65	0.71
All attributes	0.79	0.70	0.74	0.78

TABLE 4.4: The most important features that affect the prediction of student performance.