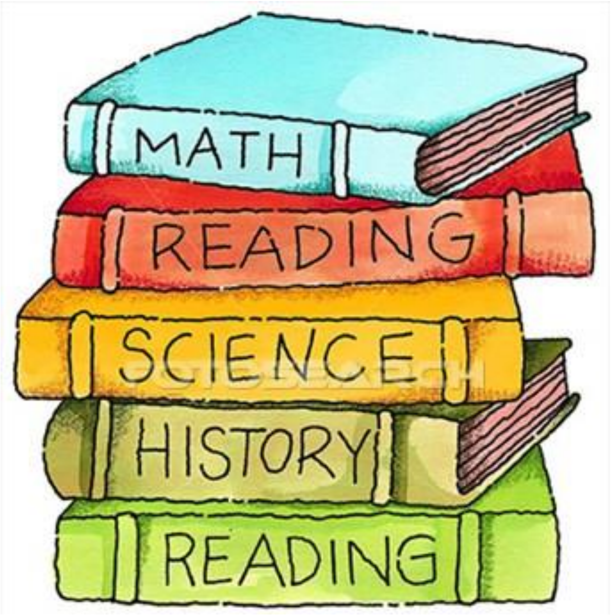
Final Grades Prediction



Classification

Model- 3

Prepared by:

Sara AlAbdulsalam

Fatimah AlShammari

Introduction:

Students, parents, and instructors seek to have an early alert of the performance of students to take actions based on it. This problem can be carried through machine learning classification model such as (KNN, Logistic Regression, and Decision Tree) to predict if the student will pass or fail which also makes the student knows whether he/she is in a position to reach his/her expectations or not. If the model shows that the student is going to fail and needs to improve then that student, with the help of the parents and instructors can prepare more for that semester to pass it. Lastly, we can get insights of what student features such as demographic, social, and school related attributes can affect student grades.

Data Description:

The dataset used to solve the problem is publicly available on Kaggle [1]:

The predictors:

School: student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

Gender: student's sex (binary: 'F' - female or 'M' - male)

Age: student's age (numeric: from 15 to 22)

Address: student's home address type (binary: 'U' - urban or 'R' - rural)

Famsize: family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

Pstatus: parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

Medu: mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" higher education)

Fedu: father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" higher education)

Mjob: mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob: father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Reason: reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

Guardian: student's guardian (nominal: 'mother', 'father' or 'other')

Traveltime: home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

Studytime: weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

Failures: number of past class failures (numeric: n if 1<=n<3, else 4)

Schoolsup: extra educational support (binary: yes or no)

Famsup: family educational support (binary: yes or no)

Paid: extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

Activities: extra-curricular activities (binary: yes or no)

Nursery: attended nursery school (binary: yes or no)

Higher: wants to take higher education (binary: yes or no)

Internet: Internet access at home (binary: yes or no)

Romantic: with a romantic relationship (binary: yes or no)

Famrel: quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

Freetime: free time after school (numeric: from 1 - very low to 5 - very high)

Goout: going out with friends (numeric: from 1 - very low to 5 - very high)

Dalc: workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

Walc: weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

Health: current health status (numeric: from 1 - very bad to 5 - very good)

Absences: number of school absences (numeric: from 0 to 93)

the response variable will be to predict whether the student will pass of fail based on his/her grades for the three semesters periods which are the following:

G1 - first period grade (numeric: from 0 to 20)

G2 - second period grade (numeric: from 0 to 20)

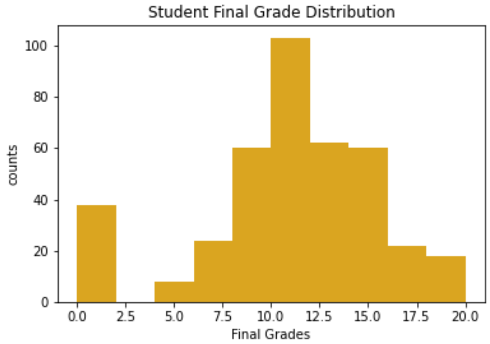
G3 - final grade (numeric: from 0 to 20)

All the above grades will be scaled to 100 .

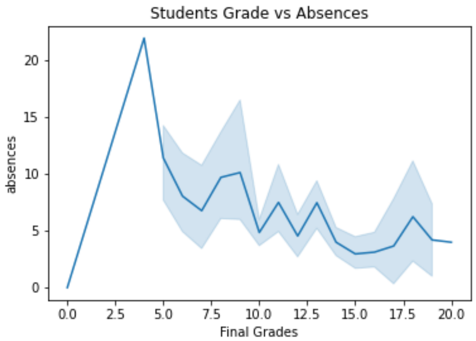
If the final grade is at least 60% of the total grade the student will be classified as pass, otherwise fail.

The dataset has 395 observations, and we performed a simple EDA:

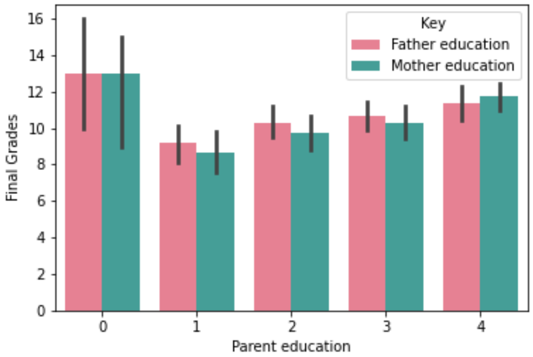
The below histogram shows the distribution of the final grades, as it shows a close to normal distribution except that there is a number of students that got zero as a final grade:



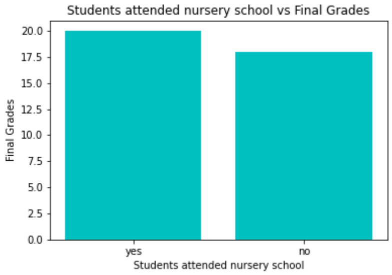
A line chart that shows the reversed relationship between final grades and students absence number:



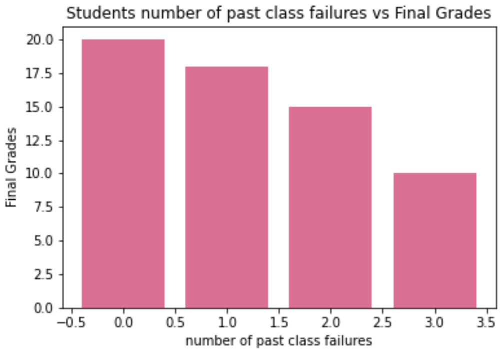
The below bar chart show the level of education of the parents against students grades, as we can see, students with parent who have lower education level got higher grades:



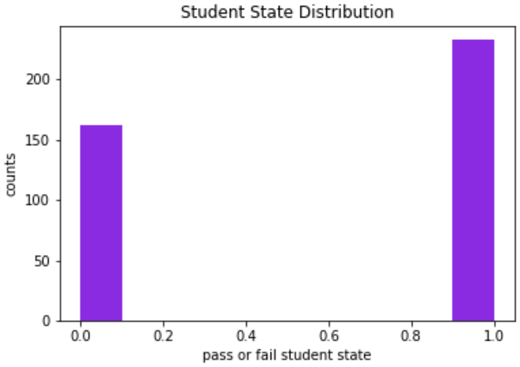
The below bar chart shows the highest grades achieved by students who went to nursery school vs who did not:



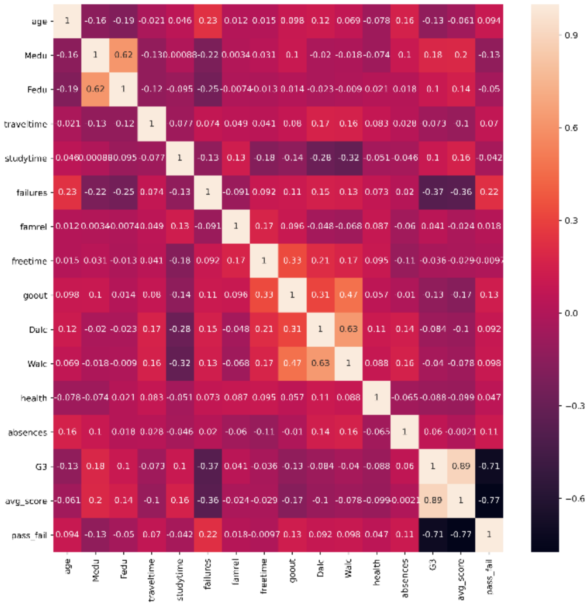
The below bar chart shows the highest grades achieved by students vs the number of past class they failed in:



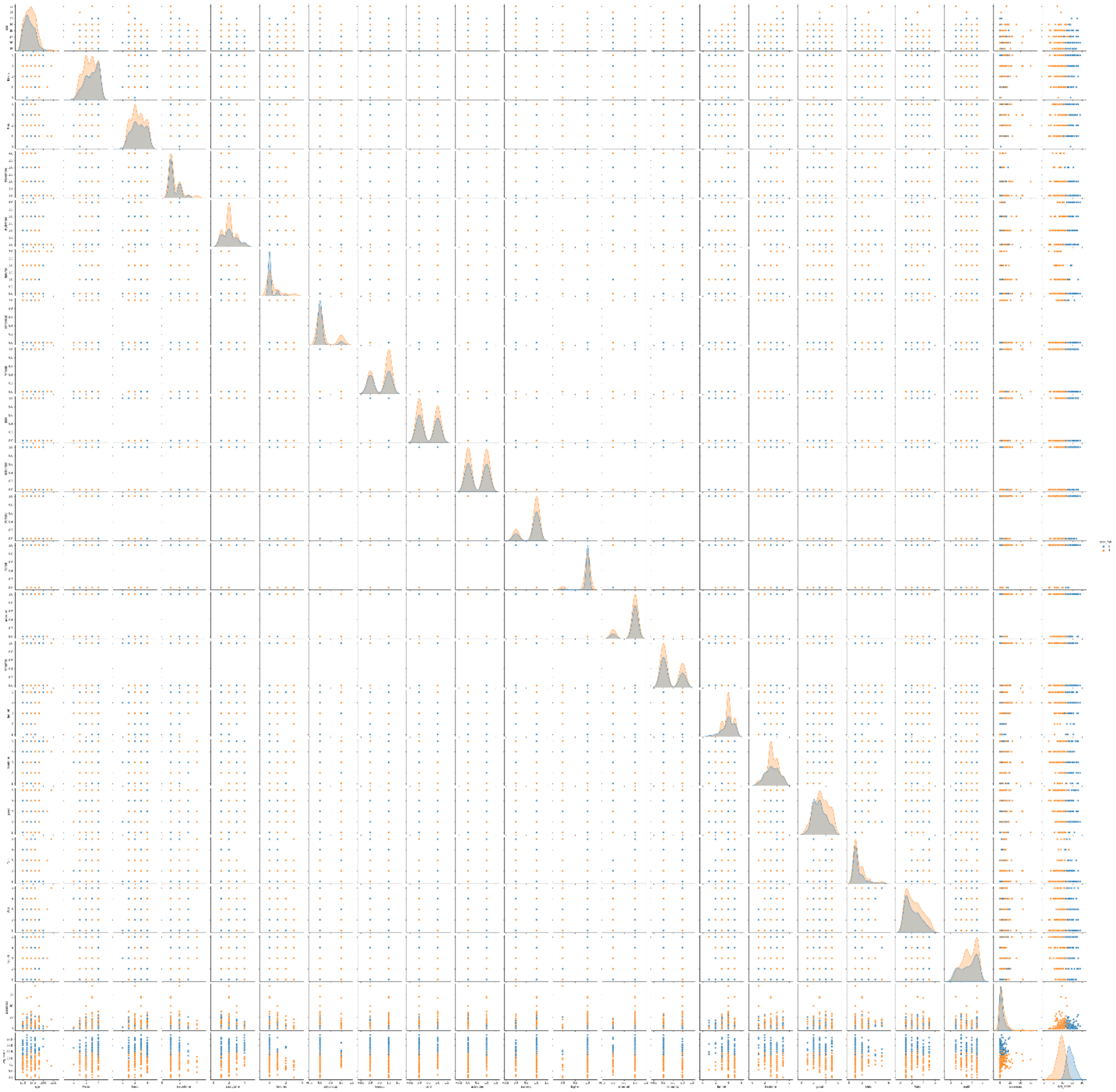
The distribution of student fail (1) or pass(0) state is:

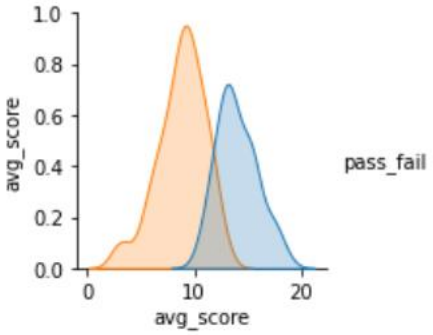


The heatmap below shows that the target ('pass_fail') has an almost strong correlation with avg_score variable which is the average of first and second periods grades:



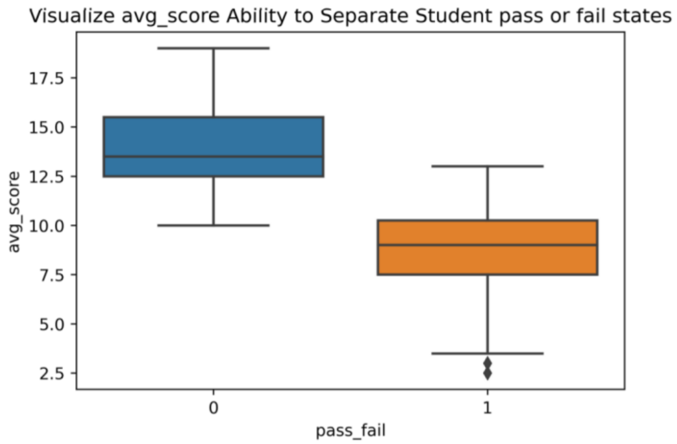
Pair plot to see what features sperate the pass state from fail state, avg score shows the best separation:





Zoom on the best observed separator:

Box plot of the avg_score with pass_fail target to visualize the separation line:



Pre-processing:

- Add pass_fail feature which is the if the final grade of the student is al least 60% then it will be pass (1) otherwise fail(0).
- Drop final grade (G3) column.
- Create a feature that is the average grade of first and second period grades. Drop G1 and G2.
- Add dummy variables for the categorical features.
- Oversampling the training set to fix class-imbalance problem.

Algorithms:

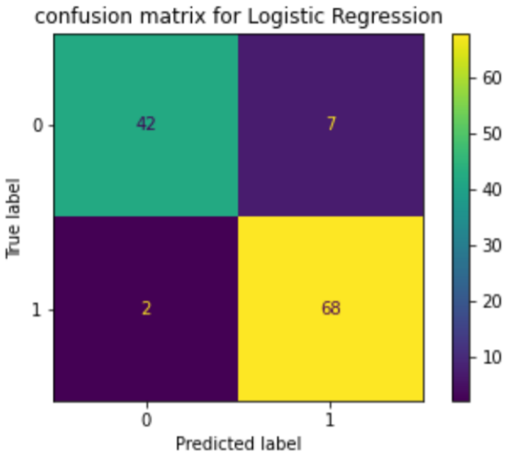
To get a solution of our problem we used three classifiers:

1. Logistic Regression
2. K Nearest Neighbor
3. Decision Tree

- Logistic Regression:
  Applied to three models each has different set of predictors, then evaluated using cross validation score to choose the best one:
  - Model 1:
    - feature: avg_score (Since it best sperate the target variable linearly)
    - Cross Validation: 10 Folds
    - Cross Validation score: 91%
  - Model 2:
    - features: avg_score, Fedu, Medu (top three separators of the target variable linearly)
    - Cross Validation: 10 Folds
    - Cross Validation score: 90%
  - Model 3:
    - features: All Features
    - Cross Validation: 10 Folds

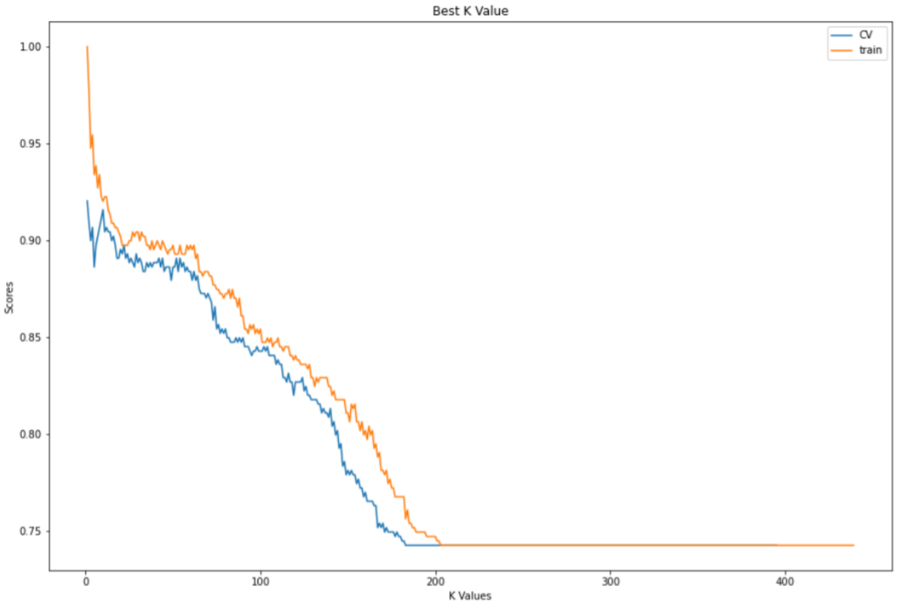
        - Cross Validation score: 93%

Based on the results having all features results best logistic regression classifier, its confusion matrix:



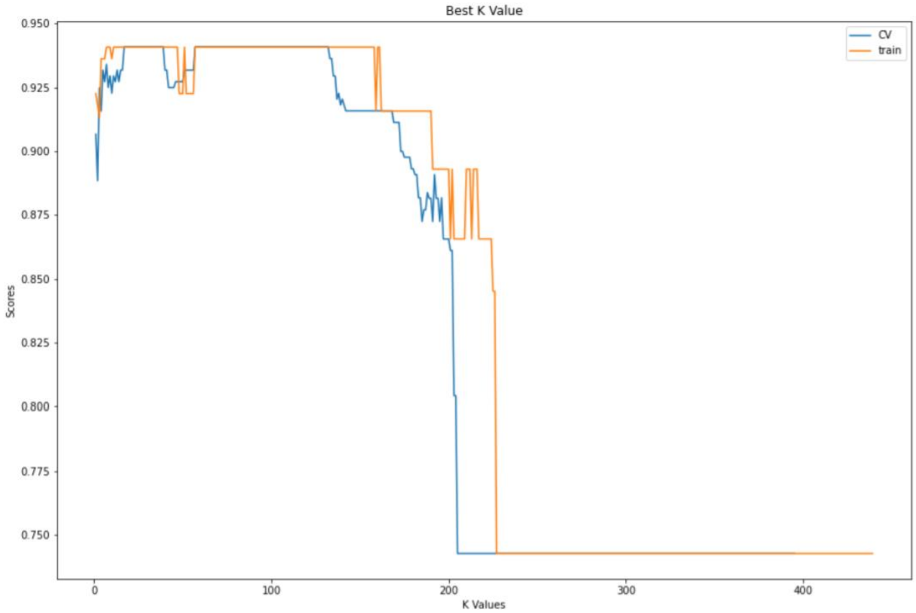
Accuracy on test set = 92%

- K Nearest Neighbor
  Applied to two models each has different set of predictors and k value, then evaluated using cross validation score to choose the best one
    - Model 1:
        - Best K = 1, as it is shown on the below line chart but therefor it has no advantages of the KNN algorithm:



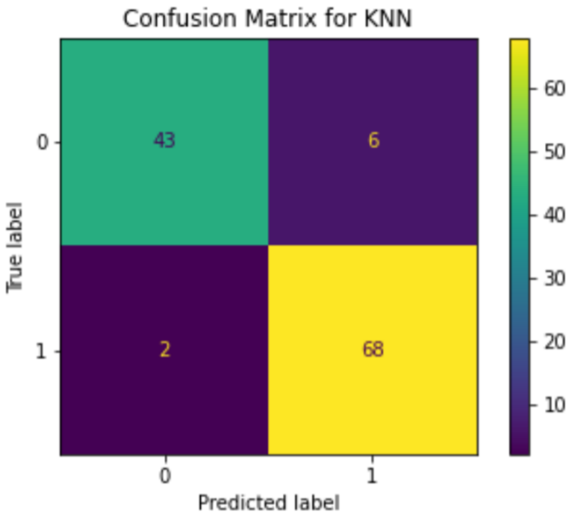
        - features: All features
        - Cross Validation: 10 Folds
        - Cross Validation score: 93%
        - Test set Accuracy: 84%

- Model 2:
  - Best K = 15, as it is shown on the below line chart:



  - feature: avg_score
  - Cross Validation: 10 Folds
  - Cross Validation score: 92%
  - Test set Accuracy: 95%

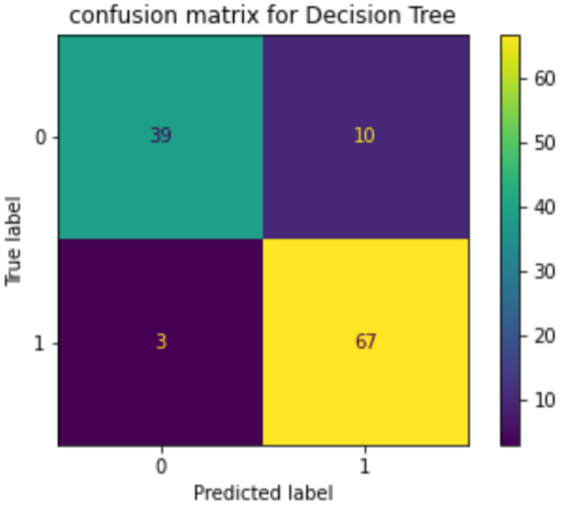
Based on the results having only the avg_score feature it results best on KNN classifier, its confusion matrix:



Accuracy on test set = 93%

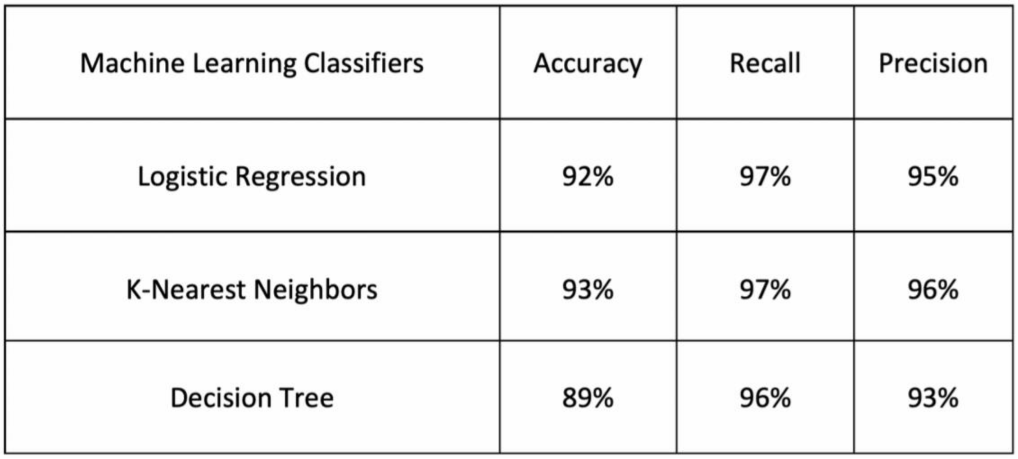
- Decision Tree:
  Applied to two models each has different set of predictors, then evaluated using cross validation score to choose the best one:
  - Model 1:
    - feature: avg_score
    - Cross Validation: 10 Folds
    - Cross Validation score: 92%
  - Model 2:
    - features: All features
    - Cross Validation: 10 Folds
    - Cross Validation score: 94%

  Based on the results having all features result best on Decision Tree classifier, its confusion matrix:



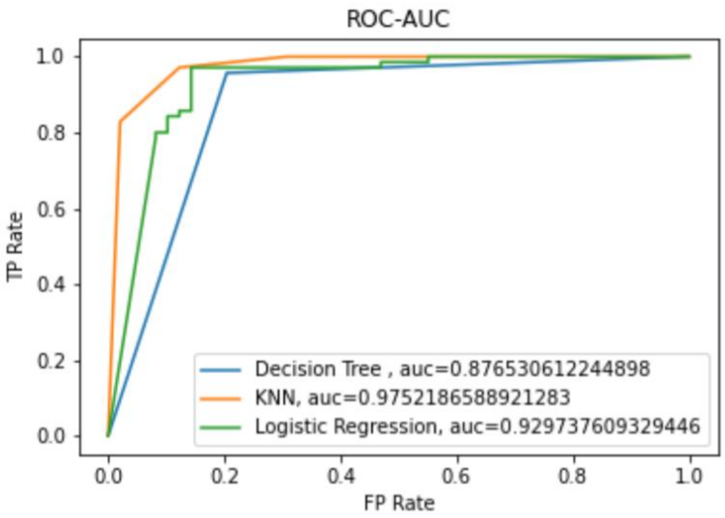
Accuracy on test set = 89%

Now let's compare the best model from each classifier:



Machine Learning Classifiers	Accuracy	Recall	Precision
Logistic Regression	92%	97%	95%
K-Nearest Neighbors	93%	97%	96%
Decision Tree	89%	96%	93%

We pay more constrain on the recall since our interest is to predict if the student will fail more than pass, since the cost of offering more constrain and support to a well performed student is less than ignoring one that has a bad performance. From the above table we can see than KNN is better in terms on Recall and overall performance metrics which is also shown on the below Area Under the Curve chart.



Tools Description:

To achieve our goal, we will analyze and explore the data in Python by using Jupyter, and we will use different packages such as: SKLearn for modeling, Pandas, Matplotlib, Seaborn, and numpy.

Conclusion:

We aim to know the students' grades based on the students' social status and external influences by using machine learning classification models to predict if the student will pass or fail. In this document, we reviewed the problem that we want to solve, a description of the data we will work on, and finally the tools that we will use. The best model resulted is the K nearest neighbors.

References:

[1] https://www.kaggle.com/dipam7/student-grade-prediction