**THE NORTHCAP UNIVERSITY**

**SECTOR-23A, GURUGRAM**

**HARYANA – 122017**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING(CSE)**

**-CSL313**

**MACHINE LEARNING PROJECT REPORT**

**ON**

**PIMA INDIANS DIABETES DATASET**

**TEAM MEMBERS:**

**ADITYA (18CSU011)**

**AMAN(18CSU015)**

**DEEPANSHU(18CSU059)**

**DEVESH(18CSU061)**

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | TITLE | PAGE NO. | REMARKS |
| 1. | Abstract | 4 |  |
| 2. | Introduction | 5 |  |
| 3. | ABOUT THE DATASET | 6 |  |
| 4. | TRAINING DESCRIPTION | 8 |  |
| 5. | DATA VISUALIZATIONS | 14 |  |
| 6. | CLASSIFIER | 25 |  |
| 7. | REGRESSION | 27 |  |
| 8. | CONCLUSION | 31 |  |
| 9. | BIBLIOGRAPHY | 32 |  |

**ABSTRACT:-**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence.The remarkable advances in biotechnology and health sciences have led to a significant production of data, such as high throughput genetic data and clinical information, generated from large [Electronic Health Records](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/electronic-health-record) (EHRs). To this end, application of machine learning and [data mining](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/data-mining) methods in biosciences is presently, more than ever before, vital and indispensable in efforts to transform intelligently all available information into valuable knowledge. Diabetes mellitus (DM) is defined as a group of [metabolic disorders](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/metabolic-disorders) exerting significant pressure on human health worldwide. Extensive research in all aspects of diabetes (diagnosis, etiopathophysiology, therapy, etc.) has led to the generation of huge amounts of data. The aim of the present study is to conduct a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management with the first category appearing to be the most popular. A wide range of machine [learning algorithms](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/learning-algorithm) were employed. In general, 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules. Support vector machines (SVM) arise as the most successful and widely used algorithm. Concerning the type of data, clinical datasets were mainly used. The title applications in the selected articles project the usefulness of extracting valuable knowledge leading to new hypotheses targeting deeper understanding and further investigation in DM.

**INTRODUCTION**

Signiﬁcant advances in biotechnology and more speciﬁcally high- throughput sequencing result incessantly in an easy and inexpensive data production, there by ushering the science of applied biology into the area of machine learning .The increasingly growing number of applications of machine learning in healthcare allows us to glimpse at a future where data, analysis, and innovation work hand-in-hand to help countless patients without them ever realizing it. Soon, it will be quite common to find ML-based applications embedded with real-time patient data available from different healthcare systems in multiple countries, thereby increasing the efficacy of new treatment options which were unavailable before.

One of the chief ML applications in healthcare is the identification and diagnosis of diseases and ailments which are otherwise considered hard-to-diagnose. This can include anything from cancers which are tough to catch during the initial stages, to other genetic diseases. IBM Watson Genomics is a prime example of how integrating cognitive computing with genome-based tumor sequencing can help in making a fast diagnosis.

Maintaining up-to-date health records is an exhaustive process, and while technology has played its part in easing the data entry process, the truth is that even now, a majority of the processes take a lot of time to complete. The main role of machine learning in healthcare is to ease processes to save time, effort, and money. Document classification methods using vector machines and ML-based OCR recognition techniques are slowly gathering steam, such as Google's Cloud Vision API and MATLAB's machine learning-based handwriting recognition technology. MIT is today at the cutting edge of developing the next generation of intelligent, smart health records, which will incorporate ML-based tolls from the ground up to help with diagnosis, clinical treatment suggestions, etc.

**ABOUT THE DATASET**

We get our data set from kaggle from a ongoing competition.

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

## Content

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes their BMI, insulin level, age, and so on.

## ACKNOWLEDGEMENTS

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). [Using the ADAP learning algorithm to forecast the onset of diabetes mellitus](http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81). In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

MOTIVATION :-

There has been drastic increase in rate of people suffering from diabetes since a decade. Current human lifestyle is

the main reason behind growth in diabetes. In current medical diagnosis method, there can be three different types of

errors-

1. The false-negative type in which a patient in reality is already a diabetic patient but test results tell that the person

is not having diabetes.

2. The false-positive type. In this type, patient in reality is not a diabetic patient but test reports say that he/she is a

diabetic patient.

3. The third type is unclassifiable type in which a system cannot diagnose a given case. This happens due to

insufficient knowledge extraction from past data, a given patient may get predicted in an unclassified type.

However, in reality, the patient must predict either to be in diabetic category or non-diabetic category. Such errors

in diagnosis may lead to unnecessary treatments or no treatments at all when required. In order to avoid or reduce

severity of such impact, there is a need to create a system using machine learning algorithm and data mining

techniques which will provide accurate results and reduce human effort

**TRAINING DESCRIPTION:-**

About the attributes

Age:-

The risk of type 2 diabetes increases as you get older, especially after age 45. That's probably because people tend to exercise less, lose muscle mass and gain weight as they age. But type 2 diabetes is also increasing dramatically among children, adolescents and younger adults

BMI:-

When you're talking about diabetes and weight, your doctor will likely refer to your body mass index (BMI), which is a measure of how much body fat you have. It's an important measurement for diabetes and weight management, according to the American Diabetes Association (ADA), but it does have its limitations

Insulin:-

Insulin helps control blood glucose levels by signaling the liver and muscle and fat cells to take in glucose from the blood. Insulin therefore helps cells to take in glucose to be used for energy. If the body has sufficient energy, insulin signals the liver to take up glucose and store it as glycogen.

Insulin resistance occurs when your cells stop responding to the hormone insulin. This causes higher insulin and blood sugar levels, potentially leading to type 2 diabetes.

Skin:-

People who have diabetes tend to get skin infections. If you have a skin infection, you'll notice one or more of the following: Hot, swollen skin that is painful. An itchy rash and sometimes tiny blisters, dry scaly skin, or a white discharge that looks like cottage cheese

Diastolic blood pressure:-

The diastolic reading, or the bottom number, is the pressure in the arteries when the heart rests between beats. This is the time when the heart fills with blood and gets oxygen. A normal diastolic blood pressure is lower than 80. A reading of 90 or higher means you have high blood pressure..

## Load libraries

We will be sticking to Python  . The very first step is to load or [import](https://docs.python.org/2.0/ref/import.html) the all the libraries and the packages required to get the results we want. Some very primary and almost necessary packages for Machine Learning are — [NumPy](https://numpy.org/doc/), [Pandas](https://pandas.pydata.org/docs/), [Matplotlib](https://matplotlib.org/contents.html)and seaborn.

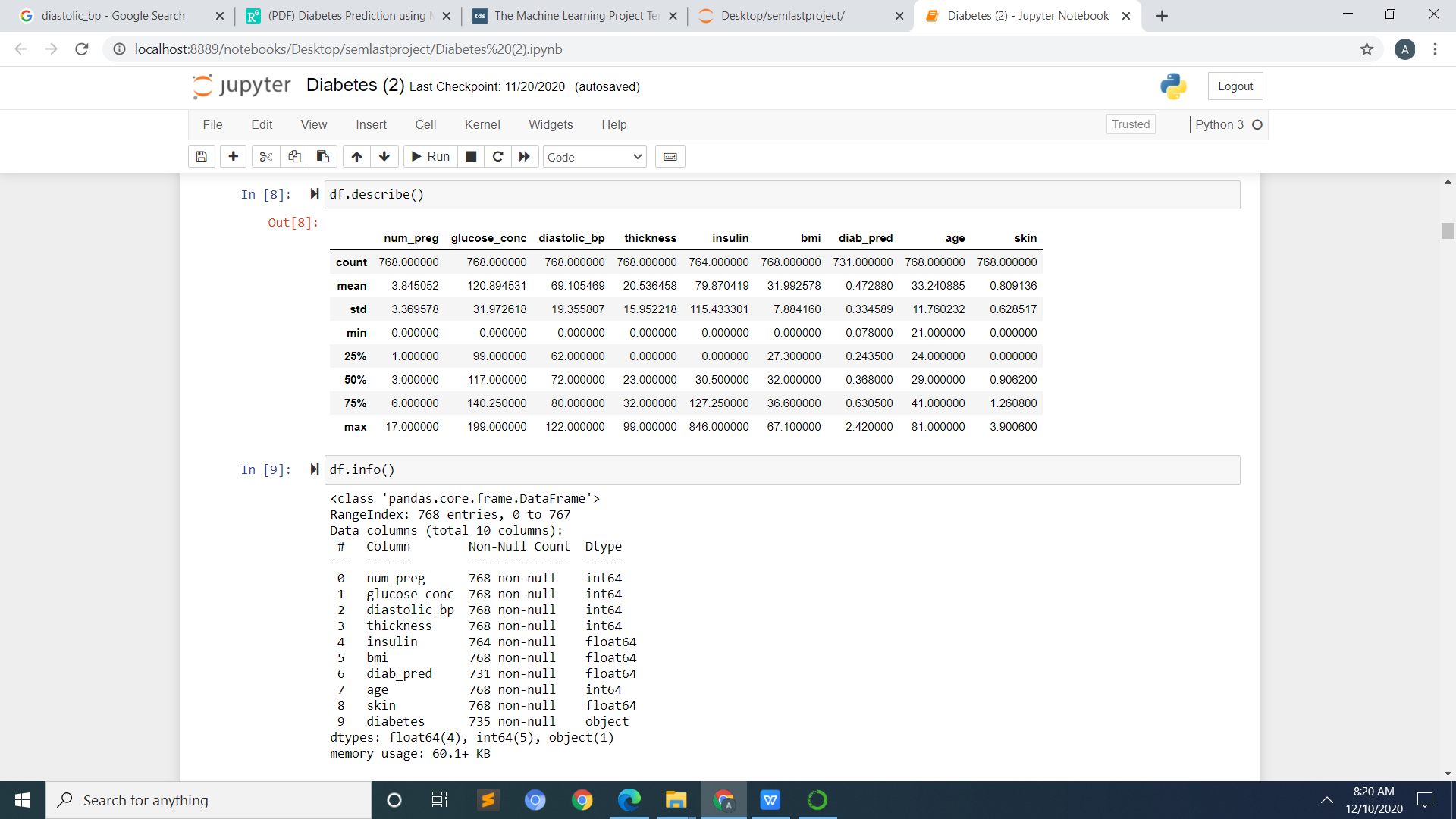
## Load dataset

Once the libraries are loaded, we need to get the data loaded. Pandas has a very straightforward function to perform this task — [pandas.read\_csv](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html). The read.csv function is not just limited to csv files, but also can read other text based files as well. Other formats can also be read using pandas read functions like html, json, pickled files etc. One thing which needs to be kept in mind is that your data needs to be in the same working directory as your current working directory or you will need to provide the complete path prefixed with a ‘/’ within the function.

# Summarize Data

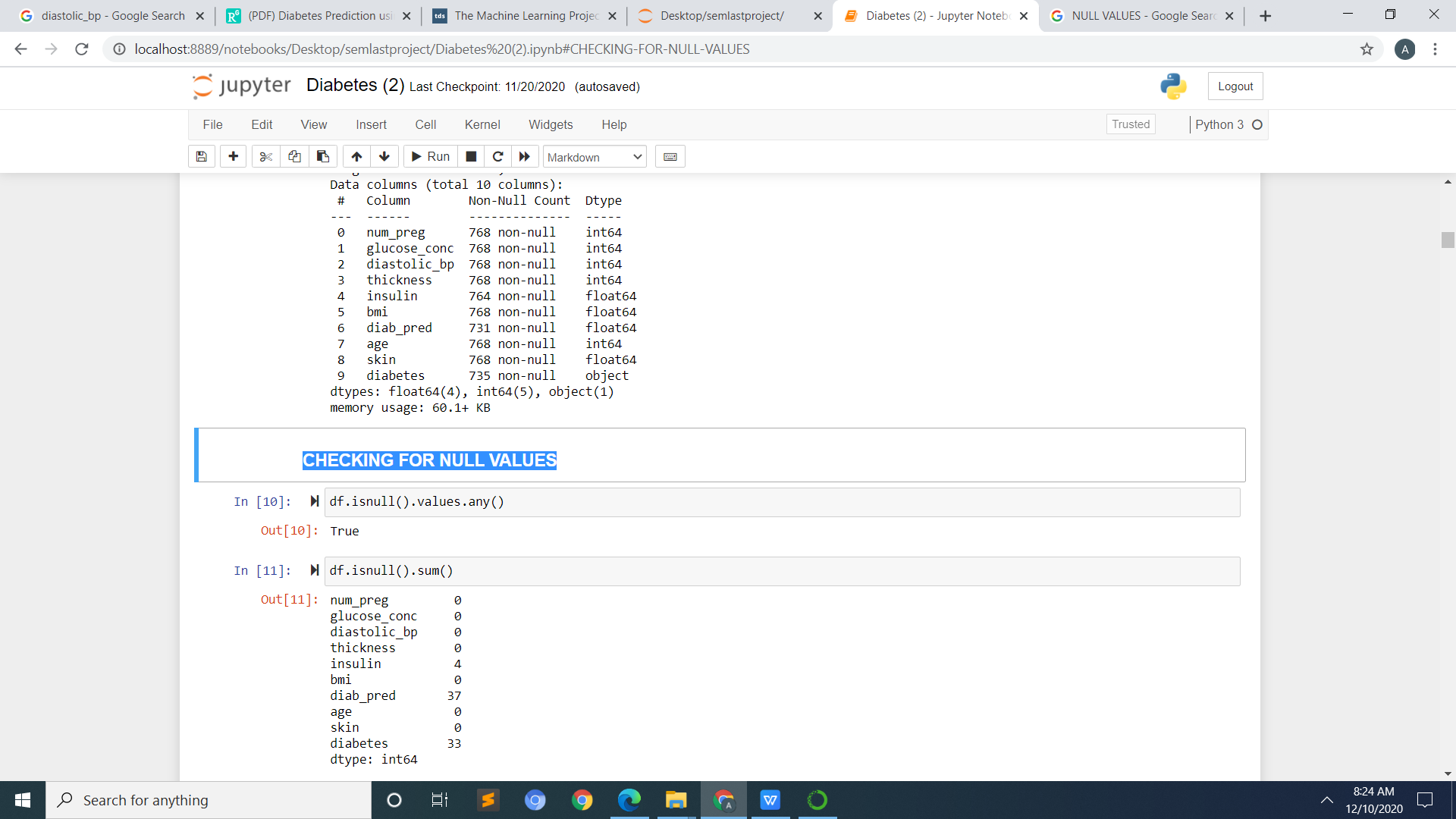
Okay, so the data is loaded and ready to be actioned upon. But we first need to check how the data looks and what all does it contain. To begin with, you would want to see how many rows and columns does the data have and what all are the data types of each column .

A quick way to take a look type and shape of your data is — [pandas.DataFrame.info](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.info.html). This tells you how many rows and columns your dataframe has and what data types and values do they contain.



### CHECKING FOR NULL VALUES

A field with a **NULL value** is a field with no **value**. It is very important to understand that a **NULL value** is different than a zero **value** or a field that contains spaces.



### REMOVING NULL VALUES:-

Removing null values from the dataset by filling in mean

## Data Cleaning

Real life data is not arranged and presented to you nicely and in a dataframe with no abnormalities. Data usually has a lot of so called abnormalities like missing values, a lot of features with incorrect format, features on different scales etc. All this needs to be handled manually which takes a lot of time and coding skills .

Pandas has various functions to check for such abnormalities like [pandas.DataFrame.isna](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.isna.html) to check for values with NaNs etc. You might as well need to transform the data format in order to get rid of useless information like removing ‘Mr.’ and ‘Mrs.’ from names when a separate feature for gender is present. You might need to get it in a standard format throughout the dataframe with the function [pandas.DataFrame.replace](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.replace.html) or drop irrelevant features using [pandas.DataFrame.drop](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html).

## DATA VISUALIZATIONS

Data Visualizations are very important as they are the quickest way to know the data and the patterns — if they even exist or not. Your data may have thousands of features and even more instances. It is not possible to analyze the numeric data for all of them. And if you do that, then what point is to have such powerful visualization packages like **Matplotlib**and **Seaborn**?

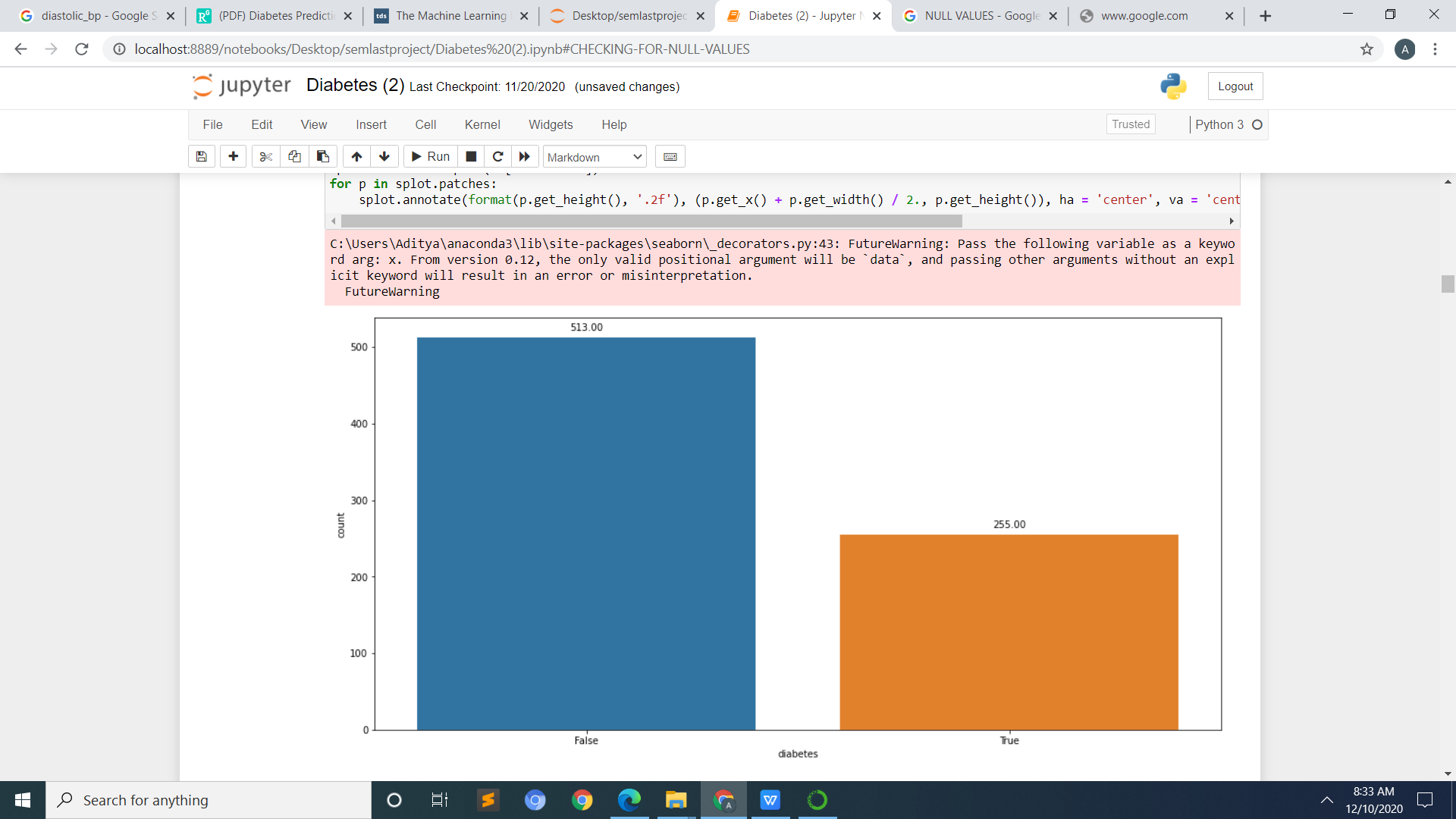
Visualizations using Matplotlib, Seaborn can be used to check the [correlations](https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22)within the features and with the target, scatter plots of data, histograms and [boxplots](https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51)for checking the spread and [**skewness**](https://www.youtube.com/watch?v=XSSRrVMOqlQ)and much more. Even pandas has its own built in [visualization](https://towardsdatascience.com/a-guide-to-pandas-and-matplotlib-for-data-exploration-56fad95f951c) library — pandas.DataFrame.plot which has bar plot, scatter plot, histograms etc.

**Seaborn**is essentially a transformed matplotlib as it is built on matplotlib itself and makes the plots more beautiful and the process of plotting much quicker. [Heatmap](https://seaborn.pydata.org/generated/seaborn.heatmap.html) and [pairplot](https://seaborn.pydata.org/generated/seaborn.pairplot.html) are examples of power of Seaborn to quickly plot the visualization of the whole data to check [multicollinearity](https://towardsdatascience.com/multicollinearity-why-is-it-a-problem-398b010b77ac), missing values etc.

One very efficient way to get most of the above descriptive and inferential statistics of the data is through [**Pandas Profiling**](https://pandas-profiling.github.io/pandas-profiling/docs/). Profiling generates a beautiful report of the data with all the details mentioned above to let you analyze it

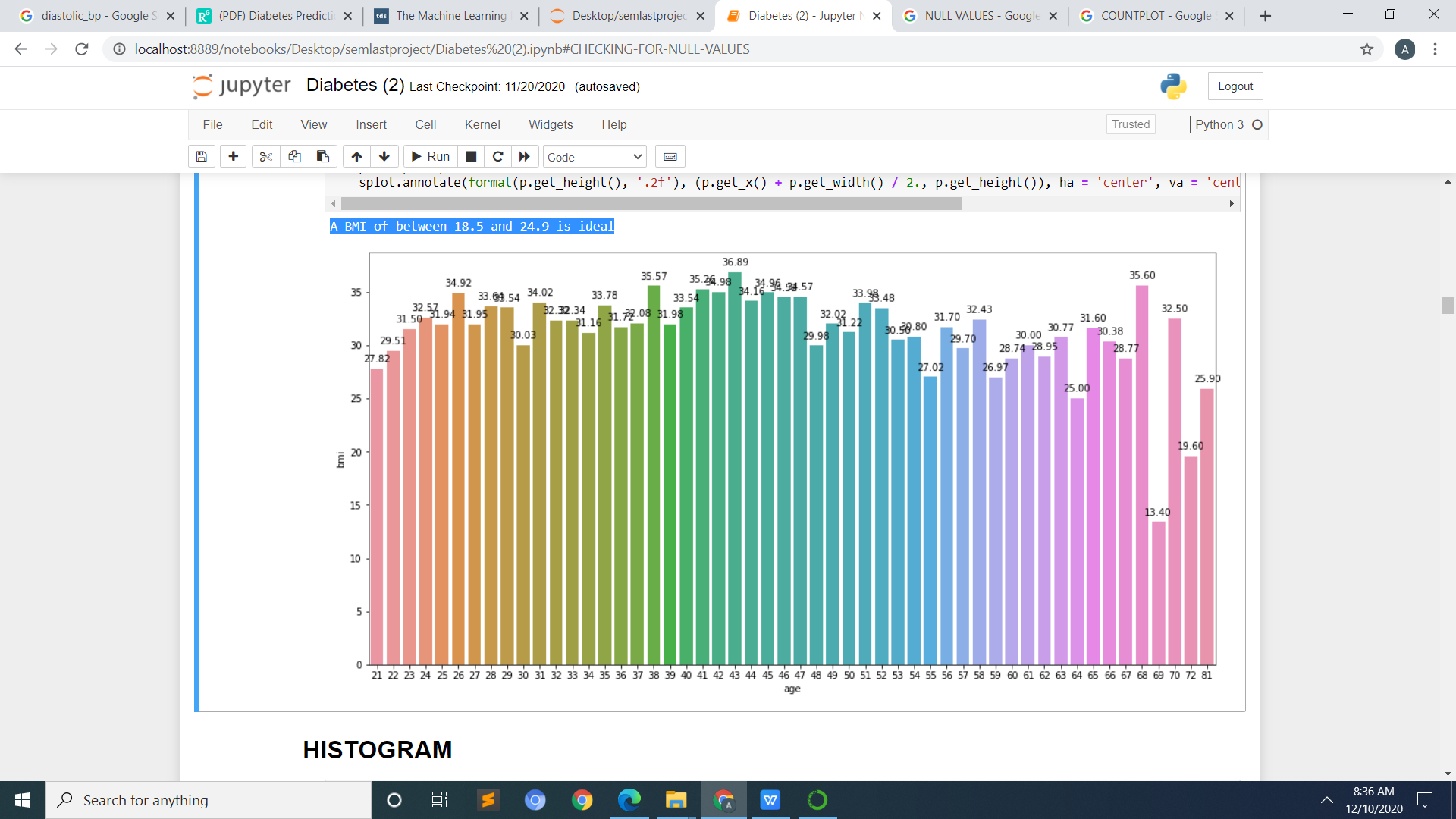
### COUNTPLOT:-

A **countplot** is kind of like histogram or a bar graph for some categorical area. It simply shows the number of occurrences of an item based on a certain type of category.



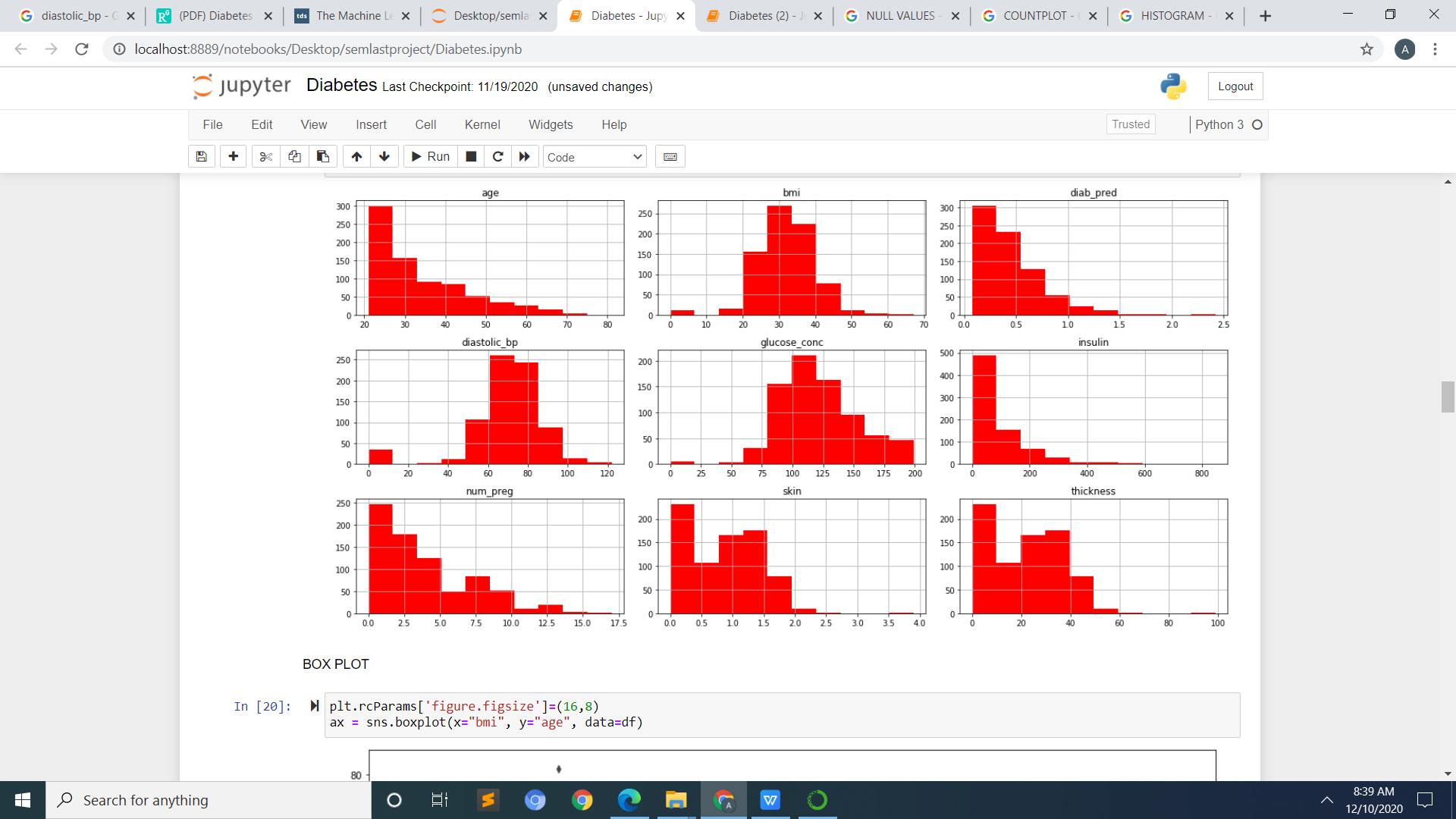
# BAR GRAPH ( BETWEEN AGE AND BMI):-

A BMI of between 18.5 and 24.9 is ideal



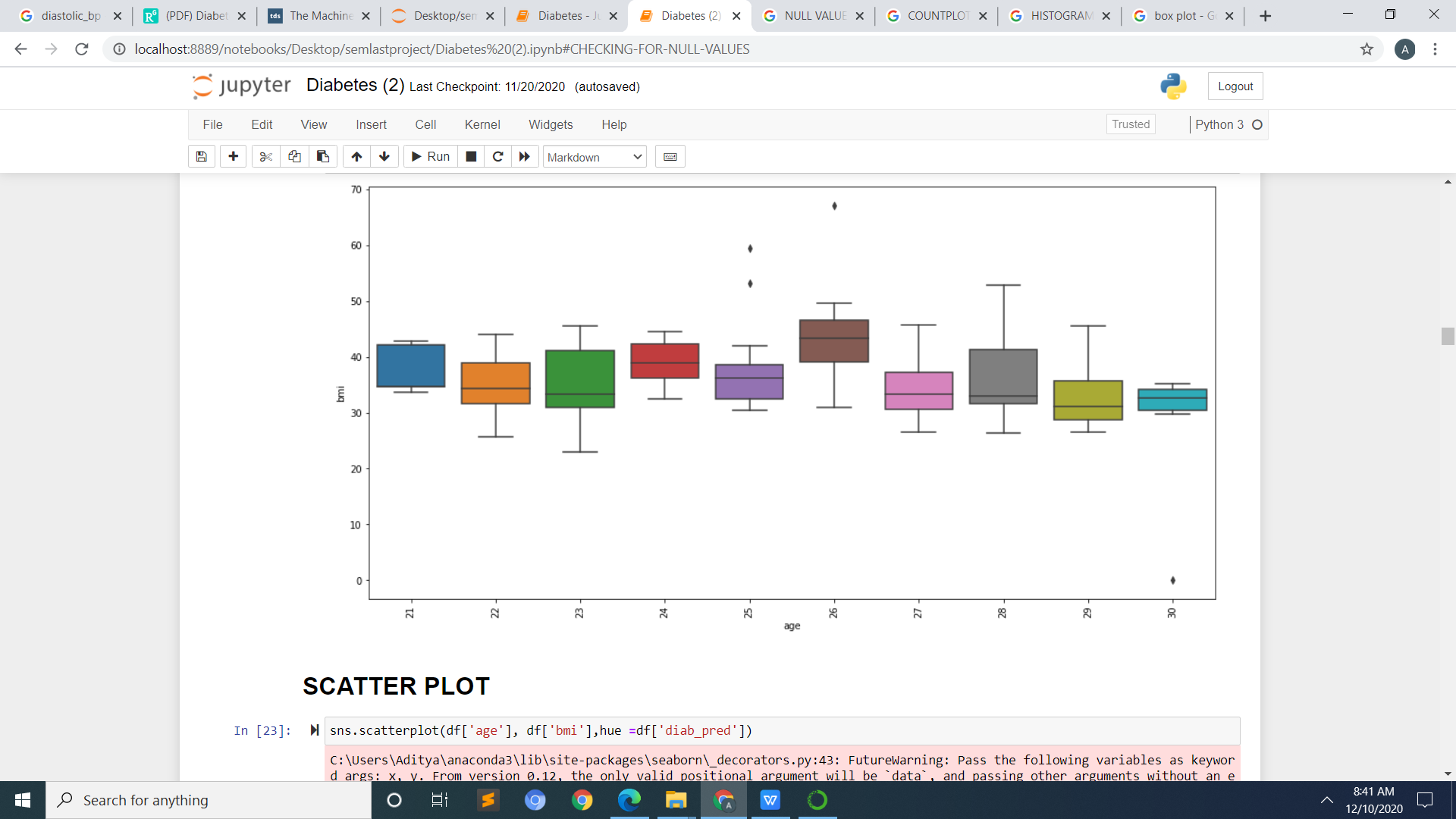
**Histogram**:-

A graphical display of data using bars of different heights. It is similar to a Bar Chart, but a **histogram** groups numbers into ranges . The height of each bar shows how many fall into each range.



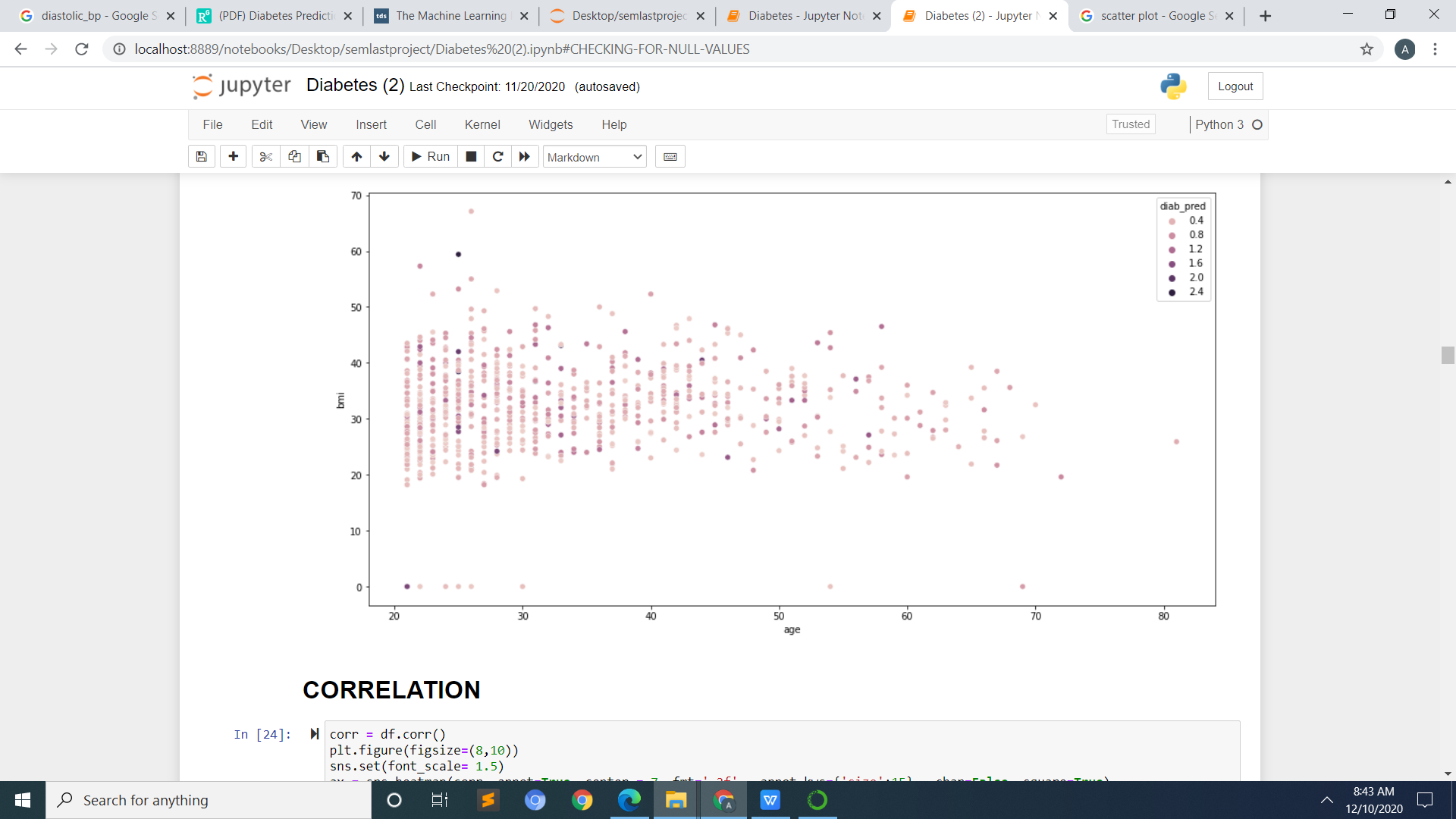
Box Plot

A box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram.



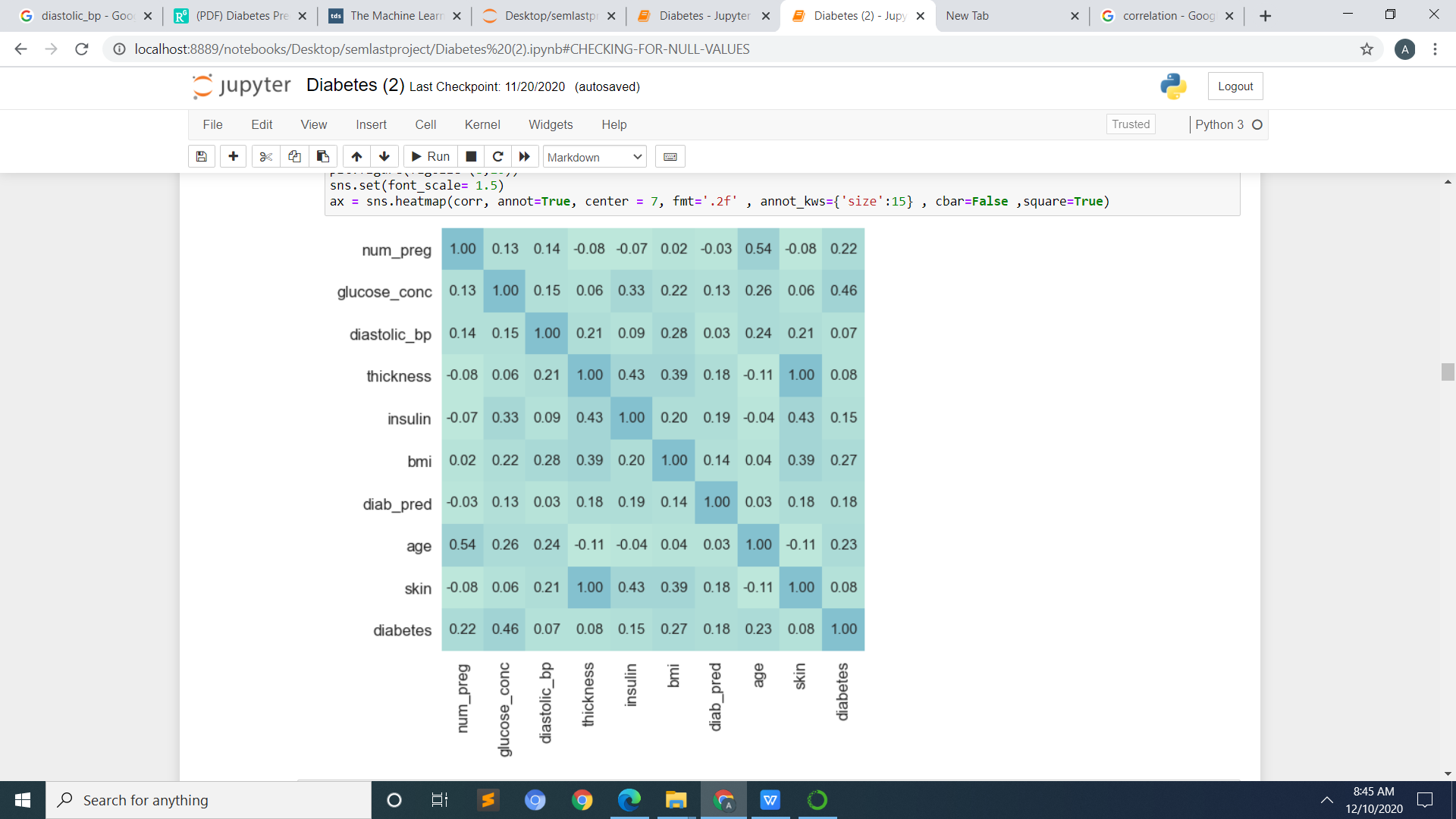
**Scatter plot:**-

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.



**Correlation**:-

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related.



# LABEL-ENCODING:-

**Label Encoding** is a popular **encoding** technique for handling categorical variables. In this technique, each **label** is assigned a unique integer based on alphabetical ordering. Let's see how to implement **label encoding** in Python using the scikit-learn library and also understand the challenges with **label encoding.**

## Feature Selection

Feature selection is the process of selecting a certain number of most useful features which will be used to train the model. This is done in order to reduce the dimensionality when most of the features are not contributing enough to the overall variance. If there are 300 features in your data and 97% of variance is explained by top 120 features, then it makes no sense to pound your algorithm with so many useless features. Reducing features not only saves time but costs as well.

Some of the popular feature selection techniques are [SelectKBest](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html), Feature elimation methods like [RFE](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html)(recursive feature elimination) and embedded methods like [LassoCV](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html).

# TRAIN - TEST SPLITING DATA

# The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset.

# Evaluate Algorithms

Once our data is ready, proceed to check the performance of the various regression/classification algorithms (based on the type of problem). we can first make a base model to set a benchmark to compare against.

## Split-out validation dataset

Once the model is trained, it needs to be validated as well to see if it really generalized the data or it over/under fitted. The data in hand can be split up beforehand as training set and validation set. This split-out has various techniques — Train Test Split, Shuffle split etc. You can also run [Cross Validation](https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6) on the entire data set for a more robust validation. KFold Cross Validation, Leave-One-Out-CV are the most popular methods.

## Test options and evaluation metric

The models need to evaluated based on a certain set of evaluation [**metrics**](https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234)which need to defined. For regression algorithms, some of the common metrics are — MSE and R Square.

Evaluation metrics pertaining to classification are a lot more diverse — Confusion Matrix, F1 Score, AUC/ROC curves etc. These scores are compared for each algorithm to check which ones performed better than the rest.

## Spot Check Algorithms

Once the data is split and the evaluation metrics are defined, we need to run a set of algorithm, say, in a for-loop to check which one performed the best. It is a trial and error to discover a short list of algorithms that do well on your problem so that we can then double down on and tune them further.

A [**pipeline**](https://towardsdatascience.com/pre-process-data-with-pipeline-to-prevent-data-leakage-during-cross-validation-e3442cca7fdc) can be made and a mixture of linear and non-linear algorithms can be set to check the performances.

## Compare Algorithms

Once we have spot run the test harness, we can easily see which ones performed the best for your data. The algorithms giving consistently high scores should be your target. we can then take the top ones and tune them further to improve their performance.

# CLASSIFIER

In statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known

**Random Forest Classifier:-**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

**XgBoost Classifier:-**

**XGBoost** provides a wrapper class to allow models to be treated like **classifiers** or regressors in the scikit-learn framework. ... The **XGBoost** model for classification is called XGBClassifier. We can create and and fit it to our training dataset. Models are fit using the scikit-learn API and the model.

**SVM**:-

In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis

**Logistic Regression:-**

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc

**KNN Classifier:-**

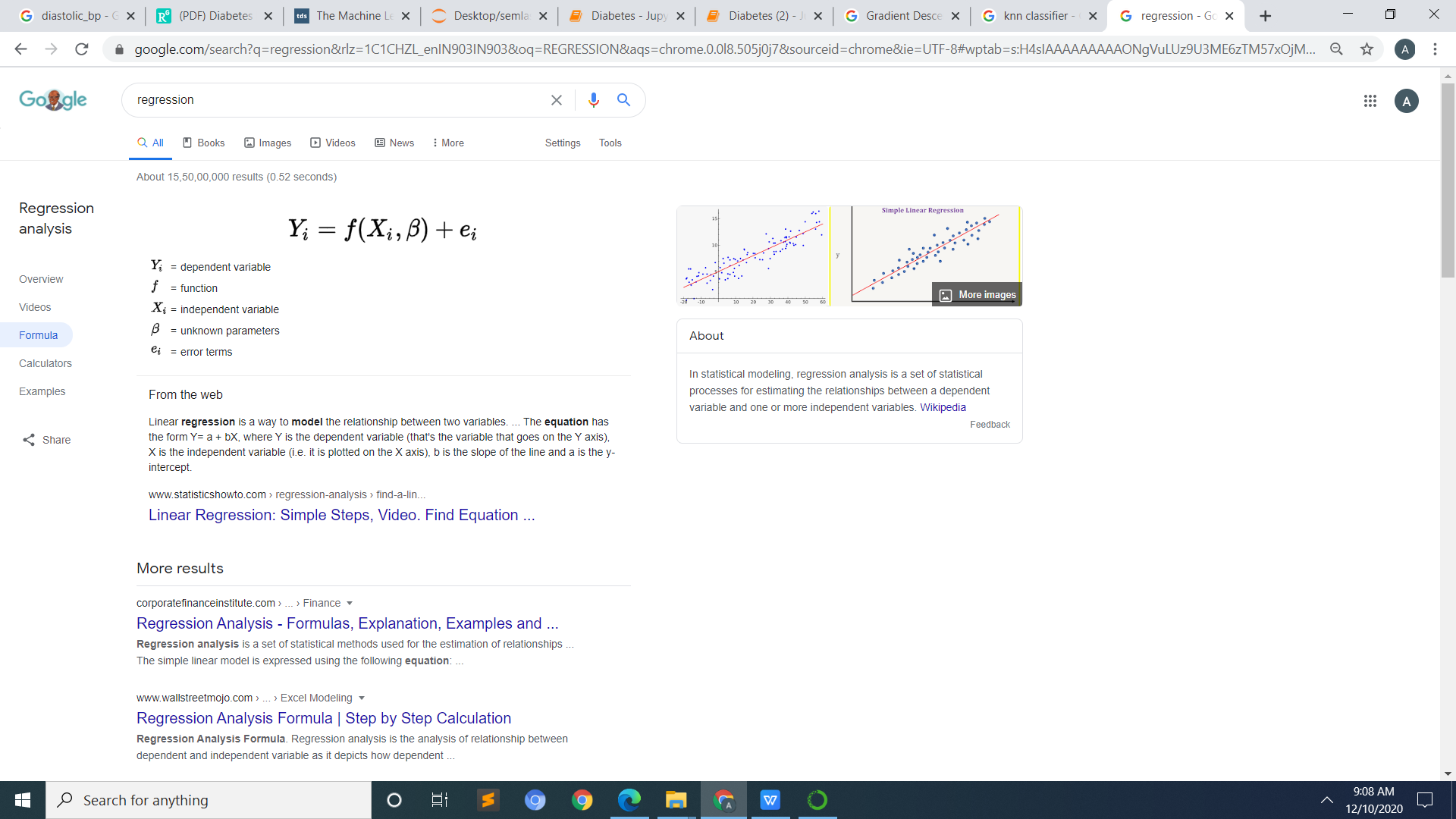
In statistics, the k-nearest neighbors algorithm is a non-parametric method proposed by Thomas Cover used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space.

**Gradient Descent:-**

Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. The idea is to take repeated steps in the opposite direction of the gradient of the function at the current point, because this is the direction of steepest descent.

**REGRESSION**:-

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.



# SIMPLE LINEAR REGRESSION:-

**Simple linear regression** is a **regression model** that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

# Polynomial Regression:-

In statistics, **polynomial regression** is a form of **regression** analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree **polynomial** in x.

# Decision Tree Regression:-

**Decision tree regression** observes features of an object and trains a model in the structure of a **tree** to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

# Random Forest Regressor:-

A **random forest regressor**. A **random forest** is a meta estimator that fits a number of classifying **decision** trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. ... The maximum depth of the tree.

**K-Means Clustering:-**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

# Improve Accuracy

After we have the best performing algorithms with , their parameters and the Hyperparameters can be tuned to give maximum results. Multiple algorithms can be chained as well.

.

## Ensembles

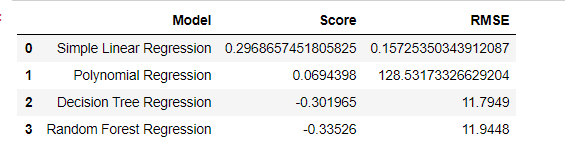
Multiple Machine Learning algorithms can be combined to make a more robust and optimal model that gives better predictions than the single algorithm. This is known as an [ensemble](https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f).

There are basically 2 types of ensembles — [Bagging (Bootstrap-Aggregating) and Boosting](https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205). Random Forest, for example, is a type of Bagging ensemble which combines multiple Decision Trees and takes the aggregate of the total output.

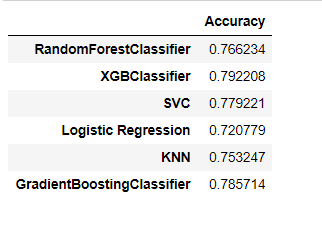
Boosting, on the other hand, combines a set of weak learners by learning them in an an adaptive way: each model in the ensemble is fitted giving more importance to instances in the dataset that had big errors by the previous models in the sequence. XGBoost, AdaBoost, CatBoost are some examples.

**CONCLUSION:-**

* **Regression – Best Regression is**

****

* **Classification – Best classification is**

****

**BIBLIOGRAPHY:-**

* .com/uciml/pima-indians-diabetes-database

.