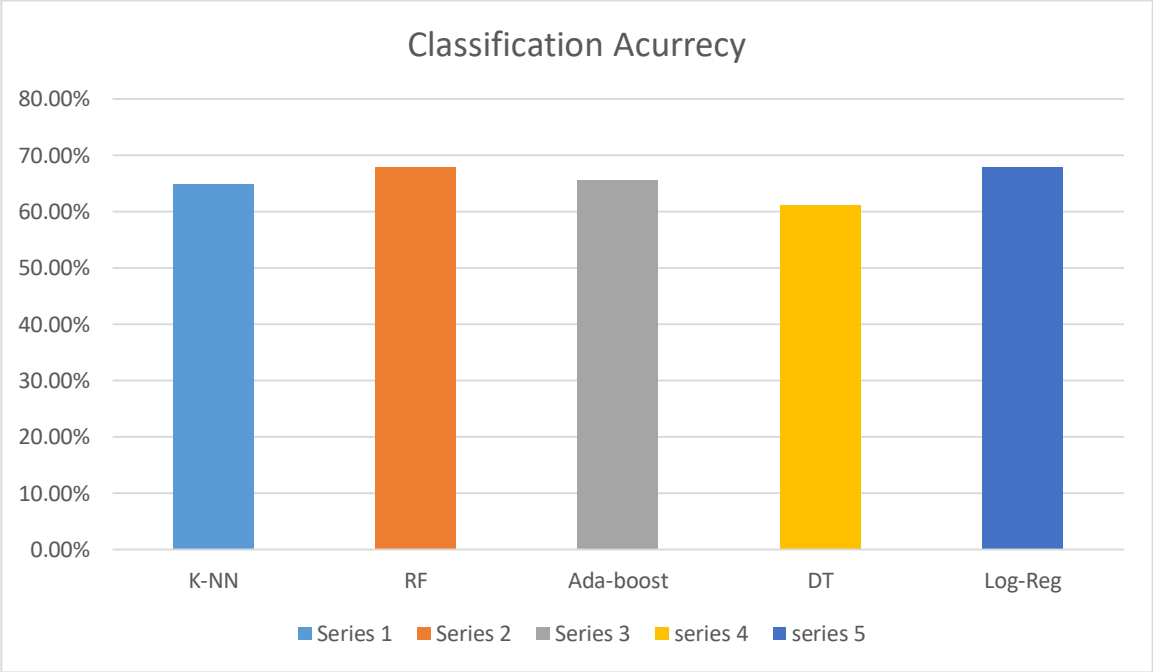
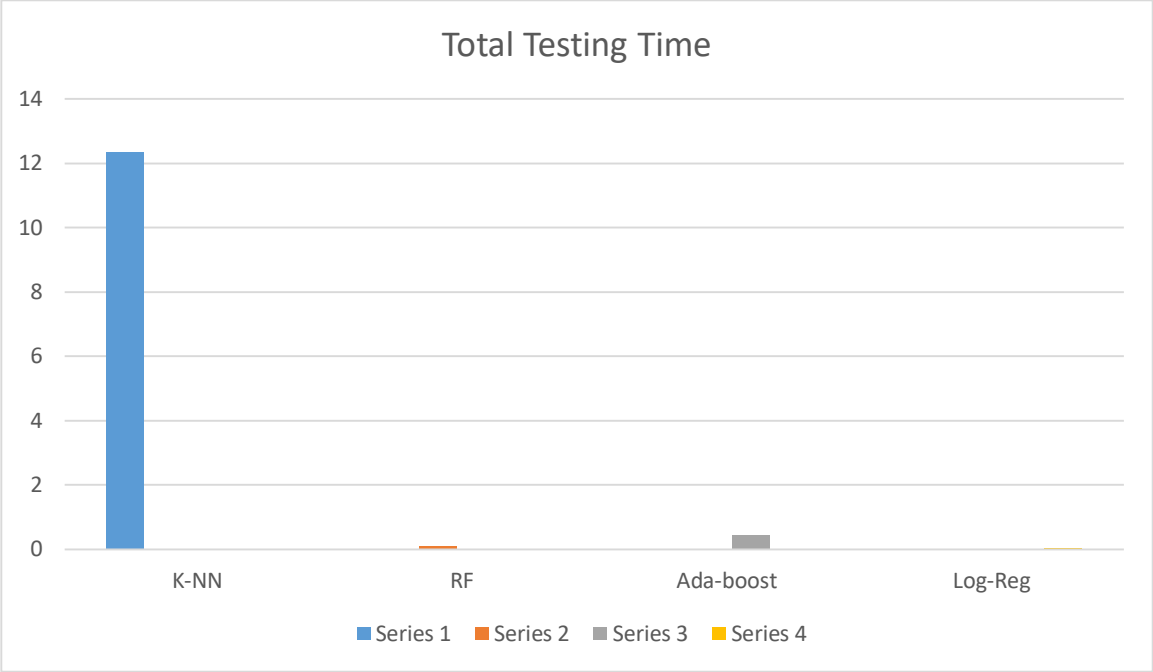
Milestone 2 Report

- Point 1:





- Point 2:

  Since we have numerical features and categorical predicates so we use “Non parametric rank correlation” to determine the correlation between numerical features and how they affect the predictions.

- Point 3:
  - First: K-NN:-
    In this model we use ‘K’ as a “hyper parameter”
    we pick the best ‘K’ value that gives us the best accuracy from 1 to 10 when we increase its value it gives better accuracy after the value of 10 the accuracy increases slightly, so we have the optimal “hyper parameter”.

  - Second: Random Forest:-
    In this model we use “number of estimators” and “max-leaf-nodes” as a “hyper parameter” we pick the best values of our parameters that gives us the best accuracy and time, we choose “number of estimators”=50 and “max-leaf-nodes” = 100.

  - Third: Ada-Boost:-
    In Ada-Boost we use “decision tree” as a model for it and its “hyper parameter” max-depth and we set it by 1 the second “hyper parameter” is “number of estimator” we set it by 100 we pick this values according to the best accuracy.

  - Fourth: log-Regression:-
    In this model we use “Inverse Regularization strength C” as a “hyper parameter”, the smallest value of ‘C’ the highest “Regularization parameter”, so we set it by 0.08.

- Point 4:
  - Conclusion:
    - Different classification techniques creates different accuracy levels using same features and same preprocessing techniques that concludes classification techniques depends a lot on “hyper parameters” to perform accurately .
    - We expected to have an 80% accuracy but we get 67% accuracy.
    - That was disapproved because of difficult preprocessing steps due to a lot of missing data and the massive range of difference between some features values.