Diabetes Prediction Using Machine Learning

INTERNSHIP PROGRAM:

Infosys Springboard AIML Internship

NAME: K.V.S.MRUDULA

Table of Contents

INTRODUCTION:

- This project aims to build a machine learning model to predict whether an individual is diabetic, pre-diabetic, or healthy.
- Diabetes is a major chronic disease affecting millions worldwide, posing significant challenges to individuals and healthcare systems.
- Early detection and intervention are essential for effective diabetes management and prevention of complications.
- The focus is on analyzing healthcare statistics and lifestyle factors to create a model that can predict diabetes risk accurately.
- By identifying at-risk individuals sooner, healthcare professionals can implement timely lifestyle changes and preventive care.
- The outcome will be an advanced tool that supports proactive management, improving patient outcomes and reducing healthcare burdens.

Project Focus and Objectives:

Project Goals:

- Understand the interplay between lifestyle, healthcare statistics, and diabetes risk.
- Develop a reliable classification model using advanced machine learning techniques.
- Provide actionable insights through feature analysis and evaluation metrics.

Objectives:

- Identify significant predictors of diabetes risk.
- Create a robust and scalable model that minimizes errors and enhances interpretability.
- Provide clear performance metrics to validate the model's effectiveness.

The project is motivated by the need for a predictive tool that can aid in the early identification of diabetes risk factors. This tool aims to enhance current diagnostic methods by incorporating machine learning algorithms that process healthcare data to detect patterns linked to diabetes. The model development involves exploring different algorithms and evaluating their performance based on metrics like accuracy, precision, recall, and AUC-ROC. The insights gained will guide healthcare professionals in identifying key risk factors and developing preventive measures.

DATASET DESCRIPTION:

The dataset diabetesInfosys.csv is designed to aid the development of a machine learning model for predicting diabetes risk. It includes various features that provide a comprehensive view of individual health and lifestyle, contributing to accurate risk assessment.

Key Features:

- **Demographics**: Age, gender.
- **Clinical Indicators**: Polyuria (excessive urination), polydipsia (increased thirst), sudden weight loss, weakness, polyphagia (excessive hunger), genital thrush, visual blurring, delayed healing, partial paresis (partial muscle weakness), muscle stiffness, alopecia (hair loss), obesity.
- **Lifestyle Factors**: Itching, irritability.

Dataset Characteristics:

- **Diverse Representation**: The dataset includes data from various age groups and demographics, ensuring its applicability to different populations.
- **Self-Sufficiency**: All data points are evaluated based on simple, human-readable checks, requiring no additional expertise to determine diabetes risk.

The dataset, derived from healthcare surveys and records, provides insights into how demographic and clinical indicators relate to diabetes risk. It includes symptoms like polyuria, sudden weight loss, and muscle stiffness to inform risk levels. Lifestyle factors such as irritability and itching further enhance the individual's risk profile for better prediction accuracy. Data augmentation techniques, including normalization and addressing class imbalances, bolster the dataset's robustness. This ensures the model can generalize well to real-world scenarios by exposing it to varied patterns and data variability.

METHODOLOGY:

❖ Data Preprocessing:

- **Data Cleaning**: Handled missing data by filling missing values using mean or median imputation, depending on feature type. Columns with significant missing values were removed to maintain dataset quality and integrity.

- **Data Augmentation**: Applied techniques like normalization to standardize numeric values and scaling to bring features to a uniform range, ensuring no single feature disproportionately influenced model training.
- **Feature Extraction**: Used feature extraction techniques, such as identifying key health indicators (e.g., polyuria, sudden weight loss, muscle stiffness), to enhance the model's understanding of significant patterns in the data.

Data preprocessing was essential for accurate model training. Handling missing values and scaling features allowed the models to process data effectively. Key features were extracted to ensure the model focused on the most relevant information for diabetes prediction.

❖ Exploratory Data Analysis (EDA):

- **Correlation Analysis**:
  - ***Strong Positive Correlations***: Features such as polyuria and sudden weight loss were highly correlated with diabetes, emphasizing their importance in the model.
  - ***Weaker Correlations***: Features like muscle stiffness showed weaker relationships with diabetes, indicating less significance for predictive accuracy.
- **Visualization**:
  - ***Heatmaps*:** Used to identify strong correlations between features and the target variable.
  - ***Boxplots***: Displayed feature distributions across diabetic and non-diabetic groups, showcasing clear patterns and differences.

EDA provided key insights into the relationships between features and diabetes status, aiding in the feature selection process and model input decisions. Visualization tools like heatmaps and boxplots helped highlight the most relevant and significant features

❖ Feature Selection & Model Development:

- **Feature Selection Techniques**:
  - **Chi-Square (Chi2) Test**: Used to evaluate the importance of categorical features and select those with strong associations to diabetes.

- ***Random Forest (RF)*:** Analyzed feature importance and helped identify key predictors by evaluating feature interactions.

- **Dimensionality Reduction**:
  - Although *PCA (Principal Component Analysis)* was considered, it was deemed unnecessary for this dataset due to its size (16 features). Removing it ensured feature interpretability and retained essential information.
- **Model Development**:
  - **Models Evaluated**: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Classifier (SVC), Extra Trees, and Decision Tree.
  - **Feature Scaling**: Applied to numerical features like age and BMI to ensure consistent treatment by the models.

Feature selection and model development were critical to building an effective model. Chi-Square and Random Forest techniques identified key predictors and interactions, while PCA was omitted to maintain interpretability. The models were evaluated based on established performance metrics to determine the best option for diabetes risk prediction.

❖ Model Evaluation:

- **Metrics Used**: Accuracy, Precision, Recall, F1-Score, and AUC-ROC.
- **Primary Metric**: AUC-ROC was chosen for its comprehensive ability to evaluate a model's capability to distinguish between diabetic and non-diabetic cases across different thresholds.
- **Best Model**: *Extra Trees Classifier* was the most effective, achieving an AUC-ROC of 0.9839, indicating superior performance in differentiating between the classes. Precision and Recall values (0.9219 and 0.9216, respectively) were balanced, ideal for healthcare application reliability.

The evaluation phase verified the chosen model's effectiveness, with the Extra Trees Classifier demonstrating excellent discrimination and balanced precision-recall, making it suitable for practical diabetes prediction applications.

❖ Hyperparameter Tuning:

- **GridSearchCV**: Used to fine-tune hyperparameters, ensuring the model reached optimal performance and predictive accuracy.
- **Outcome**: The Extra Trees Classifier was confirmed as the optimal model post-tuning.

Hyperparameter tuning with GridSearchCV enhanced model performance, refining it to achieve the best accuracy and adaptability for future data, ensuring robustness for diabetes risk prediction.

Technologies Used

- **Programming Language**: Python.
- **Libraries**: pandas for data manipulation, numpy for numerical operations, matplotlib and seaborn for visualizations, scikit-learn for machine learning algorithms, and XGBoost for gradient boosting models.
- **Visualization Tools**: Plotly Express for interactive visualizations.
- **Environment**: Jupyter Notebook, hosted on GitHub for version control and collaborative work.

Python was chosen for its extensive ecosystem of libraries that are well-suited for data analysis and machine learning. Libraries like pandas and numpy provided powerful data manipulation and numerical processing capabilities. Visualization tools such as matplotlib, seaborn, and Plotly Express were used to create insightful charts and graphs for EDA. Jupyter Notebook facilitated an interactive development environment, allowing for step-by-step code execution and real-time visualization. The code was version-controlled and shared through GitHub to ensure reproducibility and collaboration.

System Workflow

1. **Data Collection**: Gather healthcare and lifestyle statistics for individuals.
2. **Preprocessing**: Clean data, handle missing values, balance classes, and encode categorical variables.
3. **Feature Selection**: Identify and retain the most relevant features for diabetes prediction.
4. **Model Training**: Train machine learning models using the prepared dataset.
5. **Prediction**: Input patient data into the model to classify as diabetic, pre-diabetic, or healthy.
6. **Output**: Display results and provide actionable insights for early intervention.

Result:

- **Model Performance**:

  The **Extra Trees Classifier** emerged as the best-performing model, achieving an AUC-ROC of 0.9839, accuracy of 92.16%, and an F1-Score of 0.9200. This model outperformed other models such as Random Forest and Support Vector Classifier (SVC), which also performed well but exhibited slightly lower AUC-ROC scores and less balanced precision and recall metrics.

CHALLENGES AND SOLUTIONS

1. Feature Noise:

- **Challenge**: The dataset contained features that were non-contributory or less relevant, potentially reducing the model's efficiency.
- **Solution**: Implemented rigorous feature selection using Chi-Square and Random Forest techniques to identify and retain the most impactful feature

2. PCA Consideration:

- **Challenge**: Determining if dimensionality reduction was needed for the dataset.
- **Solution**: PCA was considered but ultimately not applied, as the dataset's 16 features were sufficient for training, and applying PCA could have led to a loss of critical feature information.

Feature noise was addressed through careful feature selection, which streamlined the dataset and focused on the most relevant predictors. This process not only improved the model's efficiency but also its interpretability. PCA was evaluated for potential use but was not necessary due to the manageable number of features. Applying PCA could have compromised the interpretability of the dataset without significantly enhancing model performance.

REFERENCES:

- diabetes.csv dataset for healthcare statistics.
- Documentation for Python libraries such as pandas, scikit-learn, XGBoost.

CONCLUSION:

This project successfully developed a robust machine learning model for diabetes prediction, leveraging advanced techniques in data preprocessing, feature selection, and algorithm optimization. The Extra Trees Classifier emerged as the most effective model, achieving high AUC-ROC and balanced precision-recall metrics, making it suitable for practical healthcare applications.

The insights gained from this project highlight the significant relationship between symptoms such as Polyuria, Polydipsia, and sudden weight loss with diabetes risk. The methodologies employed ensured high accuracy and generalizability, while addressing challenges such as feature noise and imbalanced datasets.