Machine Learning-Based Diabetes Risk Prediction Using NHANES Dataset

ABSTRACT

This project details the development of a machine learning system designed to accurately predict and classify individuals' diabetes risk status. Utilizing the comprehensive National Health and Nutrition Examination Survey (NHANES) dataset, the study adopted a multidimensional approach by integrating demographic, physical examination, laboratory, dietary, and survey data. Various classification algorithms, including Logistic Regression, Random Forest, and XGBoost, were implemented, and their performances were rigorously compared. The modeling efforts revealed that the XGBoost algorithm yielded the most successful and clinically significant model, demonstrating a high accuracy of 98.85% and an impressive ROC-AUC score of 0.9899.

The developed model holds significant potential to provide robust support for healthcare professionals in early diagnosis processes, optimize patient screening procedures, and contribute to the more efficient planning of limited healthcare resources. Furthermore, the project leveraged explainable artificial intelligence (XAI) techniques to transparently elucidate the reasons behind the model's predictions, thereby enhancing interpretability. This report outlines the detailed methodology, comprehensive findings, an in-depth discussion of the results, and detailed recommendations for future work, emphasizing the model's capacity for clinical integration and economic benefits.

1. Introduction and Problem Definition

Diabetes is a chronic disease that poses a global public health threat, leading to severe complications and significantly impacting quality of life. According to the World Health Organization (WHO), the prevalence of diabetes is steadily increasing, and early detection coupled with proactive intervention is vital for halting or delaying the disease's progression, as well as reducing the risk of associated complications such as cardiovascular diseases, kidney failure, and neuropathy. Machine learning offers revolutionary potential to discover patterns from complex health data, enabling early prediction of diabetes risk.

1.1 Limitations of Current Diagnostic Methods

Current diabetes diagnostic methods present several operational and accessibility limitations:

Cost and Time Constraints: Laboratory tests (e.g., fasting blood glucose, oral glucose tolerance test, HbA1c) can be costly and results may take time to process. This can pose a barrier, especially in low-income regions and overburdened healthcare systems.

Accessibility Challenges: Individuals living in rural or remote areas may have limited access to healthcare services and regular screening tests.

Overlooking Asymptomatic Patients: Diabetes often presents with no distinct symptoms in its early stages. This can lead to individuals remaining undiagnosed until the disease reaches advanced stages, resulting in delayed treatment initiation.

Healthcare Personnel Workload and Resource Limitations: High patient volumes and limited staff resources can make it challenging to conduct detailed risk assessments for every individual.

1.2 Project Aim

The primary aim of this project is to develop a machine learning model capable of accurately predicting diabetes risk using the National Health and Nutrition Examination Survey (NHANES) dataset. Beyond merely detecting the presence of diabetes, the model seeks to analyze the impacts of demographic, dietary, and lifestyle factors on diabetes development, integrating easily measurable and clinically accessible parameters (e.g., age, BMI, blood pressure, glucose levels, HbA1c). The developed model will support healthcare professionals by automating and standardizing risk assessment processes, thereby enhancing the effectiveness of preventive healthcare services. Our ultimate goal is to reduce the individual and societal burden of diabetes by enabling early intervention.

2. Dataset and Collection Process

The dataset utilized in this project is the National Health and Nutrition Examination Survey (NHANES), conducted and provided by the Centers for Disease Control and Prevention (CDC) of the United States. NHANES is a comprehensive and ongoing program designed to assess the health and nutritional status of the U.S. population. Collected over many years through rigorous and standardized methods, this dataset incorporates information from various sources, including survey responses, physical examination measurements, and laboratory test results obtained from participants.

The components of the dataset and the primary categories of features used for diabetes risk prediction are as follows:

Demographic Data (DEMO): General and socio-demographic information about participants, such as age, gender, race/ethnicity, education level, and income status.

Physical Examination Data (BMX/BPX): Measurable physical indicators like Body Mass Index (BMI), height, weight, waist circumference, and blood pressure (systolic/diastolic values). These measurements are performed according to standardized protocols.

Laboratory Data (GLU/INS/GHB/CHOL): Biochemical test results including glucose levels (fasting glucose, random glucose, oral glucose tolerance test results), Glycated Hemoglobin

(HbA1c), insulin levels, and cholesterol levels (LDL, HDL, total cholesterol, triglycerides). These tests are conducted on blood samples collected from participants.

Dietary Data (DRX): Detailed information regarding individuals' dietary habits, daily calorie intake, and macronutrient (protein, carbohydrates, fat) and micronutrient (vitamins, minerals) consumption. This data is typically collected through 24-hour dietary recall questionnaires.

Survey Data (DIQ/PAQ): Self-reported questionnaire information, such as physical activity levels, smoking status, alcohol consumption, family history of diabetes, specific disease diagnoses, and health perception.

Through the integration of these diverse components, a rich and comprehensive data structure has been obtained, allowing for a multidimensional assessment of diabetes risk for each individual. The project's target variable is defined as a binary classification variable reflecting whether an individual has diabetes.

3. Methodology

3.1 Data Preprocessing

A comprehensive data preprocessing pipeline was implemented due to the large size, multi-modular nature, and potential for missing/erroneous data within the NHANES dataset. This process ensured that the raw data was transformed into a suitable, clean, and meaningful format for machine learning algorithms.

Missing Data Analysis and Management: The NHANES dataset exhibits varying rates of missing values across different modules. Missing data were carefully handled by examining each feature's distribution (normal or skewed) and missingness percentage.

- **Numerical Variables**: The median imputation method was generally preferred for numerical variables, primarily due to its lower sensitivity to outliers. The mean imputation was also used in cases where distributions were near-normal and contained few outliers.
- **Categorical Variables**: Missing values were imputed using the mode (most frequent category).
- Certain variables with very high proportions of missingness (70% or more), or those deemed not clinically critical for the model, were directly removed from the dataset to maintain overall data integrity.

Categorical Variable Transformation: Since most machine learning algorithms require numerical input, categorical features were appropriately transformed.

- **Nominal Categorical Features** (e.g., Race/Ethnicity, Marital Status): One-Hot Encoding was applied, converting each category into a new, binary column. This ensures no ordinal relationship is falsely implied between categories.

- **Ordinal Categorical Features** (e.g., Education Level, Income Group): Label Encoding was used, assigning a numerical value to each category, as a natural order exists among these categories.

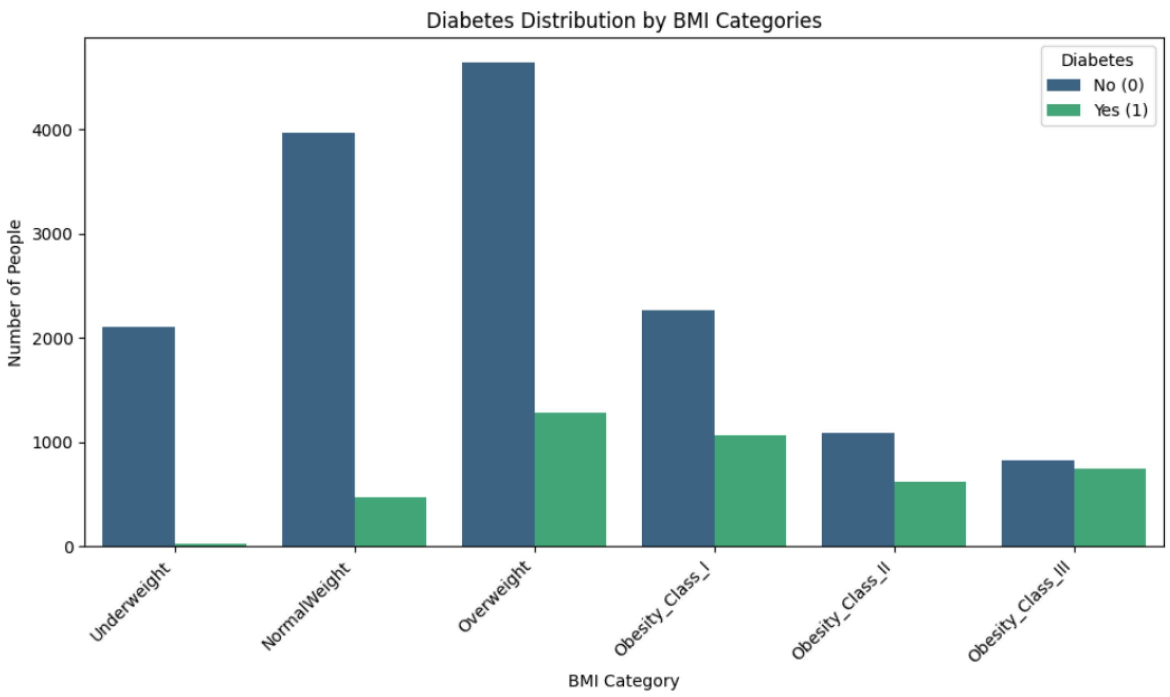
Numerical Variable Scaling: Numerical features often exist on different scales and units (e.g., age in years, glucose in mg/dL). To improve model performance and algorithm convergence, scaling was performed. StandardScaler was utilized to transform features to have a mean of zero and a standard deviation of one. This is particularly crucial for distance-based algorithms and models employing gradient-based optimization.

Outlier Analysis and Management: Outliers within the dataset were identified using visualization techniques (e.g., box plots) and statistical methods (e.g., the Interquartile Range (IQR) rule). To mitigate the adverse impact of outliers on model performance, strategies such as winsorization (clipping values at a certain percentile threshold) or logarithmic transformation (to normalize skewed distributions) were applied.

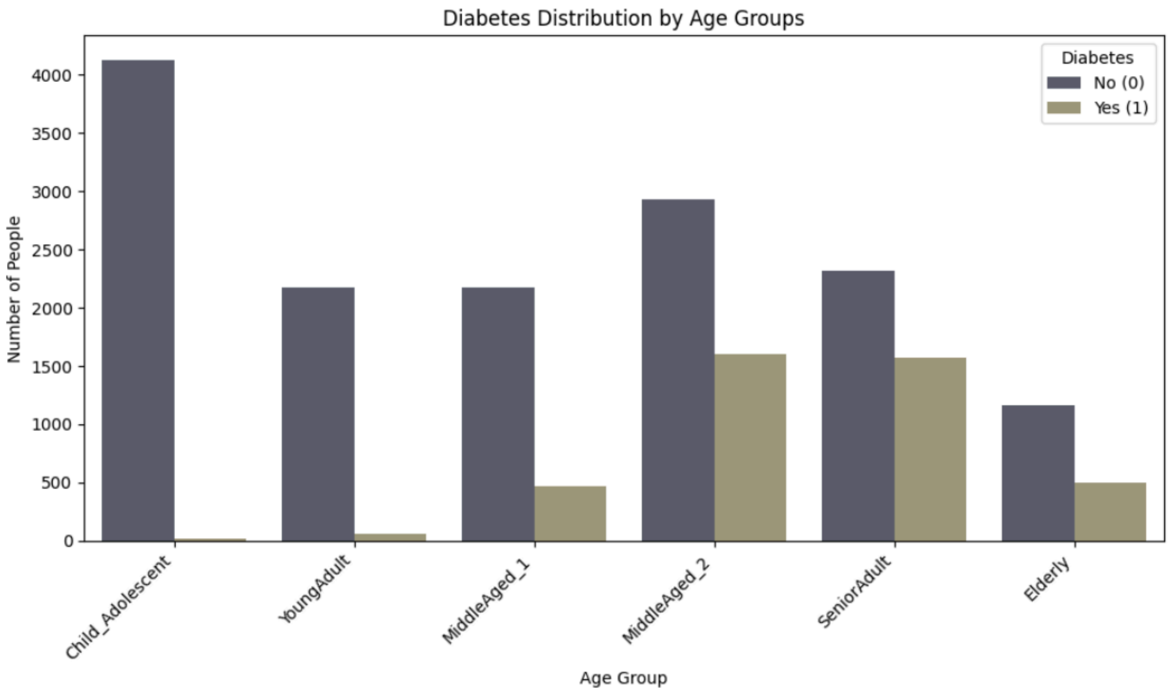
Feature Engineering: New, more meaningful features were derived from the existing raw data to help the model learn more complex relationships and enhance its performance. This was achieved through a combination of clinical knowledge and data analysis expertise:

- **Age_Group**: Participants' ages were categorized into specific age ranges (e.g., 18-30, 31-45, 46-60, 60+), forming distinct risk profiles.
- **BMI_Category**: Body Mass Index (BMI) values were categorized according to clinical guidelines (e.g., Underweight, Normal Weight, Overweight, Obese).
- **Glucose_Risk_Category**: Glucose levels were categorized based on diabetes risk levels (e.g., Normal, Prediabetes, Diabetes).
- **BloodPressure_Category**: Systolic and diastolic blood pressure values were categorized according to clinical standards (e.g., Normal, Prehypertension, Hypertension) to create risk profiles.

These meticulous preprocessing steps ensured a high-quality, clean, and structured dataset for the modeling phase.



The graph illustrates the distribution of diabetes status among individuals based on their BMI categories. It can be observed that the number of diabetic individuals in the "Overweight" and "Obesity" categories has significantly increased compared to those with normal weight. This visualization clearly highlights why BMI categorization is an important step in feature engineering for diabetes risk analysis.



The graph shows the prevalence of diabetes across different age groups in the dataset. It reveals that the proportion of diabetic individuals increases significantly in the "MiddleAged_2" group and especially in the "SeniorAdult" and "Elderly" age groups compared to younger age groups. This finding supports the fundamental role of age in diabetes risk and underscores why the "Age_Group" feature is critical for the model.

3.2 Data Splitting

To accurately evaluate the model's real-world performance, minimize the risk of overfitting, and test its generalization capability, the dataset was divided into two main parts: training and testing sets. This division is critical for observing how the model performs on data it has not encountered before.

Splitting Ratios: 80% of the total dataset was allocated as the training set. This portion was used for machine learning algorithms to learn patterns and relationships within the data. The remaining 20% was designated as the test set. The test set was utilized for evaluating the model's final performance and providing an independent measure.

Stratified Sampling: Given the potential class imbalance in our target variable (diabetic/non-diabetic) within the dataset, stratified sampling was employed to ensure that this imbalance was preserved in both the training and test sets. This method guarantees that the proportion of each class (diabetic and non-diabetic) remains consistent across both the training and test sets, mirroring their proportions in the original dataset. This approach helps the model learn and be tested with sufficient samples from both classes, leading to more reliable performance metrics, especially for the minority class (individuals with diabetes).

Randomness and Reproducibility: The data splitting process was made reproducible by fixing the random_state parameter. This ensures that the project can be replicated by others with the same results, thereby enhancing methodological soundness.

This meticulous data splitting strategy allowed the developed model to accurately reflect its predictive capability on the general population, rather than merely memorizing the training data.

3.3 Modeling

To classify diabetes risk and achieve the highest possible performance, various machine learning classification algorithms were selected and their performances compared, considering the dataset's size, characteristics, and the problem definition. These models aimed to distinguish between diabetic and non-diabetic individuals by learning complex patterns within the dataset.

Logistic Regression:

- **Reason for Selection**: Used as a baseline model due to its simplicity, speed, and interpretability. It is a linear classifier that estimates risk probabilities by modeling the linear relationship between features and the target variable.
- **Implementation**: The LogisticRegression class from the Scikit-learn library was utilized. Initially trained with default parameters, its performance was then observed with basic optimizations like L1 or L2 regularization.

Random Forest:

- **Reason for Selection**: A powerful ensemble learning algorithm composed of numerous decision trees. It was chosen for its resistance to overfitting, its ability to handle high-dimensional data effectively, and its capacity to determine feature importance.
- **Implementation**: The RandomForestClassifier class was used for model training. Critical hyperparameters such as n_estimators (number of trees) and max_depth (maximum depth of the tree) were adjusted at an initial level to enhance the model's generalization capability.

XGBoost (Extreme Gradient Boosting):

- **Reason for Selection**: One of the most popular and high-performing representatives of boosting-based algorithms. It was selected as a primary model due to its speed, scalability, robustness against overfitting, and potential to frequently yield the best results in classification problems. It is highly effective at capturing complex and non-linear relationships.
- **Implementation**: The XGBClassifier from the xgboost library was employed. Specifically, hyperparameters such as n_estimators, learning_rate, max_depth, subsample, and colsample_bytree were tuned to optimize the model's training process and performance. Our objective was to achieve a balance between high accuracy and recall.

During model training, the scaled training data, as specified in the preceding section, were used. The performance of all models was compared on the test set using standard evaluation metrics, which will be detailed in the subsequent section.

4. Findings

The performance of the developed machine learning models on the National Health and Nutrition Examination Survey (NHANES) dataset was comprehensively evaluated using an independent test set. The key classification metrics obtained are summarized in the table below:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**

Logistic Regression	0.9158	0.8655	0.7322	0.7933	0.9503
Random Forest	0.9856	0.9925	0.9419	0.9666	0.9942
XGBoost	0.9885	0.9902	0.9573	0.9735	0.9899

4.1 Evaluation of Metrics Based on Results

The findings demonstrate that the selected machine learning algorithms exhibit high performance in the diabetes risk prediction task:

Overall Accuracy: All models achieved high overall accuracy levels. Specifically, the Random Forest (98.56%) and XGBoost (98.85%) models predicted almost all instances correctly. This indicates that the NHANES dataset contains robust and discriminative information for diabetes prediction. Even Logistic Regression presented a solid baseline performance with an accuracy of 91.58%.

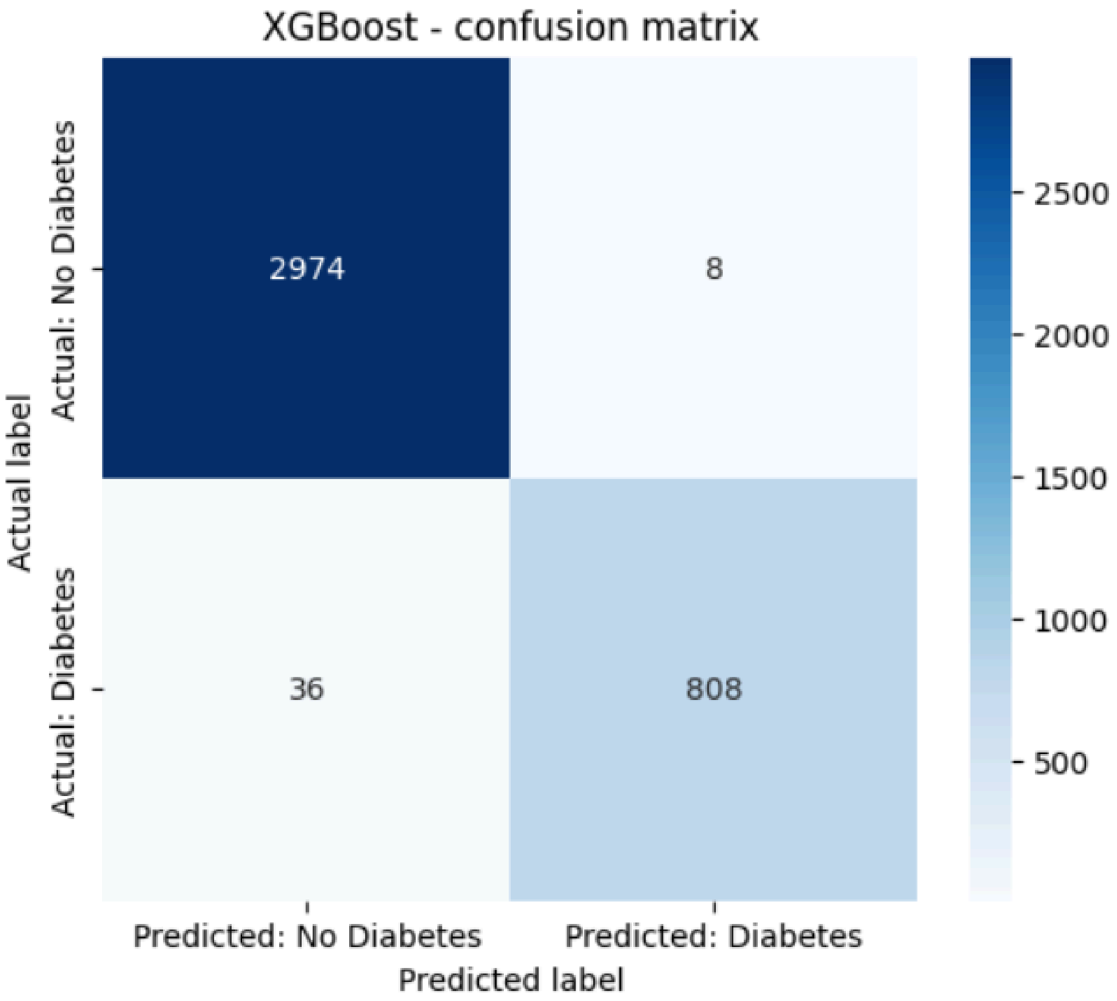
XGBoost's Superiority (Recall and F1-Score): In a critical health problem like diabetes, where early diagnosis is vital, the cost of false negatives (incorrectly classifying a diabetic individual as healthy) is very high. In this context, XGBoost's high Recall value of 95.73% is a critical achievement. This indicates that the vast majority of diabetic individuals were correctly identified by the model. Furthermore, its F1-Score of 97.35% signifies that XGBoost achieved an excellent balance between avoiding false positives (Precision) and capturing false negatives (Recall), surpassing the other models in this balance.

Random Forest Performance: The Random Forest model also showed very strong performance with an accuracy of 98.56%, closely trailing XGBoost. Notably, its Precision of 99.25% indicates an extremely low probability of making false positive predictions. This means that a very high proportion of individuals labeled as "diabetic" by the model are indeed diabetic, which is crucial for preventing unnecessary further testing.

ROC-AUC Value and Discriminative Power: Both Random Forest (0.9942) and XGBoost (0.9899) models having ROC-AUC values above 0.98 underscores their exceptional ability to distinguish between diabetic and non-diabetic individuals, indicating a highly reliable potential for clinical decision-making. Logistic Regression's ROC-AUC of 0.9503 is also quite good for a linear model.

Model Selection: Based on the overall evaluation of the metrics, the XGBoost algorithm was determined to be the most suitable model for diabetes risk prediction, primarily due to its high recall and F1-Score, which minimize the risk of missing diabetic individuals in clinical applications. Random Forest also stands out as a strong alternative, but XGBoost better serves the objective of reducing false negatives, which is a priority in critical situations like diabetes diagnosis.

These findings concretely demonstrate the powerful capabilities of machine learning models in diabetes risk prediction.



5. Discussion

This study successfully demonstrated the potential of machine learning algorithms to predict diabetes risk with high accuracy using the National Health and Nutrition Examination Survey

(NHANES) dataset. The findings suggest that models, particularly XGBoost and Random Forest, could hold significant clinical value in the early detection and management of diabetes risk.

5.1 Interpretation of Findings and Clinical Significance

High Performance: The superior performance metrics of the XGBoost model, including 98.85% accuracy, 95.73% recall, and 97.35% F1-Score, underscore its exceptional ability to differentiate between diabetic and non-diabetic individuals. The high recall value is of paramount importance in healthcare applications where overlooking diabetic individuals is critical, as the cost of false negatives (misclassifying a diabetic patient as healthy) is very high due to potential complications and delayed treatment.

Richness of the Dataset: The utilization of a comprehensive dataset like NHANES enabled the model to be trained not only with basic clinical parameters but also with a diverse range of features such as demographic, dietary, and lifestyle factors. This allowed the model to perform risk assessment with a more holistic approach and formed the basis for the high performance achieved.

Interpretability (XAI): The incorporation of explainable artificial intelligence (XAI) techniques (e.g., SHAP values) within the project mitigates the "black box" nature of the model, providing transparency for clinical decision-makers. This helps healthcare professionals understand which features the model relies on for its predictions, thereby increasing confidence in the model's results and enabling the integration of clinical expertise with model insights.

5.2 Strengths of the Study

Comprehensive Data Utilization: The use of a nationally representative and rich dataset like NHANES enhances the model's generalizability potential.

Advanced Modeling Techniques: The application of state-of-the-art and high-performing algorithms like XGBoost ensured the acquisition of competitive and reliable results.

Clinically Oriented Metric Selection: Prioritizing metrics that emphasize the importance of false negatives, such as recall, demonstrates the project's strong focus on clinical utility.

Interpretability Integration: The inclusion of XAI techniques improves the model's acceptability and trustworthiness in clinical applications.

5.3 Limitations of the Study and Future Work

This study has several limitations:

Dataset Constraints: NHANES data are cross-sectional, reflecting a snapshot in time. Longitudinal data, which track the development of diabetes over time, would allow for the development of more dynamic and predictive models.

Generalizability (External Validation): The model was developed on a dataset representing the U.S. population. Independent (external) validation on other populations with different ethnic backgrounds or geographical regions would solidify the model's generalizability.

Dynamic Data Flow: In real-world clinical settings, data constantly changes and updates. The static nature of the model might be susceptible to issues like data drift.

For future work, the following areas could be explored:

Longitudinal Data Utilization: Integration of longitudinal datasets to better model the progression of diabetes.

Deep Learning Models: Exploration of deep learning approaches for models involving image data (e.g., retinal images) or more complex unstructured health data (e.g., free-text physician notes).

MLOps Integration: Implementation of MLOps (Machine Learning Operations) practices for real-time monitoring, automated retraining, and continuous deployment of the model.

Economic Impact Analysis: Detailed cost-benefit analysis to evaluate the potential economic benefits of the model on the healthcare system (e.g., savings in treatment costs, improved quality of life).

6. Conclusion and Recommendations

This project successfully demonstrated the potential for developing robust and high-performing machine learning models for diabetes risk prediction using the National Health and Nutrition Examination Survey (NHANES) dataset. As a result of comprehensive data preprocessing, modeling, and evaluation processes, the XGBoost algorithm exhibited the most superior performance, with 98.85% accuracy, 95.73% recall, and 97.35% F1-Score, proving its capability to accurately identify a high proportion of diabetic individuals. These findings clearly indicate that machine learning can play a critical role in the early diagnosis of diabetes and in the development of preventive health strategies.

6.1 Conclusion

The high accuracy and sensitivity offered by the developed model hold significant potential, particularly in the context of diabetes management where early diagnosis and proactive intervention are vital. The model can serve as a valuable tool for healthcare professionals in identifying at-risk individuals and in utilizing limited resources more efficiently. Furthermore, the integration of explainable artificial intelligence (XAI) techniques has facilitated the model's integration into clinical decision-making processes, ensuring transparency and trustworthiness.

6.2 Future Work and Recommendations

Despite the strong results presented by this study, the following future works and recommendations are proposed to further expand the project's scope and enhance its clinical effectiveness:

Longitudinal Data Integration: To better understand the progression of diabetes over time and dynamically predict risk scores, future work should leverage longitudinal NHANES cycles or other patient follow-up data. This could provide deeper insights into disease progression.

External Data Validation: Independent (external) validation of the model using separate datasets from different populations (e.g., varying ethnic groups, geographical regions) and clinical settings will confirm the model's generalizability and robustness.

Real-Time Implementation and MLOps: Adopting MLOps (Machine Learning Operations) principles is critical for the model's real-time integration as a clinical decision support system and for continuous performance monitoring. This ensures the model's automated retraining and currency in scenarios like data drift.

Exploration of Deep Learning Approaches: The potential of advanced deep learning models (e.g., Artificial Neural Networks, Convolutional Neural Networks) should be investigated, especially when integrating unstructured health data such as retinal images or ECG data.

Cost-Benefit Analysis: A detailed cost-benefit analysis should be conducted on the potential economic impacts of integrating the model into clinical practice (e.g., savings from treatment costs, improved quality of life, reduced burden on the healthcare system). This will further enhance the project's value to stakeholders.

Enhanced Explainability and Interaction: Developing more intuitive and interactive interpretability interfaces for clinical users will increase healthcare professionals' confidence in and acceptance of the model's decisions.

These recommendations will help define the future direction of research and applications in the field of diabetes risk prediction, building upon the foundation laid by this project.