5/13/2023

**Business Analytics Assignment**

Section: 8-B

Instructor: Dr. Rauf Ahmed Shams

Name: Saman Khan

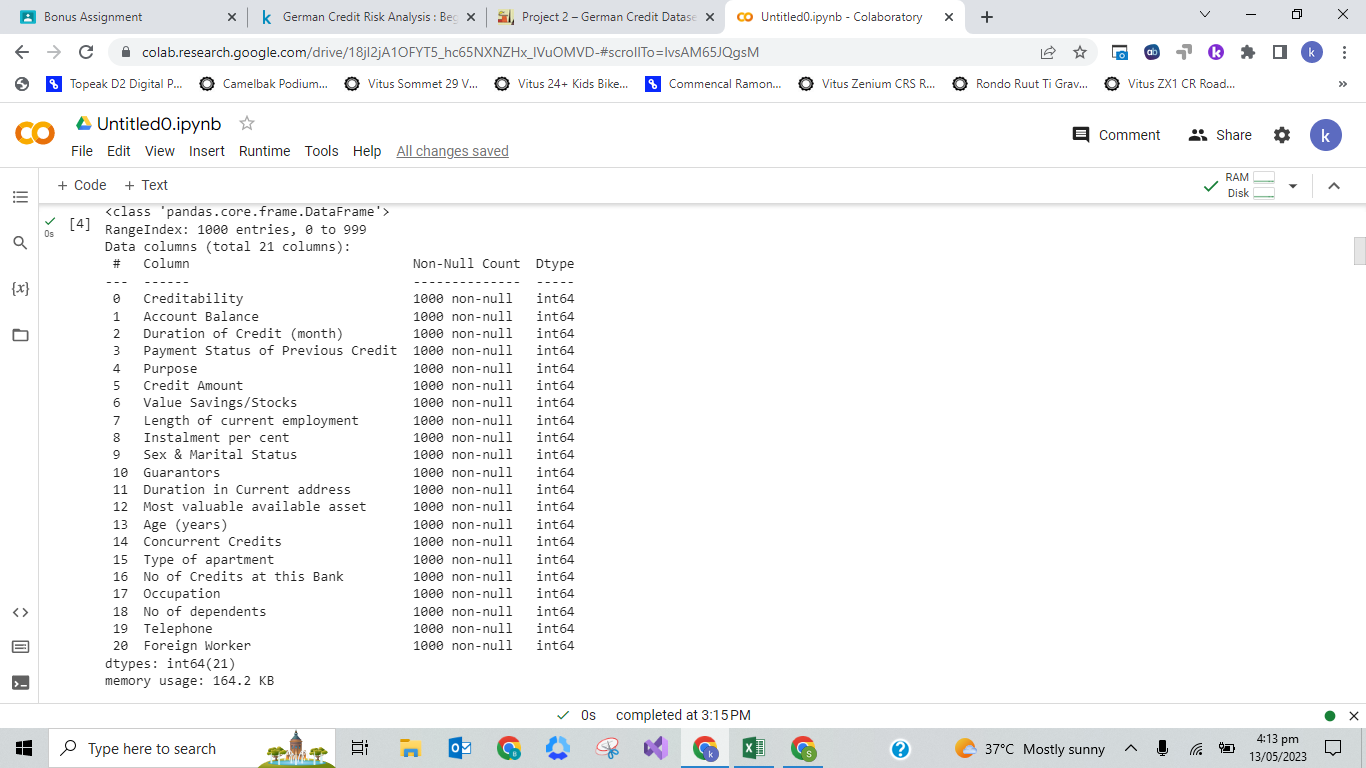
ID: 19K-0354

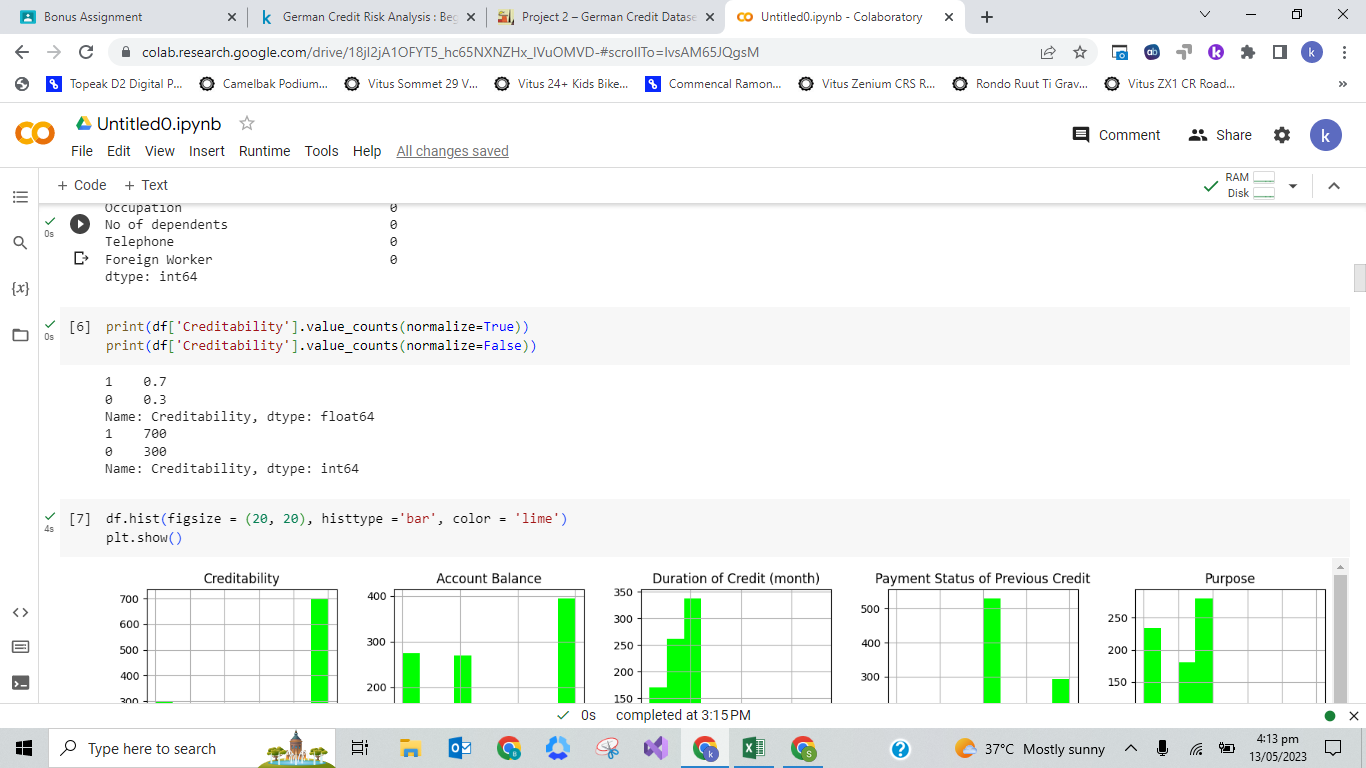
# **Introduction**

# The data offers information about the likelihood of loss due to a borrower's failure to repay a loan or meet contractual commitments. If a corporation extends credit to its customers, there is a danger that the customers would not pay their bills. Risks are classified into two types: good risk and bad risk. A good risk is an investment that one believes will be profitable. Good risks are thought to be extremely likely to be repaid. A bad risk loan is one that is unlikely to be repaid due to a poor credit history, low income, or other factors. A negative risk raises the lender's risk and the possibility of the borrower defaulting. The goal of this programme is to determine if a person is a good or bad credit risk depending on their attributes. There are 1000 entries in the dataset, with 20 independent variables and 1 target variable. Each entry indicates a person who accepts a bank's credit. According to the collection of attributes, each person is classed as a good (1) or bad (0) credit risk.

# **Data Overview**

The data given is already preprocessed and cleaned and contains no null values. Out of 1000 rows, 700 contains info about credible and 300 contains info regarding non credible.





# **Exploratory Data Analysis (EDA)**

# 

# 

# 

# 

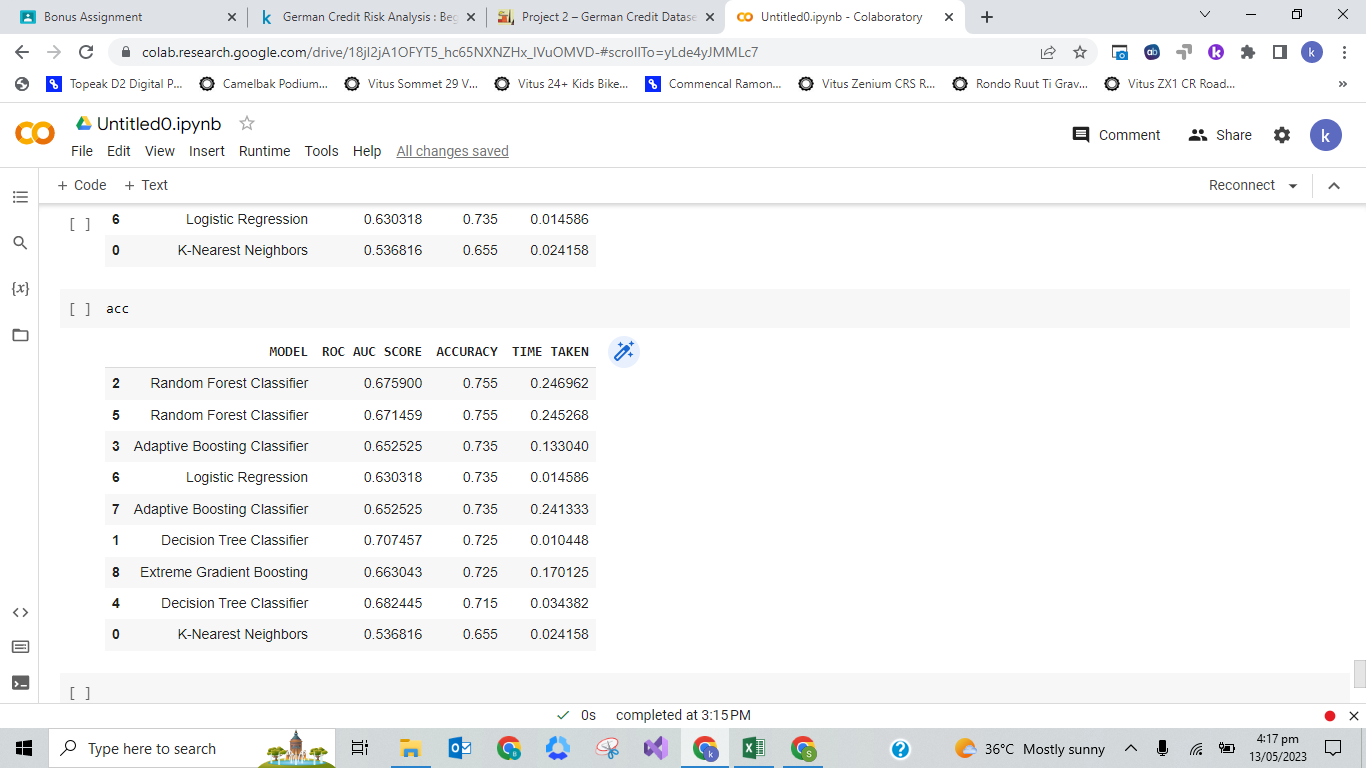
# 

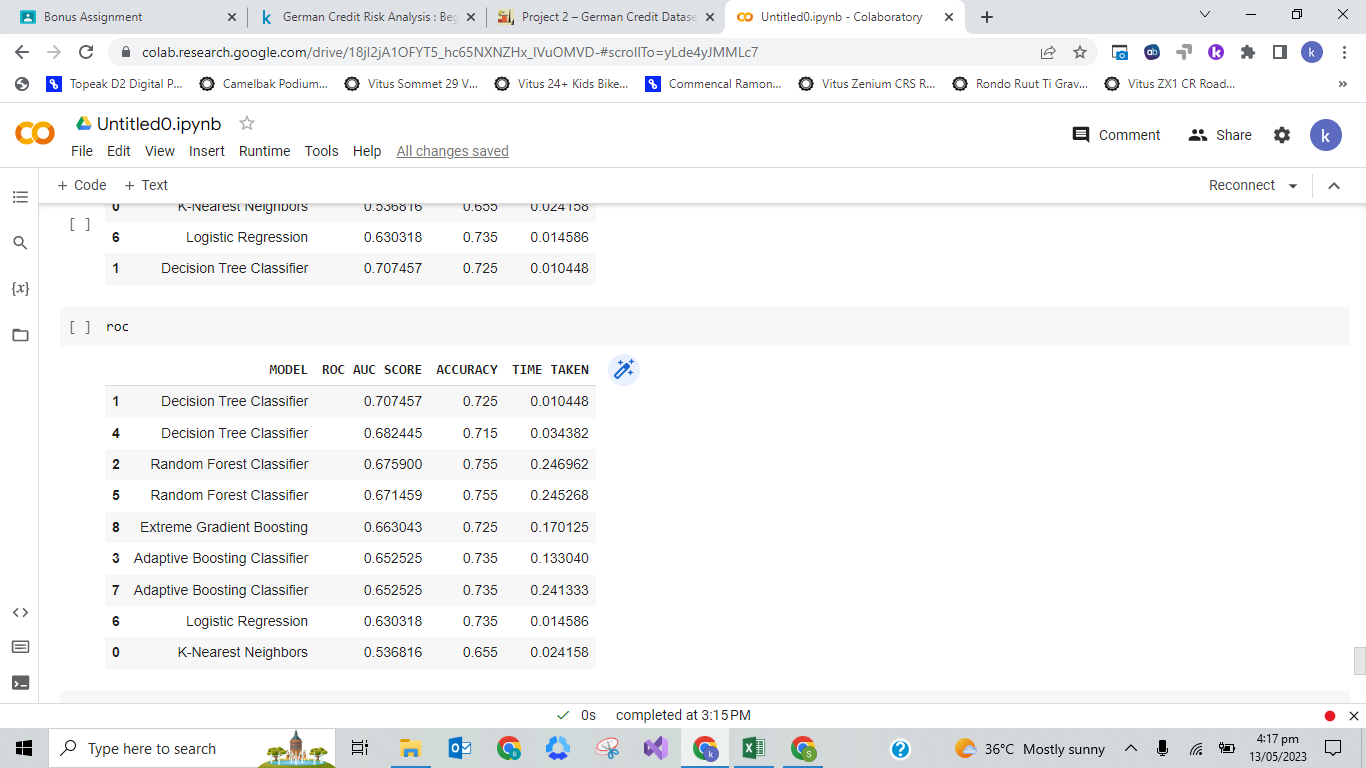
# **Model Implementation**

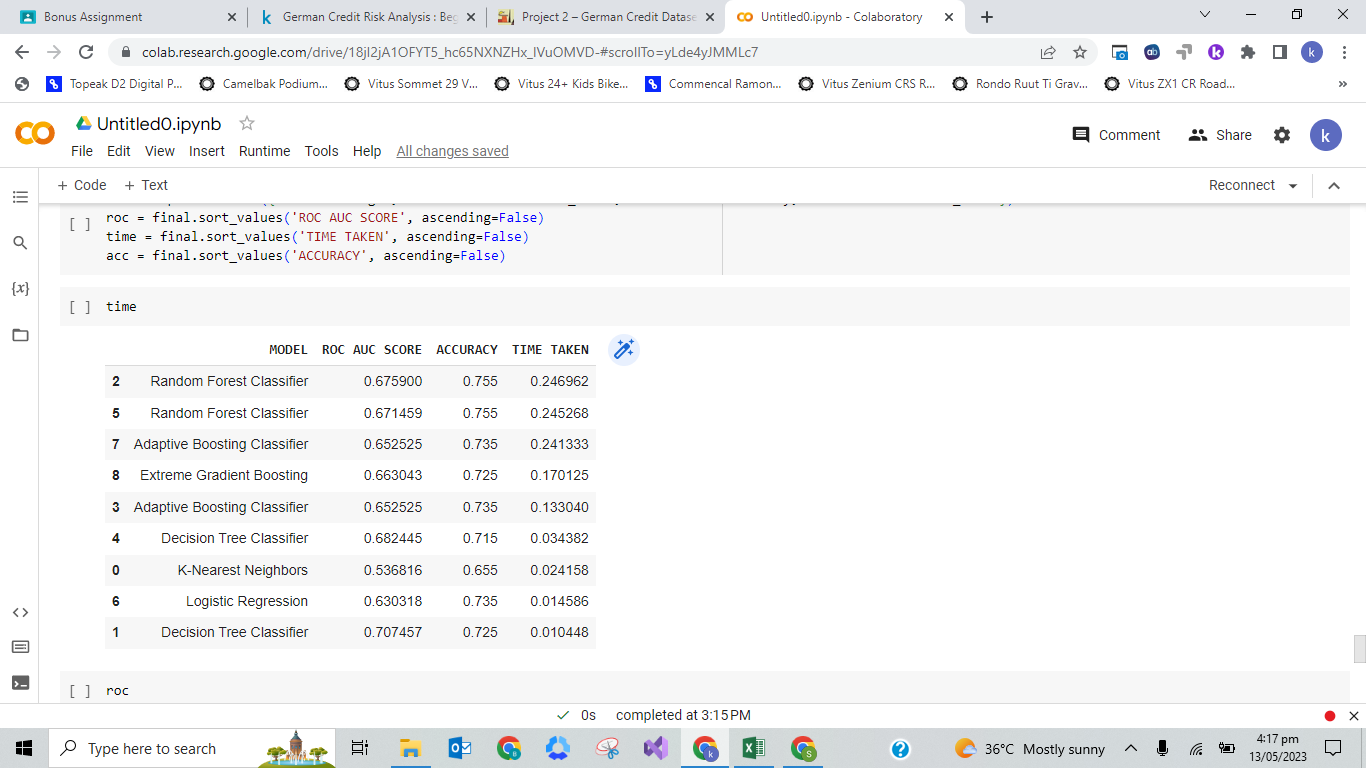
The models implemented are: Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Ada Boost, and XGB. To satisfy both sensitivity and specificity requirements of evaluation of model, ROC score is a good measure. It is essentially used to depict the trade-off between sensitivity and specificity. The ROC score gives a value b/w 0 to 1. The higher the score, the better the model performs. A total of six models are implemented in this system.

Since area under the ROC curve (also termed as AUC score) was our primary measure to determine the best model, Decision Tree Classifier and Random Forest Classifier seems to offer the best results. With AUC score of 0.71 and 0.68 respectively, these outperform the rest. On the other hand, K-Neighbors Classifier can be termed as a model that fails to perform because its AUC score lies near 0.6.

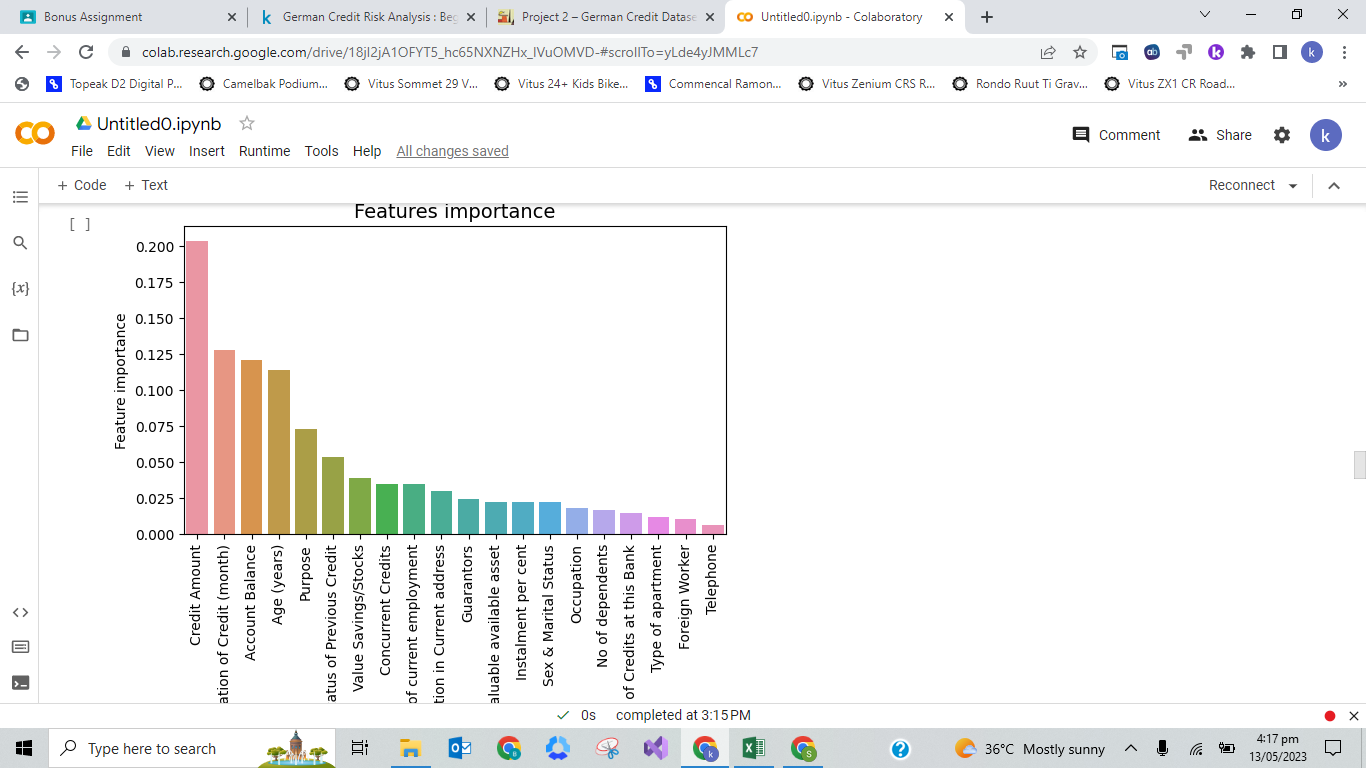
Accuracy is also an equally good measure to determine the performance of models. As per accuracy obtained, Random Forest Classifier give the highest value that is 0.76. On the basis of time elapsed during training and testing of models, Logistic Regression performs the best. However, time alone is not a good measure to obtain performance of the models. Often, complex models with good results require time to train without over fitting or underfitting. The time taken by Random Forest Classifier is about 0.25 seconds which is the minimum time taken by any model. Hence, we can say that Random Forest Classifier outperforms the rest of the models.



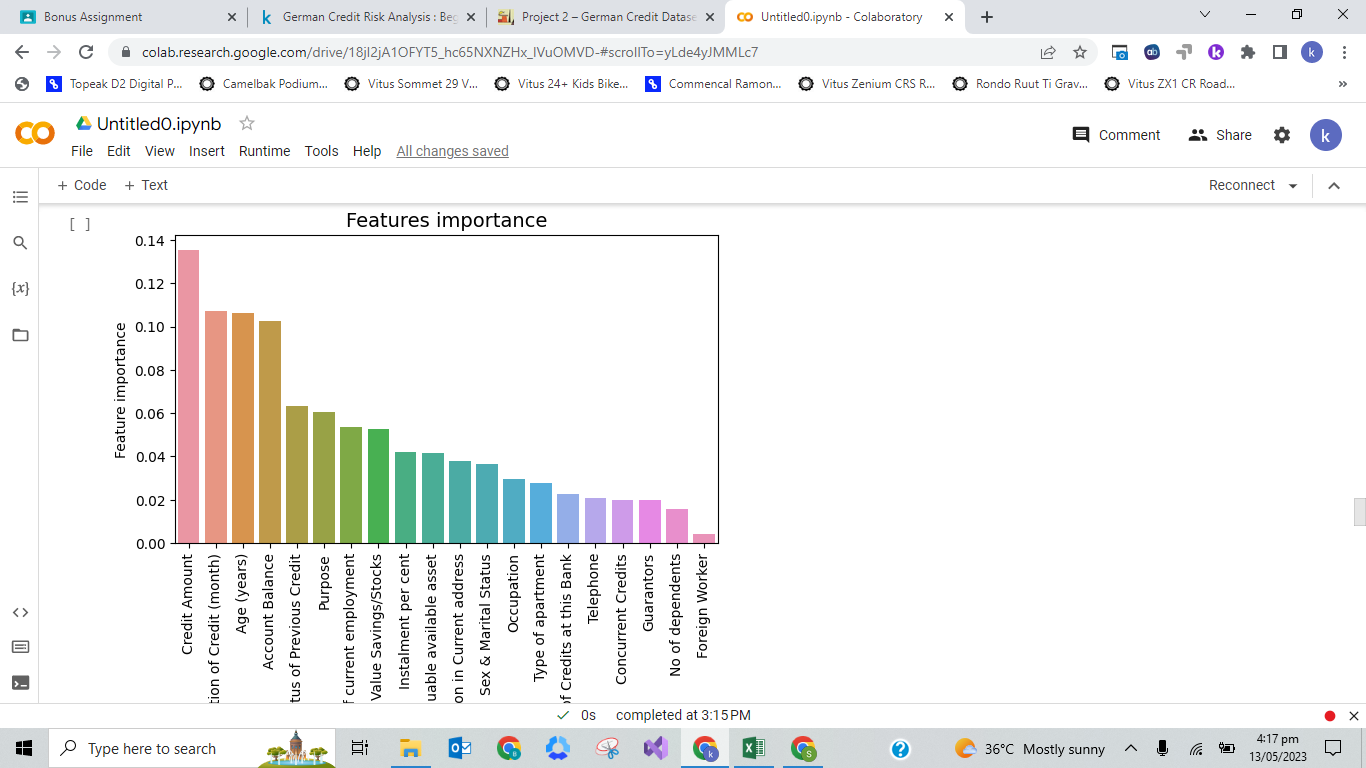




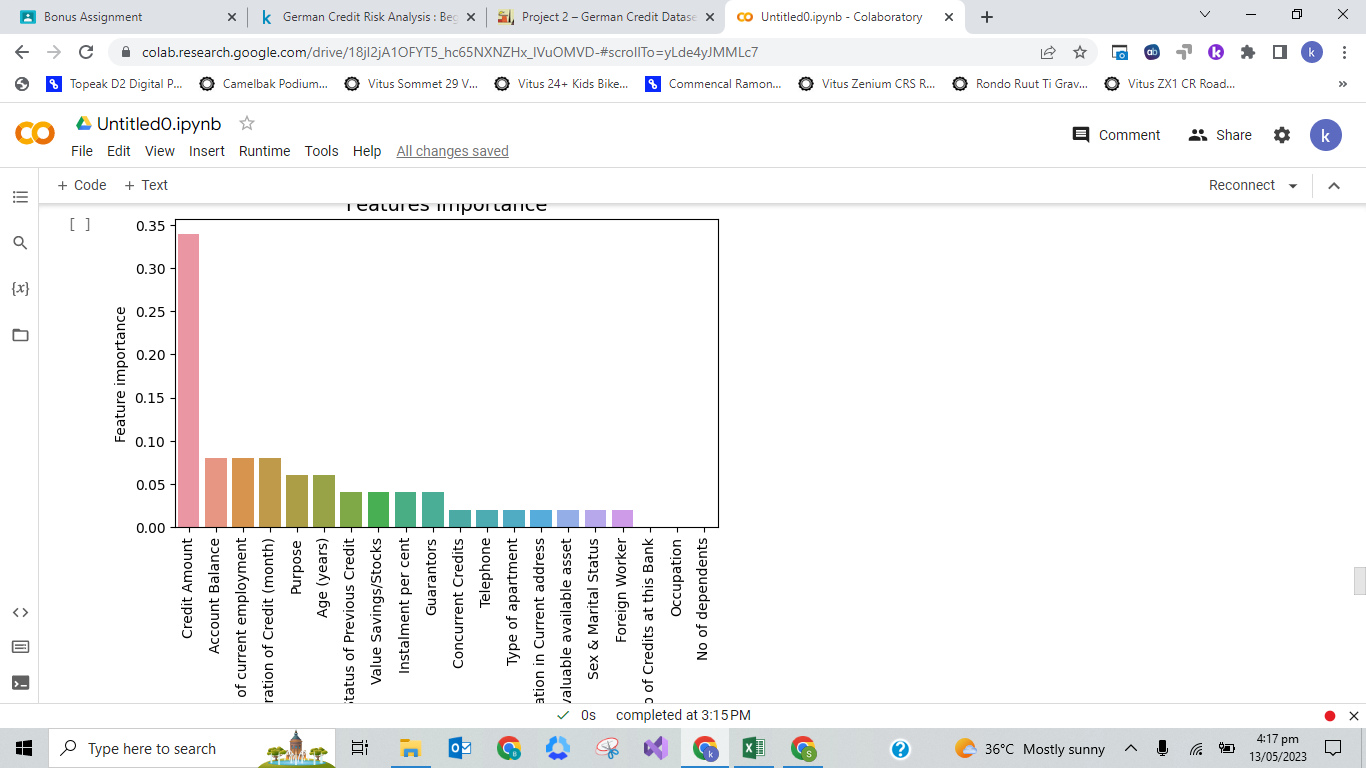
* **Feature Importance of Decision Tree Classifier**

****

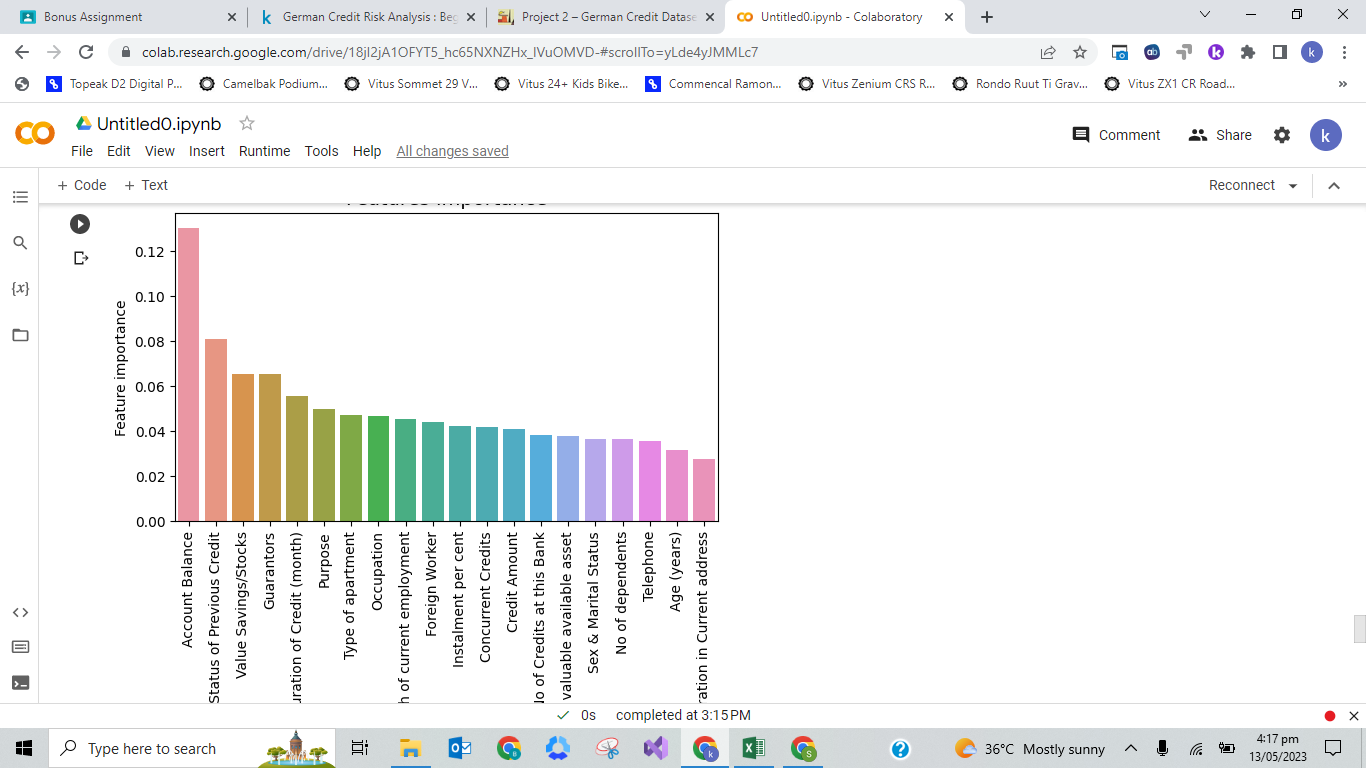
* **Feature Importance of Random Forest Classifier**

****

* **Feature Importance of Ada Boost Classifier**

****

* **Feature Importance of XGB Classifier**

****