Shri Vaishnav Institute of Science and Technology, Indore

**Training Project Report**

**“Diabetes Prediction and Analysis”**

**Submitted by:-**

Rohan Bhugra

Sapan Saxena

COMPUTER SCIENCE & ENGINEERING IN BACHELOR OF TECHNOLOGY

JULY-AUGUST 2020

**A REPORT OF TWO WEEK INDUSTRIAL TRAINING AT**

**WebTek Labs Pvt. Ltd.**

**SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD**

**OF THE DEGREE OF BACHELOR OF TECHNOLOGY**

**COMPUTER SCIENCE & ENGINEERING**

****

**JULY-AUGUST 2020**

**Submitted by:**

Rohan Bhugra

Sapan Saxena

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**INDORE**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**SHRI VAISHNAV VIDYAPEETH VISHVAVIDALAYA, INDORE**

**CANDIDATE'S DECLARATION**

We hereby declare that we have undertaken industrial training at “**WEBTEK LABS PVT. LTD.**” during a period from **27th July to 15th August** in partial fulfilment of requirements for the award of degree of B.Tech (COMPUTER SCIENCE & ENGINEERING) at SHRI VAISHNAV INSTITUTE OF TECHNOLOGY, INDORE. The work which is being presented in the training report submitted to Department of COMPUTER SCIENCE & ENGINEERING at SHRI VAISHNAV INSTITUTE OF TECHNOLOGY, INDORE is an authentic record of training work.

Rohan Bhugra (1710DMBCSE01384)

Sapan Saxena (1710DMBCSE01393)

ACKNOWLEDGEMENT

It gives us great pleasure to acknowledge the guidance, assistance and support of Ms. Mousita Dhar in making the Project and this Project report successful, which has been structured under her valued suggestion.

She has helped us to accomplish the challenging task in a very short period of time.

Finally, we express the constant support of our friends, family and professors for inspiring us throughout and encouraging us.

Rohan Bhugra

Sapan Saxena

SEMESTER: VI - CSE

# CERTIFICATE OF APPROVAL

The project **“DIABETES PREDICTION AND ANALYSIS”** made by the efforts of **Rohan Bhugra, Sapan Saxena** is hereby approved as a creditable study for the **Bachelor of Technology** in **COMPUTER SCIENCE & ENGINEERING** and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

Ms. Mousita

(Project In charge)

**Abstract**

Diabetes is one of deadliest diseases in the world. As per the existing system, patients have to visit a diagnostic center, consult their doctor and wait for a day or more to get their result. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. But with the rise of Machine Learning approaches, we have been able to find a solution to this problem using data mining. Data mining is one of the key areas of Machine learning. It plays a significant role in diabetes research because it has the ability to extract hidden knowledge from a huge amount of diabetes related data. The aim of this research is to develop a system which can predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treatment of the patients before it becomes critical. Currently, the models give accuracy of 79% for Logistic Regression.

**Contents:**

1. Introduction
   1. Problem Statement
   2. Data Description
   3. Exploratory Data Analysis
2. Literature Survey
   1. Technology used
   2. Training Work Undertaken
3. Source Code and Output
   1. Import Packages
   2. Load the Dataset
   3. Data Pre-Processing
   4. Data Regression
4. Results
5. Conclusion
6. Refrences
7. **Introduction**
   1. Problem Statement

As per the existing system, patients have to visit a diagnostic center, consult their doctor and wait for a day or more to get their result. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. A prediction by analyzing behavior of people will help them by saving cost and time.

* 1. Data description

There are 9 variable in the dataset. Several constraints were places on thr selection of these instances from a larger database. In particular, all patients here are females atleast 21 years old.

Variable transformation:

* Number of Instances: 768
* Number of Attributes: 8 plus class

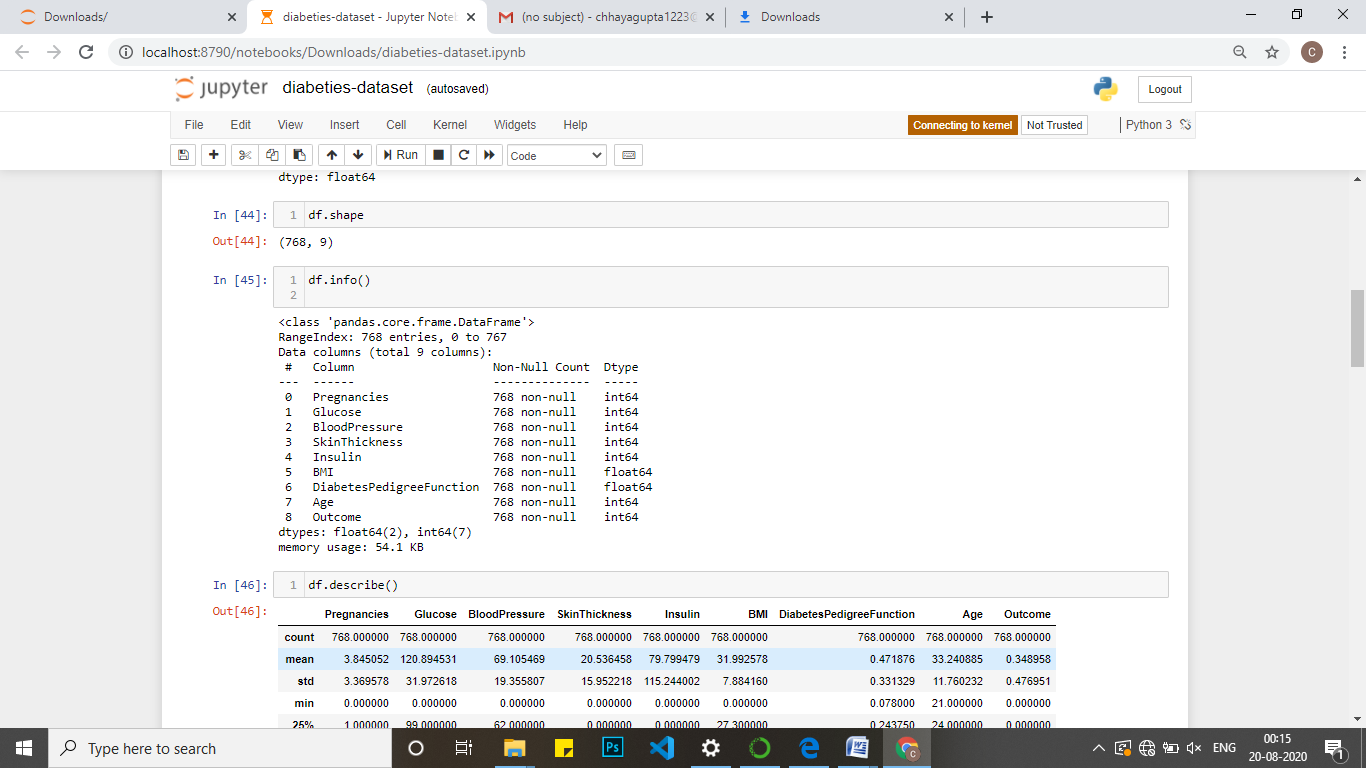
For Each Attribute: (all numeric-valued)

1. Pregnancies
2. Glucose
3. Blood pressure (mm Hg)
4. Skin thickness (mm)
5. Insulin (mu U/ml)
6. BMI(Body mass index) (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. outcome(0 or 1)

* Missing Attribute Values: Yes
* Class Distribution: (class value 1 is interpreted as "tested positive for  
  diabetes")
  1. Exploratory data analysis

Exploratory data analysis(EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given dataset there are 9 variables and data types of all variables are either float64 or int64 .there are 768 number of rows(observations) and 9 variables or columns. And missing values are present in our dataset.

Column names and their data types

Column name and their number of unique values

From the dataset we found there are 2 continuous variables and 07 categorical vars.

**2. Literature Survey**

2.1 Technology Used

2.1.1 PYTHON

Python is a high-level, general-purpose, open source, strictly typed programming language. The language provides constructs intended to enable clear programs on both a small and large scale. Python was created By Guido van Rossum. The Python Software Foundation (PSF) is the organization behind Python.

Python versions:

* First released in 1991.
* Python 2.0 was released on 16 October 2000
* Python 3.0 was released on 3 December 2008

Current Versions:

* Python 3.6.3
* Python 2.7.14

Python features:

Some of the features of python include :-

* Easy to understand
* Dynamic
* Object oriented
* Multipurpose
* Strongly typed
* Open Sourced

Python is mainly used in many domains:

* Web Development
* Data Analysis
* Machine Learning
* Internet Of Things
* GUI Development
* Image processing
* Data visualization
* Game Development

2.1.2 Anaconda

Anaconda is a open source Distribution for data science and machine learning using python. It includes hundreds of popular data science packages and the conda package and virtual environment manager for Windows, Linux, and MacOS. Conda makes it quick and easy to install, run, and upgrade complex data science and machine learning environments like scikit-learn, TensorFlow, and SciPy. Anaconda Distribution is the foundation of millions of data science projects as well as Amazon Web Service Machine Learning AMIs and Anaconda for Microsoft on Azure and Windows.

2.1.3 Packages

NumPy

NumPy is the fundamental package for scientific computing with Python.

It contains among other things:

* a powerful N-dimensional array object
* sophisticated (broadcasting) functions
* tools for integrating C/C++ and Fortran code
* useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users. 1.3.3 Scikit-learn

Scikit-learn provides machine learning libraries for python.Some of the features of Scikit-learn includes: • Simple and efficient tools for data mining and data analysis • Accessible to everybody, and reusable in various contexts • Built on NumPy, SciPy, and matplotlib • Open source, commercially usable - BSD license

Pandas

Pandas is an open source, BSD-licensed library providing highperformance, easy-to-use data structures and data analysis tools for the Python programming language.

Pandas library is well suited for data manipulation and analysis using python. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Seaborn

Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics

2.2 TRAINING WORK UNDERTAKEN

2.2.1 COLLECTING DATA FROM KAGGLE

Kaggle is a platform for predictive modelling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowd sourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know beforehand which technique or analyst will be most effective. On 8 March 2017, Google announced that they were acquiring Kaggle. They will join the Google Cloud team and continue to be a distinct brand. In January 2018, Booz Allen and Kaggle launched Data Science Bowl, a machine learning competition to analyze cell images and identify nuclei.

2.2.3 DATA SCIENCE

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

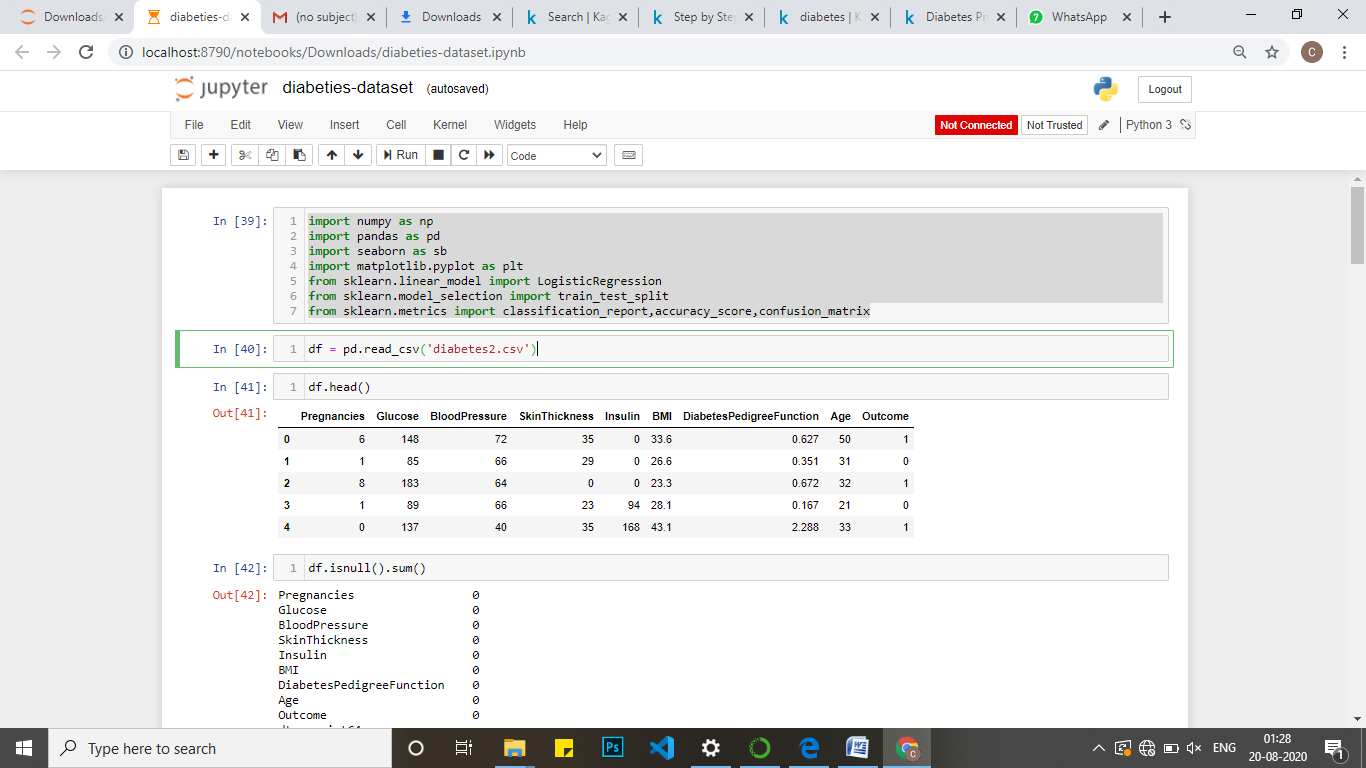
Turing award winner JiGray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. When Harvard Business Review called it "The Sexiest Job of the 21st Century" the term became a buzzword, and is now often applied to business analytics, business intelligence, predictive modeling, or any arbitrary use of data, or used as a glamorized term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness." While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents. Because of the current popularity of this term, there are many "advocacy efforts" surrounding the field. To its discredit, however, many data science and big data projects fail to deliver useful results, often as a result of poor management and utilization of resources.

**3. Source Code and Output**

3.1 IMPORT PACKAGES

* import numpy as np
* import pandas as pd
* import seaborn as sb
* import matplotlib.pyplot as plt
* from sklearn.linear\_model import LogisticRegression
* from sklearn.model\_selection import train\_test\_split
* from sklearn.metrics import classification\_report, accuracy\_score, confusion\_matrix

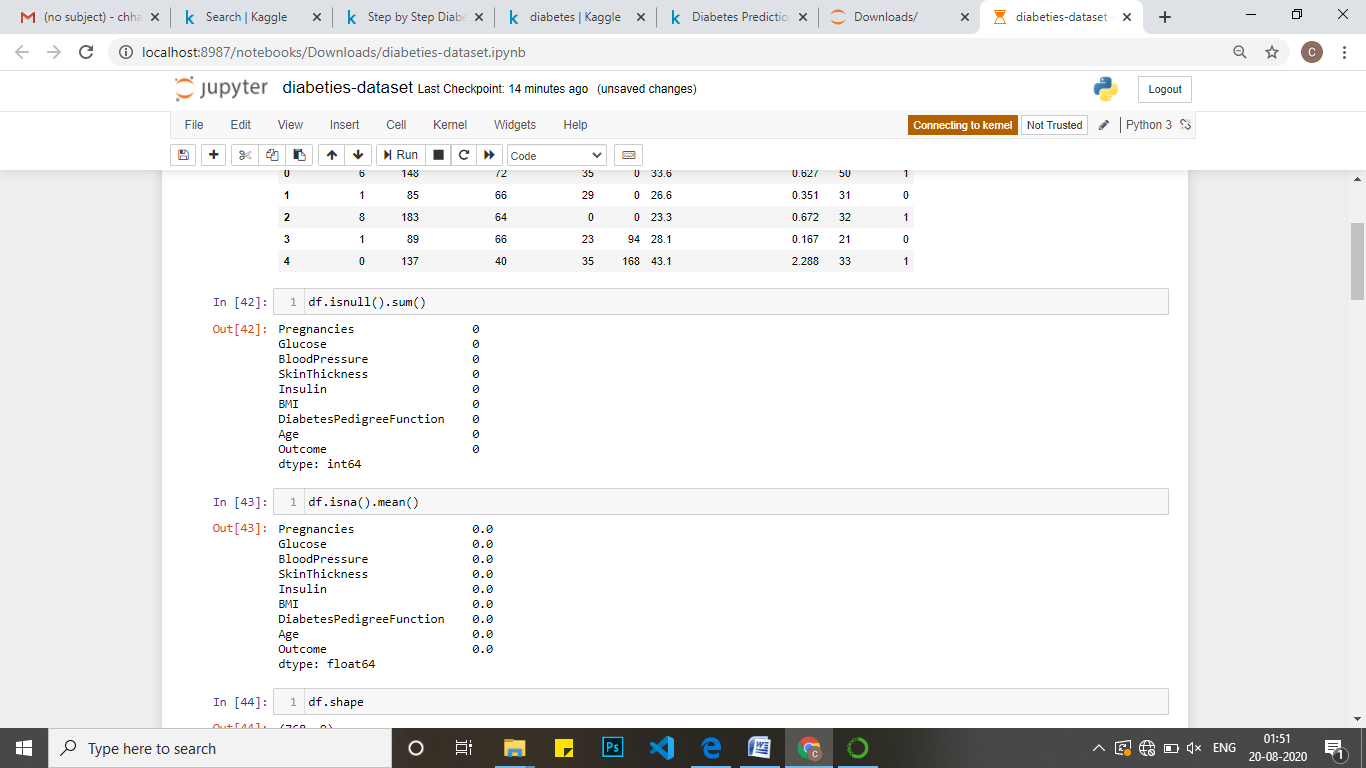
3.2 LOAD THE DATASET

* df=pd.read\_excel("diabetes.csv")
* df.head()

3.3 DATA PRE-PROCESSING

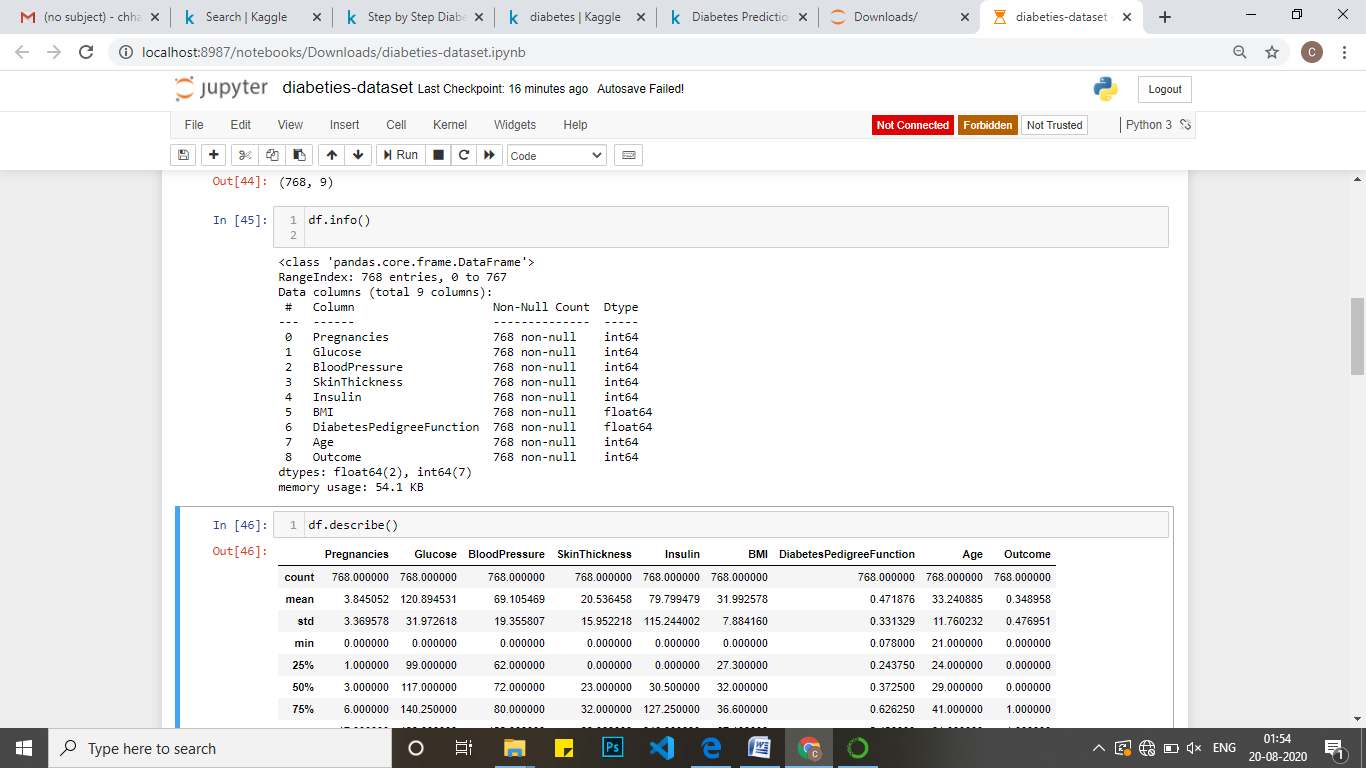
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots .this is often called as exploratory data analysis to start this process we will look at the all the probability distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

1. Checking for Missing values



No missing values found.

1. Checking Information on Original Dataset

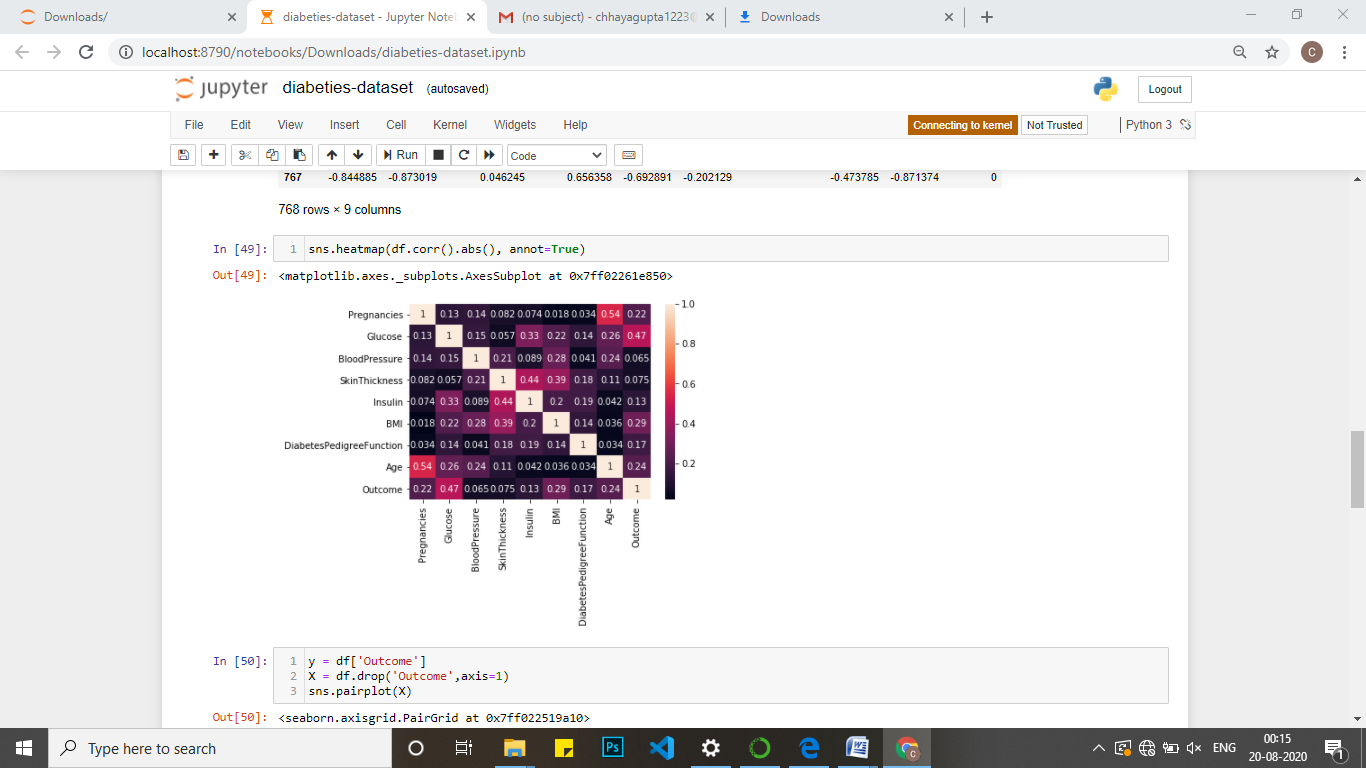


1. Dropping Un-useful Columns

* X = df.drop('Outcome',axis=1)

3.4 DATA REGRESSION (PREDICTION)

1) Checking for correlation among all the x(inputs)

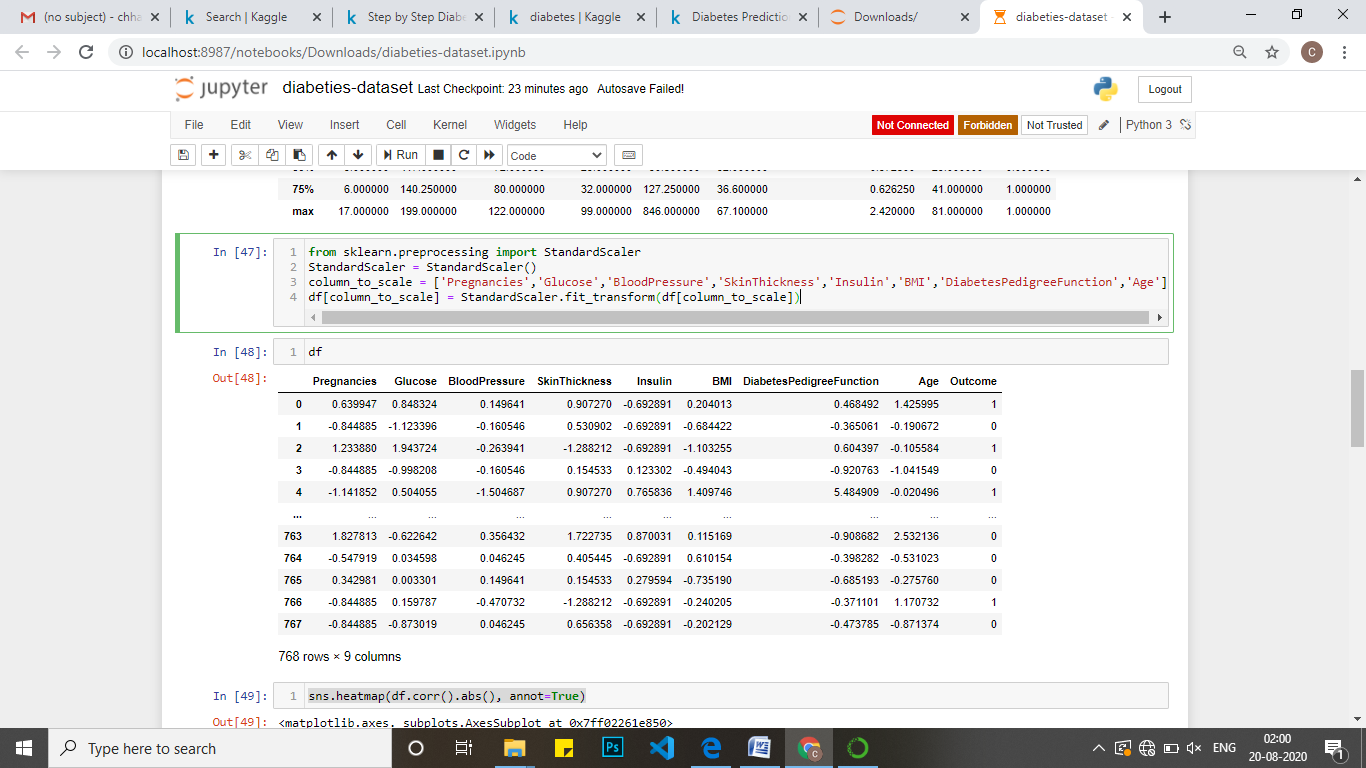
* sns.heatmap(df.corr().abs(), annot=True)

Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or feature of data. Since the range of values of raw data varies widely, in some machine learning algorithms, objective function will not work properly without normalization. For example, the majority of classiers calculate the distance between two points by the euclidean distance. If one of the feature has broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributions approximately proportionately to that distance since our data is not uniformly distributed.

In this we have used standardization method.

* StandardScaler = StandardScaler()
* column\_to\_scale = ['Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age']
* df[column\_to\_scale] = StandardScaler.fit\_transform(df[column\_to\_scale])



Feature Splitting

Before divide the dataset into the train and test case we do sampling of the data if we have long data like millions of rows but we have little data like 768 rows so there is no need to do sampling we applied splitting the original dataset and divide it into the train and test case.

* X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.33, random\_state=43)

Models

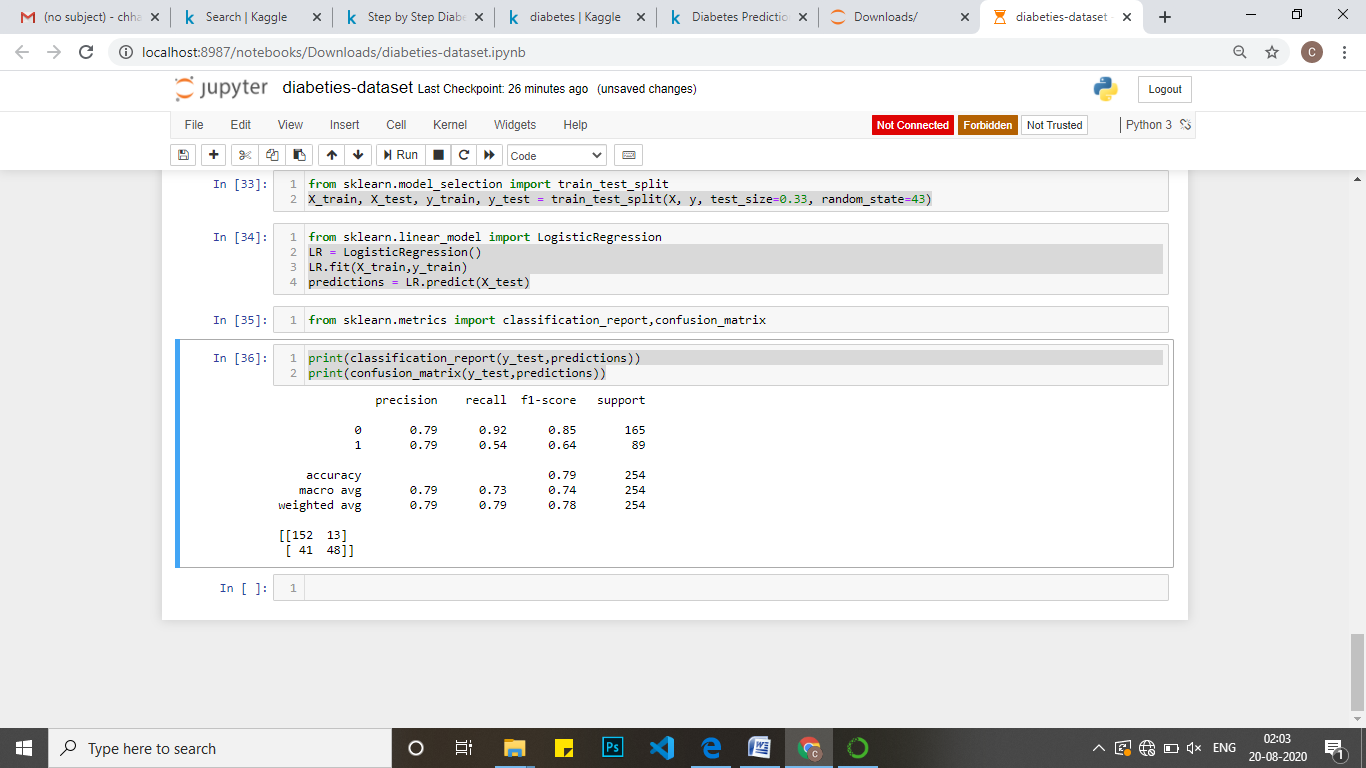
After pre processing we will be using Logistic Regressor model on our processed data to predict the target variable.

* LR = LogisticRegression()
* LR.fit(X\_train,y\_train)
* predictions = LR.predict(X\_test)

**4. RESULTS**

As mentioned above in prediction part the performance values of algorithm calculated on various measures.

On analysis the scores we can conclude that Logistic Regression algorithm is giving maximum accuracy of 79%

****

**5.Conclusion**

Analyzing and predicting diabetes according to given problem statement is done by us and systematic efforts are made in designing a system which results in the prediction of diabetes.

During this work, various machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on a Diabetes patient dataset with 768 total observations and 9 attributes.

Experimental results determine the adequacy of the designed system with an archieved accuracy of 79% using the Logistic Regression classification algorithm.

In future the work can be extended and improved for the automation of diabetes analysis and prediction of patients, including some other machine learning algorithms.

**6. References:**

* https://www.kaggle.com/
* https://www.python.org/
* <https://anaconda.org/anaconda/python/>
* http://www.numpy.org/
* <https://matplotlib.org/>
* http://scikit-learn.org/
* https://pandas.pydata.org/