ILLINOIS INSTITUTE OF TECHNOLOGY

MATH 564 - Applied Statistics

Final Project

Diabetes Disease Detection

Shahrukh Sohail
ssohail3@hawk.iit.edu

Prof. Lulu Kang

Nov 30, 202

Table of Contents 2

1. Abstract 3

2. Introduction 4

3. Data Source 5

4. Proposed Methodology 7

4.1. Machine Learning Models 7

4.1.1. Linear Regression 7

4.1.2. Ridge Regression 8

4.1.3. Lasso Regression 8

4.1.4. Best Subset Selection 9

4.1.5. Logistic Regression 10

5. Analysis and Results 11

5.1. Software & Programming Languages 11

5.2. Data Acquisition 11

5.3. Data Exploration 11

5.4. Feature Analysis 13

5.5. Linear Machine Learning Model Deployment 14

5.6. Results analysis 14

6. Conclusions 15

7. Bibliography and Credits 16

7.1. Research Papers 16

7.2. Dataset 16

1. Abstract

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. In 2019, approximately 463 million adults (20-79 years) were living with diabetes; by 2045 this will rise to 700 million. With innovation and improvement in data-warehousing, data mining and emergence of data science as an effective field of utilizing data as a powerful tool to predict useful information, many studies are being conducted to make the process effective.

In this study Linear Regression Models will be applied on the health parameters associated with diabetes disease to extract hidden patterns on which prediction will be done. These Regression ML Models will be used to predict the presence of diabetes in an individual. The models will be assessed on the basis of their classification accuracies.

Diabetes Predictive System would be developed to identify diabetes disease before time on the basics of identified attributes and algorithm. Hence precautionary measures would be taken eventually. These precautionary measures will help to decrease the death rate caused by Diabetes.

2. Introduction

Diabetes possesses a major cause for blindness, kidney failure, lower-limb amputation, heart attack and stroke. Although diabetes disease has been identified as the most chronic disease across the world, it is the most preventable one at the same time. A healthy lifestyle (primary prevention) and timely diagnosis (secondary prevention) are two main elements of diabetes control. The conventional methods are not able to predict the chance of diabetes in the patient. Data science, data mining and machine learning based techniques can be used to predict diabetes.

Machine learning has numerous objectives. Regression, classification, and clustering are the most prevalent. A machine learning system delivers continuous output in regression, such as fitting an equation to a point plot. The machine learning method divides items into groups based on their resemblance to one another in clustering. Classification is similar to clustering in that the machine learning algorithm seeks to categorize objects based on previously stated criteria supplied by training data. Clustering and classification are examples of supervised and unsupervised machine learning. Because the learning process has no prior knowledge of groups or classes, clustering is often unsupervised. However, classification is supervised.

It is also important to introduce **Parametric** and **Non-Parametric** Machine Learning Models here. Parametric models require specification of certain parameters in order for the model to train and produce results. On the other hand non-parametric models do not require any kind of parametric specifications and can train on the model. In other words parametric models assume the model function and as a result only the coefficients have to be determined, on the other hand non-parametric models do not assume the function of the model and thus have to calculate model function as well as the coefficients of the model. This project uses the Parametric approach to predict diabatic individuals, we assume that the function of the machine learning model is in fact a linear model. Therefore we apply different linear regression techniques to classify diabetic individuals and measure the accuracies of each model.

This project applied 5 Linear Regression techniques to predict the presence of diabetes in a person, the following are the classification accuracies of each model.

Table 1: Prediction Results

Model	**Accuracy**
Linear Regression	82.35%
Ridge Regression	82.35%
Lasso Regression	81.05%
Best Subset Selection	81.70%
Logistic Regression	81.70%

As a result, the Linear & Ridge Regression models had the greatest prediction accuracy of 82.35%

3. Data Source

The first step of our project work was determining the right data set. Many online resources exist with access to a plethora of classification datasets. We came across many platforms like *datahack* and *dataworld*. We selected the ***Pima Indians Diabetes*** dataset collected by the *National Institute of Diabetes and Digestive and Kidney Diseases* for our project.

Several constraints were placed on the selection of these instances from a larger database. In particular, all individuals here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above. The dataset is available on Kaggle and the link for this dataset is provided in section 7.2 Dataset. Data analysis is used to analyze and summarize the main characteristics of the data. The data analysis shows us various aspects and distributions of our data along with the analysis of different features within our data.

The 5 models were trained after pre-processing of the dataset which was obtained from Kaggle uploaded by UCI Machine Learning organization. The dataset contains 768 records against 9 attributes i.e. *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, *Age* and *Outcome*. 80% of the dataset was used for training and the remaining 20% was used for testing the trained model.

Python is chosen for EDA (Exploratory Data Analysis), this is because python libraries such as pandas, numpy and matplotlib makes it very convenient to analyze and preprocess the dataset. R being a very powerful statistical analysis language, provides better model summaries and is therefore chosen to train linear regression models. For these reasons Python and R are used for EDA and ML model training phases respectively.

The 9 columns of the dataset and what each column represents:

1. *Pregnancies*: Number of times pregnant
2. *Glucose*: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. *BloodPressure*: Diastolic blood pressure (mm Hg)
4. *SkinThickness*: Triceps skin fold thickness (mm)
5. *Insulin*: 2-Hour serum insulin (mu U/ml)
6. *BMI*: Body mass index (weight in kg/(height in m)^2)
7. *DiabetesPedigreeFunction*: It provides information about diabetes history in relatives and genetic relationship of those relatives with individuals.
8. *Age*: Age (years)
9. *Outcome*: Target

Table 2: Data types of columns

Column	Type	Data Type
Pregnancies	Integer	int64
Glucose	Double	dbl
BloodPressure	Double	dbl
SkinThickness	Double	dbl
Insulin	Double	dbl
BMI	Double	dbl
DiabetesPedigreeFunction	Double	dbl
Age	Integer	int64
Outcome	Integer	int64

The goal is to predict the ***Outcome*** target variable whether a person has a chance of having diabetes or not. The main technical challenge it poses to predicting Outcome is the high frequency of null values in data as it's medical data so those null values are crucial for out analysis. The goal of this analysis is to solve this issue by a detailed data exploration and cleaning followed by choosing a suitable machine-learning algorithm.

4. Proposed Methodology

The goal is to predict whether a person has a chance of having diabetes in the future or not, this falls under the scope of a classification problem. We intend to deploy Linear Machine Learning models in order to achieve the highest classification accuracy. The data analysis process for the deployment of linear models is based on the following steps.

- **Data Acquisition**
- **Exploratory Data Analysis**
- **Feature Engineering**
- **Linear Machine Learning Model Deployment**
- **Results Analysis**

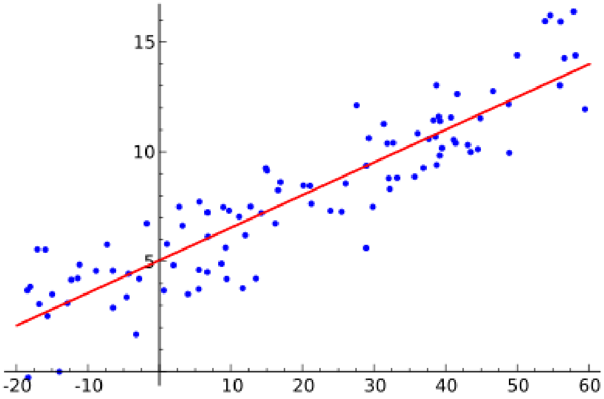
Note that the *visualization* action is performed in each step, in order to highlight new insights about the underlying patterns and relationships contained within the data.

4.1. Machine Learning Models

We will deploy 5 linear machine learning algorithms to predict the occurrence of diabetes disease.

4.1.1. Linear Regression

Linear regression makes an attempt to describe the relationship between two variables by analyzing the data and fitting a linear equation to it. Target is seen as a dependent variable, whereas Feature variables are thought to be explanatory factors. Using a linear regression model, for instance, a modeler might want to compare people's weights to their heights. It is a parametric classification method for dealing with classification and regression problems.



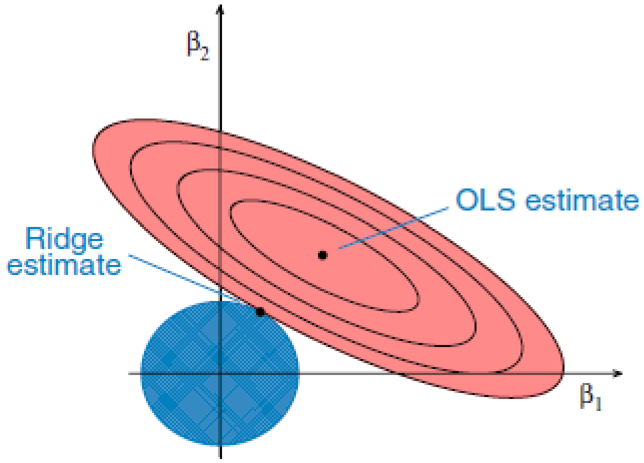
A linear regression line has an equation of the form ***Y* = *a* + *bX***, where ***X*** is the independent variable and ***Y*** is the dependent response variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

4.1.2. Ridge Regression

Any data that suffers with multicollinearity is analyzed using the model tuning technique known as ridge regression. This technique carries out L2 regularization. Predicted values differ much from real values when the problem of multicollinearity arises, least-squares are unbiased, and variances are significant. When a data set exhibits multicollinearity or when there are more predictor variables than observations, ridge regression can be used to build a parsimonious model (correlations between predictor variables). The ridge regression's cost function is:

Min(||Y – X(theta)||^2 + λ||theta||^2)

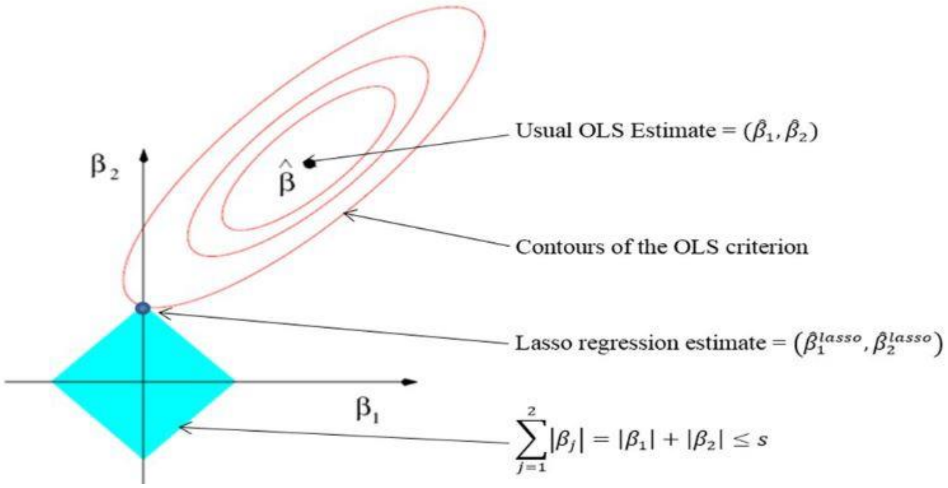
λ (lambda) given here is denoted by an alpha parameter in the ridge function, It is also called the penalty term. So, by changing the values of alpha, we are controlling the penalty term. The bigger the penalty, higher is the value of alpha and therefore the magnitude of coefficients is reduced.



4.1.3. Lasso Regression

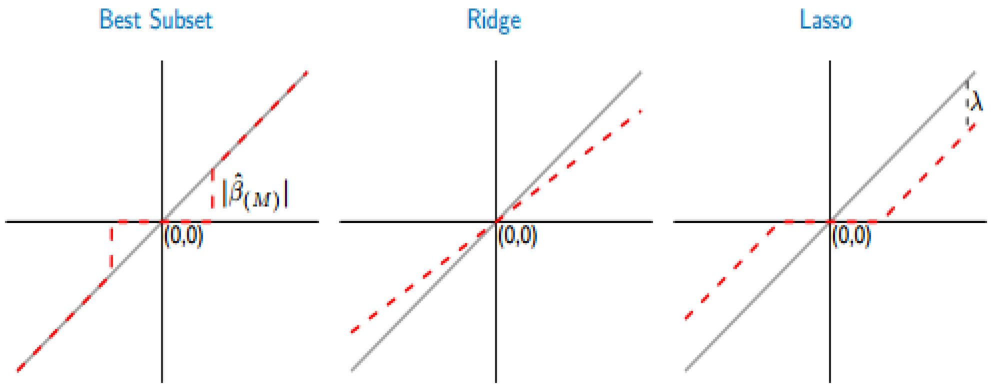
The lasso causes L1 shrinkage, resulting in "corners" in the constraint, which is equivalent to a diamond in two dimensions. If the squared sum "hits" one of these corners, the axis-specific coefficient is reduced to zero. Since the multidimensional diamond has more corners as p rises, it is very likely that some of the coefficients will be set to zero. Thus, shrinkage and (essentially) subset selection are performed via the lasso.

Lasso performs a soft thresholding in contrast to subset selection. The sample path of the estimates moves continuously to zero as the smoothing parameter is varied..


β_2
$\hat{\beta}$
β_1
Usual OLS Estimate = $(\hat{\beta}_1, \hat{\beta}_2)$
Contours of the OLS criterion
Lasso regression estimate = $(\beta_1^{lasso}, \beta_2^{lasso})$
$\sum_{j=1}^{2} |\beta_j| = |\beta_1| + |\beta_2| \leq s$

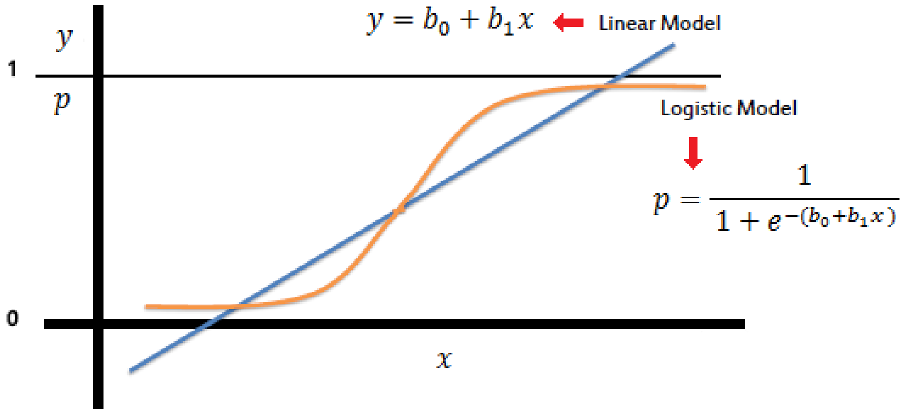
4.1.4. Best Subset Selection

Let's say we have a set of variables we can use to predict an important event, and we want to select the subset of those variables that most closely matches the outcome. One method is to fit all the potential variable combinations and select the one that meets the best criterion in accordance with specified criteria once we have chosen the type of model (logistic regression, for example). Best subset selection is what we term this. It takes a lot of calculation to use this method. We need to fit 2^p models when there are p possible predictors. Additionally, the issue becomes insurmountable very fast if we want to apply cross-validation to assess their performance. The leaps and bounds technique expedites the process of finding the "best" models and does not necessitate fitting each model individually. In any event, if there are too many predictors, even this technique is useless.



4.1.5. Logistic Regression

Logistic regression is a classification procedure that is used to forecast the likelihood of a target variable. The target or dependent variable has a dichotomous character, which means there are only two potential classes. The representation for logistic regression is an equation. To anticipate an output value, input data are linearly mixed with coefficient values. The output value is represented as a binary value, which distinguishes it from linear regression.



5. Analysis and Results

Visualizations are performed in each step, in order to highlight new insights about the underlying patterns and relationships contained within the data. The data analysis process for the deployment of linear models is based on the following steps.

5.1. Software & Programming Languages

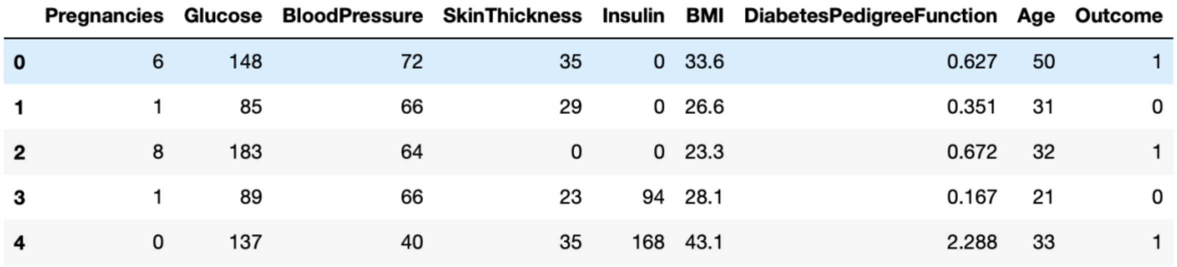
1. Python - Jupyter Notebook (EDA)
2. R - RStudio (Model Deployment)

5.2. Data Acquisition

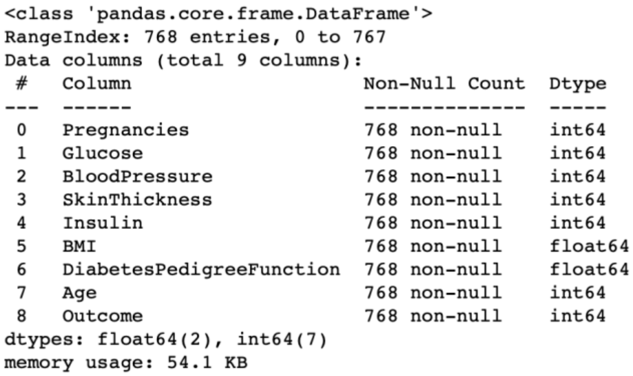
1. Download data
2. Upload data in Python environment

5.3. Data Exploration

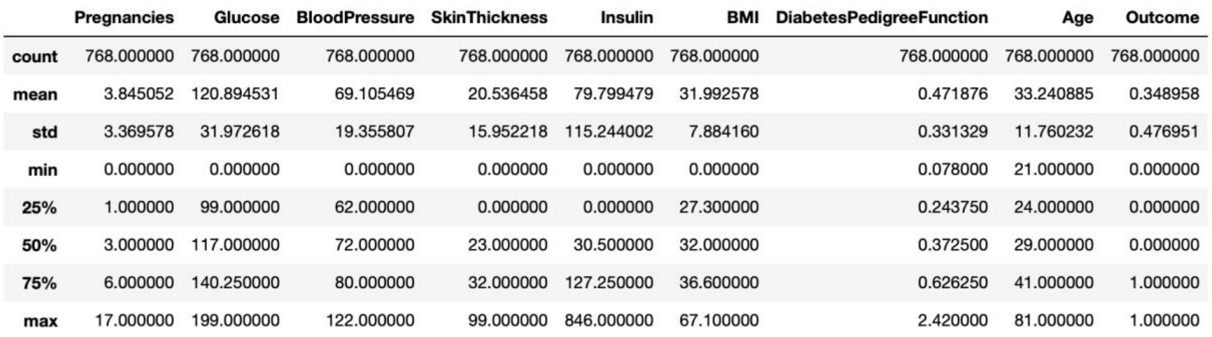
Python is used in jupyter notebook for data analysis and preprocessing.
Checking data head, info, summary statistics and null values.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
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

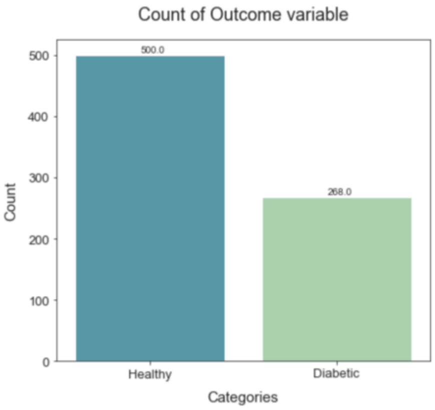


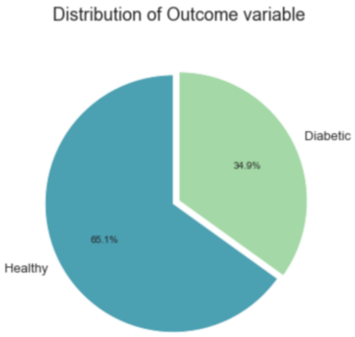
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

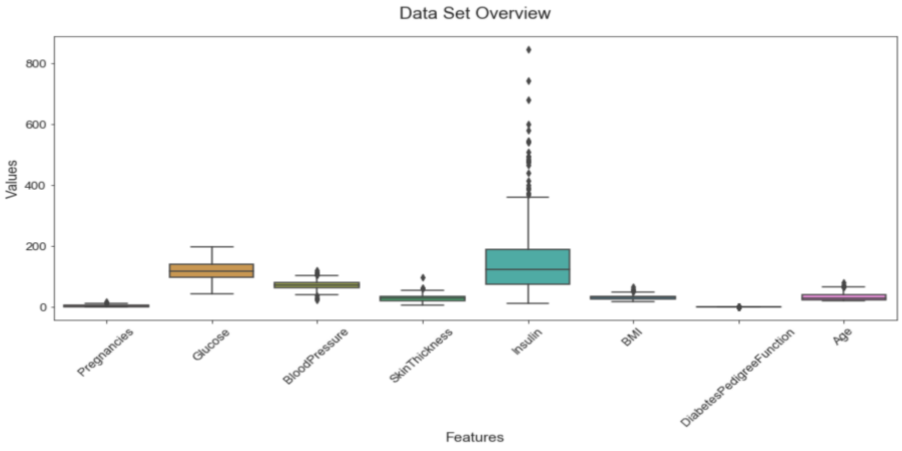
The dataset consists of several predictor variables and one target variable, Outcome. The dataset consists of 652 entries with null values.





From the above graph we can say that less than half of the females are suffering from Diabetes Disease with a percentage of 34.9%. Some features contain 0 in the data. It doesn't make sense here and this indicates that there are missing values in our data. By further looking into the data, we found that it is because there are a number of factors involved in the data to correctly assess the occurrence of diabetes, one of them being the Glucose feature.

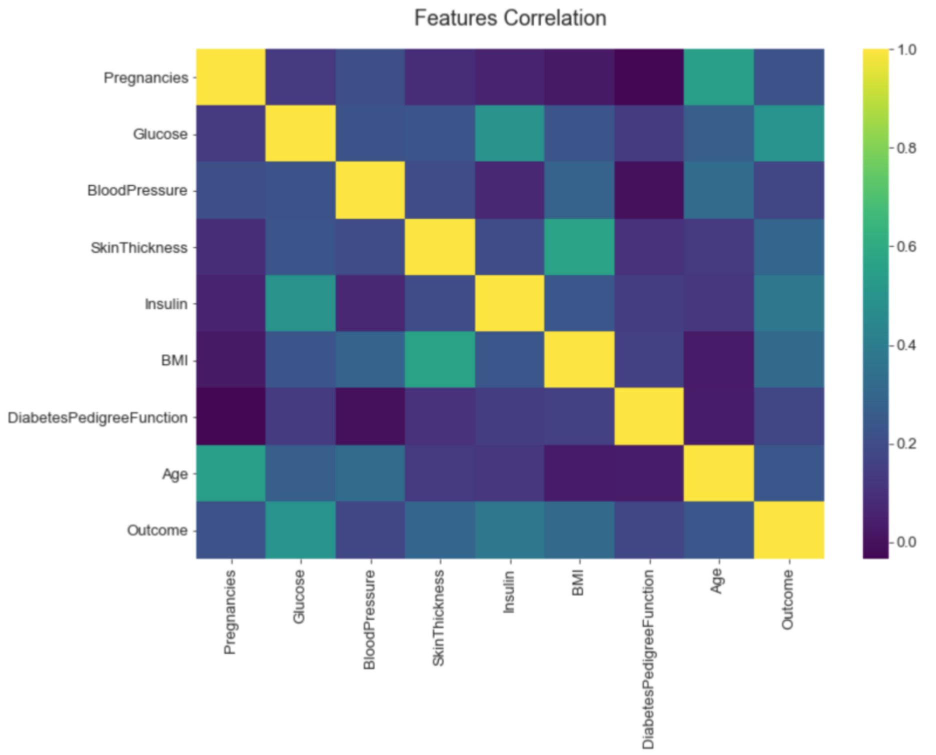
Let's get some more insights about the features with box plots.



From the above plot we can see that there are outliers in the data and those outliers might be valid values so median in this scenario is a more robust central tendency than mean and hence we will use the median by target to replace the missing data to get a much more realistic data.

5.4. Feature Analysis

For Plasma glucose concentration in 2 hours in an oral glucose tolerance test we got 107 Plasma glucose concentration level for non diabetic females and 140 for diabetic females. For Diastolic blood pressure (mm Hg) we got 70 Diastolic blood pressure for non diabetic females and 74.5 for diabetic females. For Triceps skin fold thickness (mm) we got 27 Triceps skinfold thickness for non diabetic females and 32 for diabetic females. For 2-Hour serum insulin (mu U/ml) we got 102.5 serum insulin for non diabetic females and 169.5 for diabetic females and for Body mass index (weight in kg/(height in m)^2) we got 30.1 Body mass index for non diabetic females and 34.3 for diabetic females.



From the above correlation graph we can see that highly correlated features are Pregnancies vs Age, Glucose vs Insulin and SkinThickness vs BMI.

Healthy females are concentrated with $Age <= 30$, $Pregnancies <= 6$, $Insulin < 200$, $Glucose <= 120$, $BMI <= 30$, and $BMI <= 30$ and $SkinThickness <= 20$.

5.5. Linear Machine Learning Model Deployment

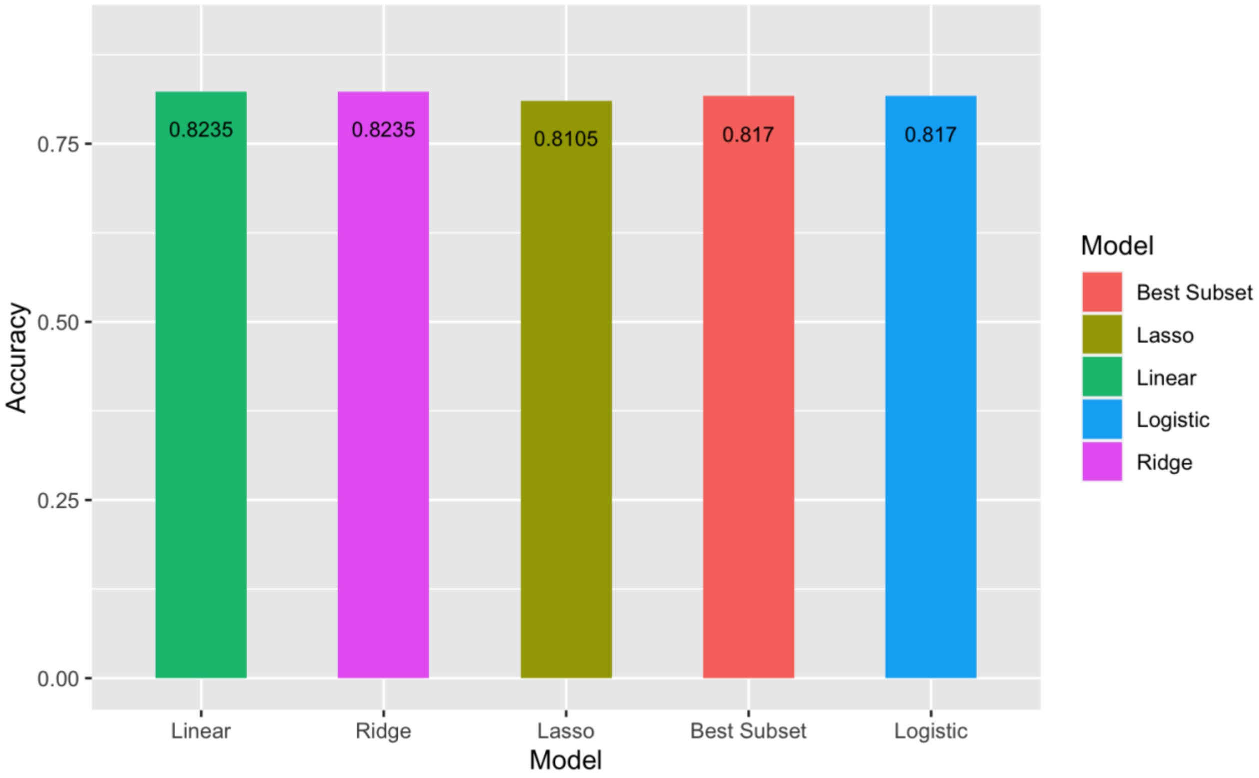
After EDA the dataset is saved as *diabetes_new.csv* and uploaded into RStudio for linear machine learning model training and testing.

1. Modeling of the linear system
2. Validation of the model
3. Visualization of the model
4. Visualization and interpretation of the results

5.6. Results analysis

1. Deployment of the models
2. Comparing prediction accuracy of ML models
3. Visualization of the results

Results were compiled and visualized, Linear model results were as follows:



6. Conclusions

The goal was to predict whether a person has a chance of having diabetes in future. Linear Regression, Ridge Regression, Lasso Regression, Best Subset Selection and Logistic Regression models were trained. Results show that **Linear** & **Ridge Regression** achieved the highest accuracy of **82.35%** after training, which is very promising considering the small size of available data.

Since we were using linear models on the data we specified a cutoff for calculating our model accuracies which has to be from 0-1 as our Outcome can either be 0 or 1. The models' accuracies change when we alter the cutoff value. We can see that even Linear Regression and Ridge Regression have the same accuracy based on a cutoff of 0.4. Linear Regression worked best based on the average of the accuracies obtained from altering the cutoff value so "Linear Regression" is selected due to its highest accuracy. The results compiled after the implementation of this algorithm are displayed as under:

Table 3: Prediction Results

Model	**Accuracy**
Linear Regression	82.35%
Ridge Regression	82.35%
Lasso Regression	81.05%
Best Subset Selection	81.70%
Logistic Regression	81.70%

As a result, the Linear & Ridge Regression model had the greatest prediction accuracy of 82.35%. Note that the best subset selection model produces a lower accuracy than linear regression model, this is because we are using linear regression with a cutoff prediction point for a classification problem. Therefore the best subset selection model does not have better accuracy than the linear regression model.

The Linear models gave us very positive results considering the size of the data. For future work if we use classification algorithms such as Random Forest, Decision Tree and Support Vector Machine, they will definitely give us better results and we can further test our models by calculating their average accuracy using k-fold cross validation technique.

7. Bibliography and Credits

7.1. Research Papers

[1] Al-Zebari, Adel, and Abdulkadir Sengur. “Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection.” 2019 1st International Informatics and Software Engineering Conference, IEEE, 2019, pp. 1–4. https://doi.org/10.1109/UBMYK48245.2019.8965542

[2] Reddy, K. V. Siva Prasad. “Prediction of Heart Disease Using Logistic Regression Algorithm.” International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 10, Oct. 2022, pp. 1352–55. https://doi.org/10.22214/ijraset.2022.47181

[3] Farbahari, Arash, et al. “The Usage of Lasso, Ridge, and Linear Regression to Explore the Most Influential Metabolic Variables That Affect Fasting Blood Sugar in Type 2 Diabetes Patients.” Romanian Journal of Diabetes Nutrition and Metabolic Diseases, vol. 26, no. 4, Dec. 2019, pp. 371–79. https://doi.org/10.2478/rjdnmd-2019-0040

7.2. Dataset

[1] https://www.kaggle.com/datasets/kandij/diabetes-dataset?select=diabetes2.csv

[2] https://drive.google.com/file/d/1wTvjt4pF1TXncZRqhC10sq8sNbjSWQIU/view?usp=share_link