**Faculty of Computers Benha University**

**& Artificial IntelligenceLogo

Description automatically generated**

7/3/2021

*Supervised By:*

**Prof. Dr. Hala Zayed**

**Eng. Abdulwahab Almestekawy**

*Team Members:*

* **Rana Farid Shawky.**
* **Rola Hossam El-Din.**
* **Rawan Hossam Ibrahim.**
* **Abdel Moniem Elsayed Abdel Moniem.**

**HEART FAILURE PREDICTION**

**Table of Contents**

1. Abstract ............................................................................................................................... 1

2. Project Description .............................................................................................................. 2

3. Running Project ………………………………………………………………………………………….…………….……. 3

*3.1 Data pre-processing ……………………………………………………………………….……………….….… 3*

*3.2 Decision Tree …………………………………………………………………………….………………….……….. 6*

*3.3 Logistic Regression …………………………………………………………………………………………………. 7*

*3.4 Support Vector Machine ………………………………………………………………………………………….….. 7*

*3.5 Ensemble Learning……………………………………………………………………………………………………. 10*

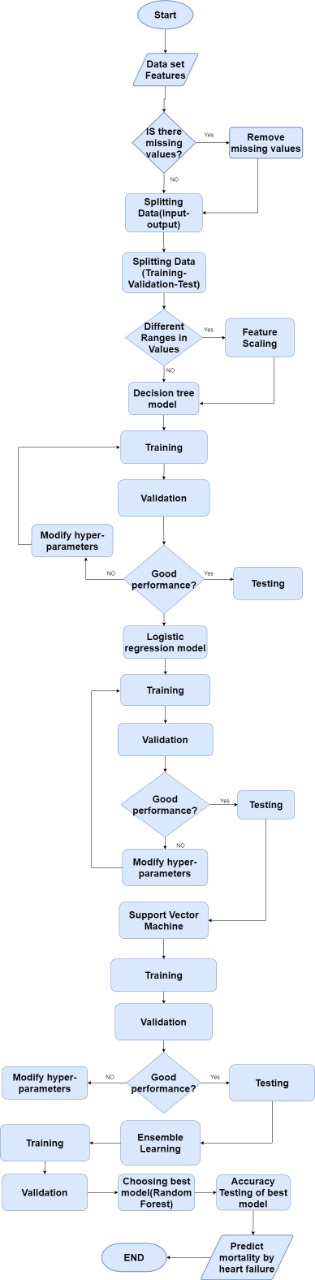
4. Results ……………………………………………………………………………….………………………………….………12

5. Conclusion ……………………………………………………………………………….………………………………….. 13

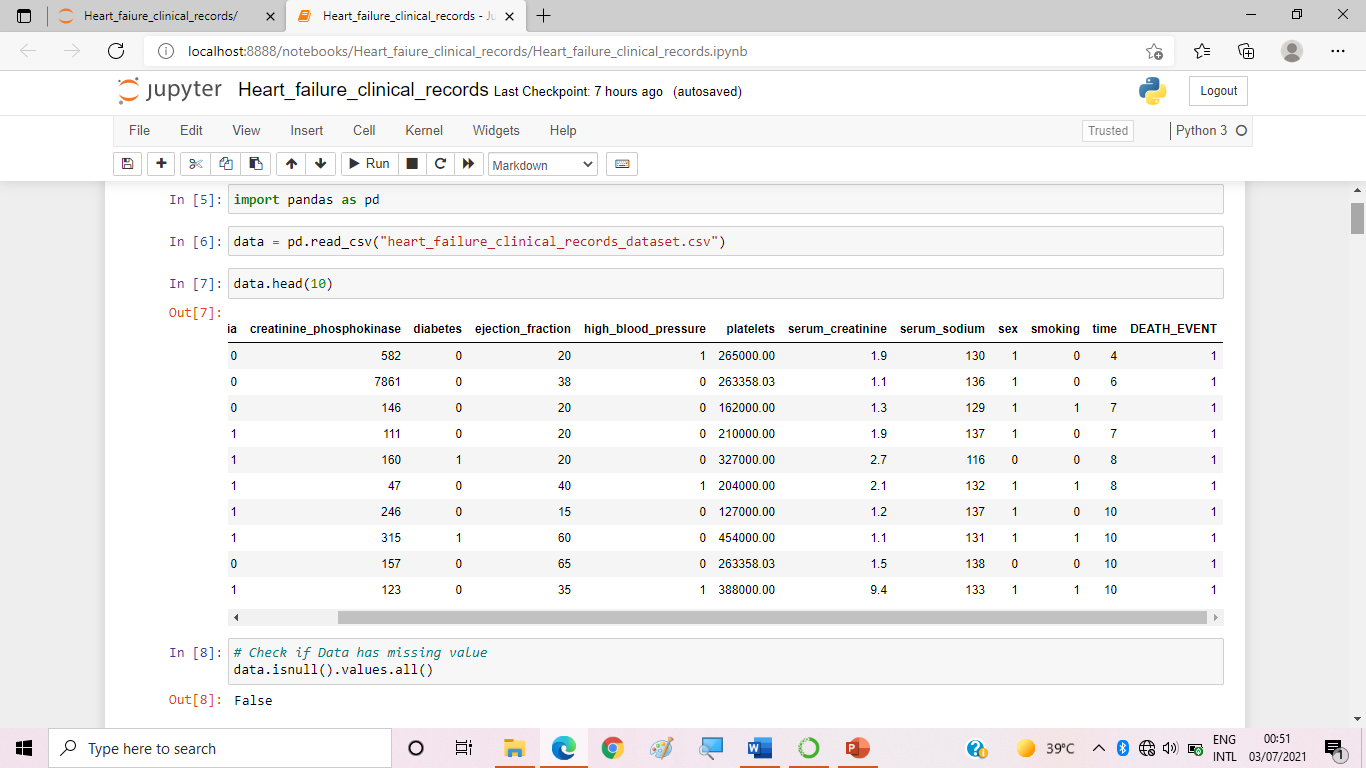
References ..............................................................................................................................14

* **Abstract**

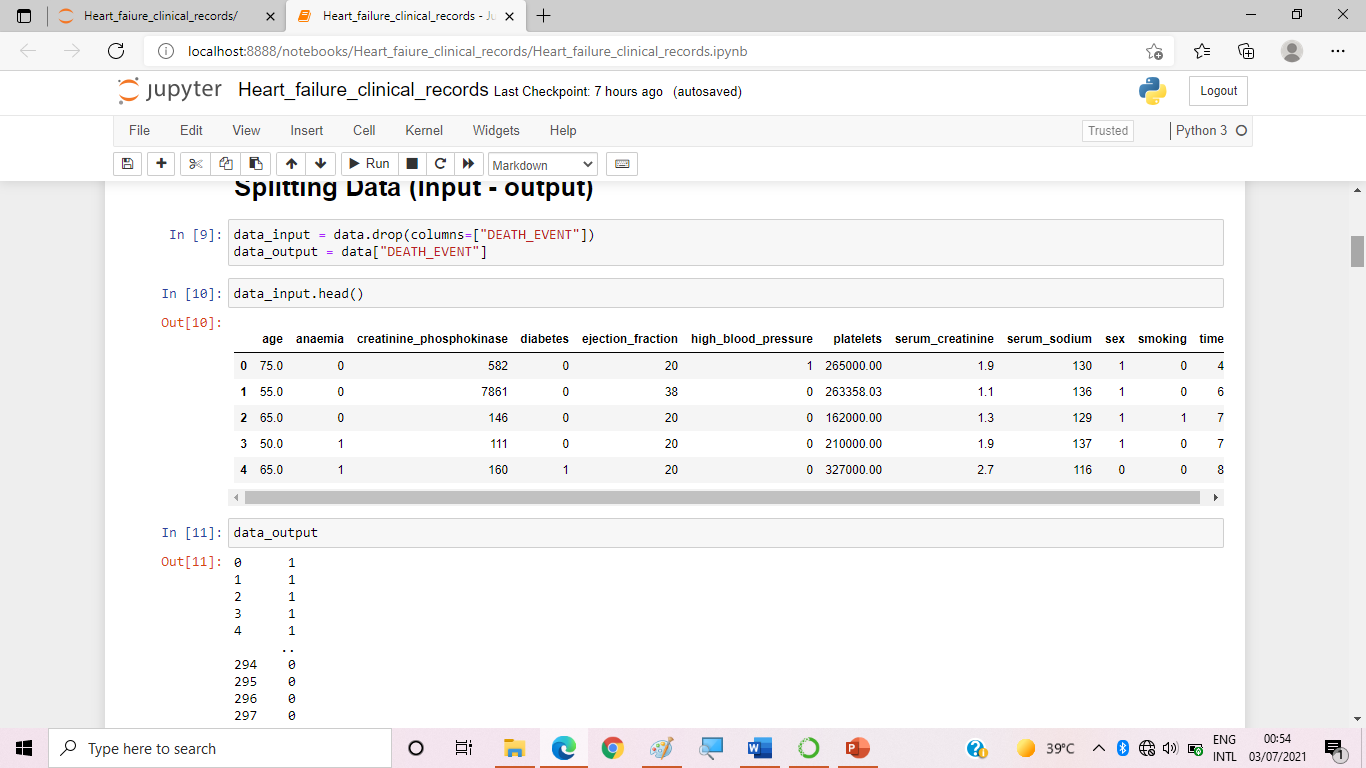
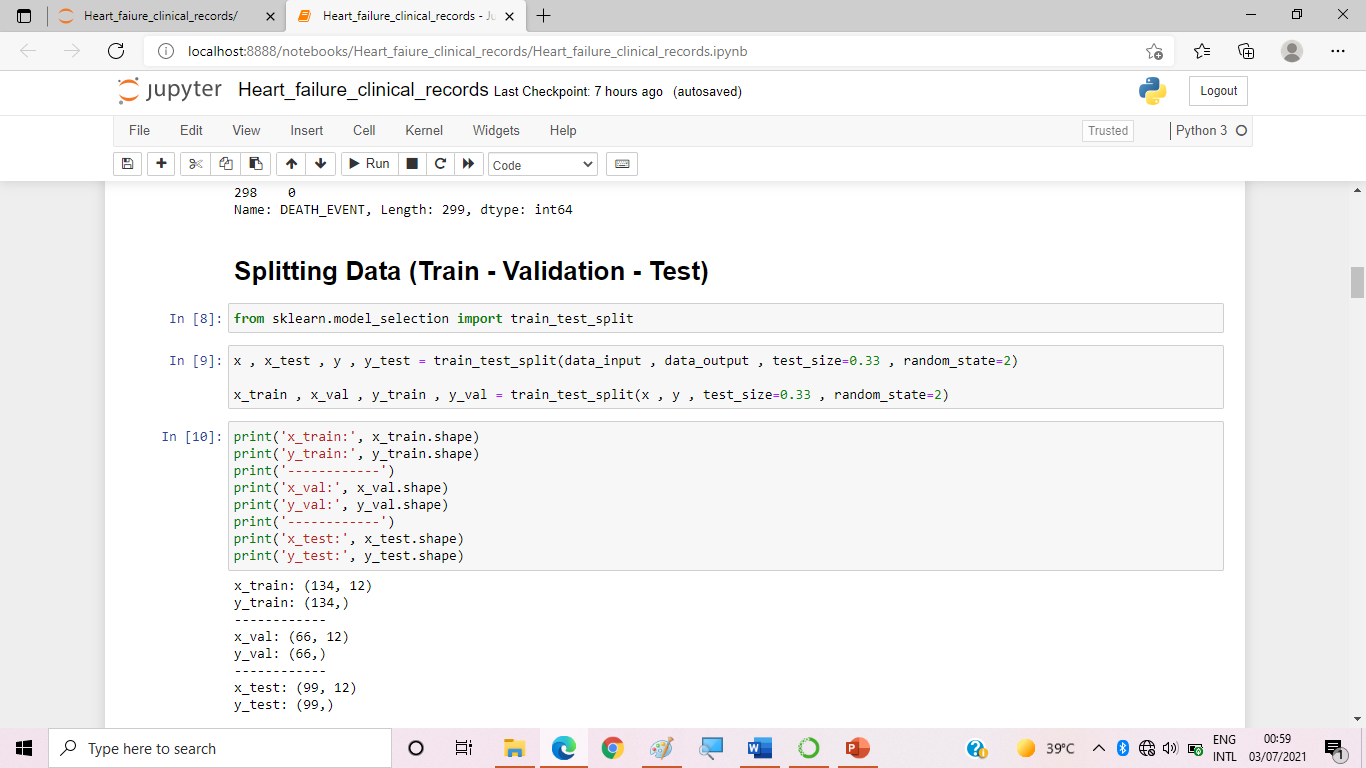
Altogether, cardiovascular diseases cause the death of approximately 17 million people worldwide annually, with fatalities figures on the rise for first time in 50 years the United Kingdom [1]. In particular, heart failure occurs when the heart is unable to pump enough blood to the body, and it is usually caused by diabetes, high blood pressure, or other heart conditions or diseases [2]. Available electronic medical records of patients quantify symptoms, body features, and clinical laboratory test values, which can be used to perform biostatistics analysis aimed at highlighting patterns and correlations otherwise undetectable by medical doctors.

* **Project Description (Flow Chart)**
* **Running Project**

1. **Data pre-processing**

Data pre-processing in Machine Learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. It refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training models. In simple words, data pre-processing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format.

**Splitting Data :**

1. Splitting data into input & output
2. Splitting data into training , validation and test

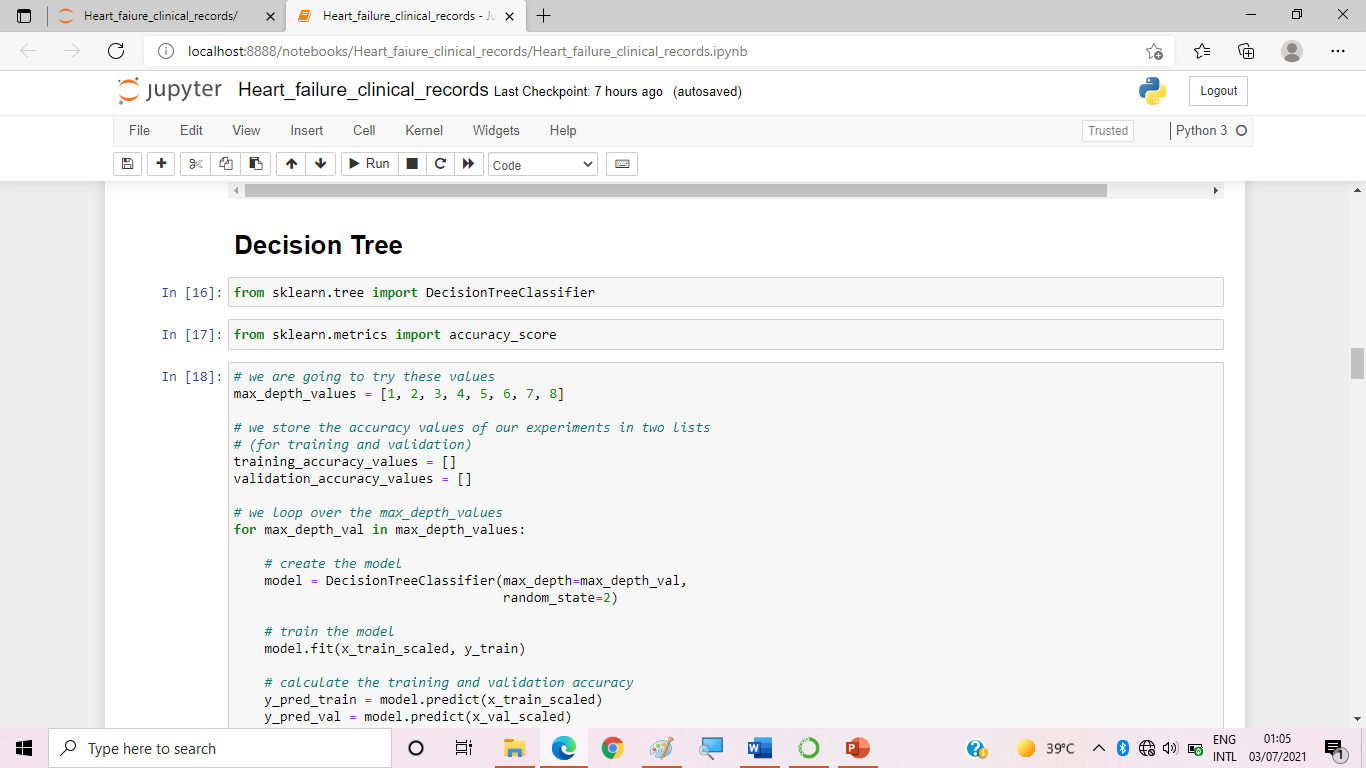
**Applied Pre-processing methods:**

Feature Scaling (Normalization)

Graphical user interface, text, application, email

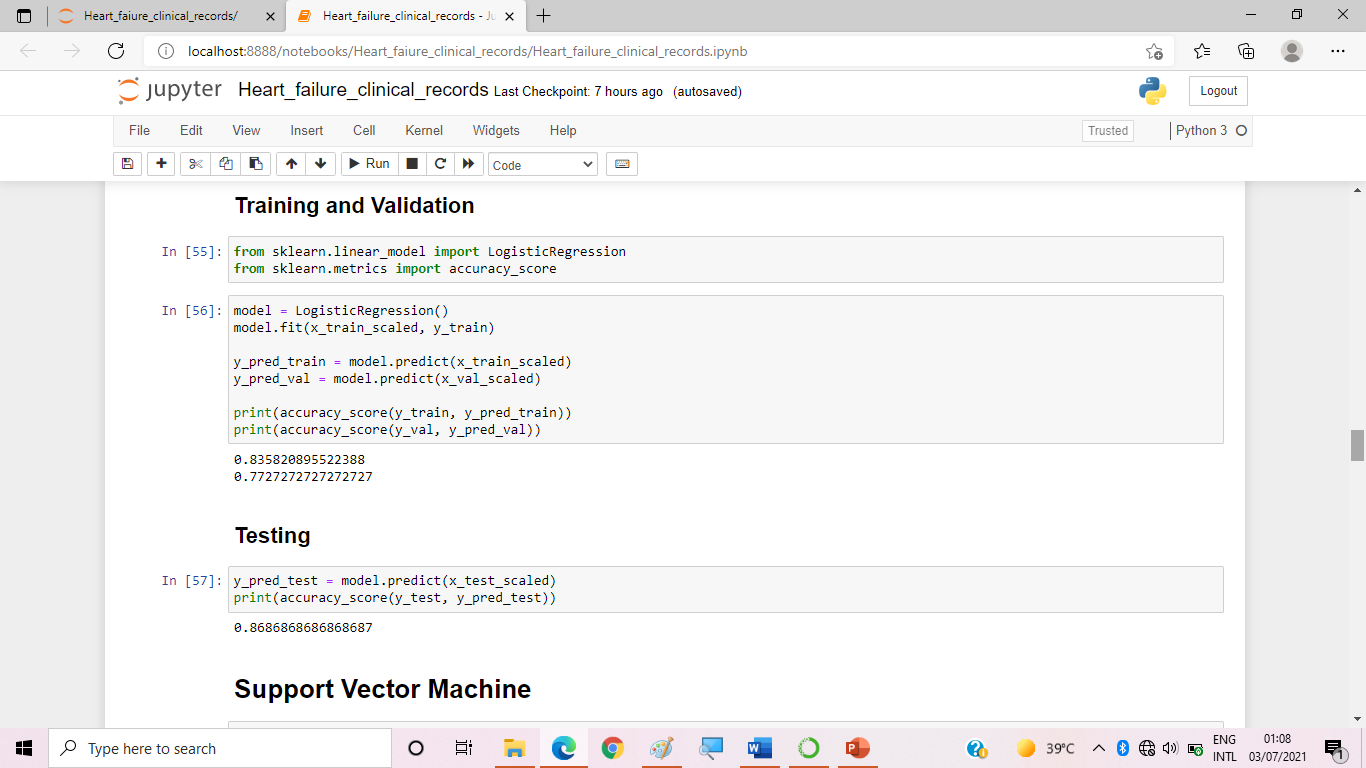
Description automatically generatedWe have only one problem which need to be preprocessed is that values of features are in very different ranges so we will apply feature scaling that limits the range of variables so that you can compare them on common grounds.

1. **Decision Tree**

Graphical user interface, application

Description automatically generatedIs a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

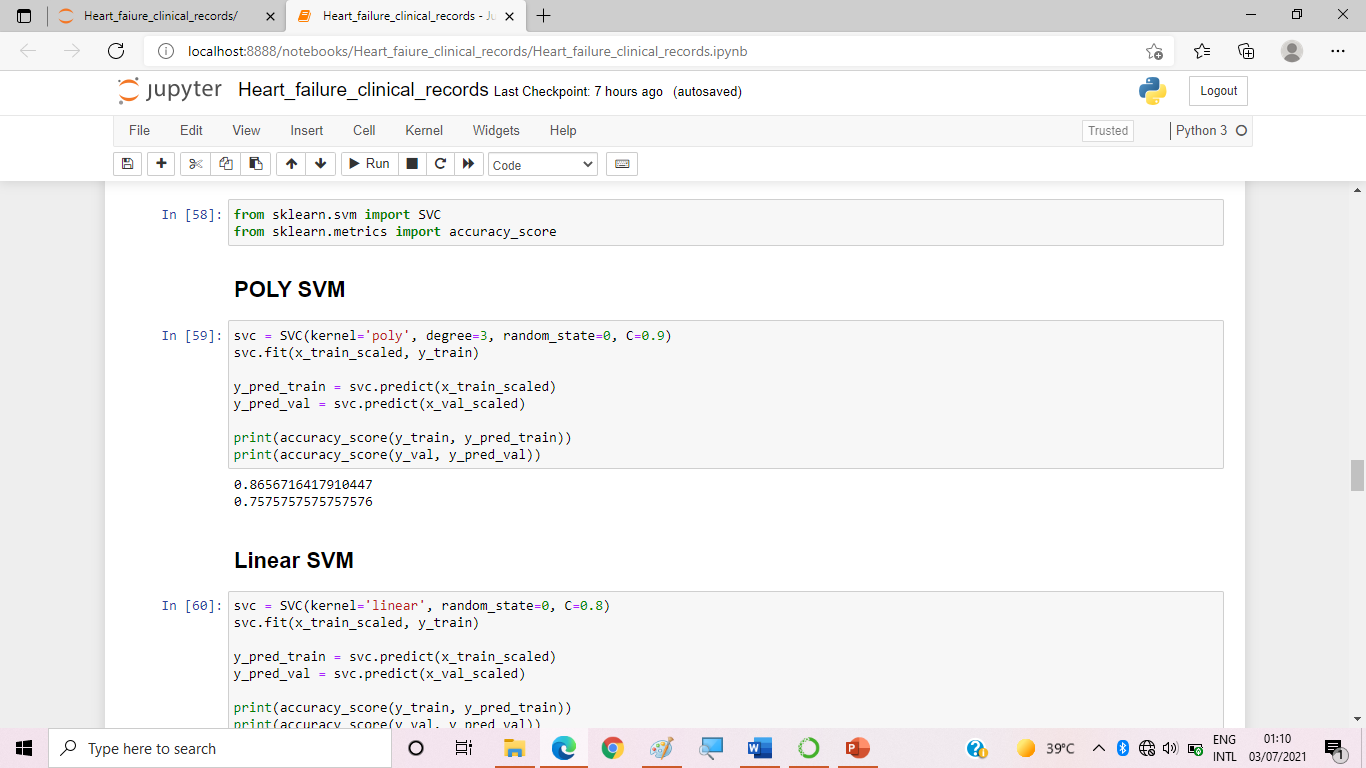
1. **Logistic Regression**

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. It is fast and relatively uncomplicated, and it’s convenient for you to interpret the results. Although it’s essentially a method for binary classification, it can also be applied to multiclass problems.

1. **Support Vector Machine**

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

SVM maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Popular Kernel Functions :

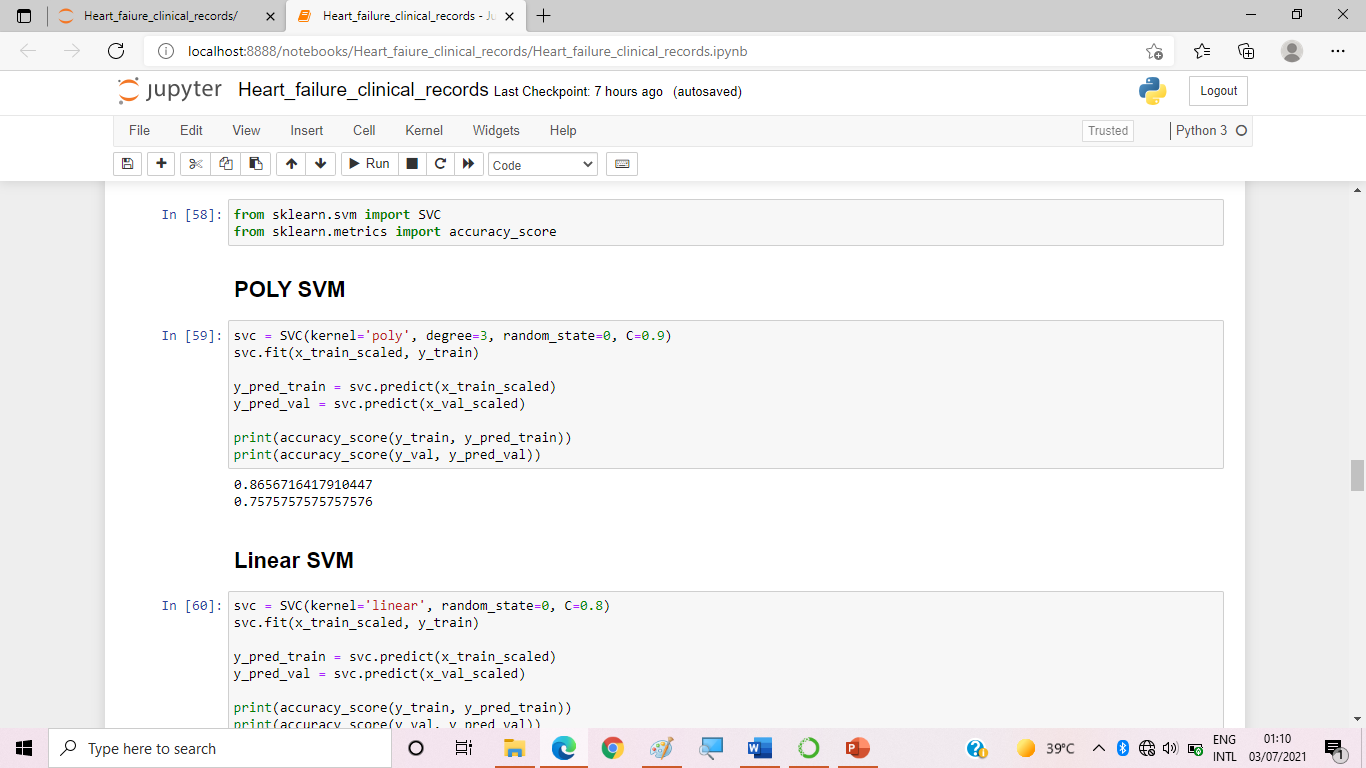
The goal of SVM is to design a hyperplane that classify all training vectors into two classes. The SVM takes these data as input and the output is a line that separates them.

For our project we used three kernels for SVM:

1. **Polynomial SVM**

A polynomial kernel is defined by the following equation:

K(x1, x2) = (x1 . x2 + 1)d,

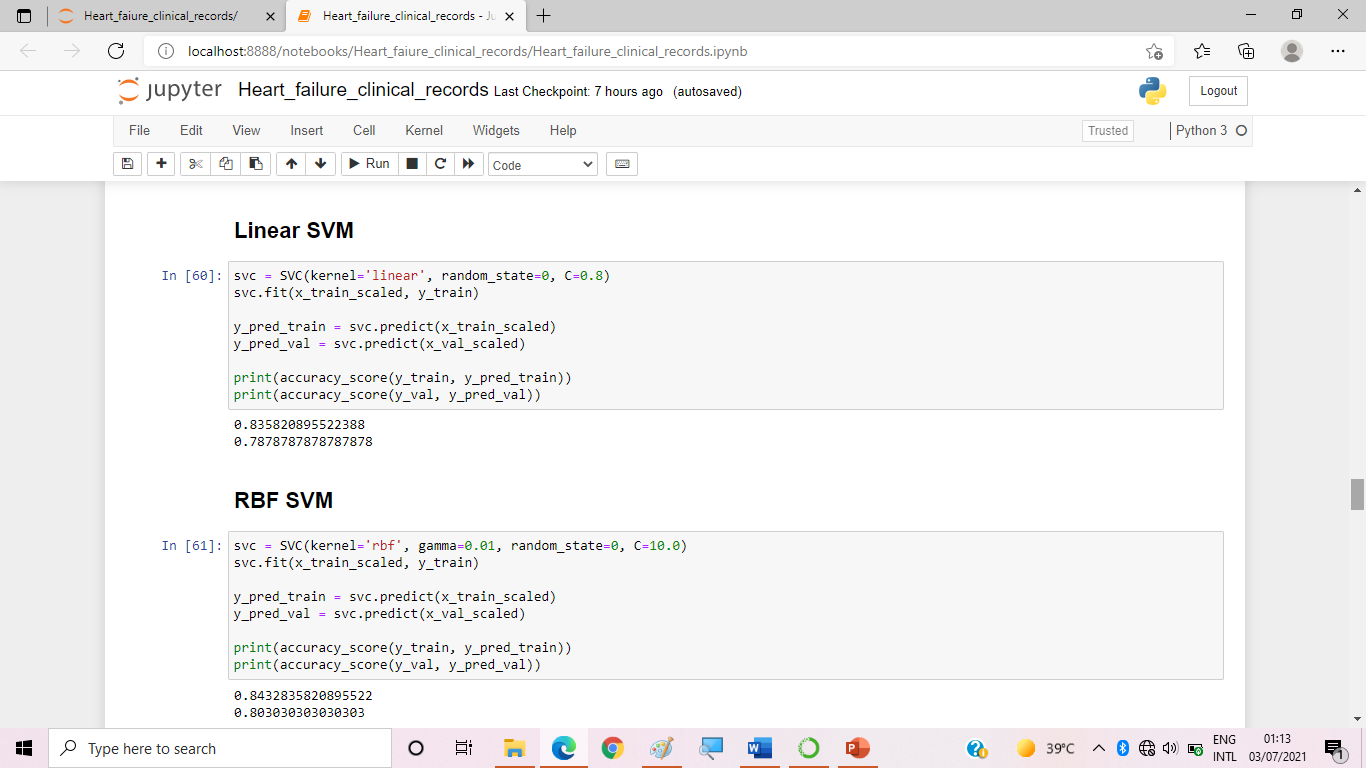
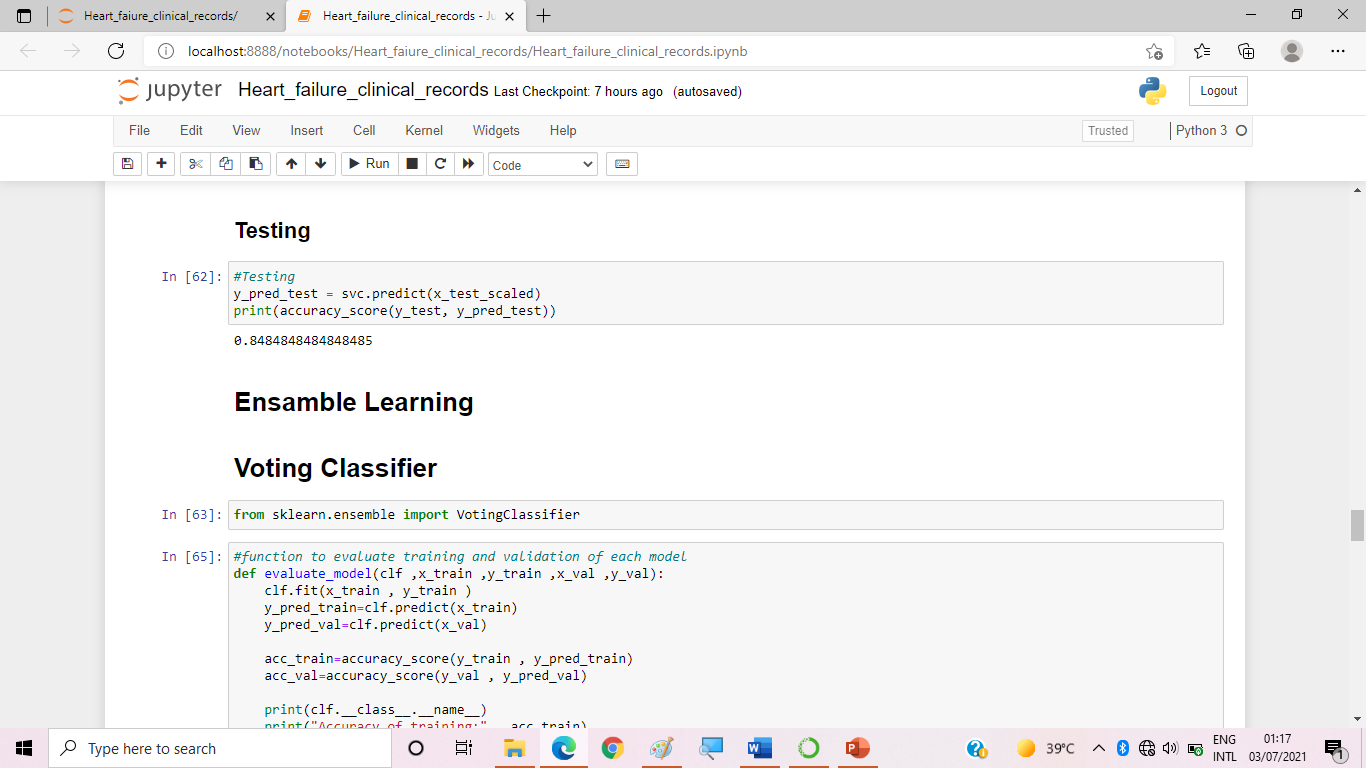
Where, d is the degree of the polynomial and x1 and x2 are vectors.

1. **Linear SVM**

Graphical user interface, text, application

Description automatically generatedLet us say that we have two vectors with name x1 and Y1, then the linear kernel is defined by the dot product of these two vectors: K(x1, x2) = x1 . x2

1. **RBF SVM**

The RBF is by far the most popular choice of kernel types used, mainly because of their localized and finite responses across the entire range of the real x-axis.

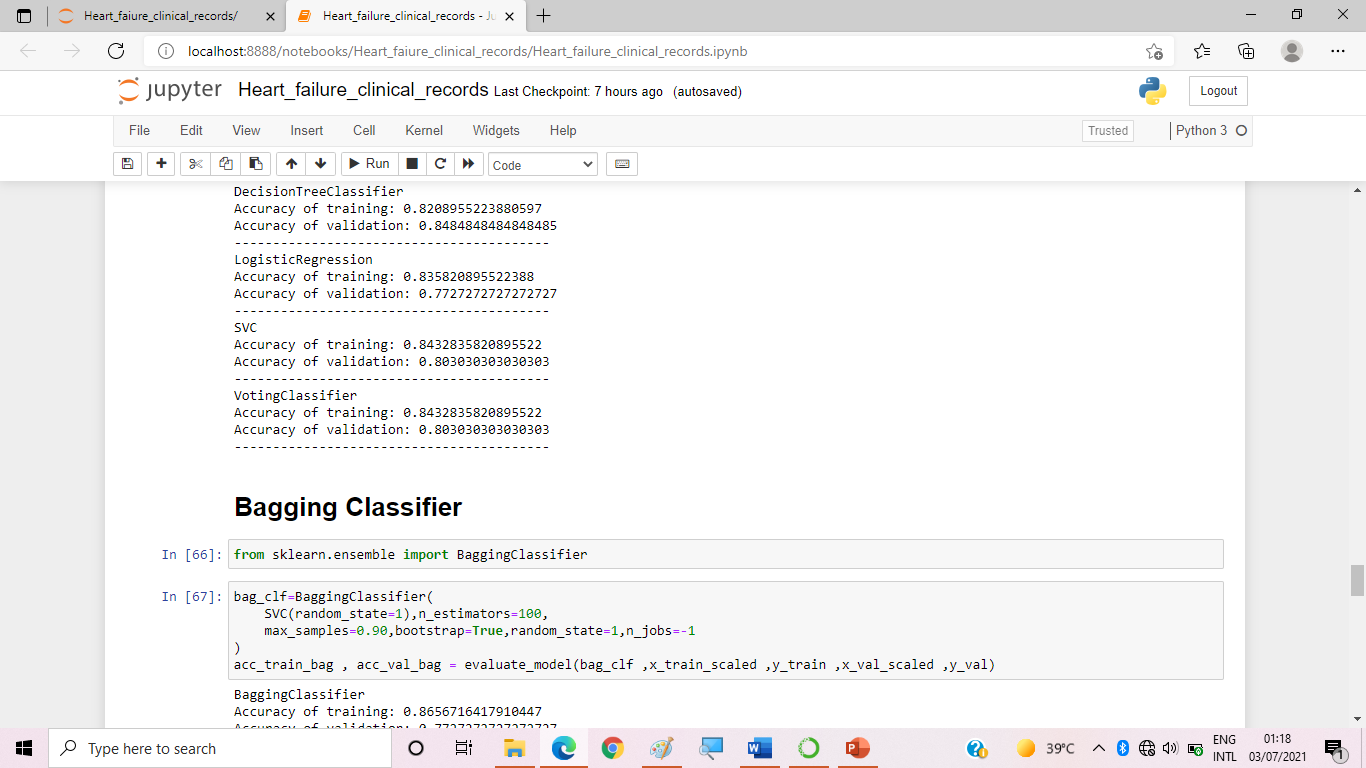
1. **Ensemble Learning**

It is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.).[3]

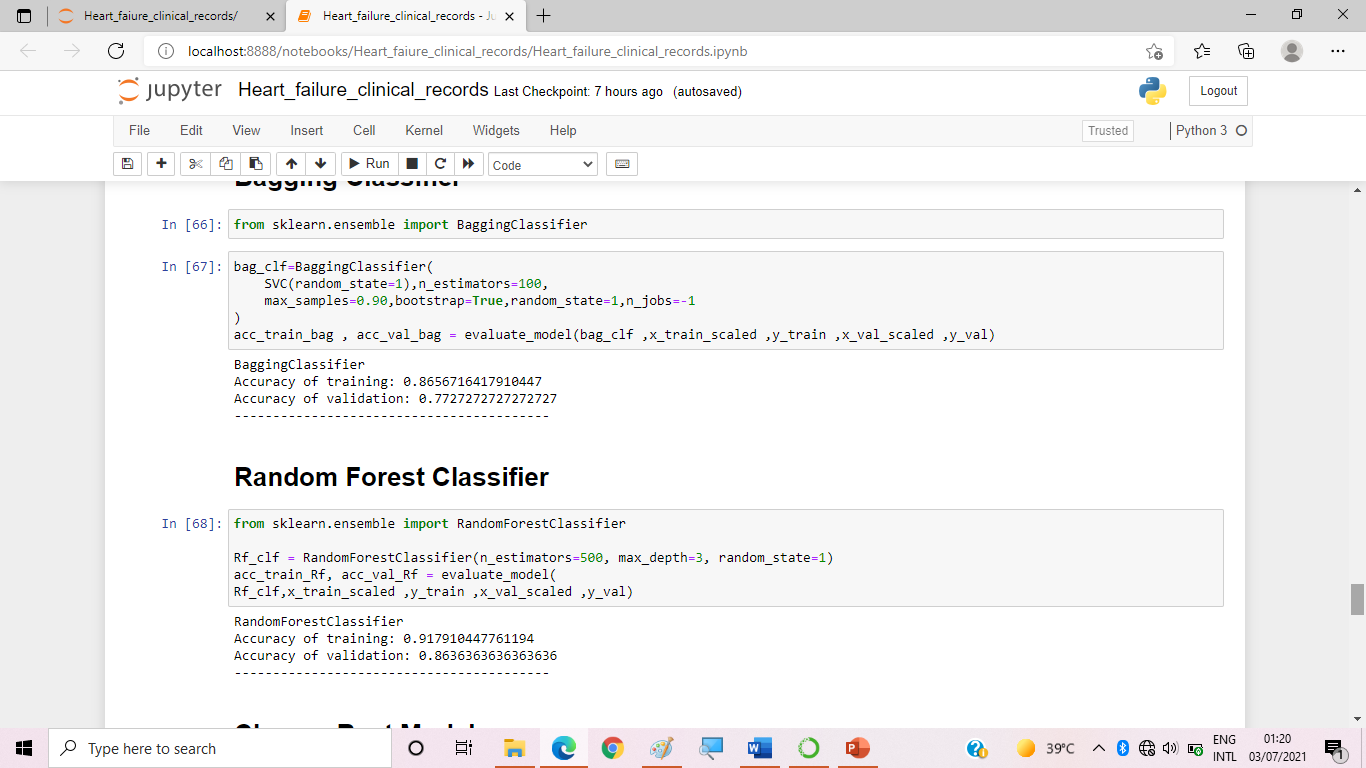
**For our project we used three models for ensemble learning:**

1. **Voting Classifier:**

We applied it to choose the best model (by majority) among three models (logistic, SVM , Decision tree).

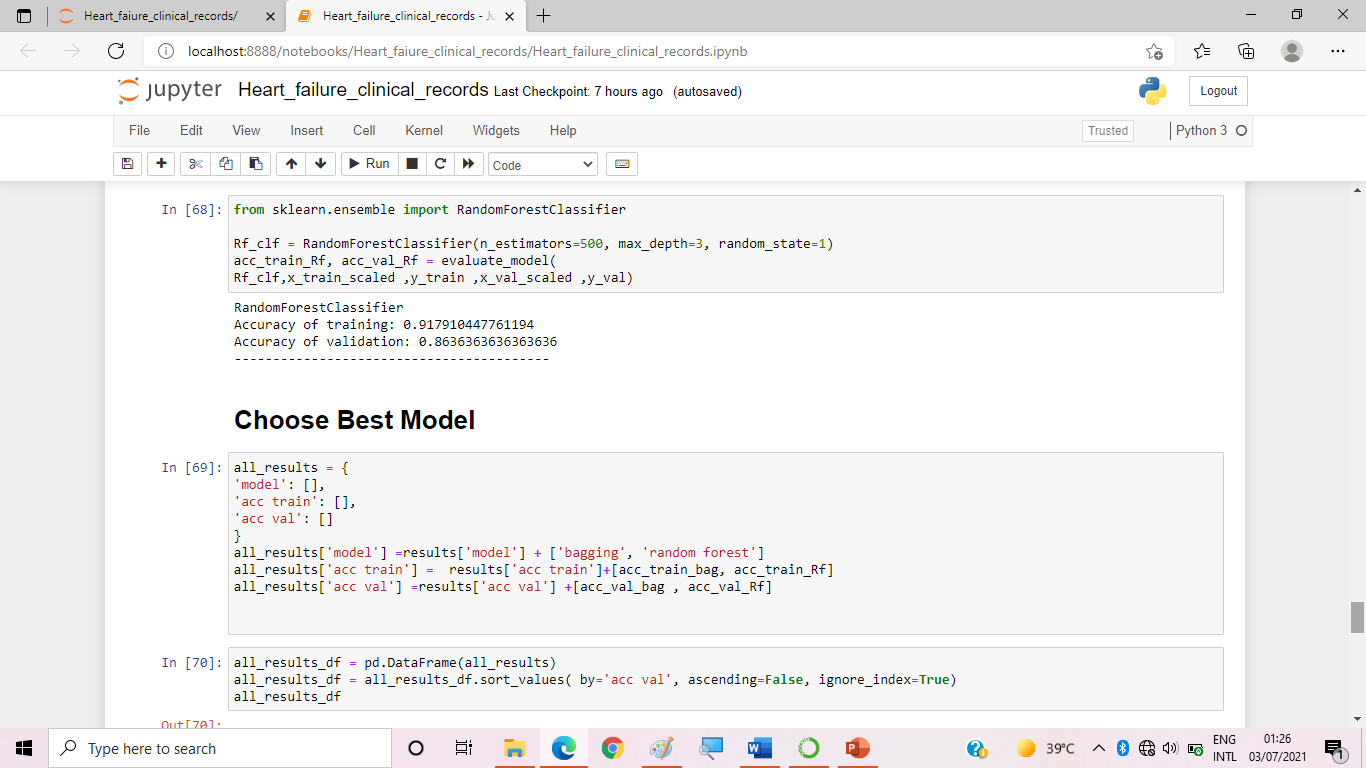


1. **Bagging Classifier:**

The same as voting but this model is applied only to one type of classifier which contains different parts of data.

1. **Random Forest Classifier:**

IN this classifier it chooses a random feature and they do the splitting operation based on it. (So decision tress will have different forms).

Random forest is the best classifier for our dataset as it gave us the best accuracy.

* **Results**

For the best model (random forest): The accuracy of training of the model is 91% and the accuracy of validation Is 86%.

* Chart, box and whisker chart

  Description automatically generated**Graphical user interface, text, application

  Description automatically generated**Accuracy score for the final Testing: 0.9191919191919192%.
* **Conclusion**

In our work, the fact that our traditional biostatistics analysis selected ejection fraction and serum creatinine as the two most relevant features confirmed the relevance of the feature ranking executed with machine learning. Moreover, our approach showed that machine learning can be used effectively for binary classification of electronic health records of patients with cardiovascular hearth diseases. Our models are Logistic regression, Data Pre-processing, Feature Scaling, Decision Tree, Support Vector Machine and Ensemble Learning . The accuracy of our model is 0.919%.

As a limitation of the present study, we have to report the small size of the dataset (299 patients): a larger dataset would have permitted us to obtain more reliable results. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases. Also, if an additional external dataset with the same features from a different geographical region had been available, we would have used it as a validation.

* **References**

[1] The Guardian. UK heart disease fatalities on the rise for first time in 50 years. [https://www.theguardian.com/society/2019/may/13/heart-circulatory-disease-fatalities-on-rise-in-uk.](https://www.theguardian.com/society/2019/may/13/heart-circulatory-disease-fatalities-on-rise-in-uk.%20) Accessed 25 Oct 2019.

[2] National Heart Lung and Blood Institute (NHLBI). Heart failure. <https://www.nhlbi.nih.gov/health-topics/heart-failure>. Accessed 20 June 2019.

[3] Polikar, R. (n.d.). *Ensemble learning*. Scholarpedia. [http://www.scholarpedia.org/article/Ensemble\_learning.](http://www.scholarpedia.org/article/Ensemble_learning.%20)