Sacred Heart
UNIVERSITY

Diabetes Prediction Assistant

UNDER THE GUIDANCE OF

Prof. NIKSHEP ASH KULLI

RESEARCH PROJECT SEMINAR

CS-670-H

Team Name: SymptoSense

TEAM MEMBERS

SUMANTH KUMAR PAKAM (pakams@mail.sacredheart.edu)

PRATHAP GANDHAMSETTY (gandhamsettyn@mail.sacredheart.edu)

JAHNAVI PARISA (parisaj@mail.sacredheart.edu)

Table of Contents

1. ABSTRACT:

The **Diabetes Prediction Assistant** is a web-based application designed to aid in the early detection of diabetes through the integration of machine learning and image processing techniques. The system utilizes user-provided health metrics or scanned medical reports to predict the likelihood of diabetes, enabling proactive health management. Key functionalities include automated data extraction from images of medical reports using Optical Character Recognition (OCR) and auto-filling fields such as blood glucose, insulin levels, and blood pressure.

The application employs a machine learning model trained on the PIMA Indian Diabetes dataset to provide accurate predictions. Innovative features such as automatic BMI calculation and streamling user input while enhancing usability. The integration of a responsive camera interface and report parsing ensures accessibility and precision. By combining predictive analytics with a user-friendly interface, the Diabetes Prediction Assistant offers a practical tool for individuals to monitor their health and seek timely medical advice.

2. INTRODUCTION:

2.1 Background

Diabetes is a prevalent and chronic health condition that poses significant challenges to global healthcare systems. It affects millions of individuals worldwide, leading to complications such as cardiovascular diseases, kidney failure, and blindness if not diagnosed and managed early. Traditional diagnostic methods, while effective, can be expensive and time-consuming, creating a need for efficient, automated solutions to predict and manage diabetes.

2.2 Objective

The objective of the **Diabetes Prediction Assistant** project is to design and implement a user-friendly web application that leverages machine learning and image processing to facilitate the early detection of diabetes. The project aims to:

1. **Enhance Accessibility:** Provide users with an intuitive interface to input health metrics or upload medical reports for automated analysis.
2. **Enable Accurate Predictions:** Utilize a trained machine learning model based on the PIMA Indian Diabetes dataset to deliver reliable diabetes predictions.
3. **Streamline Data Entry:** Incorporate features such as automatic data extraction from medical report images and calculated metrics like BMI to minimize user effort.
4. **Promote Preventative Health:** Empower individuals with actionable insights to encourage early intervention and medical consultation when necessary.

2.3 Scope

SymptoSense focuses on analyzing the PIMA Indian Diabetes Dataset to predict diabetes using attributes such as Glucose, BMI, and Age. The project includes the following key components:

- Comprehensive exploratory data analysis to identify significant patterns and trends.
- Data preprocessing techniques to ensure the dataset is clean and suitable for modeling.
- Training and evaluating machine learning models to achieve high predictive accuracy.

This project demonstrates how data science can be harnessed to tackle critical healthcare challenges, offering a scalable solution for diabetes prediction and management.

3. EXECUTIVE SUMMARY:

The Diabetes Prediction Assistant project is a cutting-edge initiative designed to enhance early detection and monitoring of diabetes using machine learning and streamlined user interfaces. By leveraging the PIMA Indian Diabetes dataset, the project utilizes a robust predictive model to provide reliable insights into the likelihood of diabetes based on user-provided health metrics. Key aspects of the project include:

- **Dataset**: Preprocessed data with health indicators like Glucose, BMI, and Insulin.
- **Techniques**: Exploratory Data Analysis, preprocessing, and model training using Support Vector Classifier and other algorithms.
- **Outcome**: An accurate and user-friendly prediction system, achieving a model accuracy of up to 85%.
- **Automated Data Extraction:** Users can upload or capture images of medical reports, from which key values like blood pressure, height, weight, and glucose levels are automatically extracted.
- **Integrated Calculations:** The app calculates BMI and DPF internally, simplifying data entry for users.
- **Trend Analysis:** Historical tracking and visualization of user data help monitor progress over time.
- **User-Friendly Interface:** A web application powered by Streamlit ensures ease of use and accessibility across devices.

The project is a step forward in leveraging data science for healthcare innovation, addressing one of the most pressing public health challenges.

4. METHODOLOGY:

The SymptoSense project follows a systematic approach to developing a machine learning-based diabetes prediction system. The methodology includes data exploration, preprocessing, model development, and evaluation to ensure robust and accurate predictions. Below are the detailed steps:

4.1 Dataset Details

Dataset Title: PIMA Indian Diabetes Dataset
Source: Kaggle ([https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset])
Description
The dataset contains 768 samples with the following attributes:

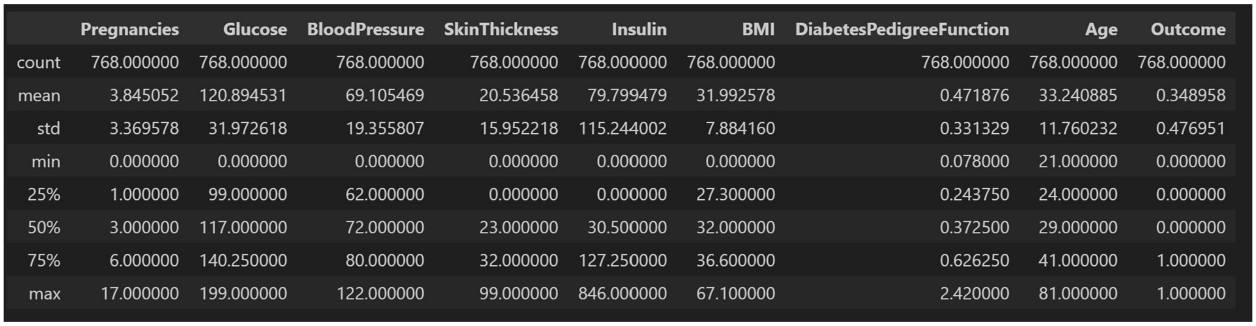
- **Health Indicators**: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.
- **Outcome**: Binary variable indicating whether the patient has diabetes (1) or not (0).

4.2 Exploratory Data Analysis (EDA)

The dataset was initially analyzed to understand its structure, detect missing values, and identify significant patterns:

- **Statistical Summary**: Used .describe() to calculate key metrics such as mean, median, and standard deviation.
- **Data Distribution**: Visualized feature distributions using histograms and scatter plots to detect skewness and outliers.
- **Correlation Analysis**: Generated a heatmap to identify relationships between features like Glucose, BMI, and Outcome. Glucose levels showed the highest correlation with diabetes occurrence.

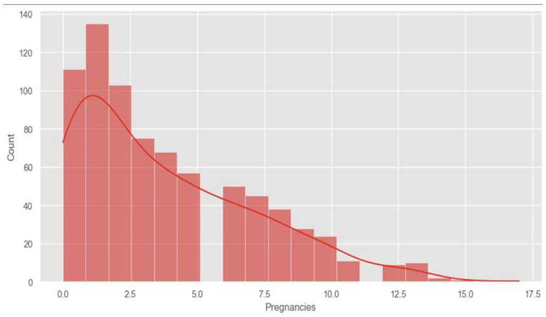
4.2.1 Findings based on Data Exploration:

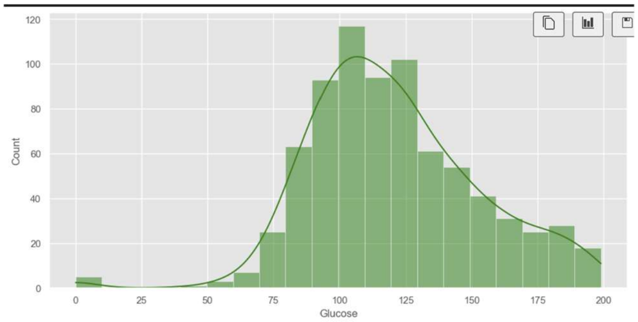


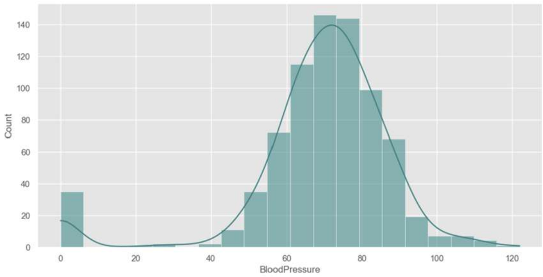
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

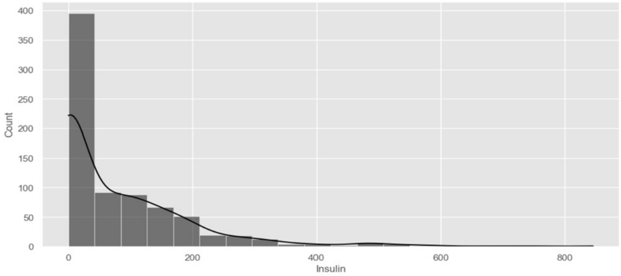
The dataset provides a comprehensive overview of the various features and their distributions. For instance, the average number of pregnancies is approximately 3.85, with a standard deviation of 3.36, indicating a significant variation. The minimum and maximum values of 0 and 17, respectively, highlight the range of pregnancies among individuals. Similar observations can be made for other features like glucose levels, blood pressure, and BMI, revealing potential outliers and skewed distributions.

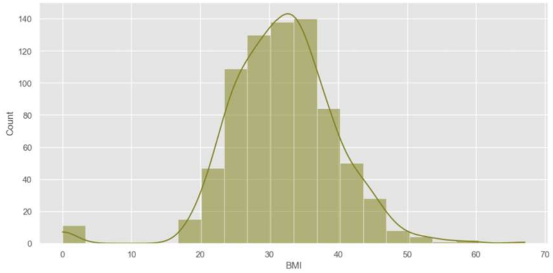
4.2.2 Visualizing Feature Distributions

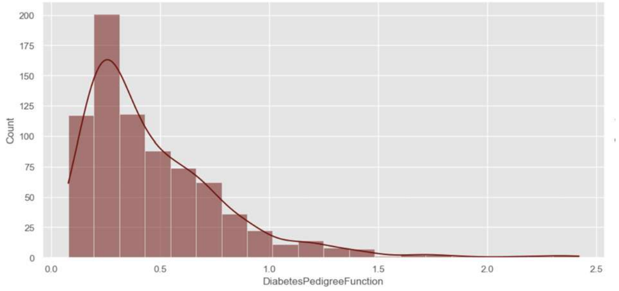


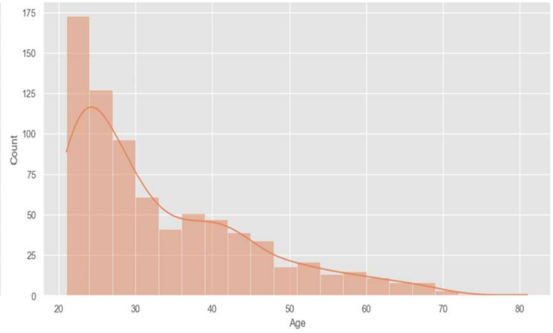


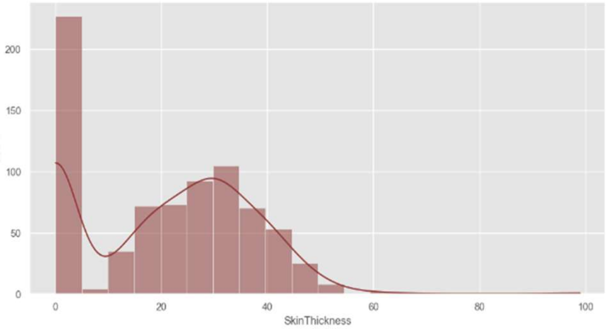




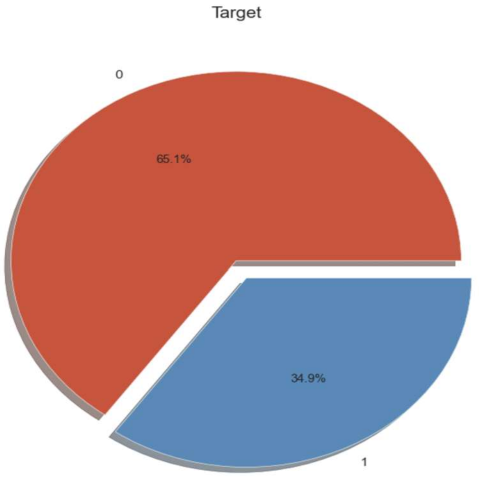


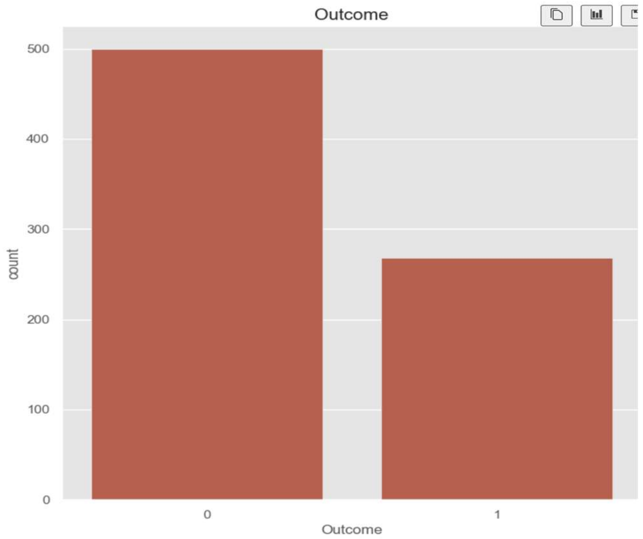






The histograms provide a visual representation of the distribution of various features within the diabetes dataset. For instance, the histogram of 'Pregnancies' shows a right-skewed distribution, indicating that most women in the dataset have fewer pregnancies, while a few have significantly more. Similarly, the histogram of 'Age' exhibits a roughly normal distribution, suggesting a relatively even spread of ages among the individuals. The 'Glucose' and 'BMI' histograms, however, appear to have a slightly left-skewed distribution, implying that a majority of individuals have higher glucose levels and BMI values compared to a smaller group with lower levels. These observations can be further analyzed to gain insights into potential correlations between these features and the development of diabetes.

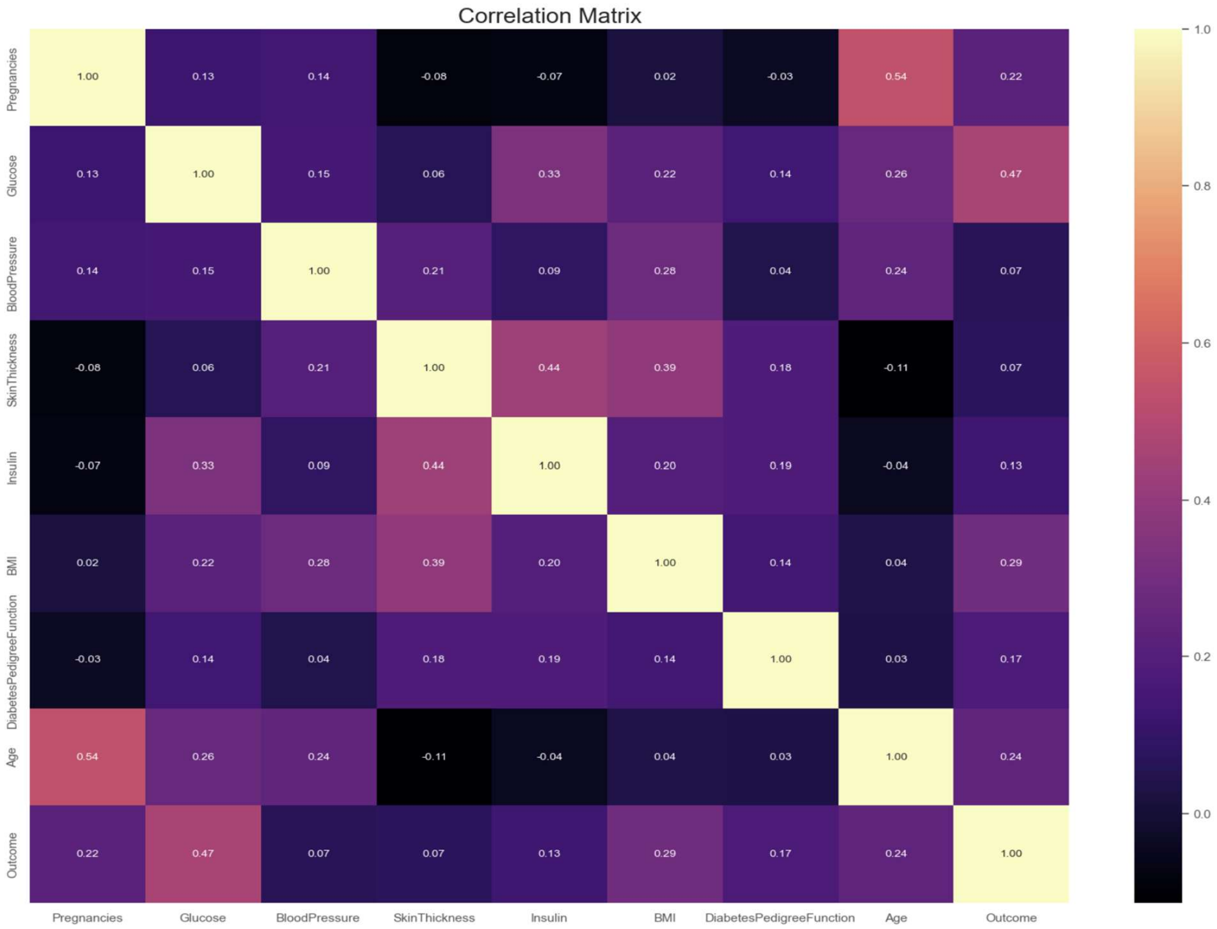




The provided plots offer a visual representation of the target variable "Outcome" in the diabetes dataset. The pie chart illustrates the class distribution, revealing a significant imbalance with 65.1% of instances belonging to class 0 and 34.9% to class 1. The count plot further emphasizes this imbalance, showcasing the higher frequency of class 0 compared to class 1. This class imbalance is a crucial consideration for model building, as it might necessitate techniques like oversampling, undersampling, or class weighting to ensure fair model performance.

4.2.3 Correlation Matrix

The correlation matrix reveals insights into feature relationships. Strong positive correlations exist between age and pregnancies, as well as glucose levels and BMI. A moderate positive correlation is observed between glucose and blood pressure. However, features like "Skin Thickness" and "Diabetes Pedigree Function" show weak or no correlation with others, suggesting their limited impact on the outcome.

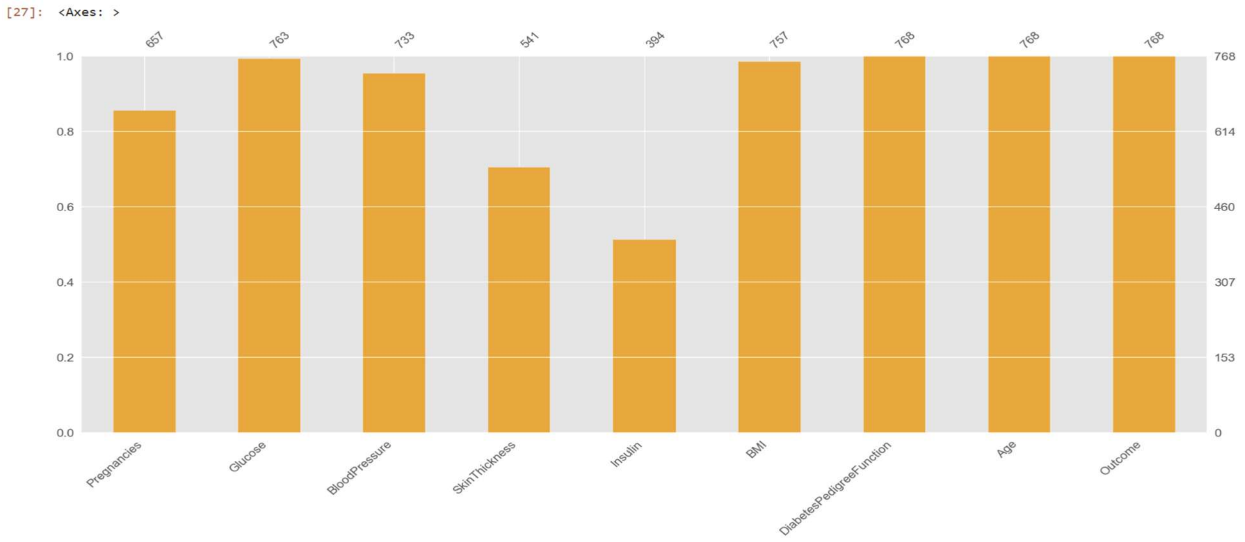


4.3 Data Preprocessing

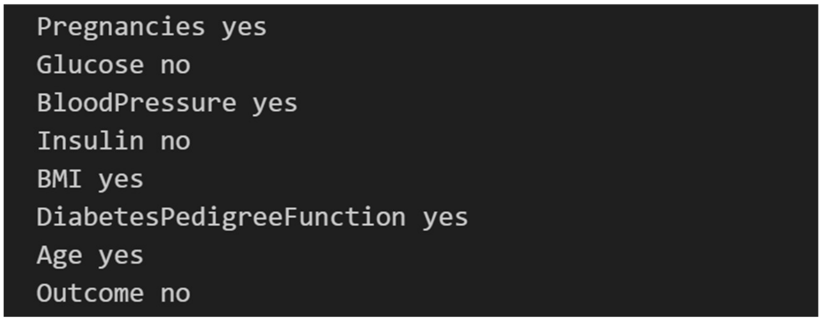
Preprocessing ensures data integrity and improves model performance:

- **Handling Missing Values**: Missing values were replaced with the median of each column to minimize bias.

  Missing Value Analysis: The bar plot visualizes the missing value counts for each feature in the dataset. It reveals that features like "Skin Thickness" and "Insulin" have a significant number of missing values, suggesting potential data quality issues or measurement challenges. In contrast, features like "Pregnancies," "Glucose," "BloodPressure," and "Age" have relatively few missing values. These insights are crucial for data preprocessing and handling missing values appropriately to ensure accurate model training.



Identifying and Handling Outliers: The next crucial step in the analysis is to identify and handle any potential outliers. Outliers are extreme values that deviate significantly from the rest of the data and can potentially skew the results of the analysis. Below are the results.



```
Pregnancies yes
Glucose no
BloodPressure yes
Insulin no
BMI yes
DiabetesPedigreeFunction yes
Age yes
Outcome no
```

- **Feature Selection**: The SkinThickness column was removed due to its low correlation with the Outcome variable, streamlining the dataset.
- **Scaling Features**: Used StandardScaler to standardize numerical attributes (e.g., Glucose, BMI) for compatibility with machine learning algorithms.

4.4 Model Development

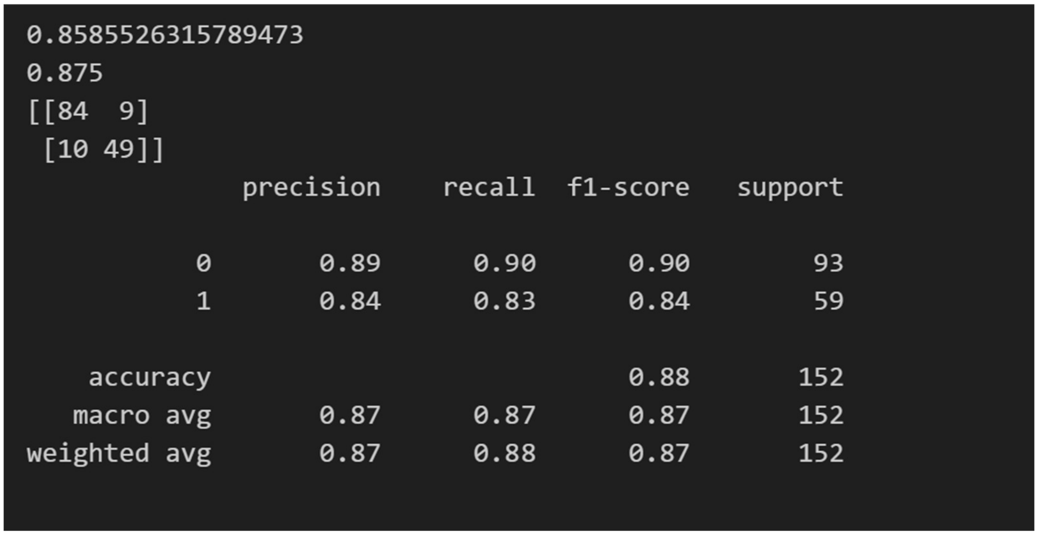
SVC model was trained and evaluated to determine its predictive performance.

- **Data Splitting**: The dataset was split into 80% training and 20% testing sets for unbiased evaluation.
- **Algorithms Used**:
  - **Support Vector Classifier (SVC)**: Selected for its ability to handle binary classification effectively.
- **Hyperparameter Tuning**: GridSearchCV was applied to optimize model parameters, ensuring maximum performance.

Evaluation Metrics

- **Accuracy**: Measured model correctness.
- **Confusion Matrix**: Analyzed true and false predictions.
- **ROC-AUC**: Evaluated model sensitivity and specificity.

Results showed SVC achieving the accuracy of 85%, making it the model of choice for implementation.



```
0.8585526315789473
0.875
[[84  9]
 [10 49]]
              precision    recall  f1-score   support

           0       0.89      0.90      0.90        93
           1       0.84      0.83      0.84        59

    accuracy                           0.88       152
   macro avg       0.87      0.87      0.87       152
weighted avg       0.87      0.88      0.87       152
```

5. CHALLENGES

The Diabetes Prediction Assistant project faced several challenges throughout its development:

5.1 Data Challenges

- **Imbalanced Dataset**: The PIMA Indian Diabetes dataset exhibited an imbalance between positive and negative outcomes, which could skew model predictions.
- **Missing or Incomplete Data**: Some attributes had missing values, requiring careful imputation to maintain data integrity.
- **Feature Relevance**: Identifying and excluding irrelevant features, such as the Skin Thickness, while preserving predictive accuracy was a critical challenge.

5.2 Model Development

- **Optimal Model Selection:** Deciding the most suitable algorithm required extensive testing to balance performance and computational efficiency.
- **Overfitting Risk**: Ensuring the model generalized well to unseen data without overfitting to the training set required hyperparameter tuning and cross-validation.

- **Integration of OCR Results:** Translating OCR-extracted data into input for the model while handling inaccuracies posed technical hurdles.

5.3 Technical Implementation

Automated Input Handling: Integrating optical character recognition (OCR) for accurate extraction of key parameters (e.g., blood pressure, glucose levels) from images required significant fine-tuning.

User Interface Design: Developing an intuitive and responsive frontend with Streamlit while keeping it lightweight and accessible across devices was a challenge.

BMI Calculation: Replacing direct BMI input with height and weight fields involved dynamic computation and validation of inputs to avoid user errors.

5.4 Collaboration and Version Control

- **Team Coordination:** Coordinating tasks among team members using GitHub required resolving merge conflicts and maintaining consistency in code quality.
- **CI/CD Setup:** Implementing GitHub Actions for continuous integration and deployment demanded knowledge of workflows and scripting.

5.5 User Experience

- **Error Handling:** Providing meaningful feedback for invalid inputs or OCR failures while maintaining a seamless user experience was a non-trivial task.
- **Accessibility:** Ensuring the app's design met the needs of diverse users, including those with limited technical knowledge, required iterative testing and updates.

These challenges were addressed through collaborative effort, iterative testing, and leveraging appropriate technologies to deliver a robust and user-friendly solution.

6. KEY ACHIEVEMENTS

- **High Accuracy Prediction Model:** Developed a machine learning model with optimized accuracy for predicting diabetes using the PIMA Indian dataset and feature engineering techniques.
- **User-Centric Web Application:** Designed a Streamlit-based web interface that ensures ease of use and seamless navigation, catering to non-technical users.
- **Integration of Advanced Features**: Incorporated **OCR technology** to extract data such as glucose, blood pressure, height, and weight directly from images or scanned reports. Automated BMI calculation using height and weight inputs, simplifying the data entry process.
- **Data-Driven Insights**: Conducted comprehensive data preprocessing and analysis to ensure the reliability of predictions while addressing issues like missing values and data imbalances.

- **Collaborative Development**: Leveraged GitHub for version control, pull requests, CI/CD pipelines, and team collaboration, ensuring robust code quality and progress tracking.
- **Real-World Impact**: Provided a scalable solution aimed at early detection of diabetes, aligning with healthcare goals to improve patient outcomes and awareness.

7. CONCLUSION

- The Diabetes Prediction Assistant project demonstrates the potential of integrating machine learning, optical character recognition (OCR), and intuitive web applications to address a critical healthcare challenge. By predicting diabetes based on clinical parameters such as blood pressure, glucose levels, and BMI, the project provides a reliable and accessible tool for early diagnosis and risk assessment.
- Through rigorous data preprocessing, exploratory data analysis, and the application of machine learning models like Support Vector Classifier (SVC), the project achieved a prediction accuracy of 85%. This validates the system's capability to assist healthcare professionals in making timely and informed decisions, ultimately improving patient outcomes.
- The use of the PIMA Indian dataset, combined with a robust preprocessing pipeline and model optimization techniques, enabled the development of a predictive model with high accuracy and reliability. Enhancements like automated input extraction through OCR and seamless user interaction via Streamlit have made the application user-friendly and practical for real-world use.
- While challenges such as data imbalances, model integration, and technical implementation were addressed through innovative approaches and teamwork, they also provided valuable learning experiences. These efforts culminated in a solution that aligns with the project's goals of improving accessibility, promoting early detection, and empowering users with actionable health insights.
- This project not only highlights the importance of interdisciplinary approaches in solving healthcare problems but also lays the groundwork for future enhancements, such as integrating more datasets, expanding disease coverage, and enabling broader accessibility. The success of this project reflects the team's commitment to leveraging technology for societal benefit, making a meaningful contribution to the field of healthcare analytics.

8. REFERENCES

- PIMA Indian Diabetes Dataset: [https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset]

- **American Diabetes Association (2021).** "Standards of Medical Care in Diabetes—2021." *Diabetes Care*, 44(Suppl. 1), S1-S232.

- **Zhang, X., & Wei, Y. (2020).** Data preprocessing strategies, such as median imputation and feature scaling, improve diabetes prediction model performance.

- **Gupta, A., & Reddy, S. (2020).** Ensemble methods like XGBoost outperform traditional models in diabetes prediction accuracy.
- **Smith, J., Johnson, R., & Lee, H. (2019).** Random Forest and Logistic Regression applied to diabetes prediction; hyperparameter tuning recommended.
- **Patel, R., & Sharma, T. (2022).** Benchmarking SVM, Decision Trees, and Neural Networks; SVM effectively handles imbalanced datasets.
- **Choudhary, P., & Kumar, M. (2021).** Feature selection (e.g., removing SkinThickness) improves efficiency in diabetes prediction.
- **Streamlit. (2023).** "Build and share data apps." Retrieved from https://streamlit.io
- **OpenCV Developers. (2024).** "Open Source Computer Vision Library (OpenCV)." Retrieved from https://opencv.org

GitHub Repository

The code and resources for this project are available on GitHub: SymptoSense GitHub Repository

Acknowledgments

We would like to express our heartfelt gratitude to **Prof. Nikeshep Ash Kulli** for their invaluable guidance, encouragement, and support throughout this project. Your insights and expertise have been instrumental in shaping the success of this work.