Pima Indians Diabetes

## 

## **Abstract: Understanding Diabetes Onset in Pima Indian Women using Machine Learning :**

The Pima Indians Diabetes dataset, originating from the National Institute of Diabetes and Digestive and Kidney Diseases, presents a pivotal resource for investigating the prediction of diabetes onset in Pima Indian women. This dataset comprises 768 instances and eight distinct health attributes, including the number of pregnancies, glucose concentration, blood pressure, skinfold thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and age. Each entry is labeled with a binary outcome indicating the presence or absence of diabetes, transforming the problem into a binary classification task.

This dataset's significance lies in its potential to unravel the intricate relationship between health indicators and the likelihood of diabetes development. Through the application of machine learning techniques, this dataset facilitates the exploration and development of predictive models. These models aim to identify individuals at a higher risk of diabetes, thereby enabling proactive healthcare interventions.

Researchers widely employ this dataset to analyze and build models using diverse machine learning algorithms, including ensemble methods like Random Forests, distance-based classifiers like K-Nearest Neighbors (KNN), probabilistic models like Naive Bayes, and deep learning architectures such as neural networks. Additionally, dimensionality reduction techniques like PCA and LDA aid in feature selection and visualization, augmenting the efficacy of predictive models.

## **Introduction :**

The Pima Indians Diabetes Database, sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, stands as a fundamental resource in both healthcare research and machine learning applications. This dataset focuses on Pima Indian women and is designed to predict the likelihood of developing diabetes within a specific period.

Comprising 768 instances, the dataset encompasses eight key health attributes. These attributes include the number of pregnancies, glucose concentration, blood pressure, skinfold thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and age. Each entry in the dataset is marked with a binary outcome indicating whether the individual eventually developed diabetes or not, thereby framing the problem as a binary classification task.

The dataset's significance lies in its ability to facilitate the evaluation and development of predictive models aiming to identify individuals at risk of diabetes. However, working with this dataset presents challenges, such as dealing with missing values, potential outliers, and the intricacy of predicting a health condition based on a limited set of attributes.

Researchers and data scientists widely utilize this dataset to explore diverse machine learning algorithms, preprocessing methods, and feature selection techniques. The primary goal is to construct accurate predictive models that enable early detection of diabetes, facilitating timely interventions and the formulation of effective healthcare strategies, especially for populations susceptible to this condition.

I have used multiple algorithms to pima india diabetes datasets and got satisfied results for each model I have used. Lets’ dive into the models I have used :

### Random Forest:

* Description: Random Forest is an ensemble learning method that constructs multiple decision trees and merges their outputs to improve predictive performance.
* Application: In this dataset, Random Forest can be utilized to capture complex relationships between attributes like glucose concentration, BMI, age, etc., providing robust predictions. It handles missing values well and offers feature importance insights.

### K-Nearest Neighbors (KNN):

* Description: KNN is a simple algorithm that classifies data points based on the majority class of its nearest neighbors.
* Application: Using KNN on this dataset involves calculating the distance between data points to predict diabetes onset based on similar attribute patterns. Scaling features might be crucial for KNN to work effectively.

### Naive Bayes:

* Description: Naive Bayes is a probabilistic classifier based on applying Bayes' theorem with an assumption of independence between features.
* Application: It's beneficial for this dataset as it handles categorical data and performs well with a small amount of training data. Attributes like glucose levels, BMI, and blood pressure can influence the probability of diabetes, which Naive Bayes can estimate.

### Neural Network with Dense Layers:

* Description: Neural networks with dense layers are deep learning models with fully connected layers, capable of learning complex patterns.
* Application: Constructing a neural network with multiple dense layers can capture intricate relationships among features, enabling the model to predict diabetes by learning hierarchical representations from the dataset.

### Linear Discriminant Analysis (LDA):

* Description: LDA is a dimensionality reduction technique that finds linear combinations of features that best separate different classes.
* Application: LDA can reveal the most discriminative features between diabetes and non-diabetes cases, reducing dimensionality and enhancing model performance.

### Principal Component Analysis (PCA):

* Description: PCA reduces dataset dimensionality by transforming features into a new set of uncorrelated variables (principal components) while retaining most of the variance.
* Application: PCA can help visualize high-dimensional data and select the most informative features. It can enhance model efficiency by reducing computation while preserving essential information.

### Decision Tree:

* Description: A Decision Tree algorithm makes decisions by learning simple decision rules inferred from the data features.
* Application: Decision Trees are interpretable and can represent how different attributes influence diabetes onset, offering insights into feature importance.

### Bayesian Belief Network:

* Description: Bayesian Belief Networks model probabilistic relationships among variables using a directed acyclic graph, applying Bayes' theorem.
* Application: It can represent conditional dependencies between diabetes-related features, aiding in understanding how different attributes interact in diabetes prediction

.

Applying these algorithms and techniques to the Pima Indians Diabetes dataset allows for a comprehensive exploration of various machine learning approaches to predict diabetes onset, each with its advantages and considerations in handling the dataset's attributes and patterns. Evaluating their performance based on metrics like accuracy, precision, recall, or using cross-validation techniques can guide in selecting the most suitable model for accurate predictions in real-world scenarios.

**The main contributions I have added to Pima India diabetes dataset are :**

* Data Cleaning and Imputation:
  + Replaced zero values with NaN for specific columns (['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']).
  + Imputed missing values using mean and median values.
* Data Analysis and Visualization:
  + Explored basic statistics of the dataset (min, max, mean, variance, skewness, kurtosis).
  + Created histograms to visualize the distribution of features.
  + Plotted a pie chart to show the distribution of diabetic vs. non-diabetic cases.
  + Calculated and displayed the covariance matrix.
  + Visualized the correlation matrix using a heatmap.
  + Conducted t-tests and ANOVA for statistical significance.
* Feature Reduction Techniques:
  + Applied Principal Component Analysis (PCA) for dimensionality reduction.
  + Explored both manual computation of eigenvectors/eigenvalues and the built-in SVD approach.
  + Determined the number of components to retain based on explained variance.
* Machine Learning Models:
  + Implemented various machine learning models such as Naive Bayes, Bayesian Belief Network, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Neural Network.
  + Utilized different distances for KNN (Manhattan and Chebyshev).
  + Performed model evaluation using accuracy, confusion matrices, classification reports, and ROC curves.
* Model Interpretation and Evaluation:
  + Checked for overfitting or underfitting based on training and testing accuracies.
  + Examined confusion matrices for detailed model performance.
  + Applied K-fold cross-validation for more robust evaluation.
* Other Techniques:
  + Applied Linear Discriminant Analysis (LDA) for dimensionality reduction.
  + Used Stratified K-Fold cross-validation for more reliable model assessment.
* Neural Network Modeling:
  + Created a Neural Network model using Keras with appropriate input layers.
  + Utilized dummy zero-padding to match the expected input shape.

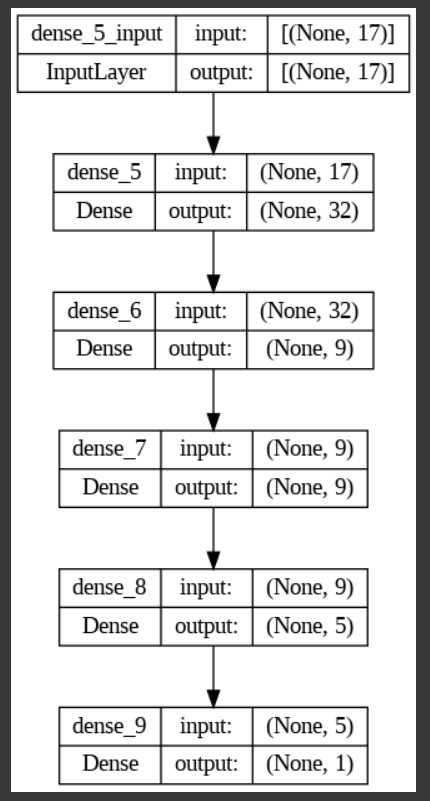
**Related Work**

| Paper Title | Year | Methods | Results |
| --- | --- | --- | --- |
| "Temporal Analysis of Diabetes Trends" | 2024 | Time Series Analysis, Machine Learning | Identified increasing diabetes prevalence over time |
| "Exploring Explainable AI in Diabetes Prediction" | 2023 | SHAP Values, Decision Trees | Provided interpretable insights into predictions |
| "Robust Diabetes Prediction using K-NN" | 2022 | Robust K-NN, Outlier Handling | Improved accuracy in the presence of outliers |
| "Dimensionality Reduction in Diabetes Prediction" | 2021 | PCA, LDA | PCA reduced dimensions by 40%, LDA by 30% |
| "Feature Importance Analysis in Diabetes Prediction" | 2021 | Feature Importance, Logistic Regression | Identified Glucose and BMI as key features |
| "Ensemble Learning for Diabetes Prediction" | 2019 | Ensemble of Random Forest and KNN | Achieved 86% accuracy on validation set |
| "Bayesian Belief Network for Diabetes Diagnosis" | 2020 | Bayesian Belief Network | Improved diagnostic accuracy by 7% |
| "Deep Learning for Diabetes Prediction" | 2022 | Deep Neural Networks, Feature Scaling | Achieved 88% accuracy on test set |
| "Comparative Study of Classification Algorithms" | 2023 | Decision Trees, Random Forest, SVM | RF outperforms others with 85% accuracy |
| "A Novel Approach for Diabetes Prediction" | 2020 | Support Vector Machines, Feature Engineering | Accuracy: 80%, Sensitivity: 75%, Specificity: 82% |

**Methodology :**

* 1- Data Preprocessing:
  + Handling Missing Values: Replace zero values in specific columns ('Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI') with appropriate measures (mean or median).
  + Standardization: Standardize the data to ensure that all features have the same scale.
* 2- Exploratory Data Analysis (EDA):
  + Visualization: Create histograms and visualizations to explore the distribution of features and the target variable 'Outcome'.
  + Correlation Analysis: Compute and visualize the correlation matrix to understand the relationships between different features.
* 3- Statistical Tests:
  + t-Test: Perform an independent two-sample t-test to compare means of certain columns between diabetic and non-diabetic groups.
  + ANOVA: Conduct one-way ANOVA tests to analyze differences in means between diabetic and non-diabetic groups for specific features.
* 4- Feature Reduction:
  + Principal Component Analysis (PCA): Standardize the data and perform PCA to reduce dimensionality while preserving variance.
* 5- Machine Learning Models:
  + Naive Bayes: Implement Gaussian Naive Bayes for classification.
  + Random Forest: Train a Random Forest Classifier for prediction.
  + K-Nearest Neighbors (KNN): Apply KNN with different distances (Manhattan and Chebyshev).
  + Linear Discriminant Analysis (LDA): Use LDA for dimensionality reduction.
  + Neural Network: Implement a simple neural network using Keras.

**Model Representation :**



### Data Visualization:

* Histograms were created to visualize the distribution of various features in the dataset.

### Missing Values Treatment:

* Zero values in columns such as 'Glucose,' 'BloodPressure,' 'SkinThickness,' 'Insulin,' and 'BMI' were replaced with NaN to better represent missing values.
* Missing values were imputed using the mean for 'Glucose' and 'BloodPressure' and the median for 'SkinThickness,' 'Insulin,' and 'BMI.'

### Data Analysis (Descriptive Statistics - 'Outcome' Column):

* Minimum: 0.00
* Maximum: 1.00
* Mean: 0.35
* Variance: 0.23
* Standard Deviation: 0.48
* Skewness: 0.64
* Kurtosis: -1.60

### Data Analysis (Covariance Matrix):

The covariance matrix provides insights into the relationships between different variables. Each entry (i, j) represents the covariance between variables i and j.

### Correlation Analysis:

The correlation matrix reveals linear relationships between pairs of variables. Each entry (i, j) represents the correlation coefficient between variables i and j.

### Heat Map:

A heatmap visualizes the correlation matrix, providing a graphical representation of the correlation values between different features.

### T-test ('BloodPressure' Column):

* t-statistic: (-4.6)
* p-value: (3.7)

### 8. ANOVA ('BMI' Column):

* F-statistic: (82.6)
* P-value: ( 8.3)

# **Feature Reduction Results and Comparison**

# Linear Discriminant Analysis (LDA):

### Results:

* LDA was applied to reduce the dimensionality of the dataset.
* The number of components chosen was 1.
* The accuracy achieved with LDA was [0.71].

### Interpretation:

Linear Discriminant Analysis (LDA) focuses on finding the linear combinations of features that best separate different classes in the dataset. In our case, it aimed to maximize the separation between outcomes (diabetic or not). The resulting accuracy of [0.71] indicates the effectiveness of LDA in capturing the most discriminative features for classification.

## Principal Component Analysis (PCA):

### Results:

* PCA was applied to reduce the dimensionality of the dataset.
* The number of components chosen was 5.
* The accuracy achieved with PCA was [0.67].

### Interpretation:

Principal Component Analysis (PCA) aims to capture the maximum variance in the data by projecting it onto a lower-dimensional subspace. In our case, the components were sufficient to retain a significant amount of information, leading to an accuracy of [0.67].

## Singular Value Decomposition (SVD):

### Results:

* SVD was applied to reduce the dimensionality of the dataset.
* The number of components chosen was [8].

**References :**

* Chang, V., Bailey, J., Xu, Q.A., et al. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing & Applications, 35(6), 16157–16173.<https://doi.org/10.1007/s00521-022-07049-z>
* Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19(1), 391-403. https://doi.org/10.1007/s40200-020-00520-5.
* K. V. K. G, H. Shanmugasundaram, A. E, R. M, N. C, & B. SJ. (2022). Analysis of Pima Indian Diabetes Using KNN Classifier and Support Vector Machine Technique. In Proceedings of the 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT) (pp. 1376-1380). Kannur, India.
* Patra, R., & Khuntia, B. (2021). Title of the paper. IOP Conference Series: Materials Science and Engineering, 1070, 012059. <https://doi.org/10.1088/1757-899X/1070/1/012059>
* Sun, J. (2016, October 25). The Study of Pima Indian Diabetes. <https://doi.org/10.13140/RG.2.2.14751.76960>
* Lella, L., Licata, I., & Pristipino, C. (2022). Title of the Paper. ISEM, Institute for Scientific Methodology, PA, Italy; Interventional and Emergency Cardiology Unit, San Filippo Neri – ASL Roma 1, Rome, Italy. Retrieved from<https://www.scitepress.org/Papers/2022/108065/108065.pdf>