## **Predictive Healthcare System Using Machine Learning on Diabetes Dataset**

### **Abstract**

This paper presents a predictive healthcare system leveraging machine learning techniques to diagnose diabetes using the PIMA Indian Diabetes dataset. The system incorporates data preprocessing, outlier detection, feature scaling, dimensionality reduction using Principal Component Analysis (PCA), and class imbalance handling using Synthetic Minority Over-sampling Technique (SMOTE). Models including Linear Regression and Random Forest Classifier were trained and evaluated. Results indicate that the Random Forest Classifier outperforms with a notable accuracy, highlighting the model’s potential in real-world healthcare diagnostics.

### **1. Introduction**

The early prediction of chronic diseases such as diabetes is crucial in preventive healthcare. With the advent of data-driven techniques, machine learning (ML) models have shown promise in identifying patterns that lead to accurate predictions. This paper explores a predictive system using ML algorithms on the PIMA Indian Diabetes dataset to determine the presence of diabetes based on various medical attributes.

### **2. Methodology**

#### **2.1 Dataset Description**

The PIMA Indian Diabetes dataset consists of medical diagnostic measurements for females aged 21 and above. It includes features such as glucose levels, BMI, insulin, age, and the outcome (diabetes positive or negative).

#### **2.2 Data Preprocessing**

Missing values were handled using the **SimpleImputer** with the median strategy to ensure data consistency. A correlation matrix and box plots were used to visualize feature relationships and outliers, respectively.

#### **2.3 Feature Scaling and Dimensionality Reduction**

Data was scaled using **StandardScaler**, which standardizes features by removing the mean and scaling to unit variance. **PCA** was applied to reduce dimensionality to two principal components for visualization.

#### **2.4 Handling Imbalanced Data**

The target variable was imbalanced. To address this, **SMOTE** was used to synthesize minority class instances, ensuring balanced data for model training.

#### **2.5 Model Training**

Two models were implemented:

* **Linear Regression**: A baseline model for comparison.
* **Random Forest Classifier**: An ensemble method known for robustness and performance.

The dataset was split in an 80:20 ratio for training and testing.

### **3. Results and Discussion**

#### **3.1 Linear Regression**

The Linear Regression model was trained and evaluated using Mean Squared Error (MSE). The model produced an MSE of **0.151**, indicating poor performance for classification tasks.

#### **3.2 Random Forest Classifier**

The Random Forest Classifier achieved an **accuracy of 91.5%**, with a precision and recall of over 0.90 for both classes, according to the classification report. This showcases its effectiveness in binary classification problems, especially after balancing the data with SMOTE.

#### **3.3 Visualization and Insights**

* **PCA Plot**: Clearly depicted separable clusters between diabetic and non-diabetic instances.
* **Correlation Matrix**: Highlighted strong positive correlation between glucose and diabetes.
* **Outcome Distribution**: Post-SMOTE, class distribution was balanced, which improved model performance significantly.

### **4. Conclusion and Future Work**

This study demonstrates the feasibility of applying machine learning techniques, particularly Random Forest, to effectively predict diabetes. The preprocessing steps, including outlier removal, SMOTE, and PCA, significantly contributed to the model's accuracy. Future work could explore:

* Hyperparameter tuning for further optimization.
* Incorporation of deep learning models.
* Deployment of the model in a real-time web or mobile application for clinical use.

**References**

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825-2830.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
3. Weka Machine Learning Project. (n.d.). *UCI Machine Learning Repository: Pima Indians Diabetes Data Set*. Retrieved from<https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>