PROJECT REPORT

*on*

**Predicting Disease Severity and Outcome**

**for COVID-19 Patients**

(CSE V Semester Mini project)

(2021-2022)

A picture containing text, scene, room, gambling house

Description automatically generated

**Guided by:**  **Submitted by:**

Mr. Kireet Joshi Ashutosh Kumar Gupta

(Resource Person) Roll No: 2015202

CSE- DS: V Semester

Session: 2021-2022

**Acknowledgement**

Here by we are submitting the project report on **“Machine learning model to predict the disease severity and Outcome in covid-19 patients ”**, as per the scheme of Graphic Era Deemed to be University, Dehradun for Vth semester mini project submission.

In this connection, we would like to express our deep sense of gratitude to our beloved institution Graphic Era Deemed to be University and also I like to express my sincere gratitude and indebtedness to **Mr Kireet Joshi.**

I consider it my cardinal duty to express the deepest sense of gratitude to **Mr. Kireet Joshi**, for the invaluable guidance extended at every stage and in every possible way.

Finally I am very much thankful to all the faculty members of the Department of Computer Science and Application, friends and our parents for their constant encouragement, support and help throughout the period of project conduction.

**Ashutosh Kumar Gupta**

                2015202

Vth Sem BTech CSE(DS&AI)

**ABSTRACT**

The early prediction of severity and outcome according to conditions given of a person can reduce the risk of death by providing better facilities to serious patients. This study aims to develop and compare prognosis machine learning models on laboratory data given for patients. In this project 3 different dataset for covid-19 patients were used and on those datasets different prediction models were used (such as:- KNN,SVM ,Decision Tree, Random Forest Classifier, Logistic Regression and AdaBoost Classifier) and the accuracies of these models are compared after training and testing to get the best performing model. Datasets that were used in this contains different data about condition of patients and also they have different number of data. Preprocessing of data has been done before training in order to get better accuracy(SMOTE is used for balancing the dataset). In two of the three models the task was to predict the mortality while in one model the task was to predict the severity.

Keywords: KNN:- K Nearest Neighbor

SVM:- Support Vector Machine

SMOTE:- Synthetic Minority Oversampling Technique

**INTRODUCTION**

The COVID-19 pandemic has tremendously strained economic and healthcare infrastructures worldwide, leaving a trail of millions of deaths behind. Even after following proper Social Distancing and major COVID-19 protocols the control of this outbreak seems to be very hard. It has affected the whole earth, many has lost their close ones, economy has been badly affected by lockdown, in many countries multiple waves are predicted.

After failure of physical measure, more focus shall be given to clinical measures. For doing so early prediction early prediction can be useful, patients will need different resources according to the severity of disease.

Sometimes it has been seen that doctors were unable to properly diagnose the patients, also a COVID-19 patient may face unpredictable shifts in the condition. Here in such scenarios AI and machine learning technology can be very helpful in order to monitor patients vitals and condition and predict the severity. By doing so a patient with high severity can be detected and can be treated with better resources and hence risk of mortality will decrease.

**METHODOLOGY**

**Libraries Used:-**  1. Pandas 2. Sklearn 3. NumPy 4.Seaborn

5. Imblearn 6. Matplotlib

**Models Used:-** 1. Random Forest Classifier

2. Decision Tree Classifier

3. KNN Classifier

4. SVM Classifier

5. AdaBoost Classifier

6. Logistic Regression

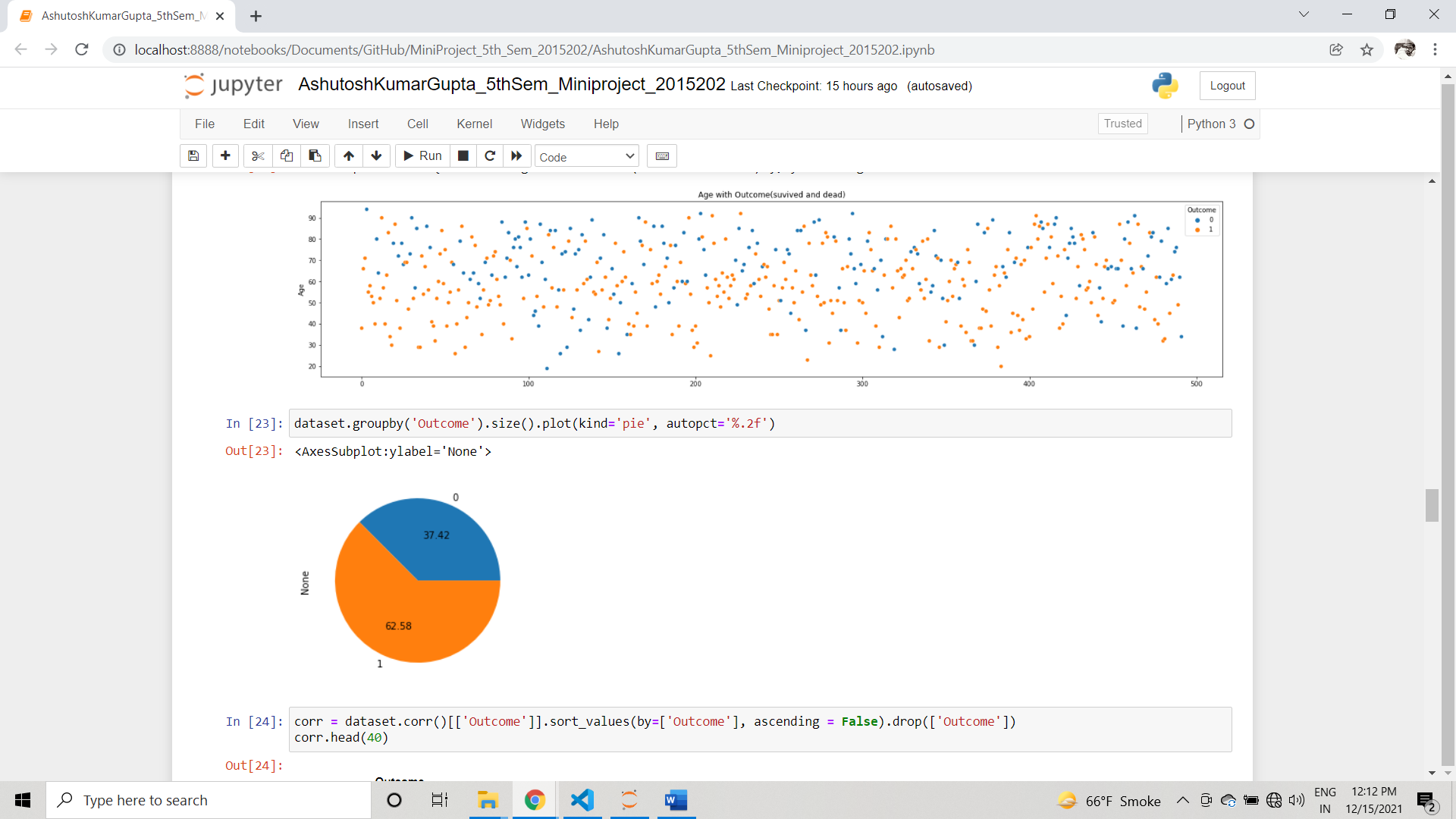
**Functionality Used:-** 1. SMOTE 2.PCA 3.Standard scaler

**Dataset:-**  3 datasets were used in this project and they were downloaded from

1. <https://www.kaggle.com/iamhungundji/covid19-symptoms-checker>
2. [www.github.com](http://www.github.com)

**RESULTS**

Training and Testing the different classification models on first dataset containing 492 rows and 39 columns. The target column in this dataset is “**Outcome**” which has 306 values of ‘1’ type and 186 of ‘0’ type. ‘1’ represent survived and ‘0’ represent dead. After applying SMOTE it has 306 values in each type.



After training and testing different models the comparison of accuracy was as follow:-

Graphical user interface, text, application

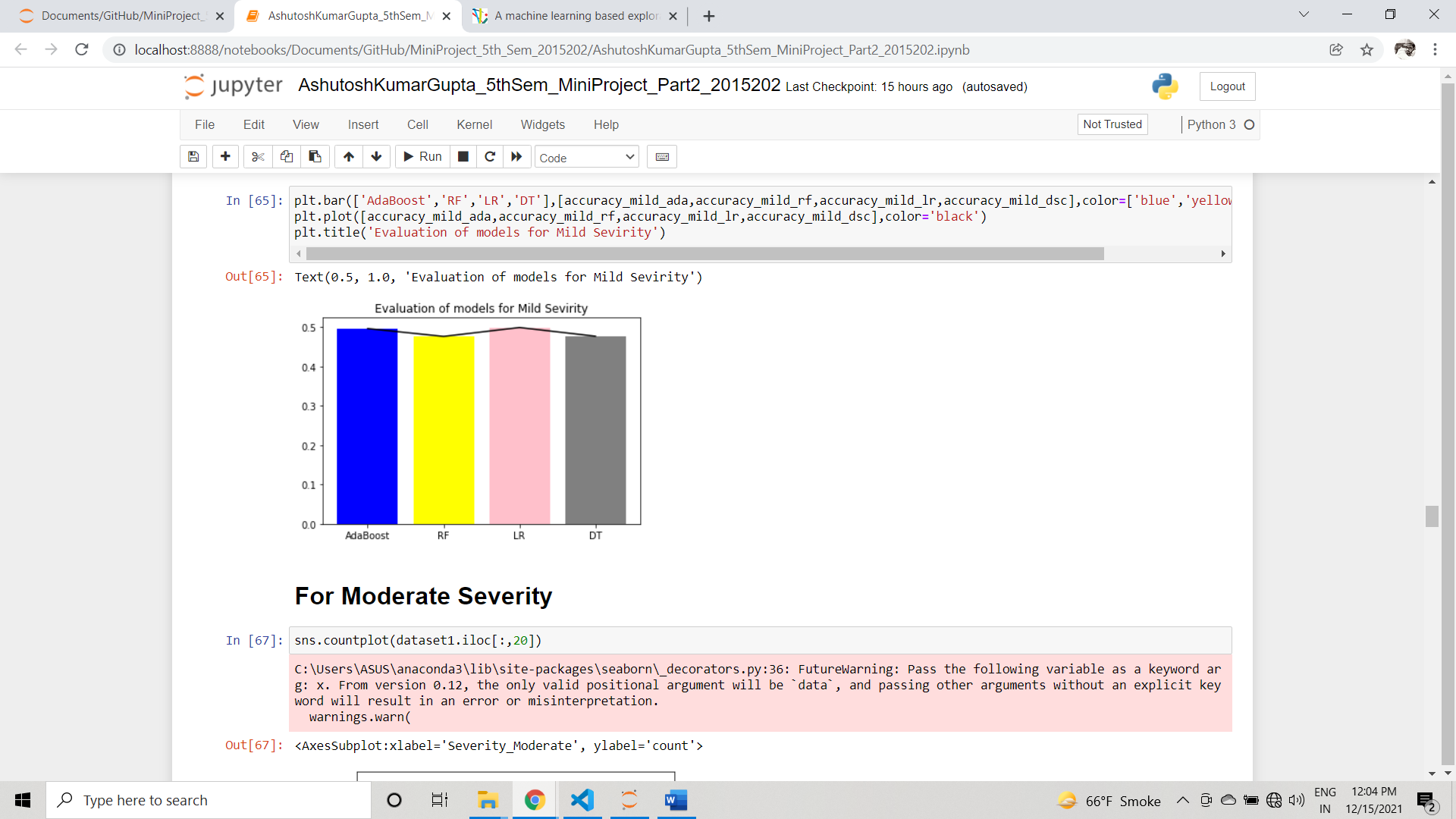
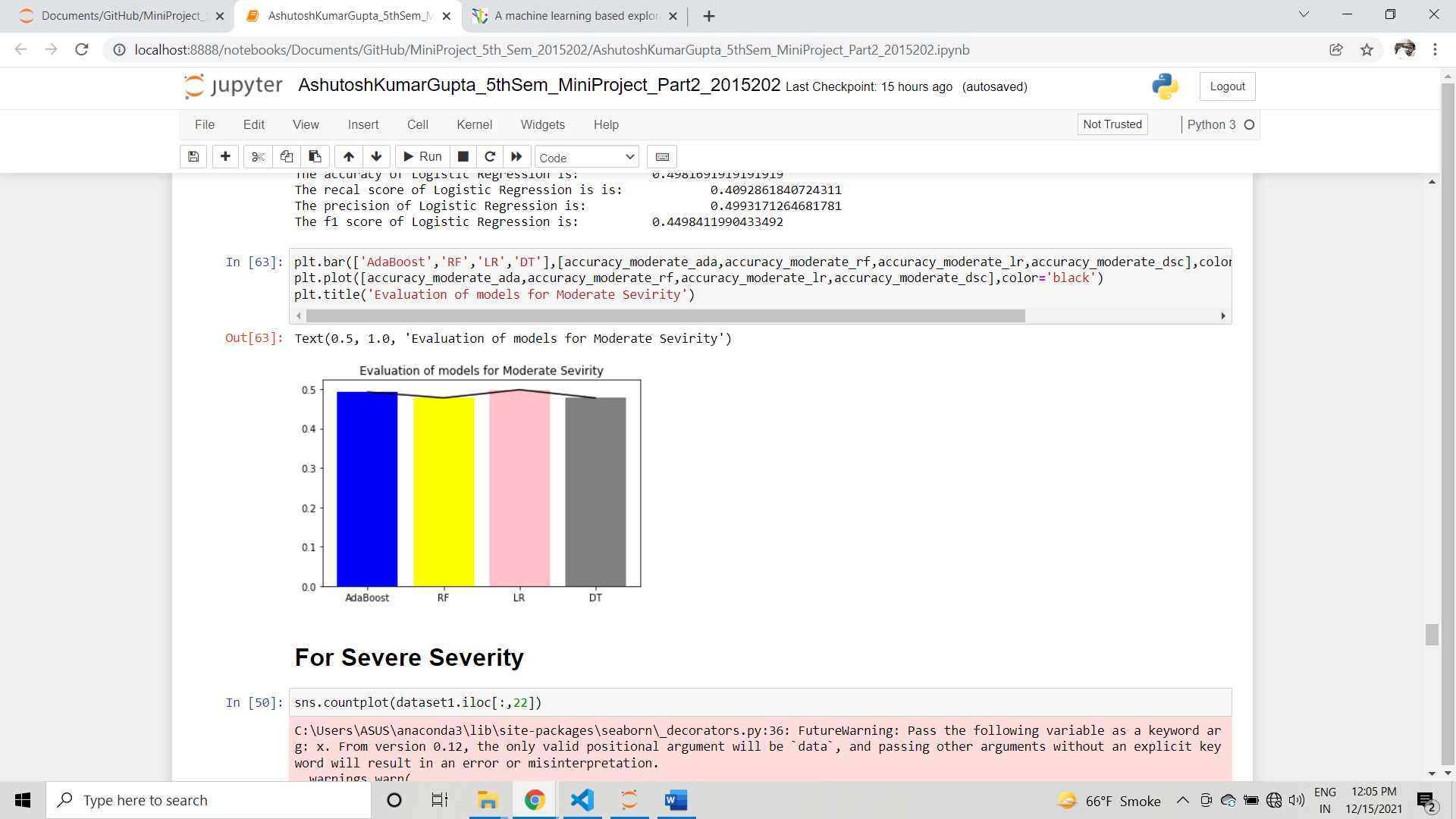
Description automatically generated

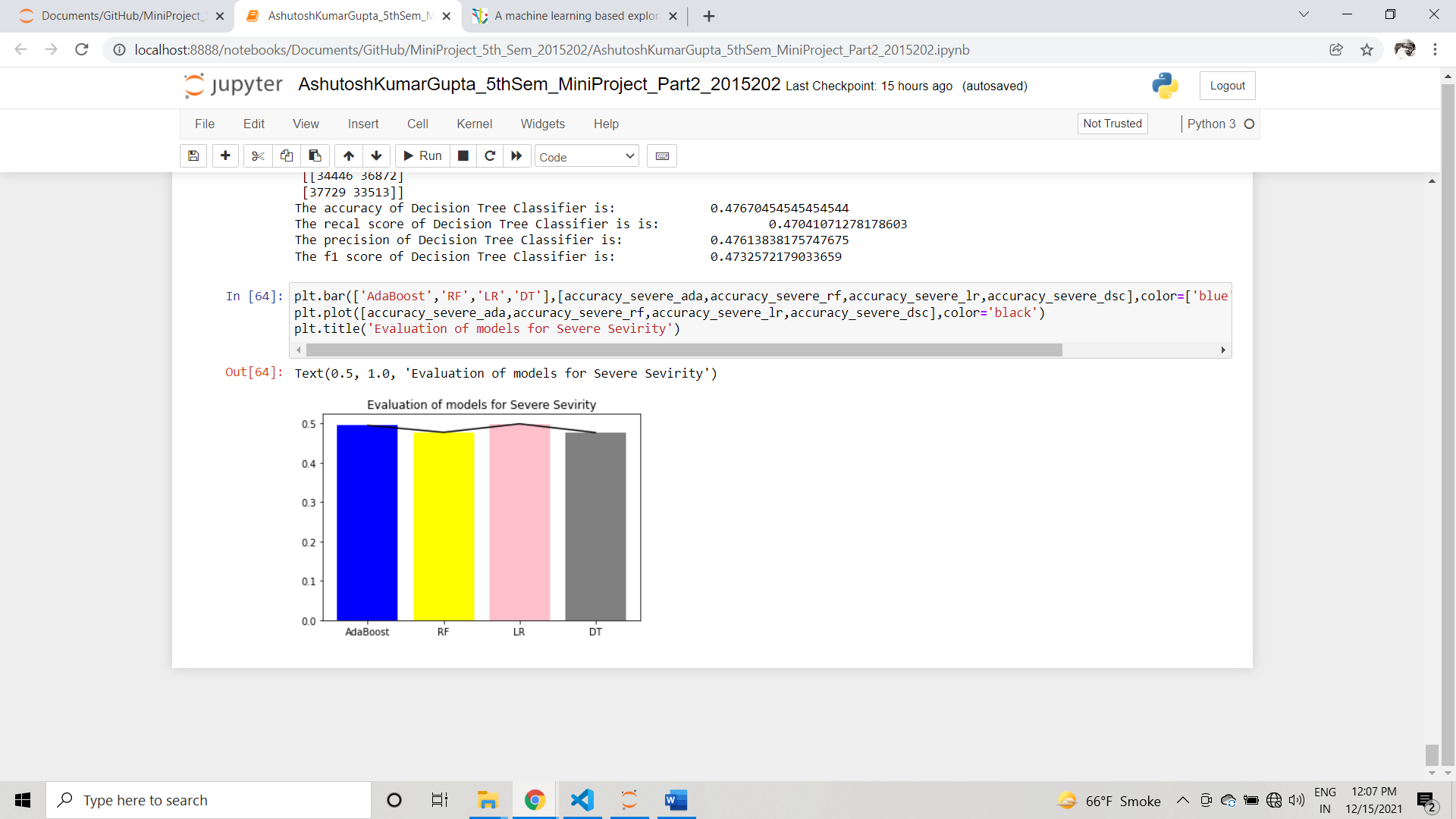
Here we got the maximum accuracy of 84% with Random Forest classifier

**FOR SECOND DATASET**

And while training the models with the dataset(second dataset) containing 316800 rows and 25 columns. The task for making models with this dataset we have to predict the severity of a patient according to the vitals or the condition given for each patient. Here we have four types of severity(mild severity, moderate severity, severe severity and none severity). And different models were trained to predict severity of each type. This dataset is highly unbalance with respect to each severity class so that was balanced by SMOTE.

The accuracies for different severity classes are:-

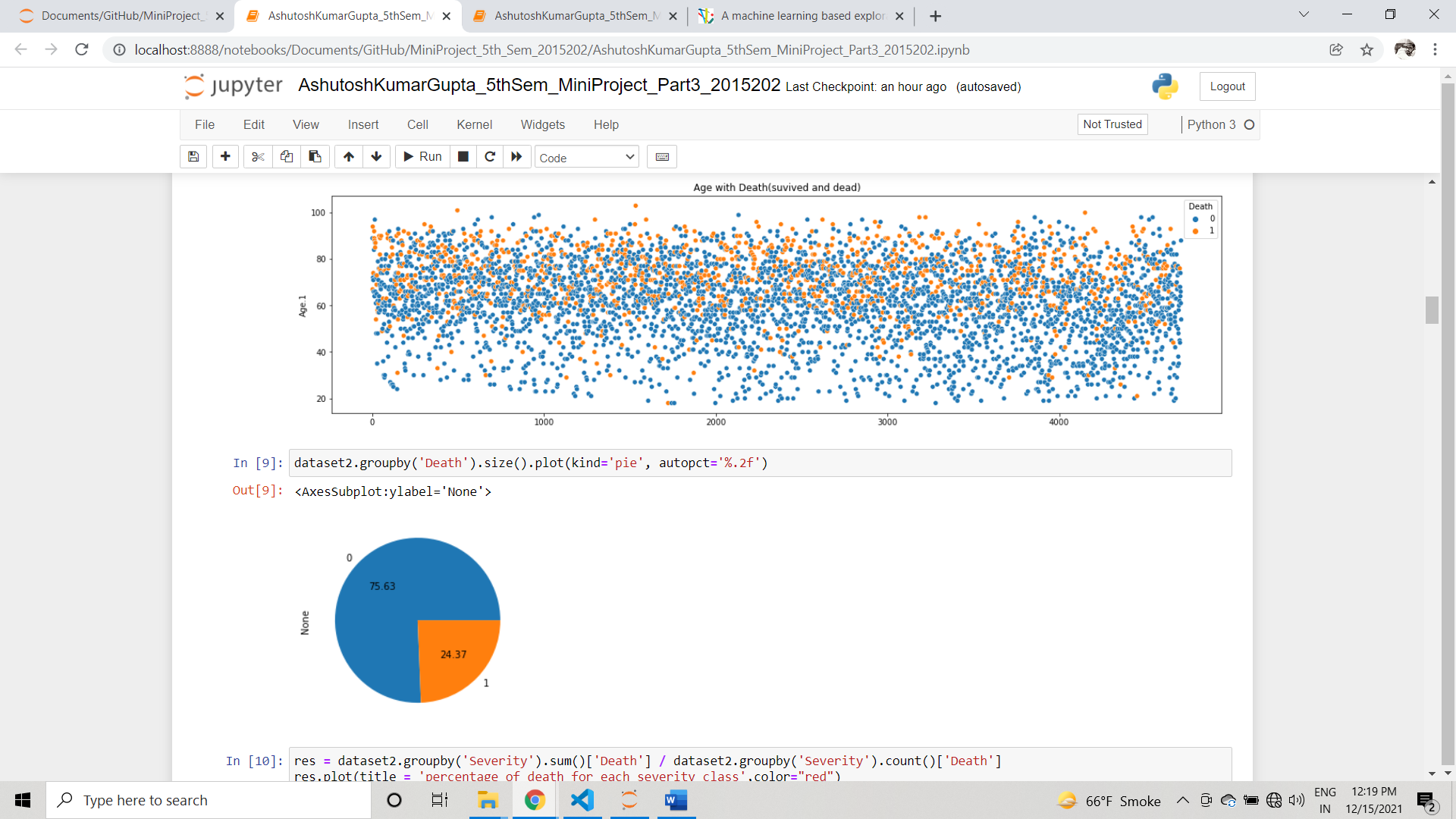


Here AdaBoost and Logistic Regression performed better than Random Forest and Decision

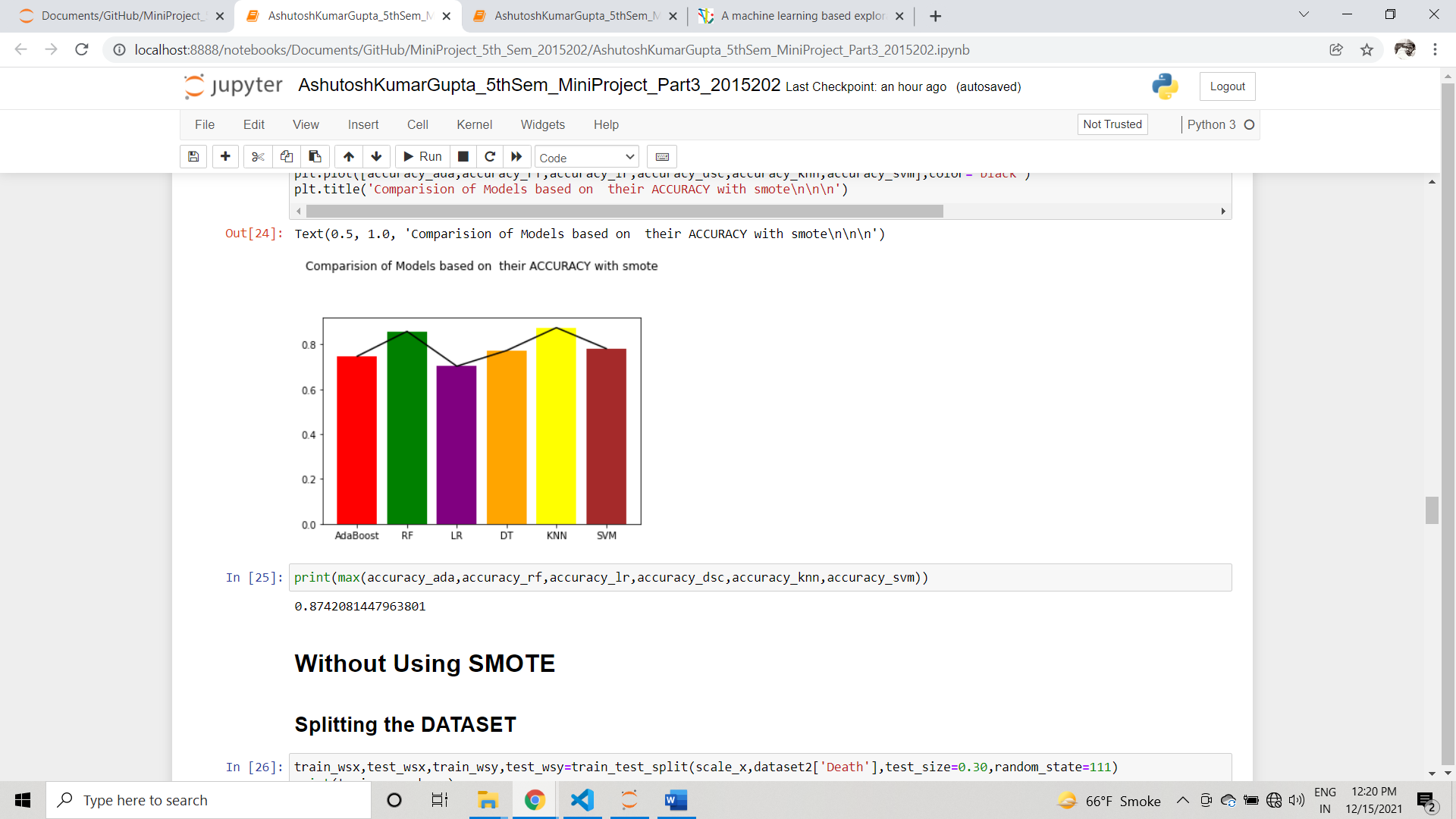
Tree, with accuracy of almost 50%. The accuracy in this dataset is low as compared to first dataset because in this dataset we have separate severity class and the data was very imbalance.

**FOR THIRD DATASET:-**

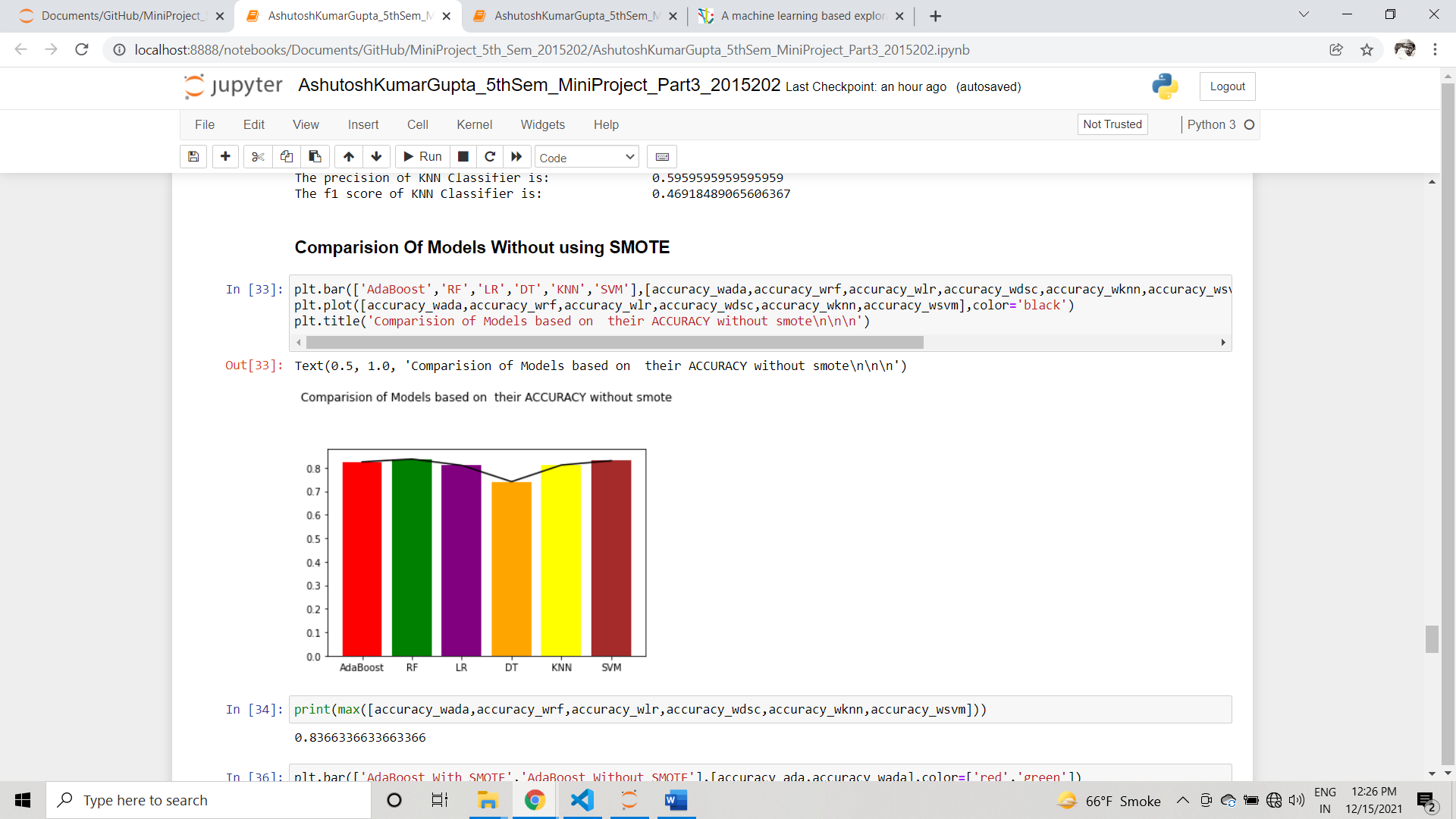
And while training the models with dataset(third dataset) that has 4711 rows and 85 columns. The task was to predict whether a person would die or not according to the condition or vitals given. The target column in this dataset is “**Death**” which is containing only two values ‘0’ and ‘1’. ‘0’ represent survived and ‘1’ represent dead. It contains 3563 of type’0’ and 1148 of type’1’.



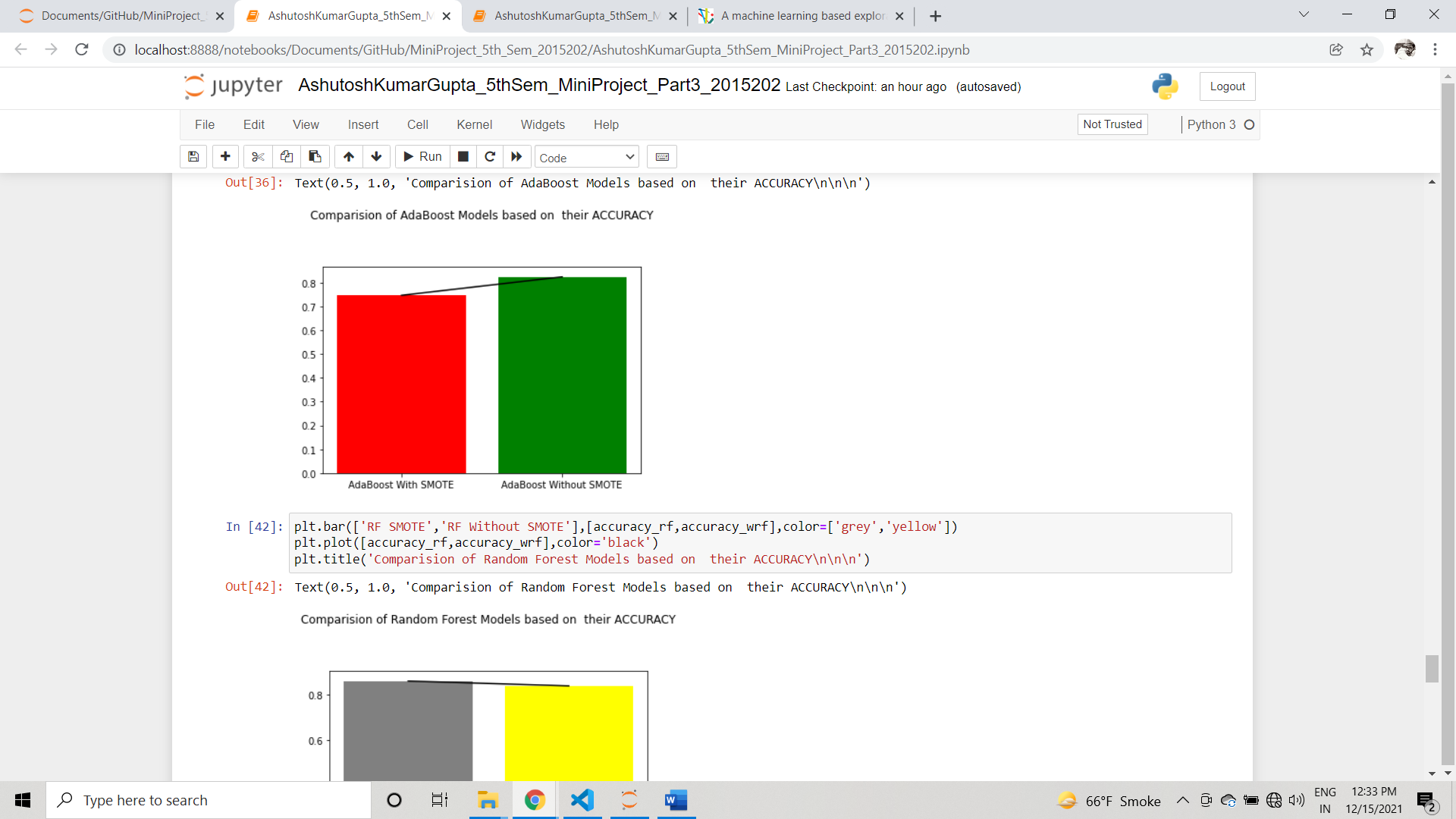
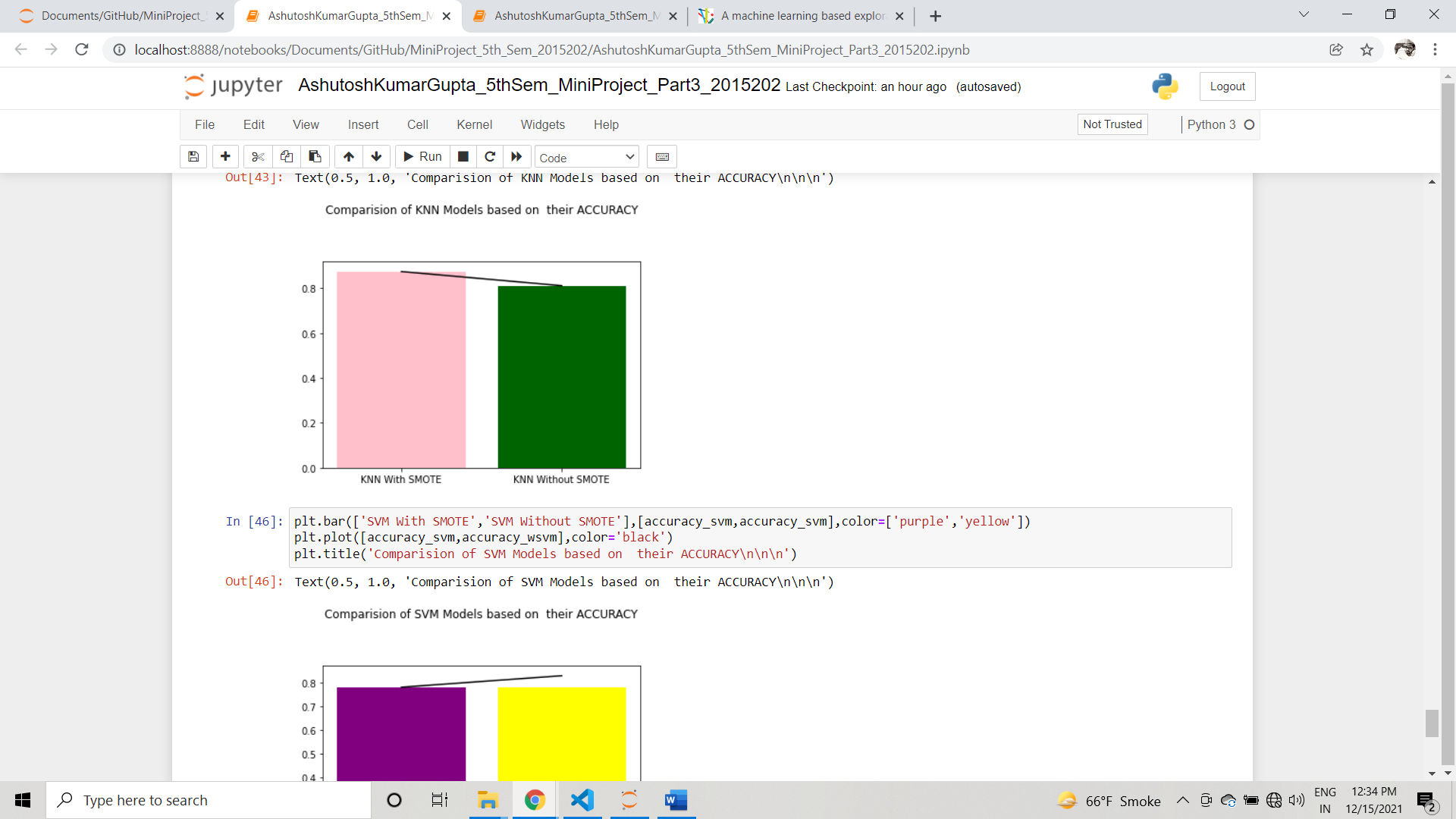
In this model we try to compare the accuracies of different classification models while using SMOTE and without using SMOTE.



Here KNN is giving the better accuracy than rest of algorithms of around 87%



Here Random Forest is giving better accuracy than the rest of the models (almost 84%) and also here the accuracy of AdaBoost is increased surprisingly to 83% as compared to while it was used in the model with using SMOTE. Also the accuracy of Logistic Regression, SVM increased.

**CONCLUSION:-**

In this project we have used various concept of machine Learning like data preprocessing,splitting the dataset, balancing the dataset , training and testing of different classification models and comparing their accuracies in different circumstances.

This project works well on determining the disease severity and outcome of covid-19 patients according to different conditions given. Also the accuracy of Random forest classifier was most consistent among all algorithms, although while using the second dataset we didn’t get very good accuracy because of faulty dataset but it can be improved if we use Deep Learning algorithms which I will be applying in future.

And with the help of these models one can easily predict the condition of a patient and this will also helps doctors to get the information of patient’s health with good accuracy and in very fast manner.

The works has been uploaded frequently on a github repository. The link is given below

<https://github.com/ashu741/MiniProject_5th_Sem_2015202>

**REFERENCES:-**

1. <https://www.kaggle.com/iamhungundji/covid19-symptoms-checker>
2. [www.github.com](http://www.github.com)
3. [www.google.com](http://www.google.com)
4. [www.towardsdatascience.com](http://www.towardsdatascience.com)
5. The information about the covid severity was also taken from multiple research papers.
6. [www.youtube.com](http://www.youtube.com)
7. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252384#sec009>