**AI PROJECT REPORT**

Diabetes Detection System

**Student’s Name:**

Muhammad Sohaib

**Roll NO:**

Su92-BSAIM-F23-067

**CLASS & SECTION:**

BSAI-3B

**TEACHER’s NAME:**

Rasik Ali

### Project Report: Diabetes Detection System

### Abstract

The **Diabetes Detection System** is a machine learning-based application designed to predict the likelihood of diabetes in individuals using medical data. It leverages the Pima Indians Diabetes Dataset to train and evaluate predictive models. The system is implemented in Python and supports two interfaces: a command-line tool for interactive use and a web application for broader accessibility. Key steps include data preprocessing, model training using ensemble methods (Random Forest and XG-Boost), and data balancing with SMOTE to address class imbalance. The system achieves approximately 80% accuracy and provides insights into the most significant predictors of diabetes. This project showcases the potential of machine learning in healthcare by offering a practical tool for diabetes risk assessment and emphasizes scalability for future enhancements.

1. **Introduction:**

The Diabetes Detection System aims to identify whether an individual is diabetic based on specific health parameters. It leverages machine learning techniques to provide predictions using real-world medical datasets. This system is implemented in Python and can function as a command-line tool and a web application.

1. **Objectives:**

* To predict diabetes using a supervised learning model.
* To preprocess raw medical data for reliable predictions.
* To enable feature analysis for insights into impactful predictors of diabetes.
* To deploy the system via a web interface for ease of access.

1. **Methodology:**

Methodology has some stages which are written below:

**Dataset Loading:**

The system uses the Pima Indians Diabetes Dataset, which includes:

* Features: **Pregnancies**, **Glucose**, **Blood Pressure**, **Skin Thickness**, **Insulin**, **BMI**, **Diabetes Pedigree Function**, and **Age**.
* Target Variable: **Outcome** (0 for non-diabetic, 1 for diabetic).

**Data Exploration:**

statistics, correlations, and visualizations (heatmaps) are utilized to understand the dataset's characteristics and relationships.

**Data Pre-processing:**

* Missing values for features like Glucose and BMI are replaced with mean values.
* Data is scaled using **Standard Scaler** for uniformity in feature magnitudes.

**Data Splitting:**

The dataset is split into training and testing sets (80:20 split).

**Model Training**

* A **Random Forest Classifier** is employed for its robustness and interpretability.
* A **Voting Classifier** combining Random Forest and XG-Boost enhances prediction accuracy in the web application.

**Model Evaluation**

Metrics like accuracy, classification reports, and confusion matrices are used to evaluate model performance. Visualizations of feature importance provide insight into the impact of each feature.

**Deployment**

* Command-line application for real-time predictions.
* Web application using Flask for user-friendly predictions via HTTP requests.

1. **Implementation**

**Key Libraries**

* **Data Handling**: Pandas, NumPy
* **Machine Learning**: Scikit-learn, XGBoost, imbalanced-learn (SMOTE)
* **Web Framework**: Flask
* **Visualization**: Matplotlib, Seaborn

**Code Structure**

1. **Command-line Application (diabetes\_detection.py)**
   * Loads, preprocesses, and trains the model.
   * Allows interactive predictions for user-provided inputs.
2. **Web Application (main.py)**
   * Offers endpoints for training (/train) and prediction (/predict).
   * Features error handling and real-time feedback.
3. **Results**

* Achieved accuracy: **~80%** on test datasets.
* Feature importance highlights:
  + **Glucose**, **BMI**, and **Age** are the most significant predictors.

1. **Features**

* **Real-time Predictions**: Accepts user input to predict diabetes status.
* **Feature Analysis**: Displays importance of each predictor for transparency.
* **Web Accessibility**: Provides REST API endpoints for predictions.
* **Balanced Dataset**: Employs SMOTE for handling imbalanced classes.

1. **Applications**

* Healthcare organizations for patient screening.
* Educational purposes in machine learning and medical informatics.

1. **Future Work**

* Integrate more advanced models like deep learning networks.
* Expand dataset with more diverse population samples.
* Implement mobile and desktop interfaces.

**Code Snippets**

**Loading and Preprocessing Data**

def preprocess\_data(data):

columns\_to\_replace = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']

data[columns\_to\_replace] = data[columns\_to\_replace].replace(0, np.nan)

data.fillna(data.mean(), inplace=True)

scaler = StandardScaler()

features = data.iloc[:, :-1]

labels = data['Outcome']

features\_scaled = scaler.fit\_transform(features)

return features\_scaled, labels, scaler

**Training Ensemble Classifier**

def train\_ensemble(X\_train, y\_train):

rf = RandomForestClassifier(random\_state=42)

xgb = XGBClassifier(use\_label\_encoder=False, eval\_metric='logloss', random\_state=42)

ensemble = VotingClassifier(estimators=[('rf', rf), ('xgb', xgb)], voting='soft')

ensemble.fit(X\_train, y\_train)

return ensemble

**Conclusion**

The **Diabetes Detection System** is a comprehensive machine learning application that effectively predicts diabetes using real-world medical data. By implementing robust preprocessing, feature scaling, and advanced ensemble models, the system achieves an impressive accuracy of approximately 80%. The inclusion of a saved model (classifier.pkl) ensures that the trained classifier can be reused without the need for retraining, significantly improving efficiency and deployment readiness.

The integration of SMOTE for handling data imbalance, detailed feature importance analysis, and user-friendly interfaces (command-line and web application) makes this system a valuable tool for both healthcare professionals and educators in machine learning. With its modular structure and persistent storage of the model, the project is primed for scalability, including further enhancements like incorporating deep learning models, expanding datasets, and deploying on diverse platforms such as mobile and desktop applications.

The system highlights the transformative potential of AI in healthcare, offering a reliable, accessible, and expandable solution for early diabetes detection and risk assessment.

**Demo Video of Project:**

https://drive.google.com/file/d/19lSoVnJsVUhWLg2xhYCJScfwr5XX3u--/view?usp=sharing