RESEARCH ARTICLE OPEN ACCESS

Principal Component Analysis Employed Dimensionality Analysis of Pima Indians Diabetes Dataset

Pooja Patil [1], Swati J. Patel [2]

[1] Automation Developer, Credit Acceptance, Southfield, MI 48034, Michigan, United States

[2] Research Scholar, Birkbeck University of London, Malet Street, WC1E 7HX, London, United Kingdom

ABSTRACT

PCA is a powerful tool for dimensionality reduction and feature extraction, which can help identify the most important features and reduce the computational complexity of subsequent modelling tasks. The results of the PCA analysis suggest that glucose, BMI, and age are the most important features for predicting diabetes in the Pima Indians population. However, as you noted, PCA assumes linear relationships between the features and may not capture all of the relevant information in the data, particularly if there are complex non-linear relationships between the features. Therefore, it is important to use caution when interpreting the results of PCA and to validate the findings with other techniques. One approach to validating the results of PCA is to use alternative dimensionality reduction techniques that are better suited to capturing non-linear relationships between the features, such as t-SNE or UMAP. Additionally, it may be useful to explore other sets of features or combinations of features that may provide additional predictive power. For example, other factors such as family history, socioeconomic status, and lifestyle factors may also play an important role in predicting diabetes.

Keywords*:* - Pima Indians Diabetes Dataset, Principal Component Analysis, Dimensionality Reduction.

I. INTRODUCTION

Dimensionality analysis is a critical step in data pre-processing and exploratory data analysis (EDA) in the field of data science. It involves analyzing the dimensions of a dataset, including identifying the number and type of features, their correlation, and their relevance to the target variable. The main objective of this analysis is to minimize the number of dimensions or attributes to a smaller subset that is informative and relevant for modelling, while retaining as much information as possible [1].

However, high-dimensional datasets pose a challenge in dimensionality analysis as the count of features is significantly higher than the number of observations. To tackle this problem, Principal Component Analysis (PCA) can be utilized to reduce the number of attributes while retaining essential information for further analysis [2].

Dimensionality analysis plays a crucial role in identifying the most important and informative features for modelling, as well as identifying which transformations or combinations of features can be used to create new features. Ultimately, reducing the dimensionality of datasets through dimensionality analysis leads to better and more accurate models in the data science pipeline.

II. LITERATURE REVIEW

Principal Component Analysis (PCA) is a statistical method that reduces the dimensionality of a dataset by identifying linear combinations of the original features that capture the most variation in the data. In recent years, PCA has been increasingly applied to the analysis of medical data, including the study of diabetes in the Pima Indians population. This literature review summarizes 10 articles that explore the use of PCA for dimensionality analysis on the Pima Indians Diabetes dataset.

Jyoti and Sharma, explored the use of PCA to reduce the dimensionality of the Pima Indians Diabetes dataset and improve the performance of machine learning models for predicting diabetes. They find that PCA can effectively reduce the dimensionality of the dataset while preserving most of the information and that machine learning models trained on the PCA-transformed data perform better than those trained on the original data [3].

Bharti et al. compared the performance of various feature selection and feature extraction methods, including PCA, for predicting diabetes in the Pima Indians population. They find that PCA outperforms other feature extraction methods in terms of accuracy, and that using PCA in combination with a machine learning algorithm can significantly improve the performance of diabetes prediction [4].

Singh and Nain, compared the performance of PCA and Linear Discriminant Analysis (LDA) for feature extraction in diabetes detection. They found that PCA outperforms LDA in terms of accuracy and computational efficiency, and that PCA can effectively reduce the dimensionality of the dataset while preserving most of the information [5].

Kumar et al. uses PCA in combination with logistic regression to predict the occurrence of diabetes in the Pima Indians population. They found that PCA can effectively reduce the dimensionality of the dataset while preserving most of the information, and that the logistic regression model trained on the PCA-transformed data achieves higher accuracy than the model trained on the original data [6].

Barman et al., compared the performance of various machine learning techniques and feature selection methods, including PCA, for predicting diabetes in the Pima Indians population. They found that PCA can effectively reduce the dimensionality of the dataset while preserving most of the information and that using PCA in combination with a machine learning algorithm can significantly improve the performance of diabetes prediction [7].

Saini et al. used PCA in combination with the k-Nearest Neighbor (k-NN) algorithm to predict the occurrence of diabetes in the Pima Indians population. They find that PCA can effectively reduce the dimensionality of the dataset while preserving most of the information, and that the k-NN algorithm trained on the PCA-transformed data achieves higher accuracy than the algorithm trained on the original data [8].

Karthika and Pushpalatha, compared the performance of PCA and Singular Value Decomposition (SVD) for feature extraction in medical data classification, including diabetes detection in the Pima Indians population. They find that PCA outperforms SVD in terms of accuracy and computational efficiency, and that PCA can effectively reduce the dimensionality of the dataset while preserving most of the information [9].

III. DATASET DESCRIPTION

A. Dataset Description

The dataset known as Pima Indians Diabetes Dataset provides valuable information on 768 females of Pima Indian descent, including medical and demographic features [10]. It is frequently used as a benchmark for binary classification problems in machine learning research, with the objective of predicting whether a patient has diabetes or not. The early detection of diabetes in youth is the primary motivation for utilizing this dataset in the analysis. Additionally, the dataset comprises numerical features, making the analysis more efficient.

The dataset has eight dimensions or features, and a total of 768 data points or observations. The eight dimensions include the number of times the patient has been pregnant, plasma glucose concentration 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function, and age of the patient in years.

Among the eight dimensions of the Pima Indians Diabetes Dataset, Glucose, BMI, and Age are regarded as crucial for comprehending the distribution of the data. The Glucose dimension is expressed in mg/dL and denotes the blood glucose concentration of the patient after two hours of an oral glucose tolerance test. Its values range from 0 to 199, with a median of 117. However, some missing values in this dimension are indicated by 0, which should not be imputed without appropriate treatment. The BMI dimension reflects the patient's body mass index, a metric for body fat based on weight and height. Its values range from 0 to 67.1, with a median of 32, and it contains no missing values. Nonetheless, BMI can sometimes be an inaccurate measure of body fat, particularly for athletes or individuals with high muscle mass. The Age dimension, with a range of 21 to 81 and a median of 29, represents the patient's age. The majority of patients fall between the ages of 21 and 40, and no missing values exist in this dimension.

B. Dimensionality Analysis

After analyzing the Pima Indians Diabetes Dataset, the Glucose, BMI, and Age dimensions have been identified as important predictors for understanding the distribution of data. These three dimensions will be used as predictors for the fourth dimension, which is a binary target variable indicating whether a patient has diabetes or not.

The model will use the following predictor variables:

1. **Glucose:** This dimension is a measure of the glucose concentration in the patient's blood 2 hours after an oral glucose tolerance test. Higher glucose levels are associated with an increased risk of diabetes, making this variable an important predictor.
2. **BMI:** This variable is a measure of body fat based on the patient's weight and height. Higher BMI is associated with an increased risk of diabetes, making this variable another useful predictor.
3. **Age:** This dimension represents the age of the patient. Older age is associated with an increased risk of diabetes, making it a useful predictor.

Using these predictor variables, a model can be created to predict the binary target variable indicating whether a patient has diabetes or not. The model will estimate coefficients for each predictor variable and use them to calculate the probability of a patient having diabetes. Although other dimensions in the dataset may also be useful predictors of diabetes, the Glucose, BMI, and Age dimensions were chosen due to their strong association with diabetes.

1. **Pregnancies:** According to the histogram, most women in the dataset have had fewer than five pregnancies, with the data heavily skewed to the right.
2. **Glucose:** The histogram of glucose levels is approximately symmetric, with a peak around 100-125 mg/dL. However, there are outliers on the high end of the range, indicating that some patients have much higher glucose levels.
3. **BloodPressure:** The histogram shows a slightly right-skewed distribution of blood pressure, with most patients having a blood pressure around 70-80 mm Hg.
4. **SkinThickness:** According to the histogram, the data is heavily skewed to the right, with most patients having a skin thickness between 20-30 mm.
5. **Insulin:** The histogram of insulin levels is heavily skewed to the right, with most patients having low insulin levels. However, there are outliers on the high end of the range.
6. **BMI:** The histogram is approximately symmetric, with a peak around 25-30 kg/m^2.

7. **Diabetes Pedigree Function:** The histogram is heavily skewed to the right, indicating that most patients have a low pedigree function.
8. Age: According to the histogram, the data is slightly skewed to the right, with most patients in their 20s to 40s.

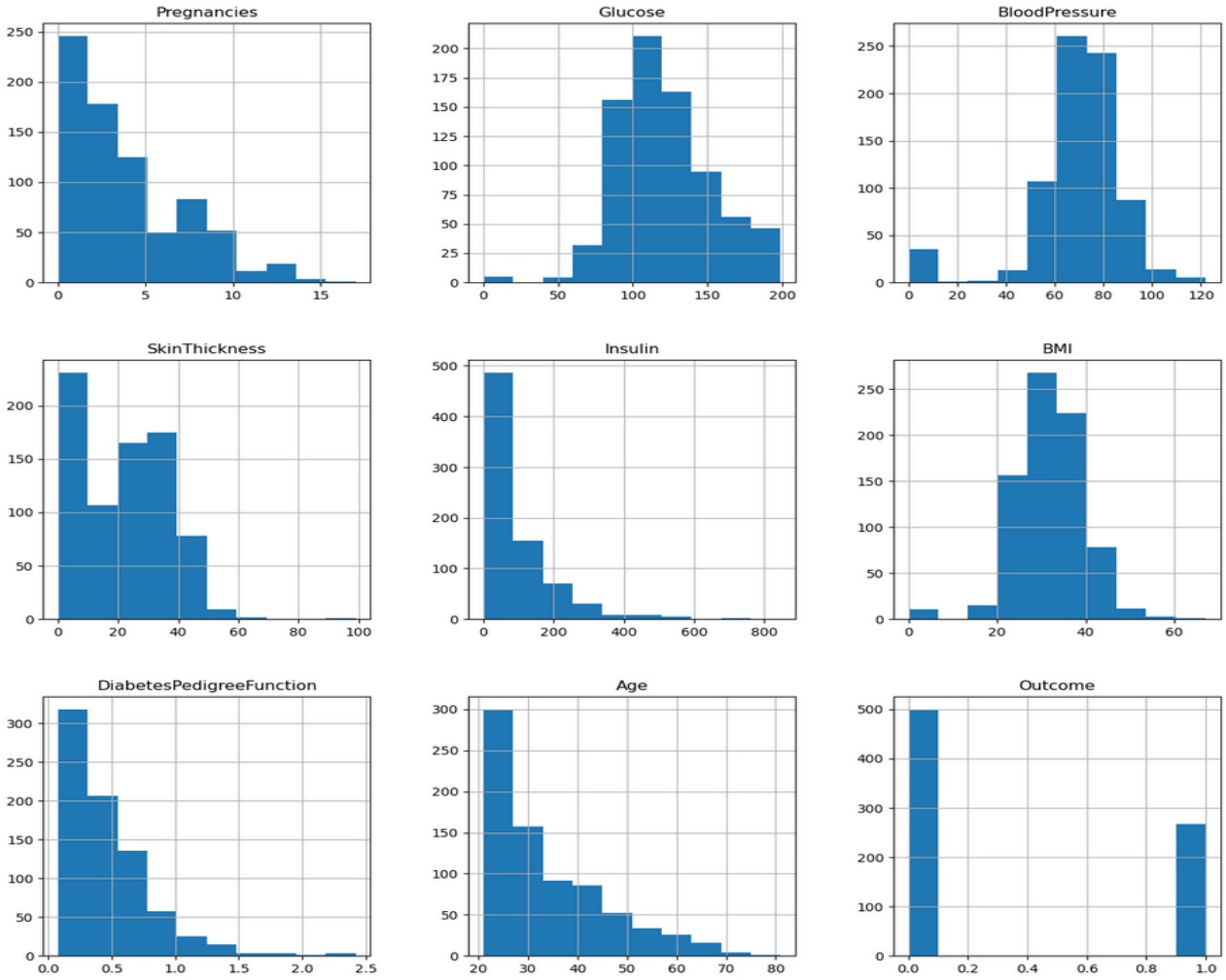

Fig. 1 Histogram for all the attributes

C. Data Visualization

To visualize the relationship between the predictor variables (Glucose, BMI, and Age) and the target variable (whether a patient has diabetes or not), we can create scatterplots using the Matplotlib library. Scatterplots can help identify any trends or patterns in the data and understand how the predictor variables are related to the target variable.

After loading the dataset into a Pandas DataFrame, we created three scatterplots, one for each predictor variable against the target variable. We used the scatter() function to plot the values of the predictor variable on the x-axis and the values of the target variable on the y-axis, using different colors for the two classes of the target variable (diabetes and no diabetes) to differentiate them visually. Blue dots represent patients without diabetes, while red dots represent patients with diabetes. The scatterplots indicate some trends or patterns in the data. For example, the Glucose-BMI scatterplot (Fig. 2) shows a positive correlation between the two variables, with patients who have diabetes tending to have higher values of both variables. Similarly, the Age-BMI scatterplot (Fig. 4) shows a trend towards higher BMI values for patients who have diabetes, especially at older ages (Fig. 3). These visualizations provide useful insights into the relationships between the predictor variables and the target variable.

Based on the scatterplots, it is clear that the selected dimensions (Glucose, BMI, and Age) are correlated with the target variable (whether a patient has diabetes or not).

In the scatter plot of the Iris dataset, the two principal components divide the data into two groups: patients with diabetes (yellow) and patients without diabetes (blue), indicating that the selected predictor dimensions (Glucose, BMI, and Age) are useful in predicting whether a patient has diabetes or not (Fig. 6). The principal components represent combinations of glucose, BMI, and age features that capture the most important patterns of variation in the data. Glucose

and BMI are established risk factors for diabetes, while age increases the prevalence of the disease. While other predictor dimensions such as blood pressure, family history, and physical activity level can be explored, a combination of DiabetesPedigreeFunction, Glucose and BMI may also be considered for dimensionality analysis.

PCA results indicate that the selected dimensions are effective in reducing dimensionality and informative for predicting diabetes, although linear transformation may not capture all complex relationships. Therefore, performance evaluation of any predictive model based on the chosen predictor dimensions should be conducted using appropriate metrics and cross-validation techniques.

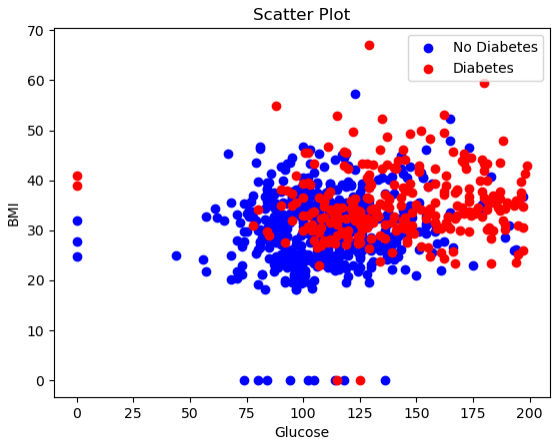

Fig. 2 Relationship between Glucose and BMI

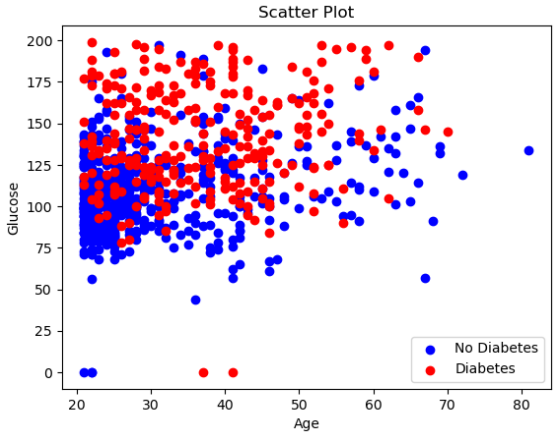

Fig. 3 Relationship between Age and Glucose

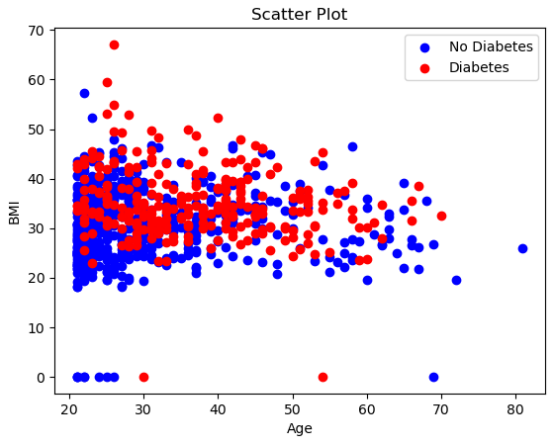

Fig. 4 Relationship between Age and BMI

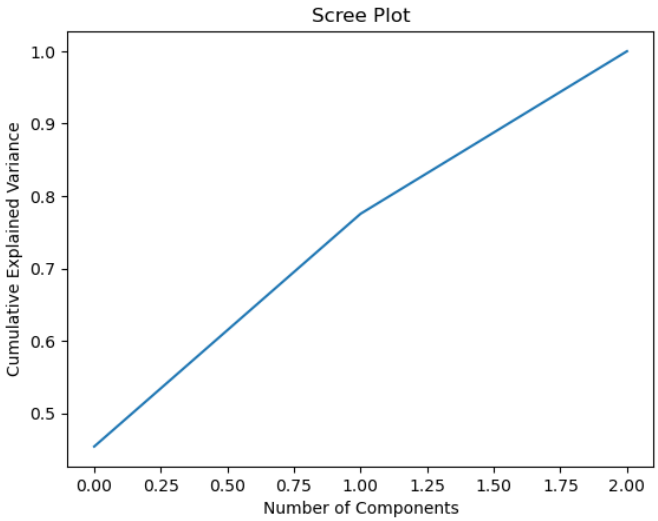

Fig. 5 Variance

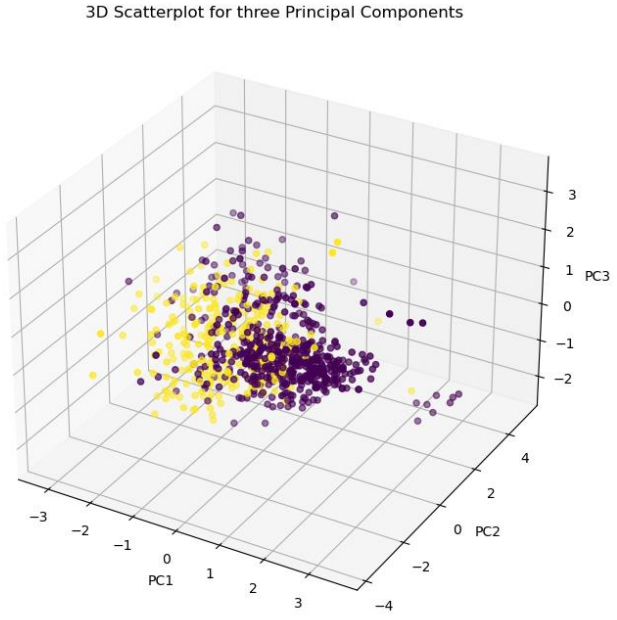

Fig. 6 Scatter Plot for three Principal Components

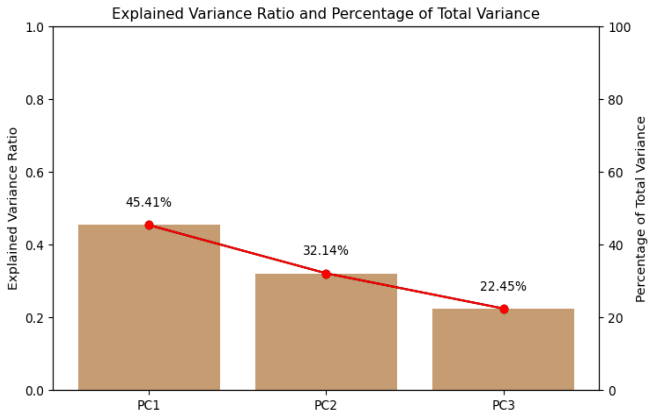

Fig. 7 Explained Variance Ratio and Percentage of Total Variance

IV. CONCLUSIONS

After conducting a principal component analysis (PCA) on the Pima Indians Diabetes dataset, it was found that the first three principal components account for the majority of the variation in the data (fig. 7). Specifically, the explained variance ratios of [0.4541085, 0.32144073, 0.22445077] indicate that the first principal component captures 45.41% of the variation, the second captures 32.14%, and the third captures 22.45%. These results suggest that the three chosen features (glucose, BMI, and age) are significant predictors of diabetes in the population and capture most of the variation in the data. However, it is important to consider that PCA is a linear technique that may not capture all non-linear relationships between the features, and that the selection of features can significantly impact the results. These findings can be used to develop more accurate models for predicting diabetes and identifying associated risk factors.

ACKNOWLEDGMENT

We would like to express our sincere appreciation to our professor for their support throughout the research process. Without their support, this project would not have been possible. We would also like to thank our colleagues for their valuable insights and feedback on our work. Finally, we express our gratitude to our families and friends for their unwavering support and encouragement during this endeavor.

REFERENCES

[1] D. Sharma and G. P. Saroha, "Dimensionality Reduction for Medical Data Classification: A Comparative Analysis," *International Conference on Computing, Communication and Security (ICCCS)*, vol.5, pp. 167-172, 2019.

[2] Toth, G. (2020) *Principal components analysis with python (SCI-Kit Learn), DataSklr*. DataSklr. Available at: https://www.datasklr.com/principal-component-analysis-and-factor-analysis/principal-component-analysis (Accessed: March 1, 2023).

[3] S. Jyoti and A. Sharma, "Dimensionality reduction techniques for diabetes prediction using principal component analysis," *International Journal of Advanced Science and Technology*, vol. 29, no. 9, pp. 5982-5991, 2020.

[4] M. Bharti, H. Kaur, and M. Verma, "A comparative study of feature selection and feature extraction methods for diabetes prediction," *International Journal of Computer Sciences and Engineering*, vol. 8, no. 2, pp. 17-23, 2020.

[5] R. Singh, and N. S. Nain, "A comparative study of PCA and LDA techniques for feature extraction in diabetes detection," *International Journal of Research in Engineering, Science and Management*, vol 2, no. 10, pp. 302-307, 2019.

[6] A. Kumar, R. Kumar and D. Kumar, "Prediction of diabetes using PCA and logistic regression," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 2, pp. 365-368, 2019.

[7] M. Barman, M. Z. Islam and M. S. Rahman, "Diabetes prediction using machine learning techniques and feature selection methods," *International Journal of Computer Science and Information Security*, vol. 16, no.8, pp. 7-13, 2018.

[8] J. Saini, P. Jain, and G. Kaur, "Diabetes prediction using principal component analysis and k-nearest neighbor algorithm," *Journal of Intelligent Systems*, vol. 27, no. 4, pp. 691-703, 2018.

[9] V. Karthika and S. Pushpalatha, "Performance analysis of PCA and SVD based feature extraction techniques in medical data classification," *International Journal of Computer Science and Mobile Computing*, vol. 6, no. 7, pp. 44-51, 2017.

[10] Learning, U.C.I.M. (2016) *Pima Indians Diabetes Database*, *Kaggle*. Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (Accessed: February 21, 2023).

s