Project Report on Diabetic Disease Detection Using Machine Learning: A Data Science Project

ABSTRACT: Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three different supervised machine learning methods including: Logistic regression, KNeighborsClassifier, RandomForestClassifier. This project also aims to propose an effective technique for earlier detection of the diabetes disease.

Introduction: Diabetes Mellitus Diabetes is one of deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most

common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L.

Machine Learning: Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term "machine learning" is identical to the term "artificial intelligence", given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience. A more detailed and formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification methods namely, Logistic regression, K-Nearest Neighbor Classifier and Random Forest Classifier algorithms.

II. Supervised Learning In supervised learning, the system must "learn" inductively a function called target function, which is an expression of a model describing the data. The objective function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values of the function, i.e. its domain, are called instances. Each case is described by a set of characteristics (attributes or features). A subset of all cases, for which the output variable value is known, is called training data or examples. In order to infer the best target function, the learning system, given a training set, takes into consideration alternative functions, called hypothesis and denoted by h. In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models try to predict distinct classes, such as e.g. blood

groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as kNearest Neighbours (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) Association Rule Mining appeared much later than machine learning and is subject to greater influence from the research area of databases. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Objectives The predominant aim of this project is to propose novel predictive model to predict diabetes mellitus using the clinical and e-diabetic Dataset. The objectives of the proposed work are formulated as below:

1. To create a diabetic disease prediction model.

2. To apply Exploratory Data Analytics (EDA) to derive patterns.

3. To predict diabetes using the generated patterns through Machine Learning.

Methodology: As we have to deal with a classification problem, so I used prominent machine learning classifiers to predict whether a patient is diabetic or not. For this I used Logistic Regression, Kn-Neighbors classifier and Random Forest Classifier. It is an end to end diabetic disease prediction model. The steps involved are:

1. Problem definition.
2. Data
3. Evaluation
4. Features
5. Modelling
6. Experimentation

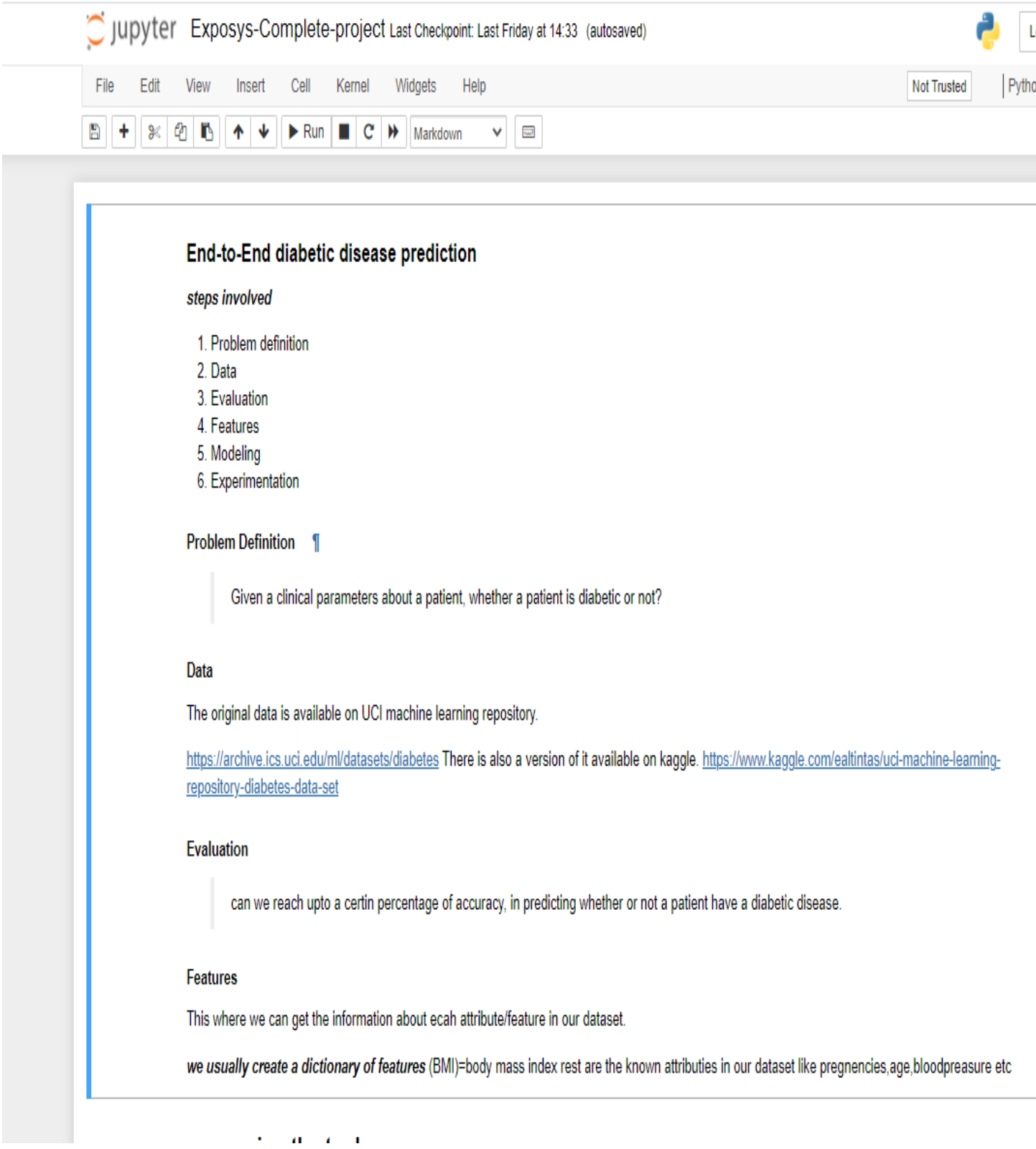


Fig1: showing steps involved .

For EDA exploratory Data Analysis (EDA) the libraries used in the project are:

1. Pandas.
2. Numpy.
3. Matplotlib.
4. Seaborn

For machine learning scikit-learn libraries are used:

1. LogisticRegression.
2. KNeighborsClassifier
3. Random Forest Classifier.

For model evaluation sckiti-learn libraries used:

1. train_test_split.
2. Cross_val_score.
3. RandomizedSearchCv.
4. GridSearchCV.
5. Confusion_matrix
6. Classification_report
7. Roc_curve

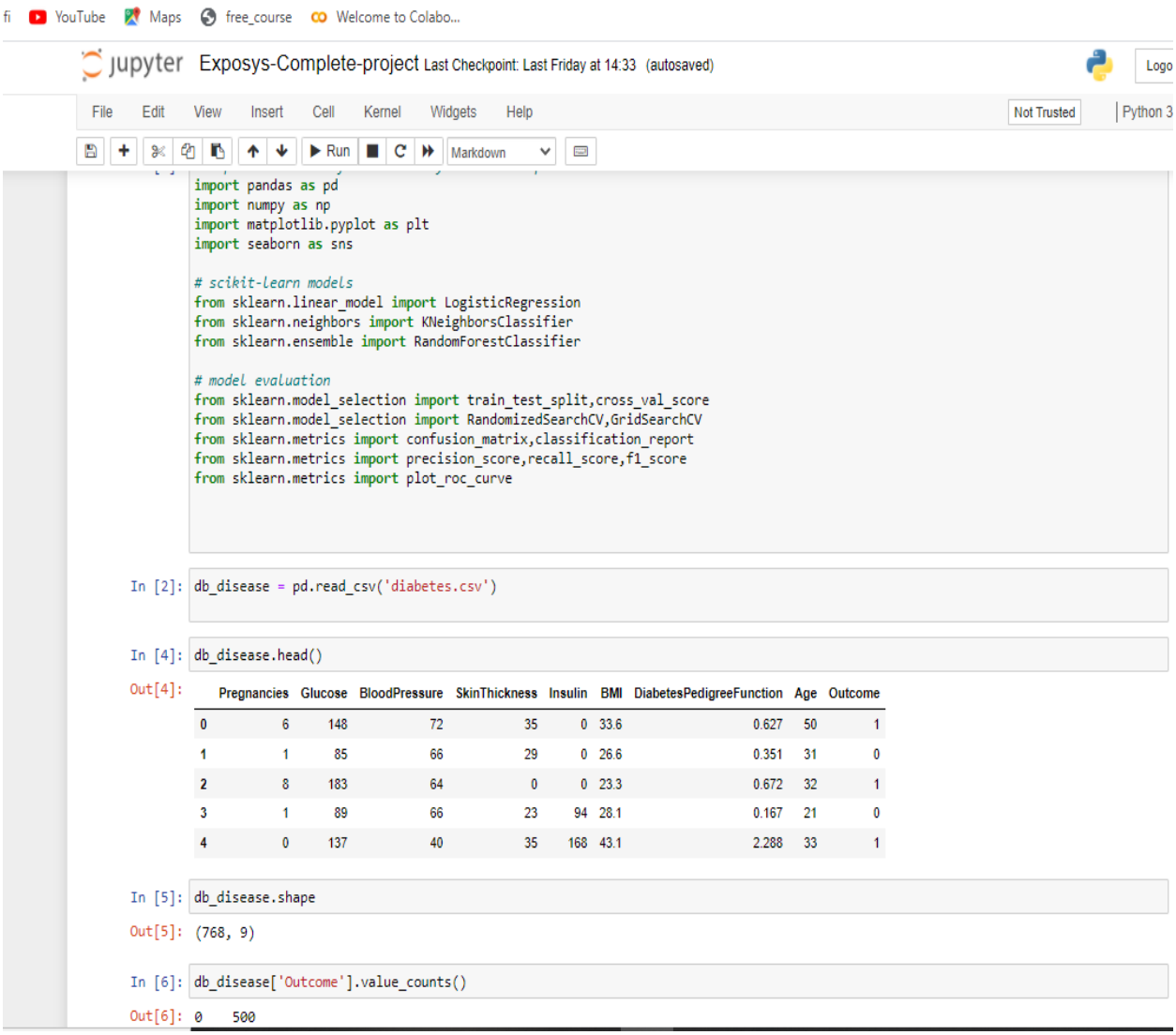


Fig2. Important libraries imported.

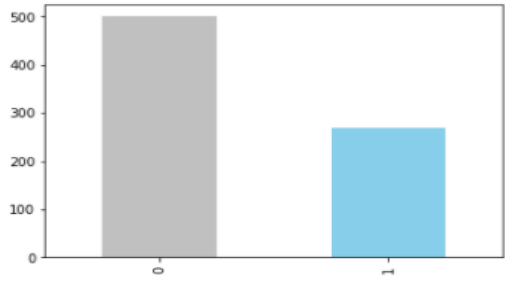
homeWifi YouTube Maps free_course Welcome to Colabo...

jupyter Exposys-Complete-project Last Checkpoint: Last Friday at 14:33 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

In [7]:
```
db_disease['Outcome'].value_counts().plot(kind='bar',color=['silver','skyblue']);
```



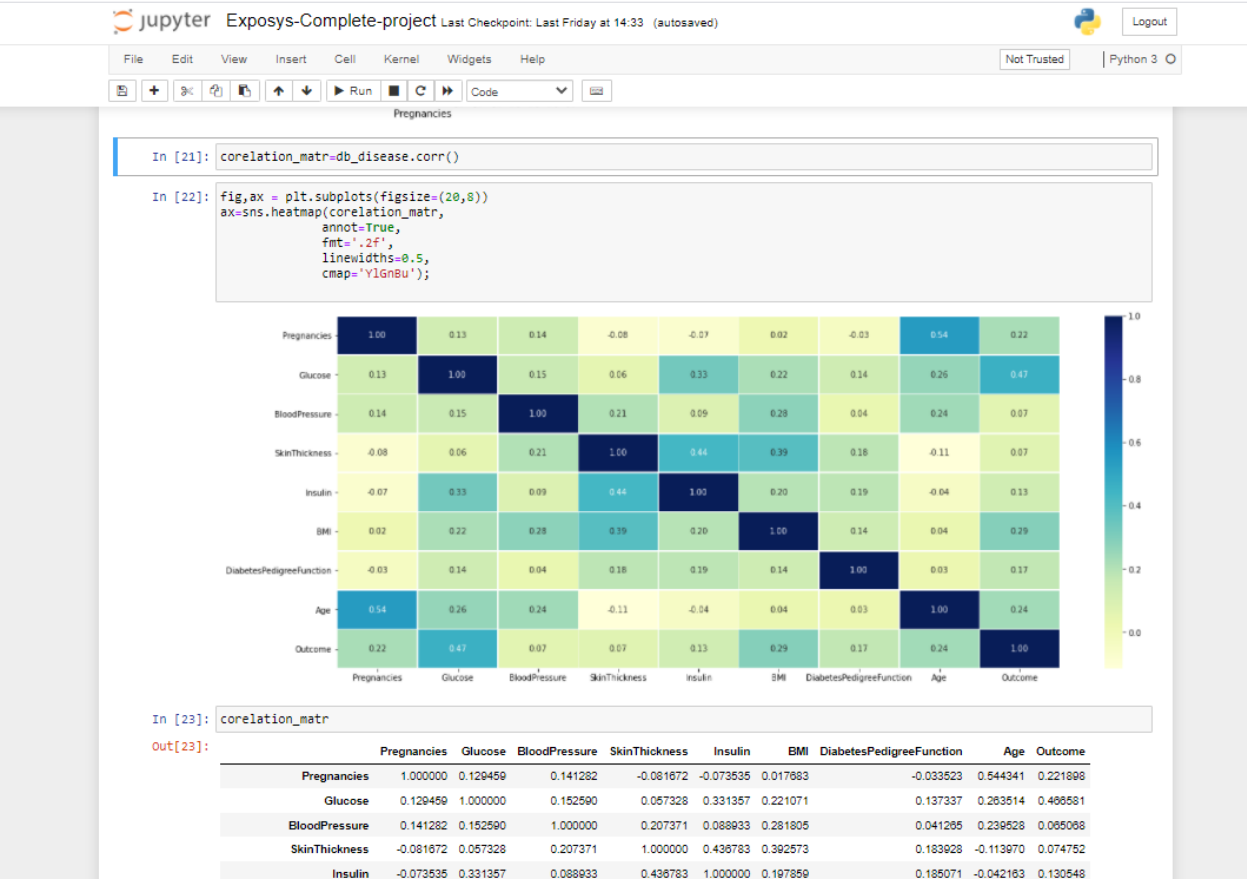
In [8]:
```
#lets get some information about our dataset and its attributes
db_disease.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

In [9]:
```
#lets check how many values are null in our dataset for each attribute
db_disease.isna().sum()
```

Out[9]:
```
Pregnancies                 0
Glucose                     0
```

Fig3. A bar graph plot showing outcome values (0&1).



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548

Fig4. Showing correlation between the attributes.

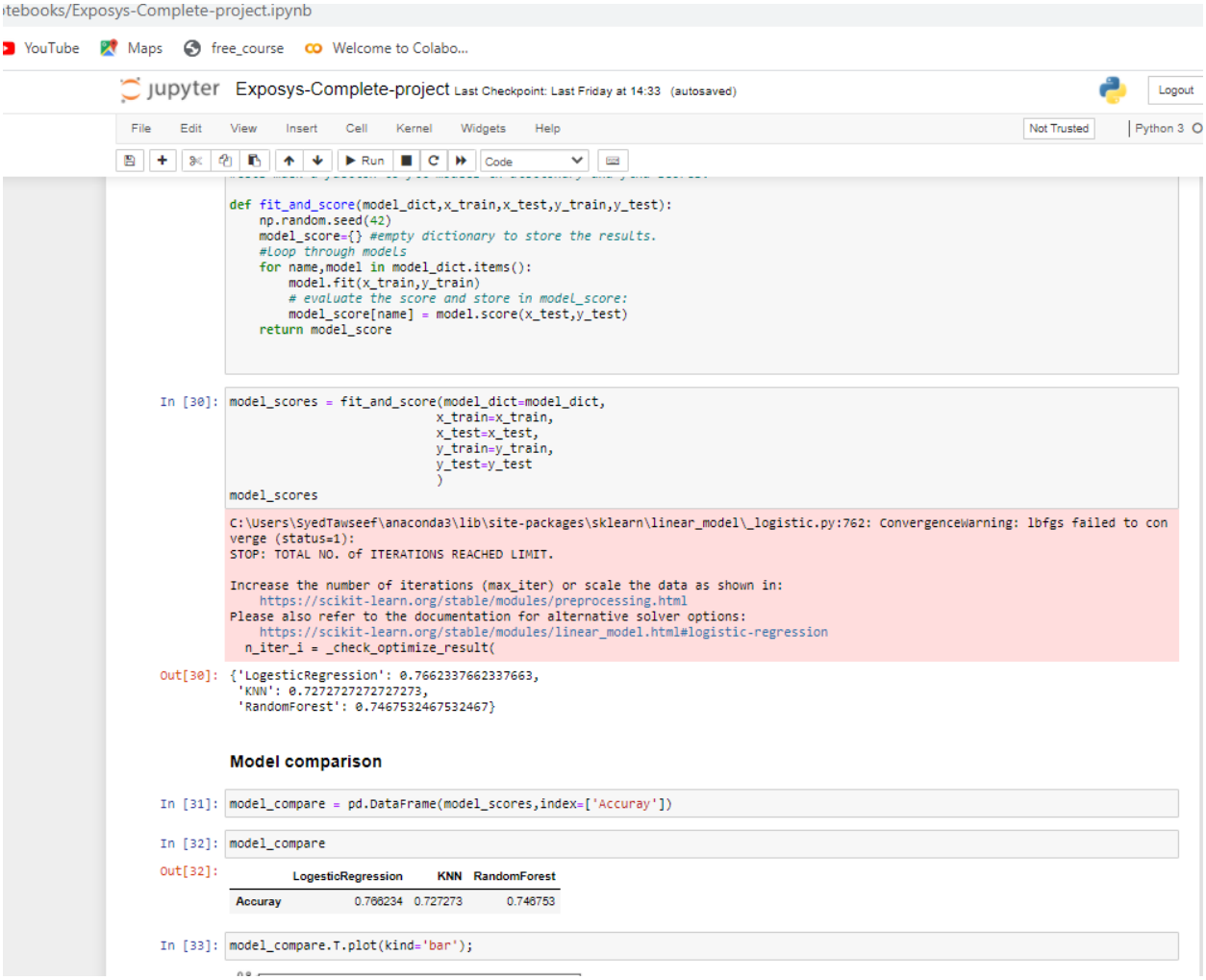


Fig5. Showing Modeling process and score for each model used.

jupyter Exposys-Complete-project Last Checkpoint: Last Friday at 14:33 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

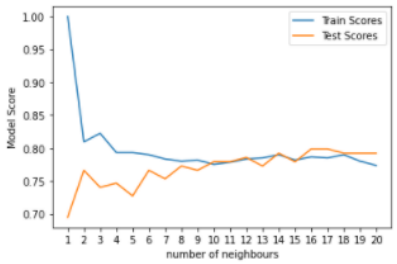
Out[35]: 1.0

In [36]:
```
max(test_score)
```

Out[36]: 0.7987012987012987

In [37]:
```
plt.plot(neighbours,train_score,label='Train Scores')
plt.plot(neighbours,test_score,label='Test Scores')
plt.xticks(np.arange(1,21,1))
plt.xlabel('number of neighbours')
plt.ylabel('Model Score')
plt.legend();
print(f'The maximum test score for knn is : {max(test_score)*100:.2f}%' )
```

```
The maximum test score for knn is : 79.87%
```



tuning hypermeters with scikit-learns randomsearchcv and gridsearchcv

In [38]:
```
#create a grid for logesticregressor parameters
lr_grid={'C':np.logspace(-4,4,20),
        'solver':['liblinear']}

#create a grid for RandomForestClassifier parameters

rf_grid={'n_estimators':np.arange(10,1000,50),
        'max_depth':[None,3,5,10],
        'min_samples_split':np.arange(2,20,2),
        'min_samples_leaf':np.arange(1,20,2)}
```

Fig6. Showing tuned hyper-parameters for kneighbor classifier.

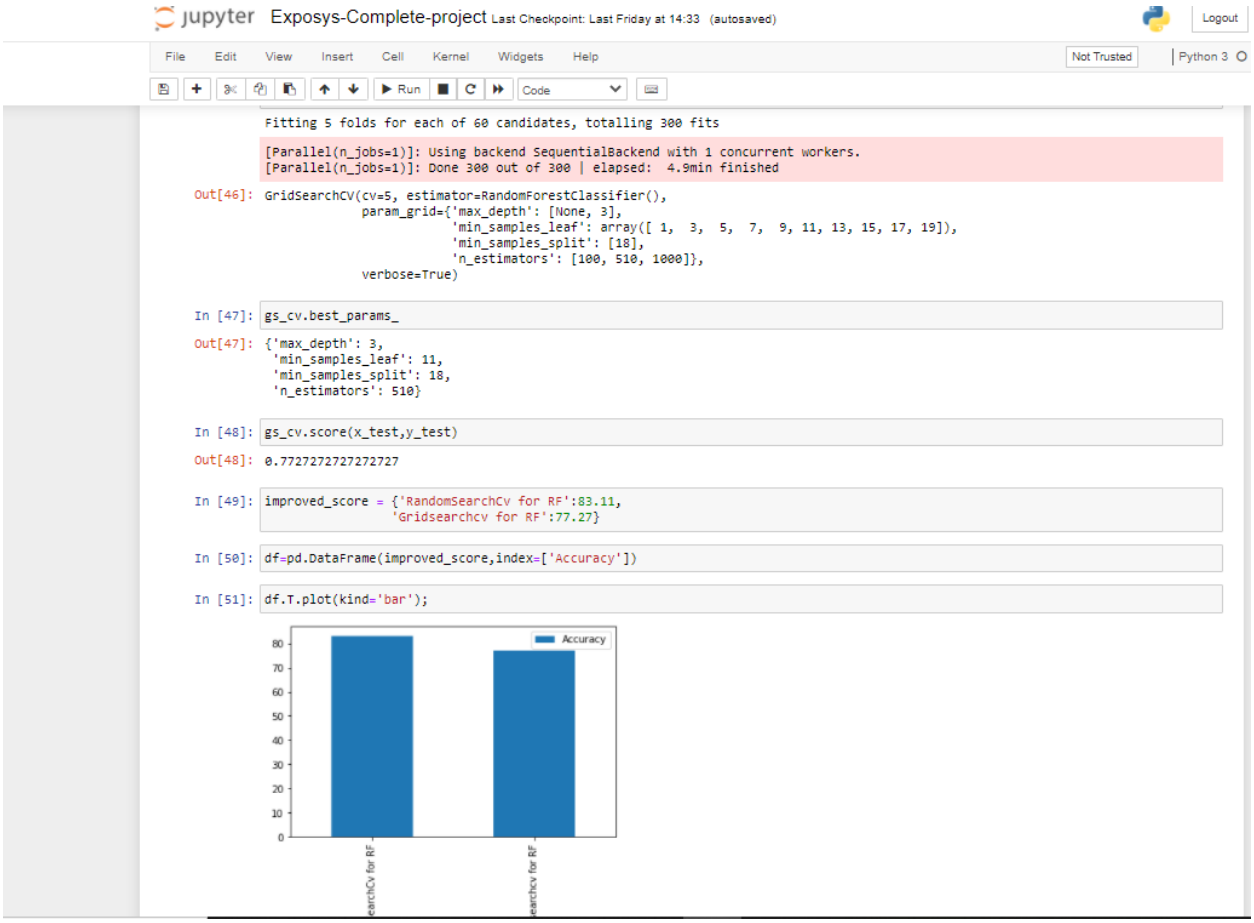


Fig7. Showing tuned parameters for randomforestclassifier and score.

jupyter Exposys-Complete-project Last Checkpoint: Last Friday at 14:33 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code

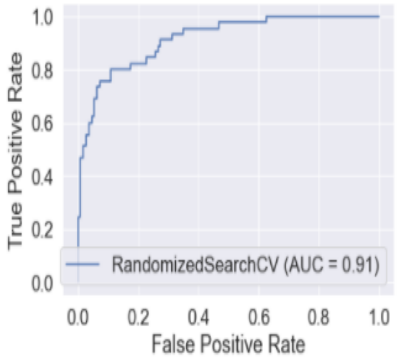
lets plot a roc curve (receiver operating cherecteristic curve)

In [89]:
```
y_pred=rf_cv.predict(x_test)
```

In [90]:
```
y_pred
```

Out[90]:
```
array([0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0,
       0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
      dtype=int64)
```

In [91]:
```
plot_roc_curve(rf_cv,x_test,y_test);
```



In [92]:
```
cf=confusion_matrix(y_test,y_pred)
```

In [93]:
```
cf
```

Out[93]:
```
array([[108,   1],
       [ 25,  20]], dtype=int64)
```

In [94]:
```
def plot_con_mat(y_test,y_pred):
    sns.set(font_scale=1.5)
    fig,ax=plt.subplots(figsize=(6,5))
```

26°C

Fig8 showing ROC curve for false positive rates and True positive rates

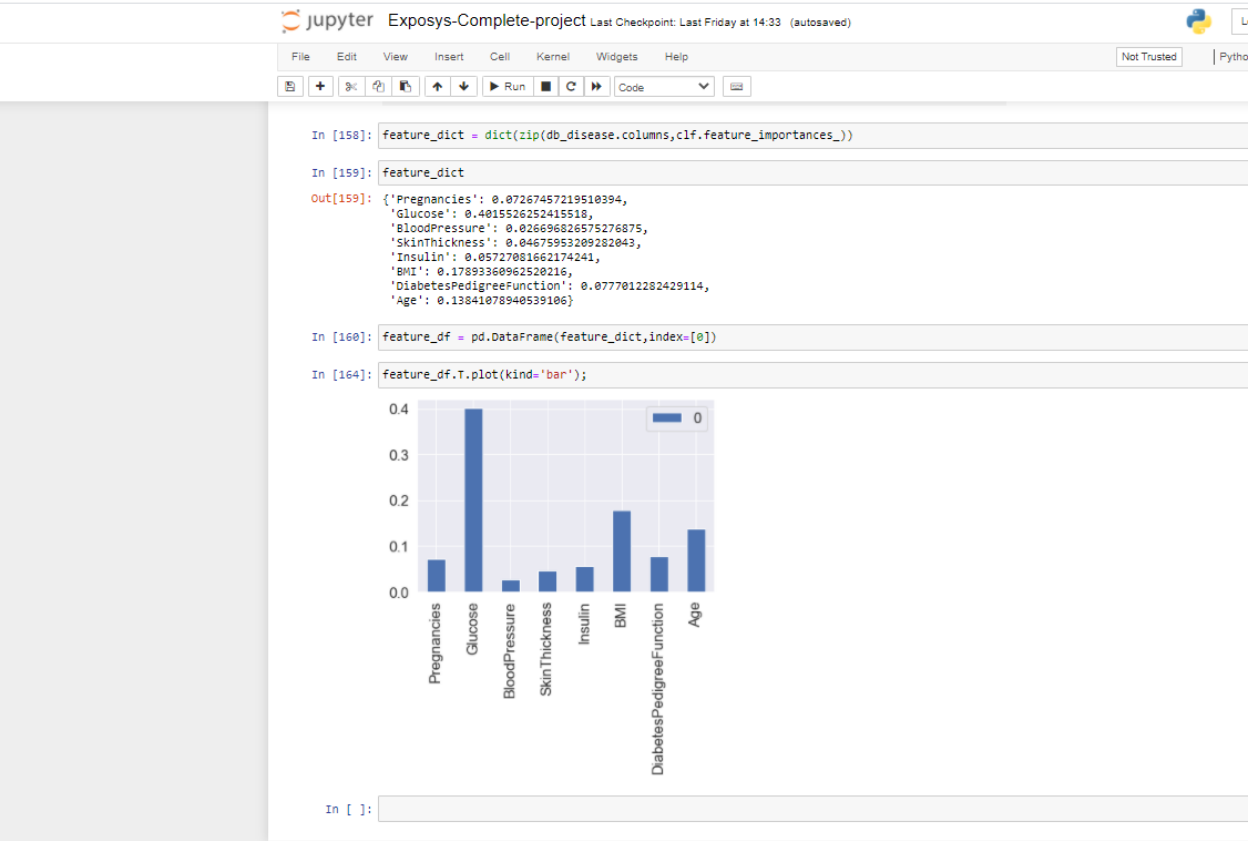


Fig9. Showing feature importance

Conclusion Further work needed to be done, as we achieved the accuracy of 83% because only. As we know Diabetes Mellitus is the most severe chronic ailment that can seriously impact the quality of living of the affected persons. In this system Data collection, Data analysis, Data modelling and improving of models has been done to achieve a good score. Since improved and tuned hyper-parameters of Random Forest Classifier showed the greater results for accuracy than other classifiers used. Hence we can say that for the dataset (diabetic-dataset/kaggle.com) used in this system Random Forest Classifier proved to be the better machine learning model.