**Diabetes Predection using python machine learning.**

**A Project Report for Winter Industrial Training**

###### ***Submitted by***

**Pranjal Chowdhury**

***in partial fulfillment for the award of the degree of* B.Tech**

**in**

**Computer Science And Engineering**

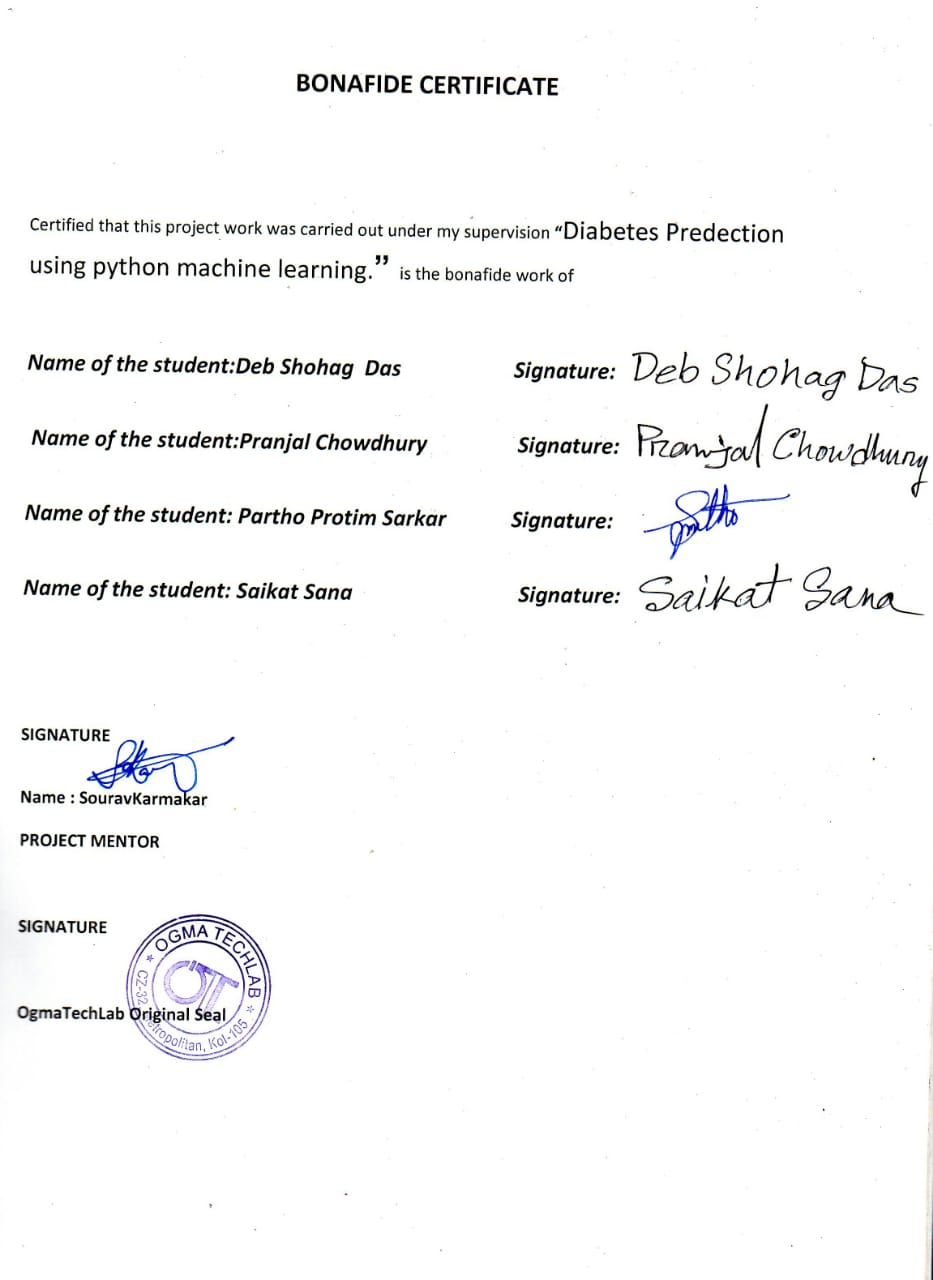
**Institute of Engineering and Management(IEM)**



**At**

**Ogma TechLab Pvt. Ltd.**



**Acknowledgement**

We take this opportunity to express my deep gratitude and sincerest thank to my project mentor, **SOURAV KARMAKAR** for giving most valuable suggestion, helpful guidance and encouragement in the execution of this project work.

We will like to give a special mention to my colleagues. Last but not the least we are grateful to all the faculty members of OgmaTechLab Pvt. Ltd. or their support.

**CONTENTS**

1. **Introduction**
2. **Method**
   1. Logistic Regression
   2. Linear regression
   3. Lasso regression

2.4 Steps Followed

2.5 Choosing ANN as classifier

2.6 Use of mu

**3. Result**

3.1 Validating the Training Accuracy

3.2 Validating the Testing Accuracy

**4. Observations**

**5. Discussion**

**6. Conclusion**

1.INTRODUCTION:

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.

2.Materials and methods:

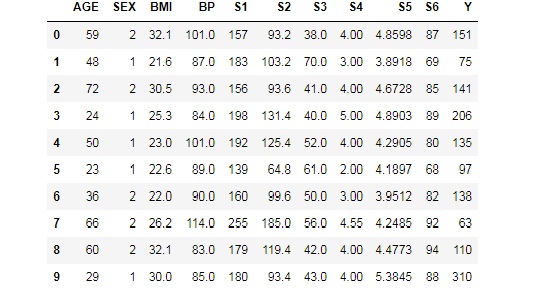
The dataset was collected from**https/scikit-learn.org/stable/modules/generated/sklearn.datasets.lod\_diabetes.html.**

**3.Steps of our project:**

* [Importing the dataset and understanding the dataset](http://localhost:8888/notebooks/Diabetes_Prediction.ipynb#Import-the-datasets-and-understanding-the-dataset)
* [Exploratory Data Analysis](http://localhost:8888/notebooks/Diabetes_Prediction.ipynb#Exploratory-Data-Analysis)
* [Modelling](http://localhost:8888/notebooks/Diabetes_Prediction.ipynb#Modelling)
* [Evaluation of the Model](http://localhost:8888/notebooks/Diabetes_Prediction.ipynb#Evaluation-of-the-Model)

### 4.Methods:

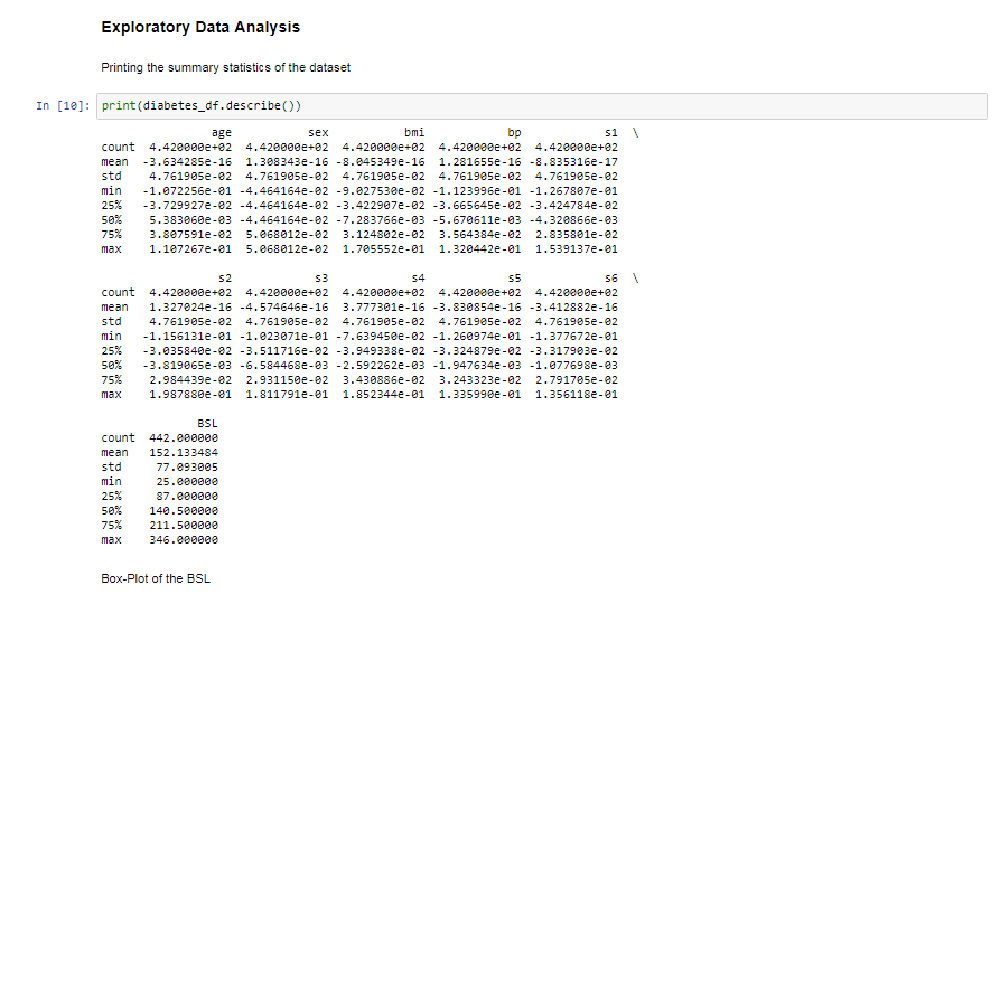
**Data preprocessing**



Number of Instances: 442, Number of Attributes: First 10 columns are numeric predictive values, Target: Column 11 is a quantitative measure of disease progression one year after baseline, Attribute Information(Age , Sex,Body mass index ,Average blood pressure,S1,S2,S3,S4,S5,S6

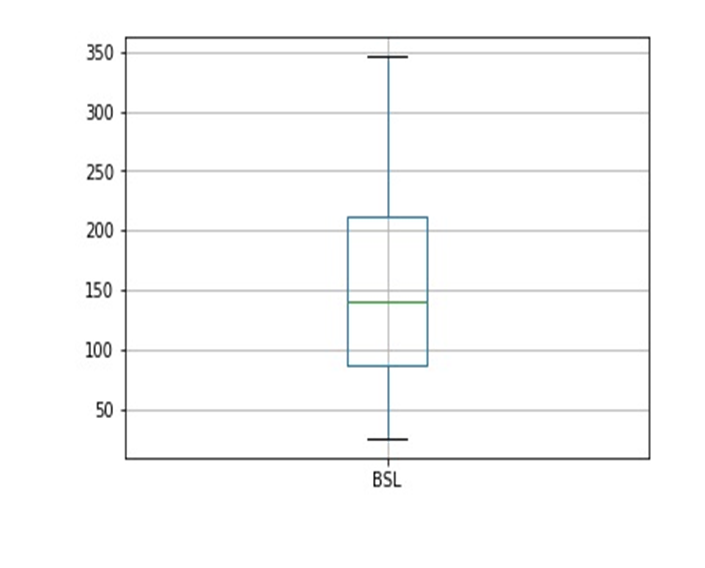
**Exploratory Data Analysis:**

Eexploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics.



**Observations**

* All the features are quantitative
* The target variable is also quantitative
* There is no outlier as revealed from the boxplot



**Logistic Regression:**

In statistics Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable-that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application is about to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing. In this paper, Logistic regression was used to predict whether a patient suffer from diabetes, based on seven observed characteristics of the patient.

**Linear regression:**

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate

linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

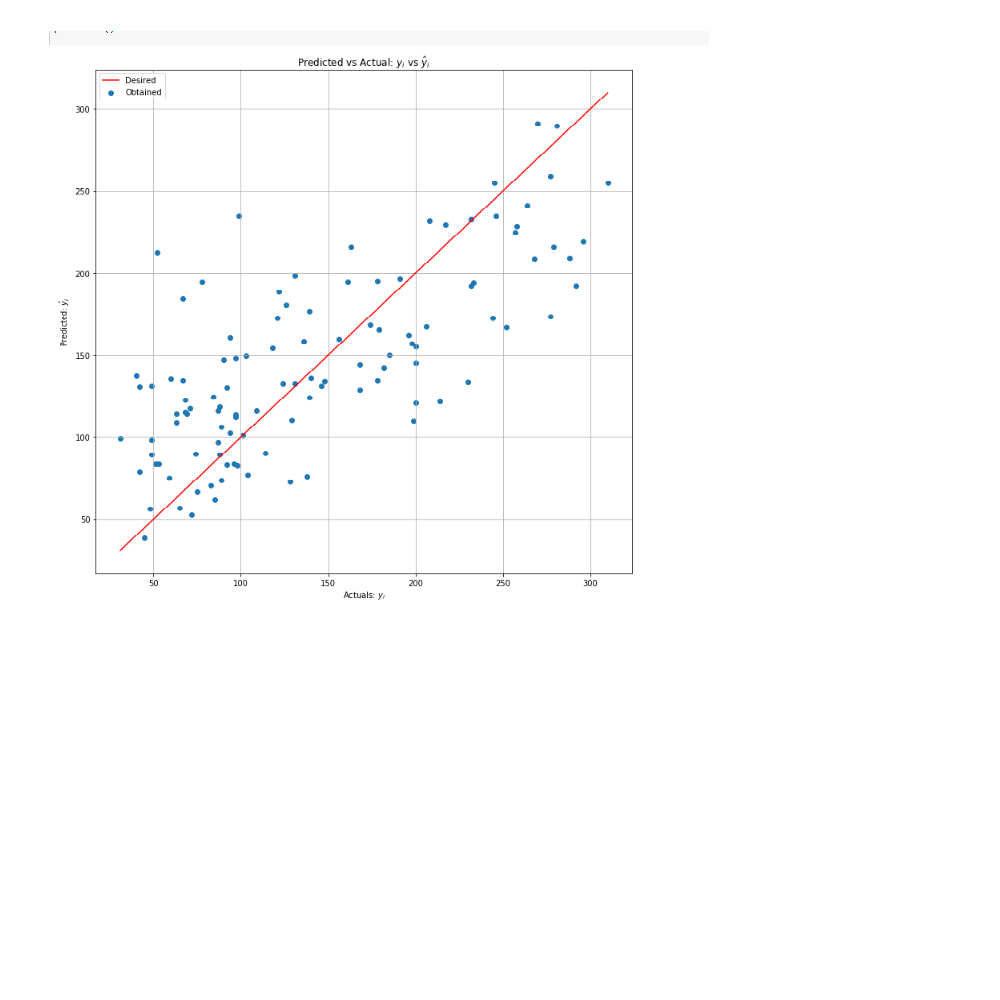
**Lasso regression:**

In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It was originally introduced in geophysics literature in 1986,and later independently rediscovered and popularized in 1996 by Robert Tibshirani, who coined the term and provided further insights into the observed performance.

Lasso was originally formulated for least squares models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates do not need to be unique if covariates are collinear.

Though originally defined for least squares, lasso regularization is easily extended to a wide variety of statistical models including generalized linear models, generalized estimating equations, proportional hazards models, and M-estimators, in a straightforward fashion.[2][3] Lasso’s ability to perform subset

selection relies on the form of the constraint and has a variety of interpretations including in terms of geometry, Bayesian statistics, and convex analysis.

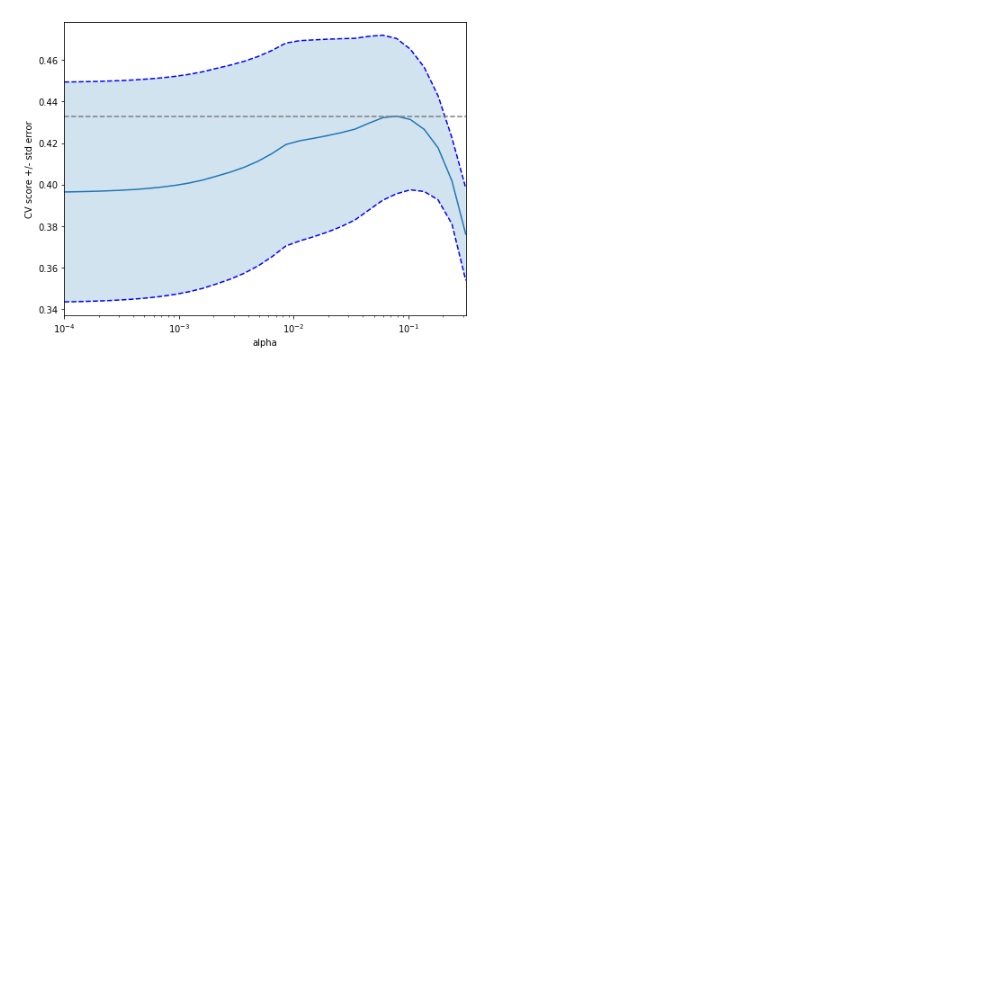


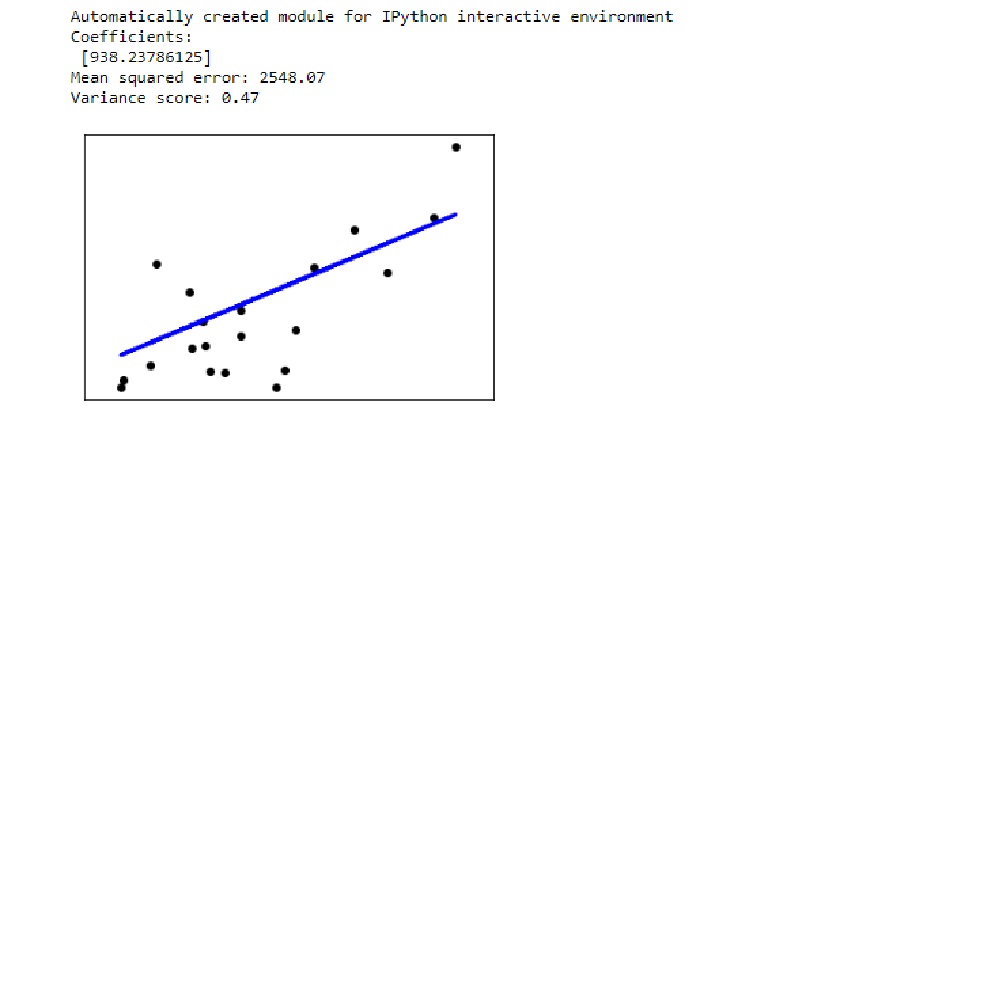
Finding Mean Square Error (MSE):

Mean Squared Error=2665.22777863786

**Observations:**

Ideally, the scatter plot of Predicted vs Actual should create a straight line. Since the model does not fit 100%, the scatter plot is not along the straight line. As, shown from the scatter plot linear regression is not a very useful model for this dataset.





**Discussion:**

To the best of our knowledge, this is the first study for predicting incident diabetes using machine learning methods based on cardiorespiratory fitness data. This study take advantage of the unique opportunity provided by our access to a large and rich clinical research dataset of the FIT project. In this study, a combination of three decision tree models.

**Conclusion:**

III. CONCLUSION Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. 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 decision about the disease status.