EXPOSYS DATA LABS

PROJECT REPORT

**PREDICTION OF DIABETES USING DATA SCIENCE**

Presented by: PALAK AGARWAL

SHAGUN GARG

**TABLE OF CONTENT**

Acknowledgment

Abstract

1. Introduction
2. Conventional Machine Learning techniques

2.1. Supervised Learning

2.2. Unsupervised Learning

2.3. Reinforcement Learning

1. Proposed Model
2. Methodology

4.1 About the dataset

4.2 Problem Statement

4.3 Dive In

1. Implementation

5.1 Data exploration and Visualization

5.2 Checking the Distribution of the Predictor variable

5.2.1 Distribution of Glucose

5.2.2 Distribution of Insulin

5.2.3 Distribution of BMI

5.2.4 Distribution of Pedigree function

5.2.5 Distribution of Age

5.3 Checking for any missing values in dataset

5.4 Plotting Relationship in the Dataset

5.5 Training the Dataset

5.6 Building the model

1. Libraries Used in the Project

Conclusion

**ACKNOWLEDGEMENT**

We would like to express our heartfelt gratitude to our Mentor for giving us the opportunity to work on the project and help us gain immensely professional experience.

It helped me in doing a lot of Research and I came to know about so many new things. I am really thankful to them*.*

The successful completion of our project would not have been possible without the dedicated support from all our mentors, family and friends.

PALAK AGARWAL

SHAGUN GARG

**ABSTRACT**

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to the 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 increased 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 a 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: SVM, Logistic regression, ANN. This project also aims to propose an effective technique for earlier detection of the diabetes disease.

Keywords: Machine Learning, Supervised, Svm, Ann, Logistic Regression

**1. INTRODUCTION**

Diabetes Mellitus Diabetes is one of the 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,

1. type 1 diabetes (T1D) and
2. 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 autoimmune logical 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.

**2. CONVENTIONAL MACHINE LEARNING TECHNIQUES**

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 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, KNN method, Logistic regression and Artificial Neural Network algorithms.

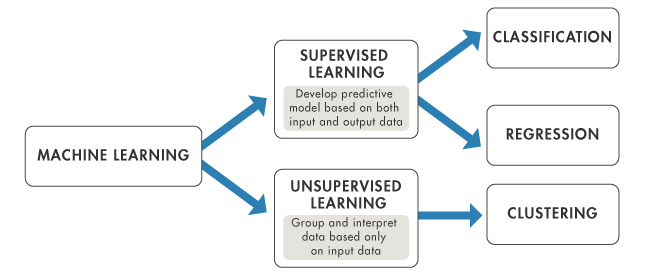


fig 1: Types of machine learning

**2.1. 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.

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).

**2.2. UNSUPERVISED LEARNING**

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels. 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.

**2.3. REINFORCEMENT LEARNING**

The term Reinforcement Learning is a general term given to a family of techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward. It is important to mention that the system has no prior knowledge about the behaviour of the environment and the only way to find out is through trial and failure (trial and error). Reinforcement learning is mainly applied to autonomous systems, due to its independence in relation to its environment.

**3. PROPOSED SYSTEM**

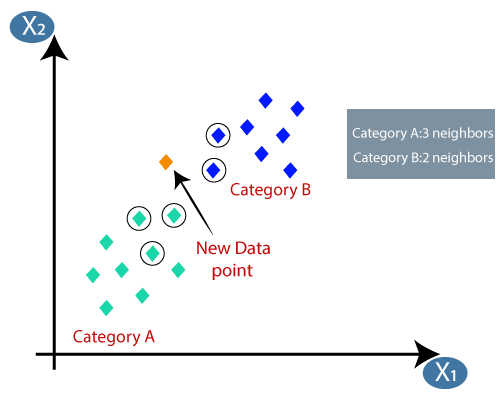
Classification is one of the most important decision making techniques in many real world problems.

In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classification problems, the higher number of samples chosen doesn't lead to higher classification accuracy.

In many cases, the performance of algorithms is high in the context of speed but the accuracy of data classification is low.

The main objective of our model is to achieve high accuracy. Classification accuracy can be increased if we use much of the data set for training and few data sets for testing.

This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data. Thus, it is observed that techniques like KNN etc for implementing the Diabetes prediction system.



**4. METHODOLOGY**

**4.1. ABOUT THE DATASET**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.

It is provided courtesy of the Pima Indians Diabetes Database . It consists of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The dataset has 9 columns as shown below:

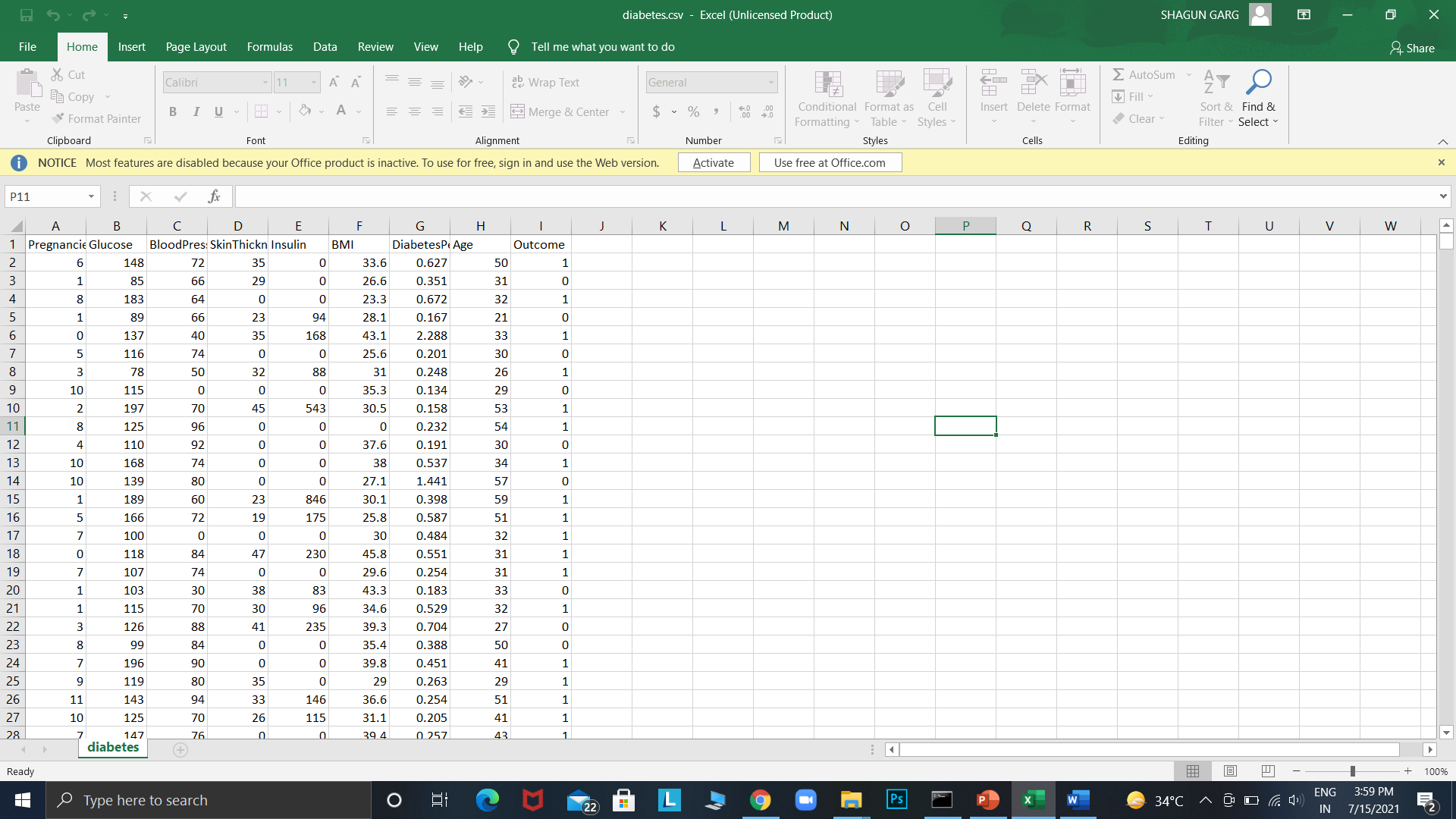


fig 2: Dataset used

**Pregnancies**              Number of times pregnant

**Glucose** Plasma glucose concentration a 2 hours in an oral glucose tolerance test

**Blood Pressure**          Diastolic blood pressure (mm Hg)

**Skin Thickness**          Triceps skinfold thickness (mm)

**Insulin**                     2-Hour serum insulin (mu U/ml)

**BMI**                         Body mass index (weight in kg/(height in m)^2)

**Diabetes Pedigree Function** Diabetes pedigree function

**Age**                           Age (years)

**Outcome**                   Class variable (0 or 1) 268 of 768 are 1, the others are 0

**4.2. PROBLEM STATEMENT**

This is a classification problem of supervised machine learning.

The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

**0 – Absence of Diabetes**

**1 – Presence of Diabetes**

**4.3. DIVE IN**

In this project, we will use python and Jupyter notebook. Feel free to use your preferred IDE. We will go through the project by importing the dataset, conducting exploratory data analysis to get insights and understanding on how the dataset looks like and then build the model by using KNN machine learning technique.

The following general framework of steps in supervised machine learning;

1. Data Collection
2. Data Preparation
3. Choosing a model
4. Training the model
5. Evaluating the model
6. Parameter tuning
7. Making prediction

So let us begin by importing the required libraries. We will import data analysis libraries (pandas, numpy) and visualization libraries (matplotlib, seaborn). In addition to that,

If using jupyter notebook, it is important to set %matplotlib inline to show the visualizations on the notebook.

And for the modelling we have used the Sklearn library in our project.

**5. IMPLEMENTATION**

**5.1 DATA EXPLORATION AND VISUALIZATION**

In this section, we will create graphs to display different distributions of the data and available relationships to allow us to understand it much better. This is a very critical section since it determines how the model will be built.

**5.1.1 CHECKING THE DISTRIBUTION OF TARGET VARIABLE**

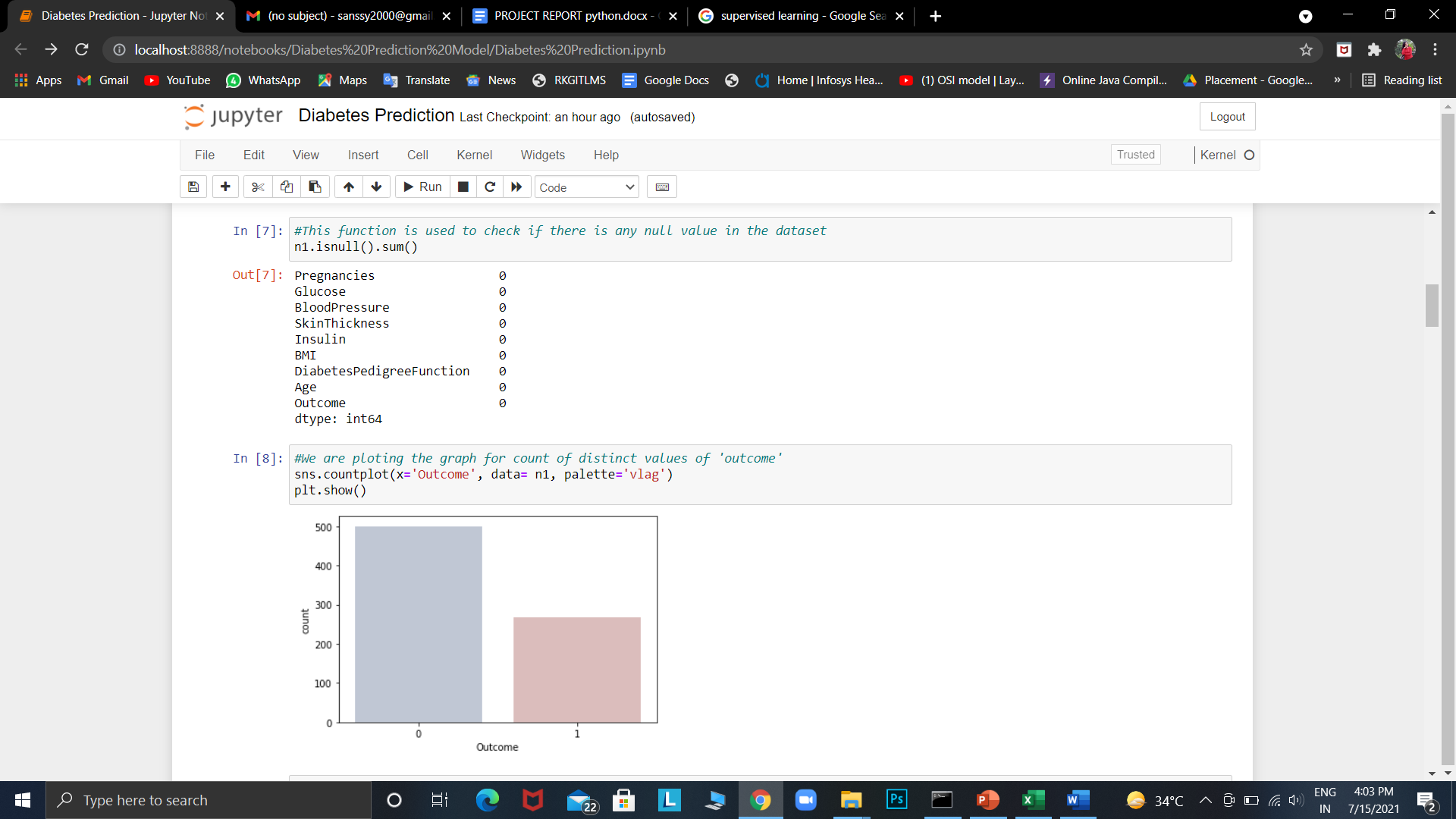


fig 3: Graph for count of distinct values of outcome

**5.2. CHECKING THE DISTRIBUTION OF THE PREDICTOR VARIABLES**

Let us plot each variable to show its distribution in the dataset.

**5.2.1. DISTRIBUTION OF GLUCOSE:**

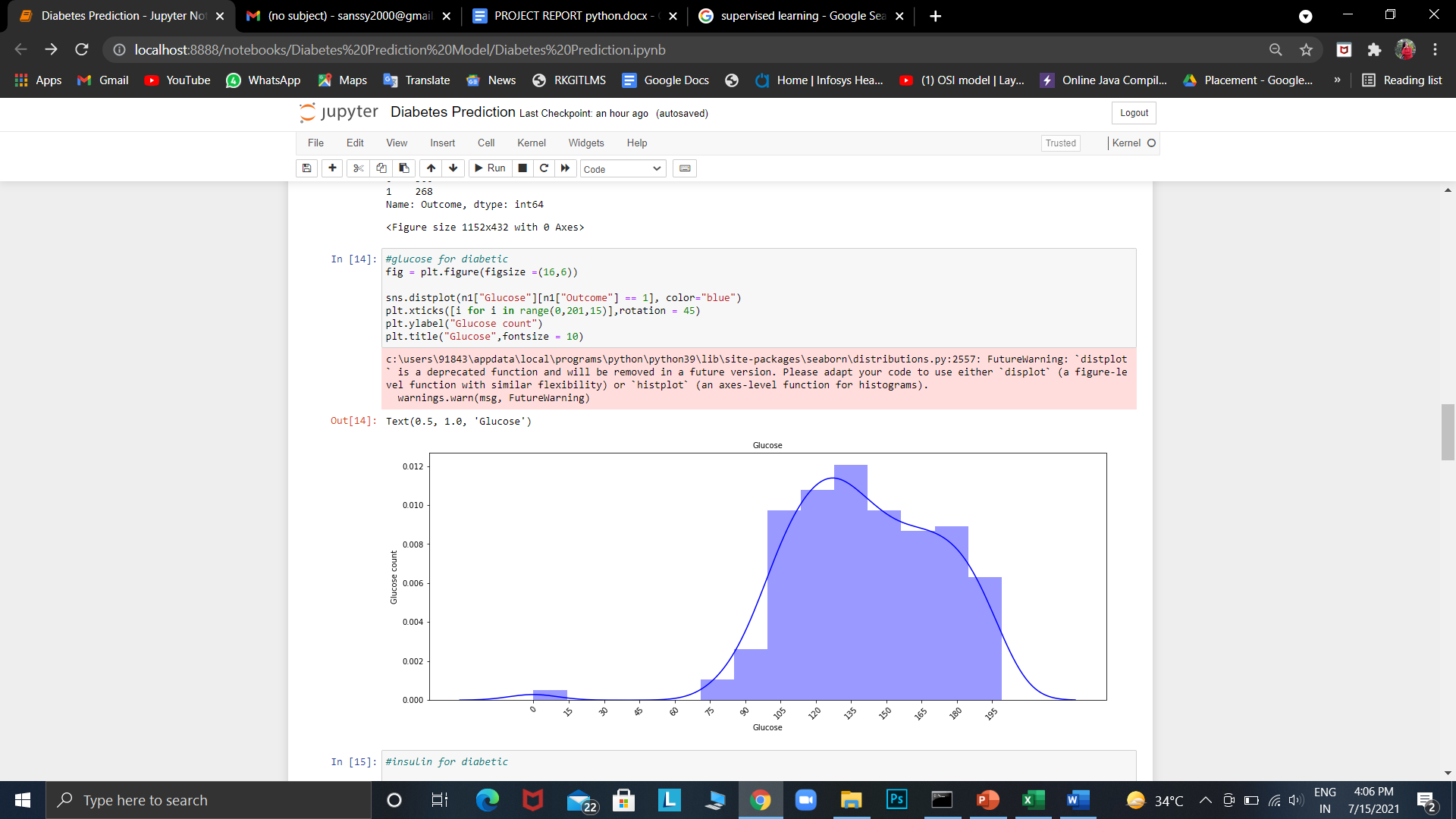


fig 4: Graph for Glucose distribution

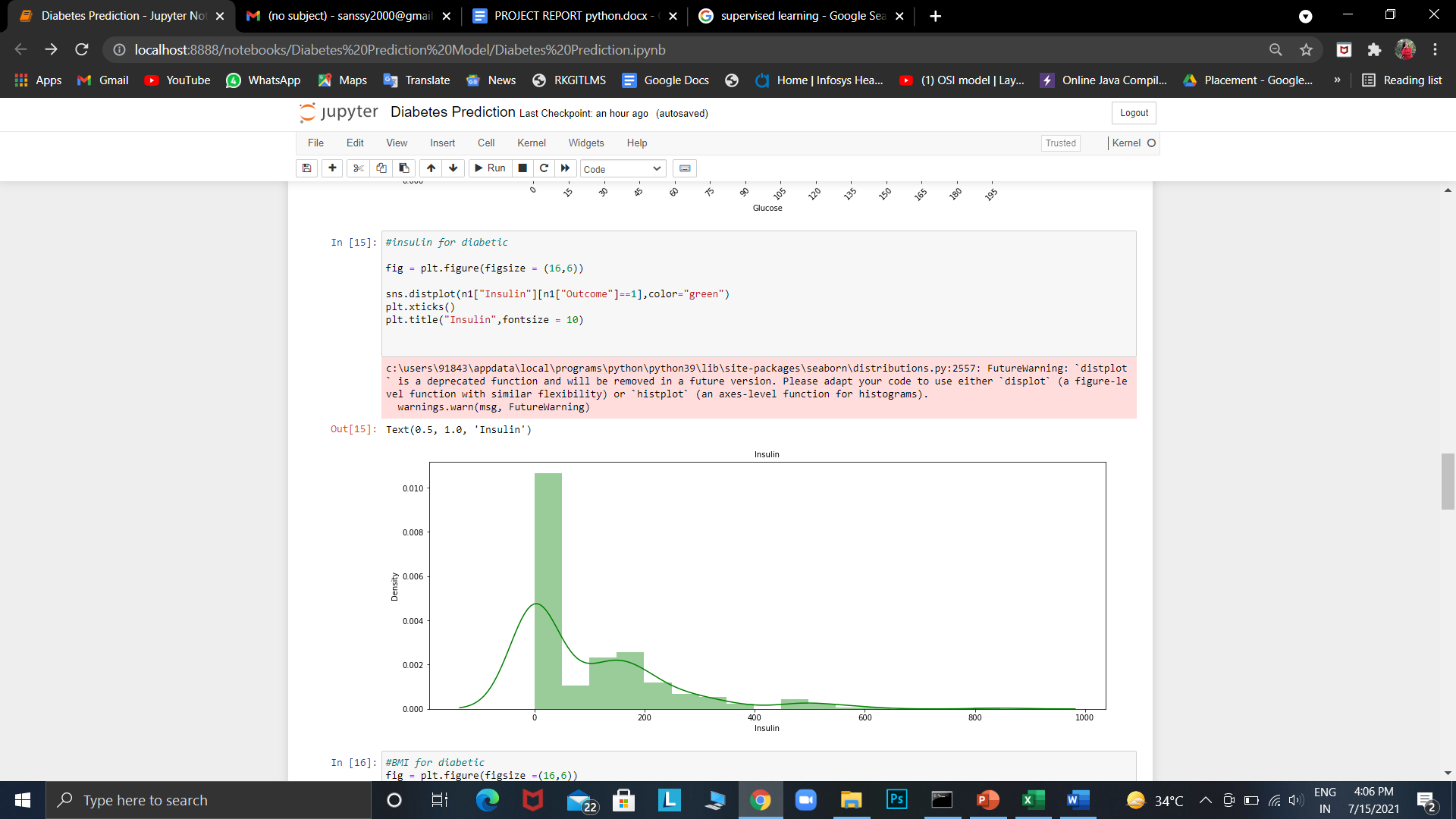
**5.2.2. DISTRIBUTION OF INSULIN:**

fig 5: Graph for Insulin distribution

**5.2.3. DISTRIBUTION OF BMI:**

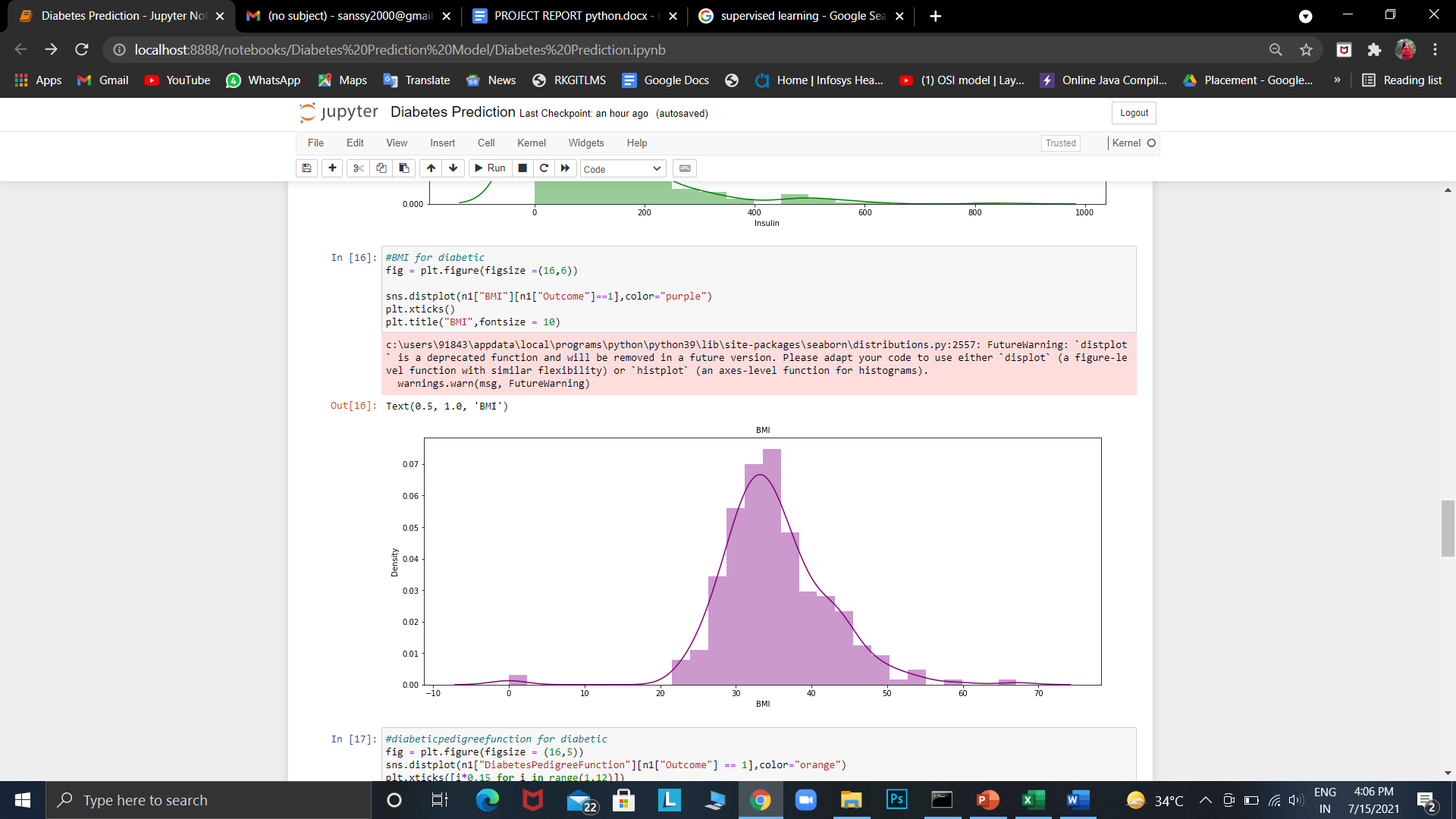


fig 6: Graph for BMI distribution

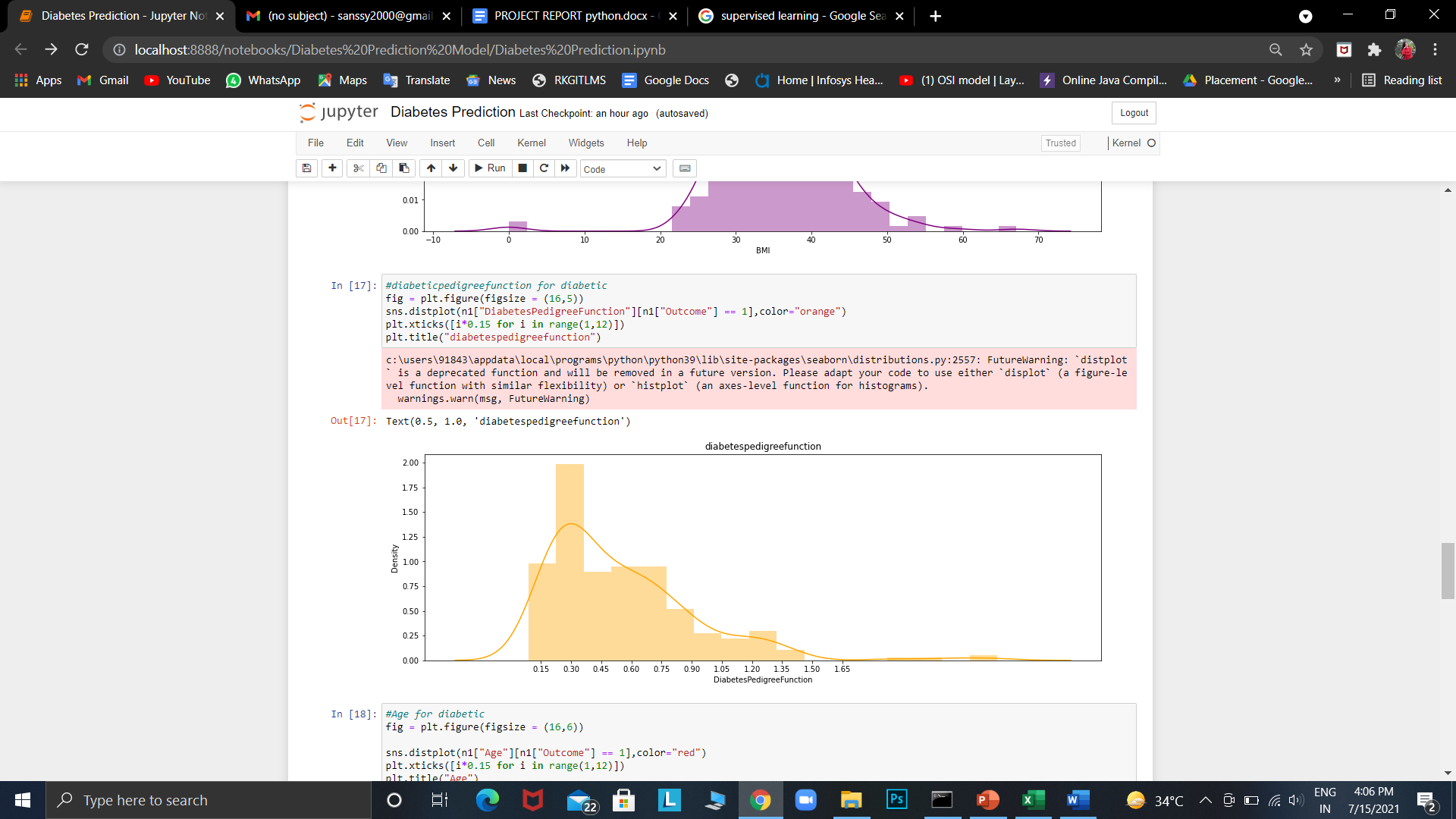
**5.2.4. DISTRIBUTION OF DIABETES PEDIGREE FUNCTION:**

fig 7: Graph for diabetes pedigree function

**5.2.5. DISTRIBUTION OF AGE:**

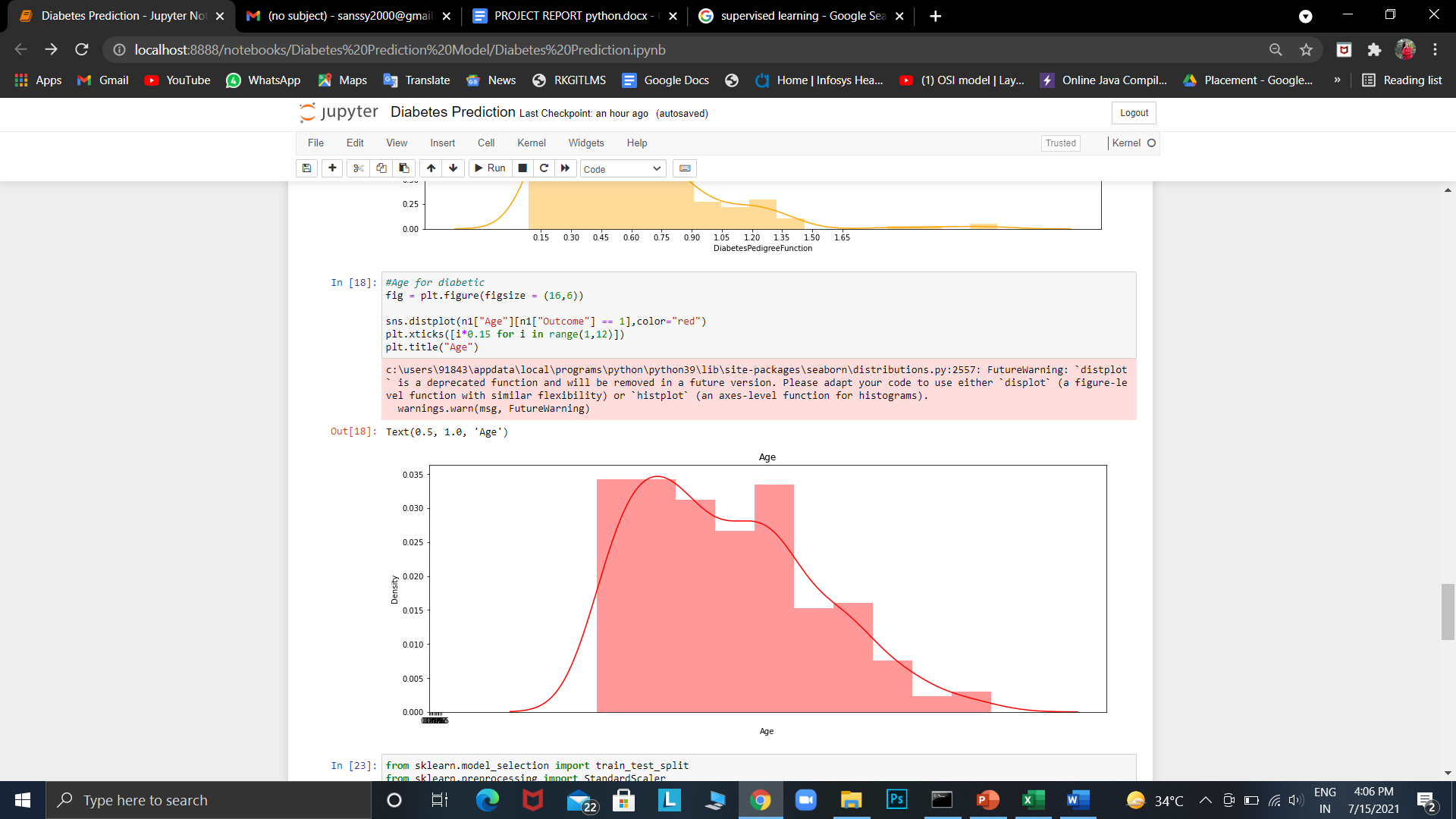


fig 8: Graph for age distribution

**5.3. CHECKING FOR ANY MISSING VALUES IN DATASET**

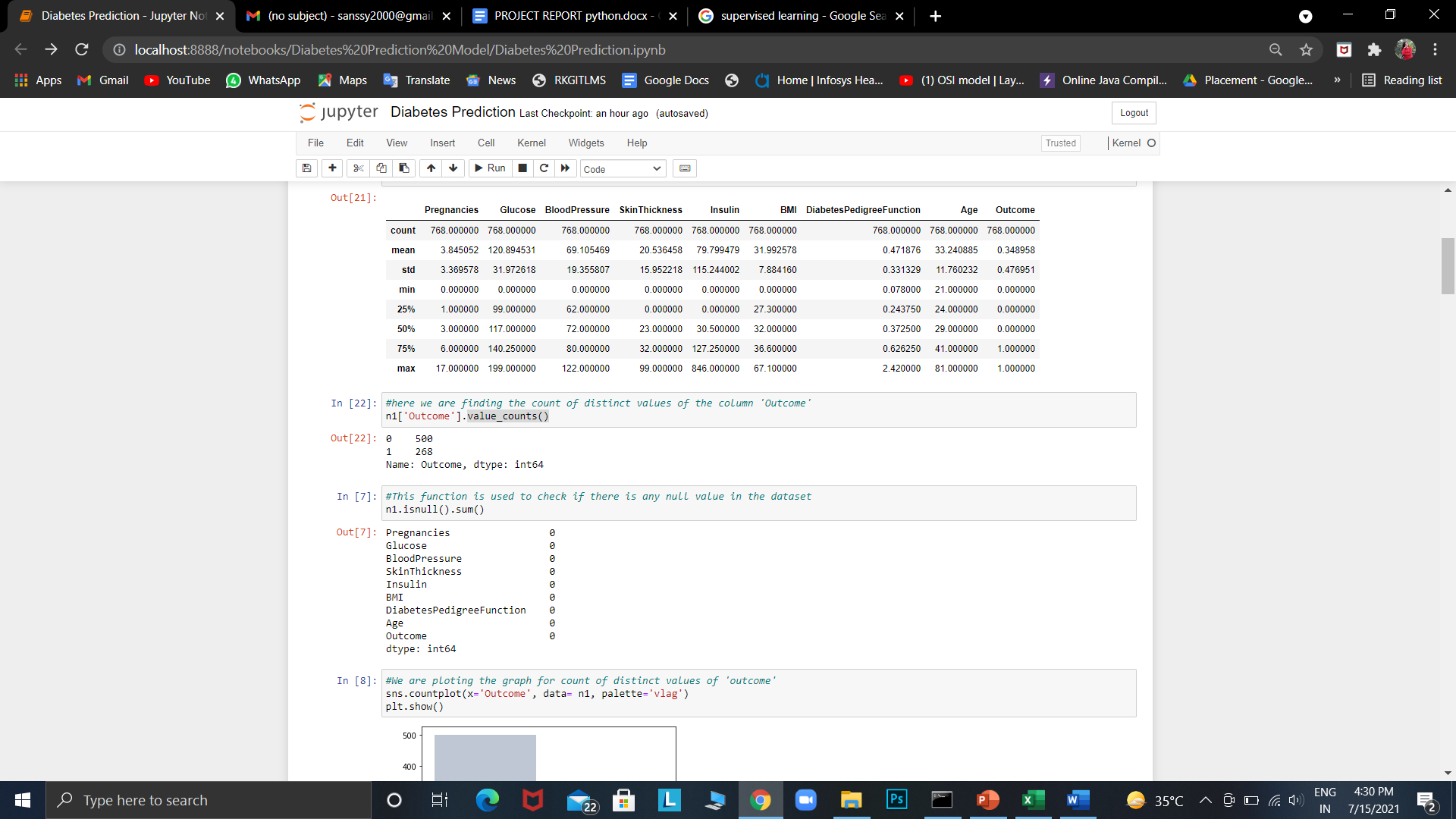


fig 9: No null values in the dataset

There are no missing values in the dataset. The dataset had already been cleaned. Sometimes while working on a project, you may find a dataset with missing values. It is important to know how to handle the missing data.

**5.4. PLOTTING RELATIONSHIPS IN THE DATASET**

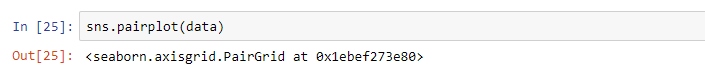


fig 10: plotting relationship

Next, we will proceed in checking the relationships by visualizing correlations as shown in the table below.

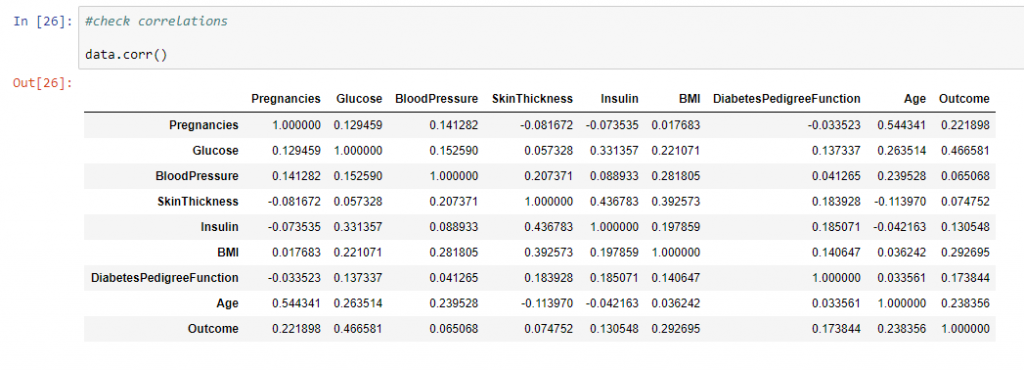


fig 11: Finding correlation

The table displays specific correlations for different variables in the dataset in probability form.

**5.5. TRAINING THE DATA**

We will now split our dataset before we train it. X will contain all the Independent variables while y will have the Dependent variable (Outcome).

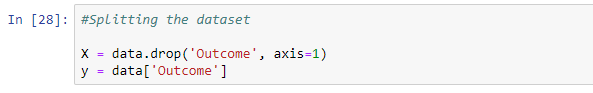


fig 12: Splitting the dataset

After successfully splitting the dataset, let us train it using train\_test\_split.

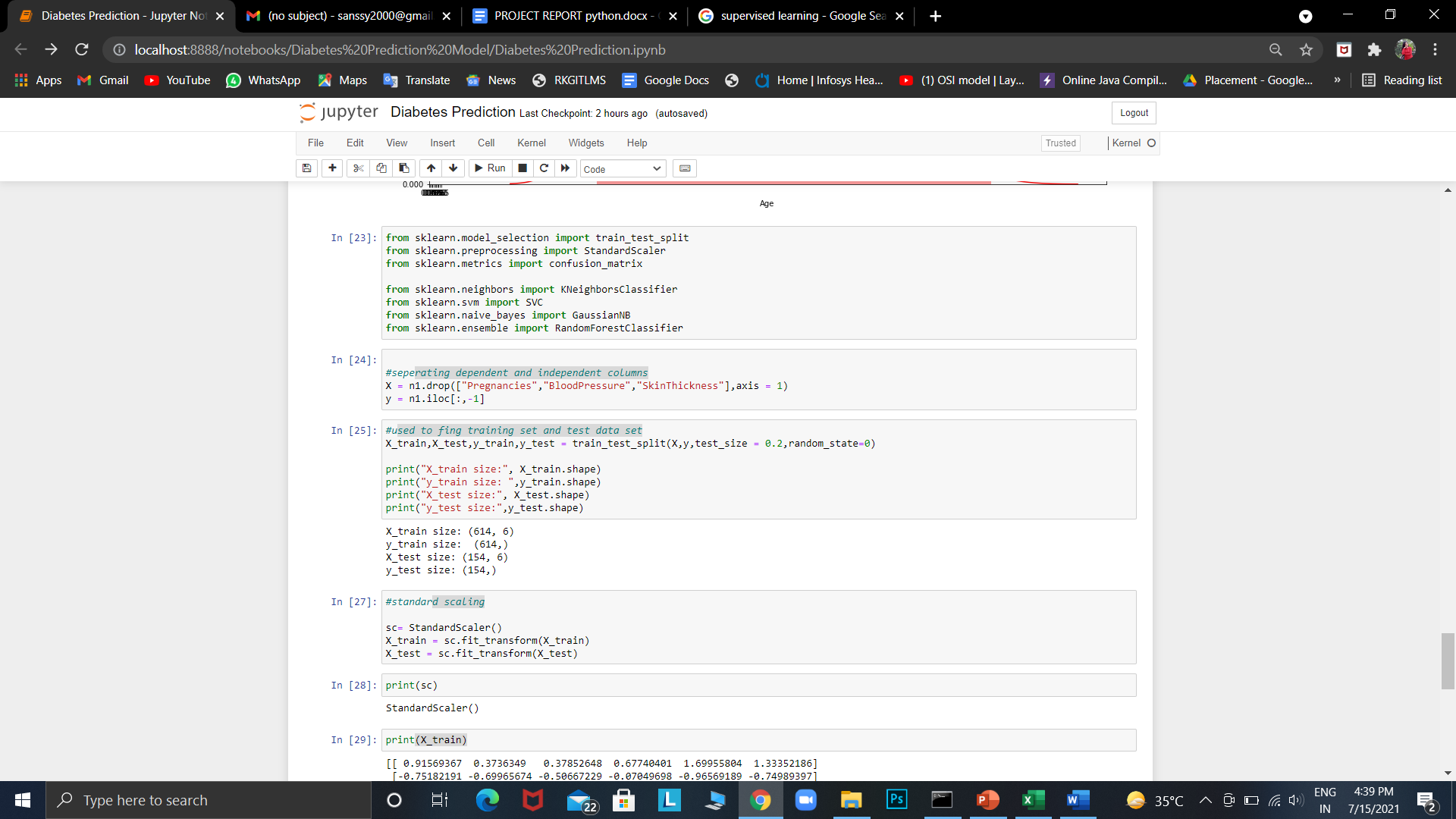


fig 13: dividing into test data and training data

Before we build the model, let us impute the zero values in our dataset. If you check the head of the dataset, you will notice that there are some independent variables with zero values. This can make our model not efficient.

We therefore need to impute the zero values by using the mean of the other values in the same column. The code below shows how we can check the zero values in the dataset by printing for each variable.



fig 14: checking column for zero value

**5.6. BUILDING THE MODEL**

For building the machine learning model we have used KNN method, in order to find the accuracy of our build model.

#### 

fig 15: Modeling through KNN technique

**6. LIBRARIES USED IN THE PROJECT**

**Numpy**: version NumPy 1.20.0

**Pandas**: version: 1.3.0

**Matplotlib:** version 3.4.2. pip

**Seaborn:** version 0.11.1

**Sklearn:** version 0.22.2

**CONCLUSION**

1. Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of a large amount of epidemiological and genetic diabetes risk dataset.

2. Detection of diabetes in its early stages is the key for treatment.

3. This work has described a machine learning approach to predicting diabetes levels.

4. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decisions about the disease status.