Diabetes Prediction Using Machine Learning

Internship Program: *Infosys Springboard AIML Internship*

"Empowering Early Detection and Management with ML"

Presented by: K.V.S.MRUDULA

Overview: The Problem and Solution

The Challenge:

- Diabetes is a chronic disease affecting millions globally, posing significant challenges for individuals and healthcare systems.
- Early detection is crucial for minimizing complications and reducing the overall burden on healthcare resources.

Our Solution:

- We have developed a machine learning model to predict diabetes risk using healthcare and lifestyle data.
- This allows healthcare professionals to proactively identify at-risk individuals and implement early interventions.

Project Objectives

Goals

- Analyze healthcare and lifestyle data to identify diabetes risk factors.
- Build a reliable and scalable machine learning model.
- Provide actionable insights through feature analysis and model metrics.

Deliverables

- A high-performing classification model for diabetes prediction.
- Detailed performance validation using metrics like AUC-ROC, accuracy, precision, and recall.

Dataset Overview

Features

- Demographics: age and gender.
- Clinical indicators: Polyuria, and Polydipsia, Sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, delayed healing, partial paresis , muscle stiffness, alopecia , obesity.
- Lifestyle factors: Itching, irritability.

Characteristics

- Captures diverse age groups, genders, and socio-economic profiles.
- Ensures inclusivity for identifying diabetes risk patterns across populations.
- Enhances real-world relevance for varied demographics and geographic regions.

Preprocessing Steps

- Imputed missing values.
- Removed outliers using statistical methods.
- Applied one-hot encoding for categorical variables.

Methodology: From Data to Insights

1 2 3 4

Data Preprocessing

We imputed missing values using mean/median techniques, scaled numerical features using Min-Max scaling, and encoded categorical variables with label/one-hot encoding. We also removed outliers using statistical methods.

Exploratory Data Analysis (EDA)

We performed correlation analysis, identifying strong predictors like Polyuria and Polydipsia. We visualized feature distributions using heatmaps and boxplots to identify patterns.

Feature Selection

We used Chi-Square tests for feature significance and Random Forest for feature importance to select the most impactful features.

Model Building

We evaluated various models, including Logistic Regression, Random Forest, Gradient Boosting, Extra Trees Classifier, and SVC. Feature interpretability was retained by avoiding PCA.

Model Evaluation: Identifying the Best Fit

92.16%

Accuracy

92.19%

Precision

92.16%

Recall

0.920

F1-Score

0.984

AUC-ROC

- We evaluated our model using various metrics including accuracy, precision, recall, F1-Score, and AUC-ROC.
- The Extra Trees Classifier achieved superior performance in distinguishing between diabetic and non-diabetic cases, achieving an AUC-ROC of 0.9839, accuracy of 92.16%, and balanced precision-recall, making it ideal for healthcare applications.

Challenges and Solutions

Feature Noise

Challenge: Non-contributory features reduced model efficiency.

Solution: Applied rigorous selection using Chi-Square and Random Forest techniques to retain impactful features.

Dimensionality Reduction

Challenge: Evaluating PCA's necessity for a manageable 16-feature dataset.

Solution: Omitted PCA to retain feature interpretability without compromising model performance.

Conclusion:

Summary

Developed a robust ML model for diabetes prediction with high accuracy and interpretability. Provides an early detection tool to assist healthcare professionals.

Impact

Highlights the significance of features like Polyuria, Polydipsia, and sudden weight loss in diabetes risk prediction.

Thank You