Principal Component Analysis for Identifying Diabetes

By SIVA TEJA NARAYANABHATLA, 40166568

***Abstract*— Principal component Analysis ((PCA) is a strategy for decreasing the dimensionality of such datasets, extending interpretability, but restricting data loss. It does as such by making new uncorrelated elements that continuously support change. A persistent sickness diabetes mellitus is expecting disease extent around the world. Subsequently pervasiveness is significant in all viewpoints. Analysts have presented different strategies, yet, the improvement is a requirement for characterization methods. This paper thinks about head part examination (PCA) methods, on a solitary stage to approaches on the polytomous variable-based characterization of diabetes mellitus and some features responsible for it. The PCA result shows eigenvalues, and the complete difference is clarified for the vital parts (PCs) arrangement. A sum of five features was dissected with the goal to exact the example of the relationship with least factors as could be expected. Ordinarily, factors with enormous eigenvalues are held. The initial two parts have their eigenvalues sufficiently huge to be held. That clarifies ~84.7% of all-out difference. We further applied K-implies bunching with the guide of the initial two PCs. Too, relationship results between diabetes mellitus and highlights liable for diabetes.it has been uncovered that diabetes patients are bound to have high glucose and insulin. Along these lines, the review approves examination of two of the standard classifiers to decide the best classifier. Finally, we reported a highest accuracy of 77% with the SVC classifier.**

***Index Terms*—pca, covariance, linear regression, diabetes.**

I. Introduction

Diabetes is an infection that happens when your blood glucose, additionally called glucose, is excessively high. Blood glucose is your principal wellspring of energy and comes from the food you eat. Insulin, a chemical made by the pancreas, helps glucose from food get into your cells to be utilized for energy. Some of the time your body doesn't make enough—or any—insulin or doesn't utilize insulin well. Glucose then, at that point, stays in your blood and doesn't arrive at your cells. Over the long run, having a lot of glucose in your blood can cause medical issues. In spite of the fact that diabetes has no fix, you can find ways to deal with your diabetes and remain solid. In some cases individuals refer to diabetes as "a bit of sugar" or "marginal diabetes." These terms recommend that somebody doesn't actually have diabetes or has a less genuine case, yet every instance of diabetes is not kidding. The most common types of diabetes are type 1, type 2, and gestational diabetes. If you have type 1 diabetes, your body does not make insulin. Your immune system attacks and destroys the cells in your pancreas that make insulin. Type 1 diabetes is usually diagnosed in children and young adults, although it can appear at any age. People with type 1 diabetes need to take insulin every day to stay alive. If you have type 2 diabetes, your body does not make or use insulin well. You can develop type 2 diabetes at any age, even during childhood. However, this type of diabetes occurs most often in middle-aged and older people. Type 2 is the most common type of diabetes. In this report we gather the data from the file called diabetes_dataset.latest.csv.

https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Fig 1 sugar reading.

In further sections we explain the methodology , description of PCA, classifiers and experimental results along with classification results and with the conclusion we make a statement about which classifier is a more accurate in detecting the diabetic respectively in II,III,IV,V and VI sections.

II. PRINCIPAL COMPONENT ANALYSIS

PCA is one of the most decrease methods that is utilized for removing significant elements (parts) from a tremendous arrangement of factors available in a dataset. It removes a low dimensional arrangement of components from a high dimensional dataset with a goal of protecting however much data as could be expected. PCA is more helpful when the information comprises of at least two or three-dimensionalities. It is constantly performed on a symmetric connection information. First head part is a direct combination unique indicator factors which addresses the most noteworthy fluctuation of the dataset. It chooses the heading of most fluctuation in the information. Higher the changeability trapped in the primary part suggests more data got by part. No other part can have changeability higher than the main head component. The primary component part brings out to be a line which is closest information for example it restricts the amount of squared distance between an element and the like an analogous way, we can likewise process the second component part. Second component part is a straight mix of unique indicators like the primary part which gets the remainder of fluctuation in the dataset and is uncorrelated with the first component result. Subsequently, between first- and second-part to be zero. The bearing of two parts are symmetrical, assuming they are uncorrelated.

III. CLASSIFICATION ALGORITHMS

To describe classification briefly it's a binary separation in layman terms it's like or dislike without having common ground. There are two ways to perform classification they are supervised learning and unsupervised learning algorithms. The supervised machine learning algorithm is executed when the preprocessed data is labelled and then the output of the test data is given based on the labelled trained data. In case of unsupervised the data is directly fed to the classifier and without any labels. After labeling the dataset the training procedure is carried out to retrieve feasible output for further knowledge discovery. The unsupervised learning algorithms deal with unlabeled dataset. This report utilizes marked information from the clinical field for characterization purposes. The classifiers used in this report are SVC (Support vector classifier) and KNN (k- nearest neighbor) classifier algorithms which are unsupervised and supervised algorithms respectively.

A. Support Vector Classification:

It is a supervised machine learning algorithm which can be used for regression and classification more often we use it for classification problems and in SVM algorithm we plot each data item in a n-dimensional space with n being the number of vectors with value of each feature being the value of particular coordinate. Then we perform the classification by calculating the hyperplane which separates the two classes with perfection and the support vectors are the coordinates of the individual observation as shown in the image.

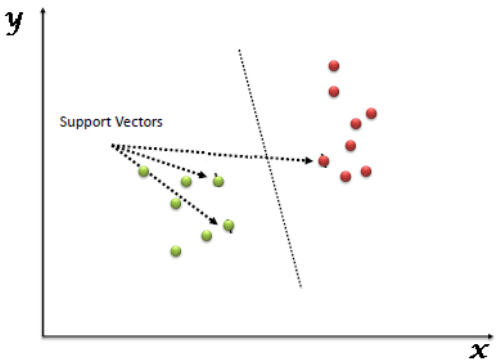


Fig2 SVM classifier

B. KNN classifier Algorithm (K – nearest neighbor)

The KNN algorithm assumes that similar data exists in nearby vicinity based on the distance between the points. We assume the K value and when we decrease the K value gradually to 1 we observe that our predictions also become more random. Conversely as we increase the K value our prediction become more stable due to majority of the points falling as per the requirement. To have more concrete decision about the K value we choose to be an odd number.

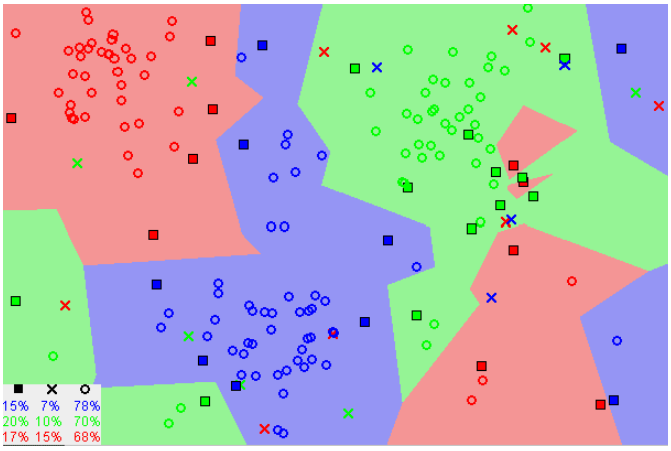


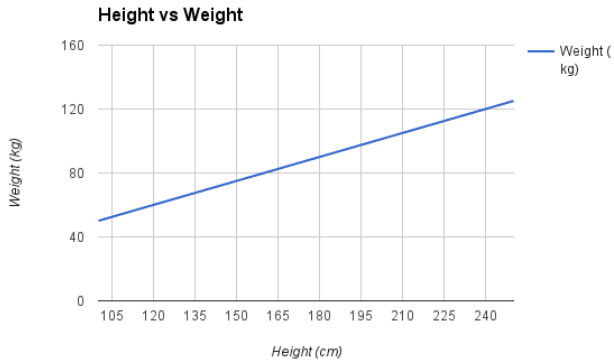
Fig 3 shows similar data points exist close to each other.

C. Linear Regression Algorithm:

It is a supervised machine learning model in which the finds the best linear fit between independent and dependent variables. The direct condition relegates one scale variable to each information worth or segment, called a coefficient, and addressed by the capital Greek letter Beta (B). One extra coefficient is likewise added, providing the line with an extra level of opportunity (for example going all over on a two-dimensional plot) and is frequently called the catch or the predisposition coefficient.

$$y = \beta_o + \beta_1 X$$

At the point when a coefficient becomes zero, it viably eliminates the impact of the info variable on the model and accordingly from the expectation produced using the model (0 * x = 0). This becomes important assuming you take a gander at regularization techniques that change the learning calculation to lessen the intricacy of relapse models by coming down on the outright size of the coefficients, driving some to nothing. A very naïve yet effective algorithm



IV. EXPERIMENTAL RESULTS

A. Data Set Description:

We have saved our data in a csv file. Then we read our data set into panda's data frame as df. We use df.head() function to show the first five row of the data which we saved. We have chosen most common features in all human beings such as age, glucose levels, BMI, blood pressure, skin thickness insulin. These factors are standard irrespective of age and sex of the person. The last column is outcome or class which is either 1 or 0. This gives us proper classification whether the person is diabetic or not. We initially preprocess the dataset

such that it does not have any null values or values which can't be converted into numerical. Then we have excluded few columns which are not related to our classification such as columns index, class and etc. Our aim is to target the class column and classify the data accordingly and index numbers in the dataset is does not impact the classification.

B. PCA to remove redundant data.

We know that PCA is used to remove the data which are mutually dependent to avoid a biased decision or incorrect classification. We apply PCA on the dataset we saved and delete those rows or columns which are dependent on the other columns who values are base for the other columns. Let's say the PCA reduces our dataset to **r** columns which are mutually independent to each other. To determine the value of **r** we used the scree plot or elbow plot.

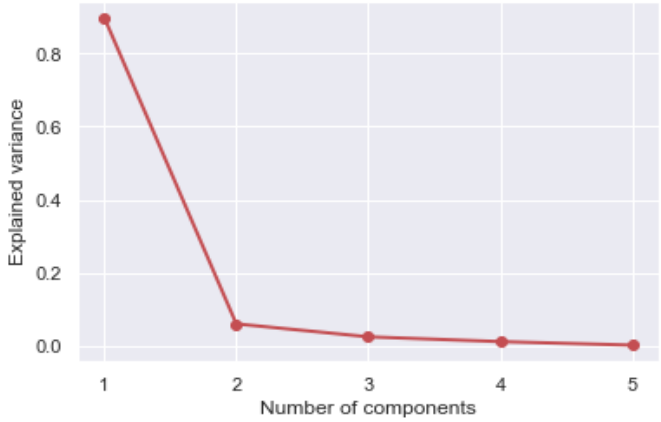


Fig 3 scree plot.

https://www.kaggle.com/uciml/pima-indians-diabetes-database/data

We used box plot *fig 4* to determine the distribution of the data and also identify which data points are out of outliers and where the data is centered and to determine the covariance of the different values of the features we have plotted the covariance matrix in *fig 5*.

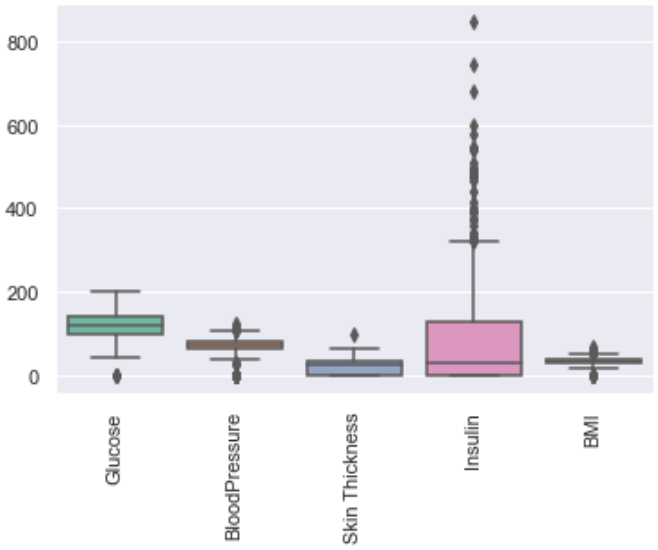


Fig 4 Box Plot

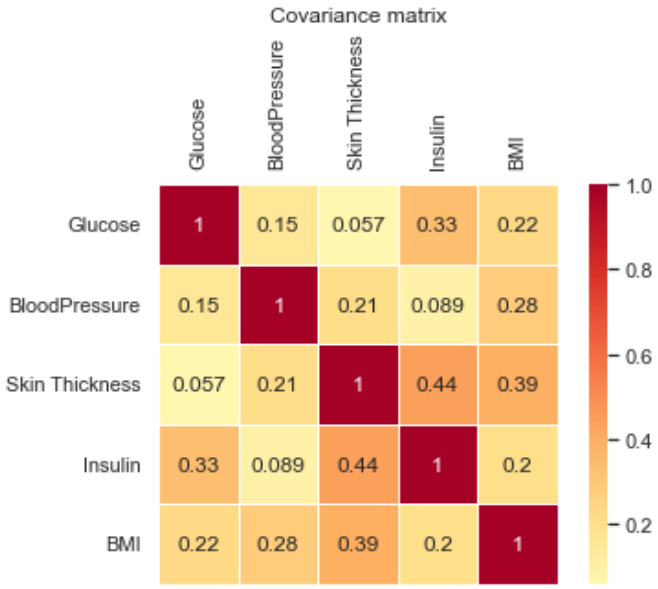


Fig 5 Covariance matrix

In-order to perform PCA on the data set we first convert the observations into a centered matrix which is subtracting each observation with the mean of the column in which the observation is present i.e. let x_i be the observation and X^- be the mean of the column the observations of the column are changed from x_i to $X_i - X^-$. Second step is calculating the covariance matrix. This can be calculated by the formula

$$S = \frac{Y'Y}{n-1}$$

We then calculate the Eigen values and vectors using python.

Principal Components:

To calculate the PC'S we compute the transformed matrix Z by multiplying the centered matrix with the Eigen vector matrix. In transformed matrix the rows of the matrix corresponds to observations while the columns are the PC's shown below.

Fig 6 principal components.

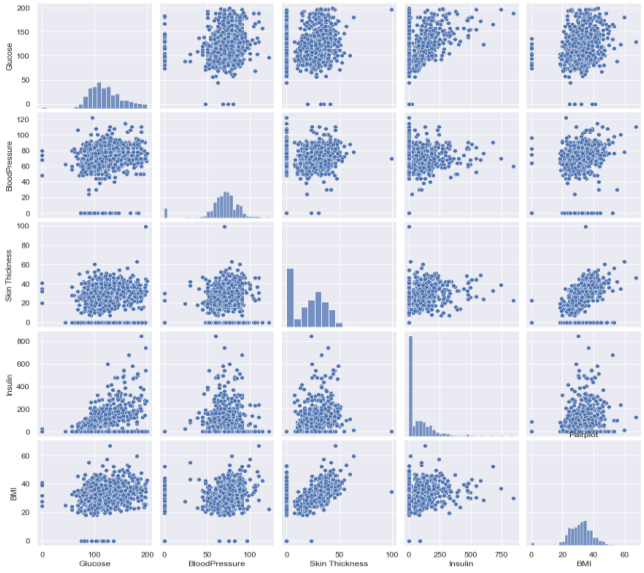


Fig 7 pair plot

The Eigen values ae then arranged in descending order of the values indicating that the correlation between the components. Based on the r values the r values are retained and the significance components are identified.

We have also plotted the scatter plot to show which variables are similar to which PC's.

The biplot helps us to visualize both the principal component coefficients for each variable and scores of the each observation. Principal components are axes of the Biplot. Whereas the observed variables are shown as vectors. The blue and orange dots represent the classes to determine as diabetic or non-diabetic.

Fig 9 biplot

PCA Control Chart:

Principal components including the correlation among variables are useful in detecting the outliers and thus represented using the PCA control chart and the scatter plot. Which shows how far the variables from each other are.

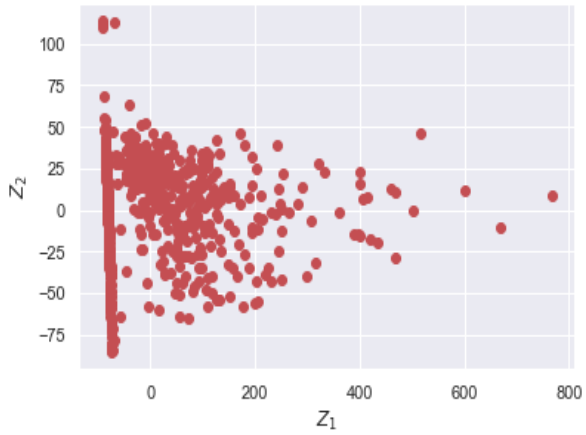


Fig 9 PCA

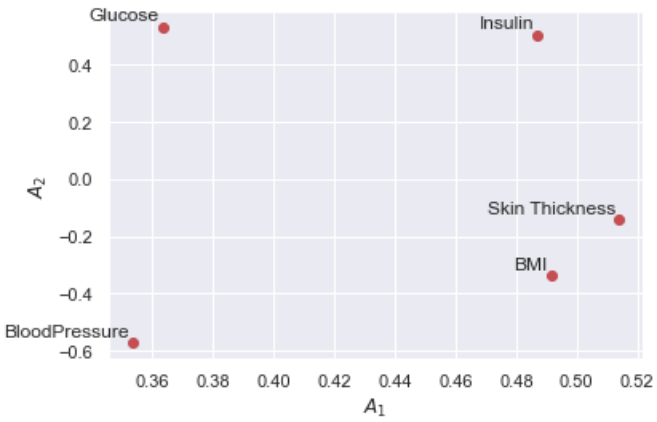


Fig 10 Scatter plot

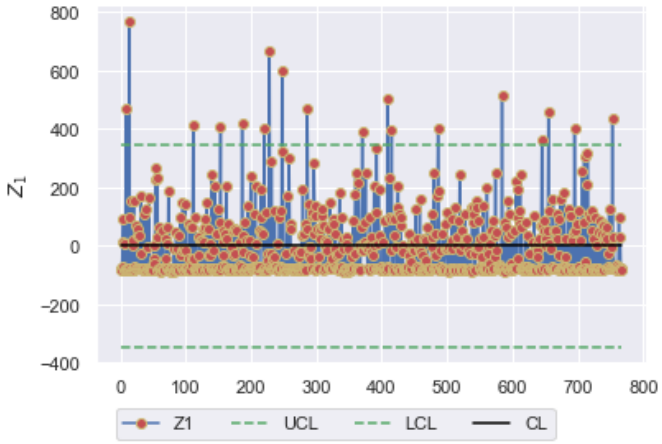


Fig 11 out of control plot

V CLASSIFICATION

In this section we apply three classification algorithms namely linear regression, svm algorithm and KNN classification on the preprocessed data set. Then we compare the accuracy of the algorithms and analyze them. The aim is to check how PCA helps or how it impacts the classifiers.

A. *Linear Regression:*
Analysis components required
Original data set
Principal components
Dominating components

B. *Svc Algorithm:*
Analysis components required
Original data set
Principal components
Dominating components

C. *KNN Algorithm:*
Analysis components required
Original data set
Principal components
Dominating components

The assessment of the classifier relies upon the dataset that is utilized. We will utilize the fit time and precision to evaluate the presentation of the classifiers. A accuracy of the classifier is determined and higher the accuracy better the algorithm in classifying the data correctly.

VI CONCLUSION

In this report, we successfully discriminated benign from malignant cancers with a high accuracy (77%) for SVM classifier. We applied PCA on the diabetes dataset with 200 benign and 563 malignant subjects. We have found that 84.7% of variance in the first two PCs components. As a result, we work with these two components. We analyzed the impact of PCA on different classifiers. We tested the PCA component's control charts and found that some of the data points were lying out of control (around 6 points). On the second part of this project we applied Logistic Regression and Naive Bayes classifiers on the resulting data. We classified data into benign and malignant. The performance of each classifier is determined by the fit time and the accuracy. We noticed that the Logistic Regression has better performance than Naïve Bayes in all cases. For Naive Bayes, it performs well in the first two components compared to all PCA components and the original dataset. For the Logistic Regression, it performed slightly well in all PCA components and original dataset compared to the first two components.

VII REFERENCES

1) H. Cheng and M. Cui, "Mass lesion detection with a fuzzy neural network," Pattern Recognition, vol. 37, no. 6, pp. 1189–1200, 2004.

2) A. Mahajan, S. Kumar and R. Bansal, "Diagnosis of diabetes mellitus using PCA and genetically optimized neural network," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 334-338, doi: 10.1109/CCAA.2017.8229838.

3) S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 141-146, doi: 10.1109/ICACCS51430.2021.9441935.

4) J. S. Parab, R. S. Gad and G. M. Naik, "Influence of PCA components on glucose prediction using non-invasive technique," 2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES), 2016, pp. 473-476, doi: 10.1109/ICAEES.2016.7888091.