Early Prediction and Risk Analysis of Type2 Diabetes Mellitus Using the Nonlinear Least Absolute Shrinkage and Selection Operator (LASSO) Regression Technique

[1] Mohammed Abraruddin, [2] Jerald Prasath G

[1] M. Tech Scholar, [2] Assistant Professor,
[1, 2] Department of Computer Science and Engineering,
[1, 2] TKR College of Engineering and Technology, Hyderabad, India.

Abstract: Due to its constantly increasing occurrence, diabetes mellitus is increasingly influencing more and more families. Most diabetics know rarely about their health quality or pre-diagnosis risk factors. Based on this study, we proposed a contemporary model based on Machine Learning techniques for predicting type 2 diabetes mellitus (T2DM) and risk analysis. The key issues we are assessing to overcome are enhancing the prediction model's accuracy and minimize the prediction error. The most aim of this project obtains a subset of predictors reducing prediction that minimizes prediction error for a quantitative response variable. The Least Absolute Shrinkage and Selection Operator (LASSO) do This is by putting a limit on the parameters of the model that causes regression coefficients to shrink to zero for certain variables. Results of the study show that the proposed approach to select the most important characteristics of diabetic data is useful and accurate. This study will help to build a model using the selected features that can predict diabetes using machine learning systems.

Index Terms **–Type 2 Diabetes Mellitus (T2DM), Least Absolute Shrinkage and Selection Operator (LASSO)**

I. INTRODUCTION

Diabetes mellitus is a different word for the way the body turns food into energy. It's also called diabetes mellitus. Hormone insulin transports blood sugar into the cells to be processed or used for energy purposes. For diabetes, the body simply doesn't produce enough insulin or can't use the insulin it produces efficiently. High blood sugar can damage your nerves, Eyes, kidneys and other organ through diabetes. Based on the cause, Diabetes exists in various ways, such as Type1diabetes, Type2 diabetes and Pre-diabetes and Gestational Diabetes. Type 2 diabetes mellitus is a metabolic condition that prevents the body from taking insulin, as insulin resistance is identified to patients with type 2 diabetes. People who've been mid-life or older often have diabetes like this, so it's often called adult-onset diabetes.

However, Type2 diabetes also affects children's and teenagers, mainly due to childhood obesity. This is the most common type of diabetes. Were in type 2 diabetes, the cells of your body are not as capable of responding to insulin as they should be. Your body may also not produce enough insulin in later stages of the disease. Unrestrained type2 diabetes can cause consistently high levels of blood glucose. It causes many symptoms and can lead to extreme complications. The purpose of this work is to utilize an ML methodology, named Least Absolute Shrinkage and Selection Operator (LASSO), to detect T2D using features extracted from novel Pima Indian Diabetic Dataset.

II. LITERATURE SURVEY

Other researchers' work that is relevant to this study is presented in this section.

[1] Michele Bernardini et.al., The work mainly focused on to the group of metabolic disorders with a long period of high blood sugar levels. Subsequent urination, increased thirst, and increased hunger include symptoms of high blood sugar. The aim of this study is to utilize EHRs to diagnose Diabetes Mellitus and related diseases . The main challenges are the identification and robust diagnosis of significant features of EHR. iIn this perspective, the candidate will need machine learning approaches for data-driven diagnosis of Diabetes Mellitus to be studied, investigated and applied.

[2] Yu Wang et.al., In this paper, the use of EHR data as support for decision-making helps doctors and patients to gain more insight into current conditions of health, allowing patients to adapt knowledge to the patients, thus shared decisions making (SDM) more effective. Instead of the entire process, one or more of the SDM elements were investigated. The recent lack of research consensus on how best to develop the system to enable self-government for self-handling and decision-making.

[3] Minyechil Alehegn et.al., Author developed the framework to order to identify concepts with and without T2DM from EHR via features and machine learning, researchers implement an informed data framework. In our context we evaluate and contrast the performing identity of widely used machine-learning models, including k-Nearest-Neighbors, Naïve Bays, Decision Tree.

[4] Jeffrey P Andersonet.al., This paper makes use to created the predictive model ensembles for development to prediabetes or T2D from the aggregated EHR data sample using a new analysis platform. Explore the space of a large model and distribute risk estimates from a set of predictive models. This technique can be used in downstream use for personalized medicine and clinical

research. Electronic health record (EHR) data approaches to machine learning algorithms may provide useful insights into the disease processes. They used that approach to build predictive models for prediabetes and type 2 diabetes (T2D) progression.

[5] Amir Talaei-Khoeiet.al., There are literatures which use algorithms to predict patients who develop T2D using a machine learning classification. The aim of this paper is to classify patients at Type 2 risk of diabetes (T2D). The latest research compares the success and risk of short- to medium and long-term T2D development of these classification algorithms. Furthermore, the list of predictor variables that are important for T2D progression prediction is given. Some predictors can be identified by patients, others can be tested by doctors or ordered for further laboratory analysis, which would theoretically reduce the strain imposed on the clinical settings.

[6] Benjamin Shickel et.al., The paper proposes current research on the application of in-depth learning to clinical activities based on EHR data, where a number of in-depth learning strategies and frameworks are applied to different types of clinical applications, including knowledge extraction, representational learning, outcome prediction, phenotyping, and detection. We recognize many shortcomings of current research that include issues including model interpretability, data heterogeneity, and lack of standardized benchmarks. We conclude by summing up the state of the field and suggesting avenues for potential work on deep EHRs.

III. REGULARIZATION

This is a type of regression that restricts / regularizes or reduces the estimated coefficient to zero. This technique therefore diminishes the learning of a more complicated or versatile model in order to avoid over fitting risks. It looks a strong correlation for linear regression. Here Y represents the learned relationship, and β represents estimates of the coefficient for various variables or predictors(X).

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P}$$

It is a test of data uncertainty and an estimation model. A strong fit of the model to the data is shown by a tiny RSS. This is used in parameter selection and model selection as an optimal criterion.

$$\mathbf{RSS} = \sum_{i=1}^{n}\left(\mathbf{y_i - \beta_0 - \sum_{j=1}^{p} \beta_{j0} x_{ij}}\right)^2$$

Therefore, the coefficients are modified based on the training results. Unless the training data are noisy, the predicted coefficients do not generalize well to future data. It's where regularization occurs to shrinks or regularized the *learned estimates towards zero.*

Regularization is categories in two ways they are:
1. Ridge Regression
2. LASSO Regression

A. Ridge Regression

Ridge regression is a technique for analyzing multiple regression data that includes several linearities. When multi-collinearity exists, fewer squares are accurate predictions, but their variances increase so that they are far from true.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

In above equation RSS is getting modified by adding the shrinkage quantity, were coefficients are get minimized by this equation. Were Lambda is the tuning parameter which decides how much penalty should be added.

$$\textbf{Ridge Regression} = \textbf{Loss} + \underbrace{\boldsymbol{\alpha}\,|\boldsymbol{W}|^2}_{\textbf{(Penalty)}}$$

Where α is a constant and $\alpha|W|^2$ is a penalty as a compensation for loss or damage caused and $|W|^2$ is a absolute value which are consider as coefficients

$$|\mathbf{W}|^2 = |\mathbf{W1}|^2 + |\mathbf{W2}|^2 + |\mathbf{W3}|^2 + \ldots\ldots\ldots + |\mathbf{Wn}|^2$$

In the above equation $|W1|^2$ to $|Wn|^2$ are the absolute values of coefficients and $|W|^2$ is the vector of all coefficients.

A. LASSO Regression

In machine learning or statistics, lasso, a regression analysis method that performs both the selection and the regularizing of varieties in order to increase the predictability and interoperability for the statistical model produced by lasso (less absolute shrinking and selection operator); also Lasso or LASSO).Lasso is a good technique of regression. It perpetuates the magnitude of characteristics' coefficients and reduces the error between expected and real observations. The Lasso algorithm can be applied using the python SciKit-Learn Library and it is also called as L1 Regularization Technique, Lasso tries to keep costs to a minimum.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

The regression of Lasso is a kind of linear regression using shrinkage. Shrinkage is where data values are decreased, such as the mean, to a central point. The lasso process promotes simple and small models (i.e. less-parameter models). This special regression is suitable for models with high muticollinearity, or if certain sections, such as variable selection parameter elimination are to be automated. LASSO regression is a one type regularization technique.

$$\textbf{LASSO Regression} = \textbf{Loss} + \underset{\textbf{(Penalty)}}{\alpha\,|W|}$$

Where α is a constant and α|W| is a penalty as a compensation for loss or damage caused and |W| is a absolute value which are consider as coefficients

$$|W| = |W1|+|W2|+|W3|+\ldots\ldots\ldots+|Wn|$$

In the above equation |W1| to |Wn| are the absolute values of coefficients and |W| is the vector of all coefficients.

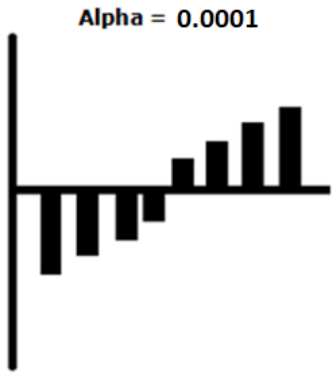


Fig : 1

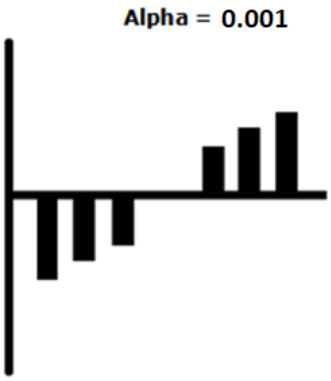


Fig : 2

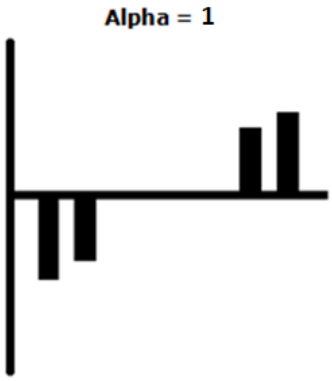


Fig : 3

In above figures we can see that however, Alpha values are getting increasing the coefficients of magnitudes are getting shrunk or scale down. In Ridge Regression it won't get zero but in LASSO Regression it get exactly zero which can see in above figures.

IV. MATERIALS AND METHODOLOGY

A. Data Set

The data collection comes from the Diabetes and Digestive and Kidney National Institute.The goal is to determine whether a patient has diabetes on the basis of diagnostic measurements. There were some limitations in choosing these cases from a larger database. Here, the majority of patients are Pima Indian women aged at least 21 years of age.
Pregnancy: number of pregnancy periods

Glucose: blood glucose levels in oral tolerance checks for 2 hours

BloodPressure: Blood pressure diastolic (mm Hg)

SkinThickness: Triceps skin fold thickness (mm)

Insulin: Serum insulin for 2 hours (mu U / ml).

BMI: composite body mass (in kg/(in m) ^ 2 body mass index)

DiabetesPedigreeFunction: history of diabetes

Age: Age (years) Outcome: Class variable (0 or 1)

B. Data Pre-processing and Visualization

The analysis aims to highlight the most important attributes for the prediction of the response variable which explanatory variability (functions). LASSO method can be use of the dataset by using the python library which is SciKit-Learn. Before creating a model, we have to analyze the dataset to better understand the data.

1. Data Pre-processing

In Machine Learning Data Preprocessing is a technique to preparing or cleaning and organizing the raw data to make it appropriate for a constructing and training Machine Learning models. The data pipeline begins with data collection and finishes with results communication, this is not as simple as it seems. There are several stages to clean and organize the data among them one of the key steps is data pre-processing. Data pre-processing itself has several phases and the number of phases depends on the data file type, the nature of the data, various types of value and more.
The data pre-processing involve data cleaning, data selection, normalization of the data, data transformation and feature extraction and many more. Let's load our diabetic data and explore whether any pre-processing of the data is required.

	pregnant	glucose	bp	skin	insulin	bmi	pedigree	age	label
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

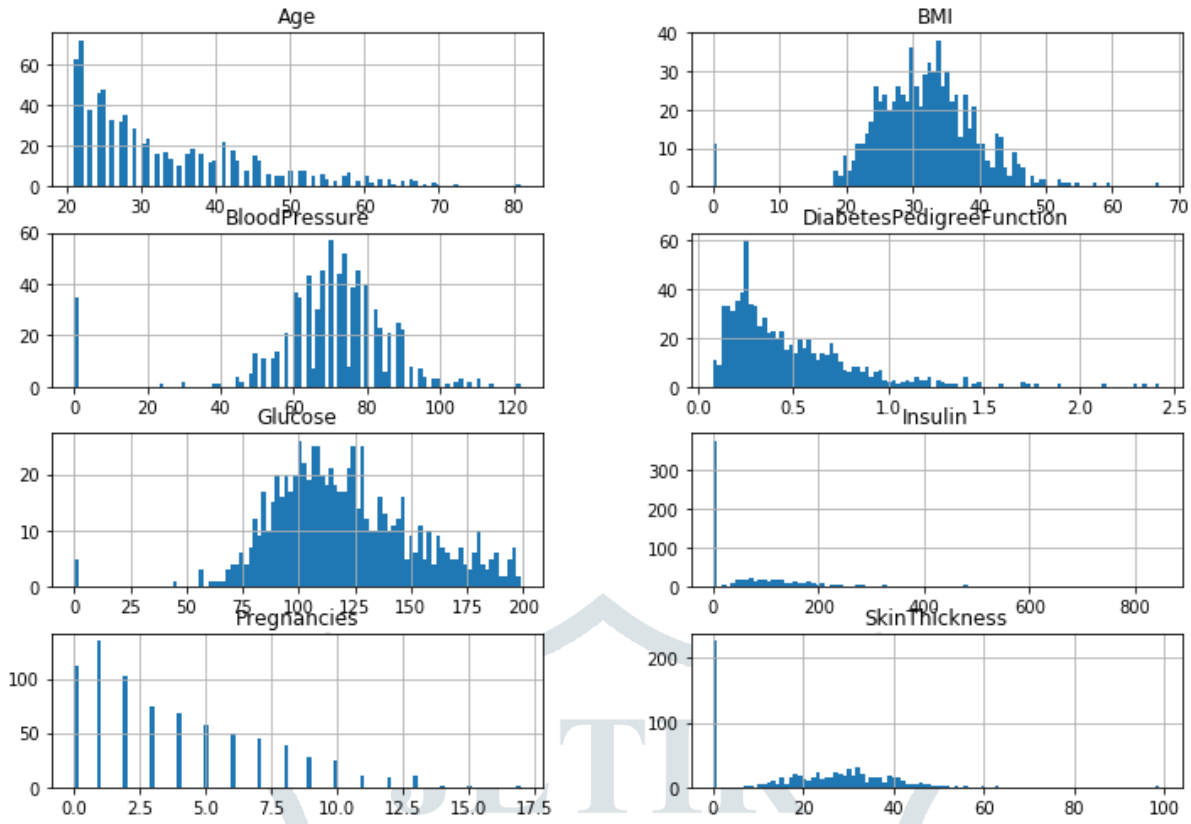
After loading data we have to check whether missing values or there or not in the data. It seems like there is no missing values are there in the data, so we doesn't required any kind of imputation or inserting values. If any missing values have been there then we have to insert value in the missing area with means of its. Were we had cleaned and formatted the data, now we want visualize the data in graphical representation by using some visualization techniques.

2. Data Visualization

Visualization of data is an important part of every project in data science. As the size of the dataset increases, understanding observational data using excel spreadsheets or files becomes harder. Let's understand the visualization and its relevance in modelling machine learning. We'll also seek to use a couple of these tools to explore the diabetic dataset.

A. Histogram Plot of the diabetic data

Histogram is a comprehensive representation of the distribution of the numeric results. The distribution of probability for a continuous (quantitative) variable is determined.



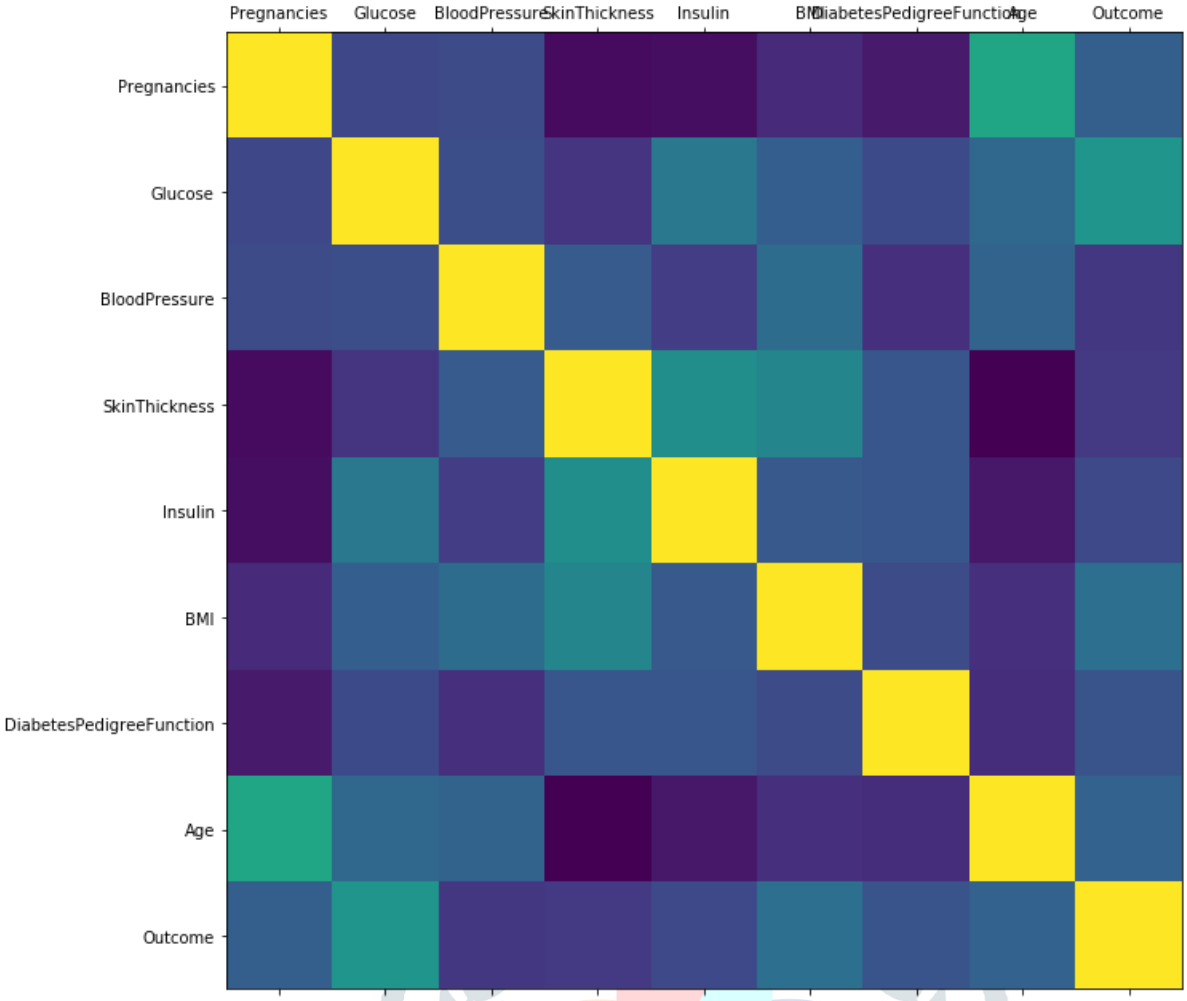
There are counting and drawing up the number of different values predicted. The bars are shown side by side because the calculated variable is constant and located on the x-axis. They may use matplotlib in the development of histograms. Where matplotlib is a python library to plot the numerical data. A histogram indicates the vertical axis frequency and another element is a horizontal axis. It normally has bins, which have a minimum and maximum value in each bin. Each bin has an infinite and an x frequency.

B. Correlation of diabetic data

The pairing of all columns in the dataframe is done by Pandas dataframe.corr(). All na values are removed automatically. It is neglected in the dataframe for non-numeric data type columns.

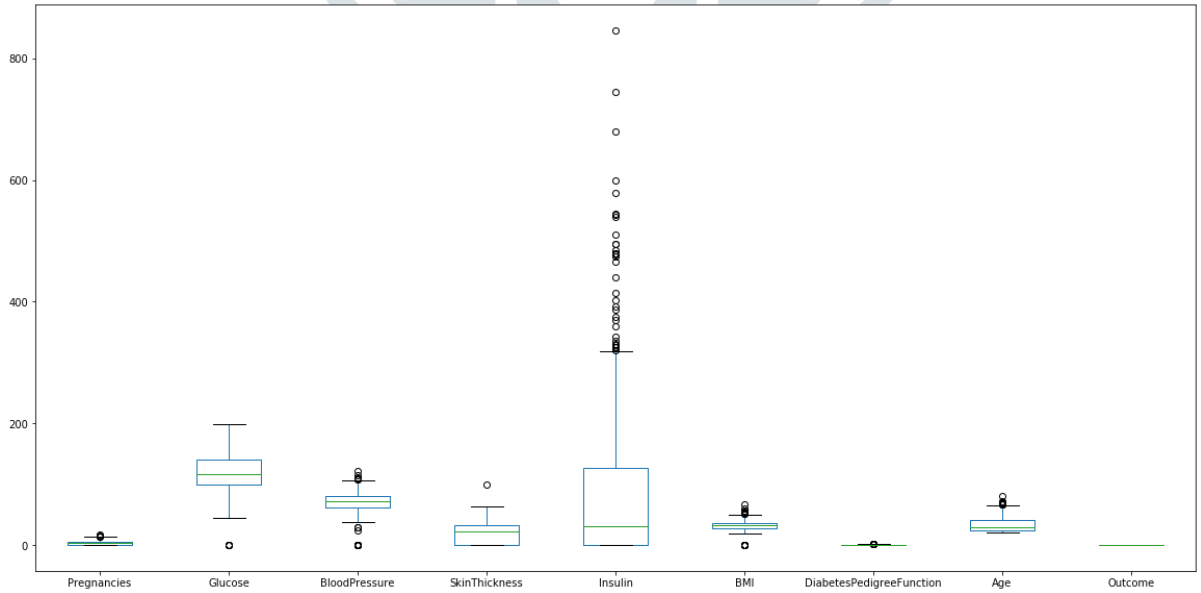
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	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
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

The data frame output can be elucidate as for any cell, correlation of row variable and column variable is the value of the cell. As previously indicated, were the correlation of variable for its self is 1 because of these reason all the diagonal values are 1.0. Now we will plot the above correlational data in graphical presentation.



C. Box Plot for diabetic data

Box plots calculate how well the data in a data set are distributed. It divides the set of data into 3 quartiles. It represents the min, max and median data in the dataset by first quadrilles and third quadrilles. A box-plot is a way of displaying numerical data groups through their quadrilles graphically in descriptive statistics. Box plots can also have lines that run vertically from the boxes to the upper and lower quadrilateral.



In the above plot, we can see some outlier in Insulin column and SkinThickness column which we have to remove from the dataset to increase accuracy. If we doesn't remove the outlier from the data then we doesn't the appropriate result. Removing the outlier and replacing zero's with mean.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.000000	35.000000	78.033987	33.600000	0.627	50	1
1	1	85.0	66.000000	29.000000	78.033987	26.600000	0.351	31	0
2	8	183.0	64.000000	20.406536	78.033987	23.300000	0.672	32	1
3	1	89.0	66.000000	23.000000	94.000000	28.100000	0.167	21	0
4	0	137.0	40.000000	35.000000	168.000000	43.100000	2.288	33	1
5	5	116.0	74.000000	20.406536	78.033987	25.600000	0.201	30	0
6	3	78.0	50.000000	32.000000	88.000000	31.000000	0.248	26	1
7	10	115.0	69.115033	20.406536	78.033987	35.300000	0.134	29	0
8	2	197.0	70.000000	45.000000	543.000000	30.500000	0.158	53	1
9	8	125.0	96.000000	20.406536	78.033987	31.985359	0.232	54	1

Now, all the outlier in the dataset has been got removed from the dataset and all zero's are been replaced by its mean which has been seen in the above table.

D. Normalization for diabetic data

Normalization is often used as part of machine learning for data preparation. The objective of normalization is to change the numerical column values of the dataset into a common scale without distorting the values. In machine learning does not require normalization for every dataset. It required only when there are different ranges in features.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.352941	0.743719	0.590164	0.555556	0.114756	0.500745	0.259091	0.617284	1.0
1	0.058824	0.427136	0.540984	0.460317	0.114756	0.396423	0.145041	0.382716	0.0
2	0.470588	0.919598	0.524590	0.323913	0.114756	0.347243	0.277686	0.395062	1.0
3	0.058824	0.447236	0.540984	0.365079	0.138235	0.418778	0.069008	0.259259	0.0
4	0.000000	0.688442	0.327869	0.555556	0.247059	0.642325	0.945455	0.407407	1.0

V. RESULTS AND DISCUSSION

Here we have divided the dataset into two have one is training set and another is testing set. Were 80% of data is training set and 20% of the data is testing set. The Original Diabetes True Values in the dataset is 266 which are over 34.77% and Original Diabetes False Values in the dataset is 499 which are over 65.23%. From the dataset we get Training Diabetes True Values in the dataset is 222 which are over 36.27% and Training Diabetes False Values in the dataset is 390 which are over 63.73%. Were Test Diabetes True Values in the dataset is 44 which are over 28.76% and Test Diabetes False Values in the dataset is 109 which are over 71.24%. A lasso model was implemented for this where we have selected a best alpha for the model which is Alpha = 0.001 and we got the accuracy over 83.66 %.

	Feature Name	Alpha = 0.000100	Alpha = 0.001000	Alpha = 1.000000
0	Pregnancies	0.362900	0.351717	0.0
1	Glucose	1.310810	1.257582	0.0
2	BloodPressure	-0.144503	-0.000000	0.0
3	SkinThickness	-0.046653	-0.000000	0.0
4	Insulin	-0.091761	-0.000000	0.0
5	BMI	1.002982	0.846669	0.0
6	DiabetesPedigreeFunction	0.379385	0.337602	0.0
7	Age	0.134959	0.084185	0.0

In the above table we can see that, the alpha value getting increases, the more features are getting coefficients of 0.

VI. CONCULISION

The main objective of this research is to build a suitable prediction model for early prediction and risk analysis of type 2 diabetes militias. Were we proposed a novel model, to statistical modelling task and feature selection using the lasso model for machine learning. The performance analysis is better focused on the accuracy rate of regression techniques. The accuracy of experimental data is assured. The other advantage here is that the Pima Indian Diabetes Dataset and other datasets can be applied to our model. But the drawback is that the pre-processing component takes longer.

VII. REFERENCE

[1] Michele Bernardini,Micaela Morettini,Luca Romeo,Emanuele Frontoni,Laura Burattini," Early temporary prediction of type 2 diabetes risk from a multiinstance improvement approach by the general practitioner" DOI: 10.1016/j.artmed.2020.101847.

[2] Yue Wange, Chen-jie Zhang, Ben Willem Mol, Cheng Li, Lei Chen, Yan-ting Wu, Jian-zhong Sheng, Jian-xia Fan, Yi Shi, He-feng Huang, "Early prediction of high risk gestational diabetes mellitus via machine learning models.", doi: 10.1101/2020.03.26.20040196.

[3] Minyechil Alehegn, Rahul Joshi, Preeti Mulay, "Analysis and prediction of diabetes mellitus using machine learning algorithm" In International Journal of Pure and Applied Mathematics 2018.

[4] Jeffrey P. Anderson, Jignesh R. Parikh, Daniel K. Shenfeld, Vladimir Ivanov, Casey Marks, Bruce W. Church, Jason M. Laramie, Jack Mardekian, Beth Anne Piper, Richard J. Willke, and Dale A. Rublee, "Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records", DOI: 10.1177/1932296815620200

[5] Amir Talaei-Khoei, V JamesM.Wilson, "Comparing prediction analysis methods and predictive variables Identifying individuals at risk for type 2 diabetes" DOI:10.1016/j.ijmedinf.2018.08.008

[6] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, Parisa Rashidi, "Download citation Share Request full-text High ehr an update on recent developments in electronic health record machine learning techniques Analysis" DOI: 10.1109/JBHI.2017.2767063