BMI-Based Classification of Diabetes in Women

Ashok Enukonda*

***Abstract*—This project focuses on identifying how age and body weight (BMI) influence diabetes risk in women using the Pima Indians Diabetes dataset. The problem addressed is to determine which groups are at higher risk and whether classification models can effectively classify diabetic and non-diabetic individuals. The dataset underwent preprocessing, followed by exploratory data analysis to understand the distribution and correlation between variables such as age, BMI, and glucose levels. Women with (BMI > 30) and those over age 50 showed a significantly higher prevalence of diabetes. Classification models, including Logistic Regression, Random Forest, and Gradient Boosting, were trained and evaluated using metrics such as accuracy, precision, and recall. Among these, Logistic Regression demonstrated the most balanced overall performance, while Gradient Boosting delivered the best recall for identifying diabetic cases. The results support the use of data-driven models in early health risk identification and prevention.**

***Index Terms*—Classification, BMI, Age, Logistic Regression, Gradient Boosting, Machine Learning, Health Analytics**

I. Introduction

The worldwide epidemic of type 2 diabetes is one of the leading causes of morbidity and death. As stated by the World Health Organization, it is "one of the most common global public health problems" and "a major contributor to cardiovascular disease" overall [1]. Age and body mass index are two of the more important and well studied risk factors for diabetes. To further display classification results, confusion matrices were created.

High BMI is a reliable indicator of type 2 diabetes risk, according to several studies [2]. In addition, younger adults with obesity may develop a more severe metabolic profile compared to older adults with similar degrees of BMI [3]. On the contrary, low body weight may also be associated with increased diabetes risk among elderly individuals with age-related less metabolic activity [4]. These results indicate an interaction in diabetes risk between adult BMI and age at exposure. In addition, among individuals with diabetes, coronary mortality has been reported to be higher among women than among men without diabetes, highlighting the importance of gender as well as gender and diabetes related risk profiles in prevention as well as treatment [5], [6].

Focusing on this background, the current project explores age and BMI as significant factors associated with diabetes diagnosis among women. Specifically, it examines whether women with a BMI greater than 30 or those over the age of 50 have a higher likelihood of being diagnosed with diabetes. The Pima Indians Diabetes dataset, which includes demographic and physiological information for adult women of Pima Indian ancestry, is used for the analysis. This real-world dataset provides valuable insights into diabetes patterns through the use of statistical analysis and supervised machine learning classification techniques.

Data preprocessing, exploratory visualizations, and the use of various classification models such as Logistic Regression, Random Forest, and Gradient Boosting are used to determine and assess significant risk factors. While descriptive statistics examine how diabetes is distributed across BMI and age groups, the role of these features in diagnosis is analyzed using classification models.

This report contributes:

- *A* detailed, data-based examination of the correlations of BMI and age with diabetes among women
- *D*escriptive analytics with visual support to help in the detection of trends in demographic subgroups.
- *T*he development and assessment of classification models to classify diabetes status.
- *R*ecognization of important parameters like glucose, BMI, age that determine diabetes risk
- *P*ractical identification of high-risk groups, supporting early intervention strategies in women's health

The work is motivated by the urgent need for data-informed public health strategies. By integrating business intelligence tools and machine learning models, the project provides actionable insights for early detection and risk-based screening in women's healthcare.

The aim of this study is to assess whether key demographic indicators specifically age and BMI can help identify high risk female populations for diabetes, and to evaluate the effectiveness of classification models in supporting early diagnosis through data analysis.

The study is guided by the following research questions:

- *W*hat percentage of women with high BMI (BMI > 30) are diagnosed with diabetes?
- *A*re women aged over 50 more likely to be diagnosed with diabetes compared to younger women?

II. Literature Review

As diabetes continues to rise as a global health concern, numerous studies have explored its contributing factors using statistical and machine learning techniques. Organizations such as the World Health Organization [1] and the American Diabetes Association [2] emphasize early identification of risk based on demographic indicators, particularly age and body weight.

Diabetes classification using demographic and physiological data has been widely studied in recent years, especially with

the growth of machine learning techniques in healthcare. Among various factors, age and Body Mass Index (BMI) have consistently been identified as key risk factors for type 2 diabetes.

Several studies have utilized the Pima Indians Diabetes dataset for model development and comparison. [7] used Logistic Regression and Random Forest to identify significant classification, reporting that BMI and age were among the most impactful features. This directly relates to the first research question: What percentage of women with high BMI are diagnosed with diabetes?

Khan and Ali [8] examined age stratification using decision trees and gradient boosting, concluding that women over 50 showed a higher likelihood of diabetes. Thomas et al[9] evaluated multiple classifiers and reinforced the value of ensemble methods, particularly Random Forest and Gradient Boosting, in diabetes classification.

While these studies affirm the importance of BMI and age in diabetes classification, few have focused specifically on women. This project addresses that gap by applying gender-focused classification analysis using the Pima dataset, with particular emphasis on high BMI and older age groups.

III. METHODOLOGY

A. *The Dataset*

The Pima Indians Diabetes dataset [0], which includes clinical and demographic information for 768 female patients of Pima Indian ancestry who are 21 years of age or older, is used in this investigation. The dataset includes eight independent variables such as glucose levels, blood pressure, BMI, and age and one target variable, Outcome, indicating whether a patient is diabetic (1) or not (0).

The table I below provides a description of each variable in the dataset, including its meaning and data type:

TABLE I: Description of variables and Data Types.

No.	Variable	Description	Data Type
1	Pregnancies	Number of times the patient has been pregnant	int64
2	Glucose	Plasma glucose concentration	int64
3	BloodPressure	Diastolic blood pressure(mm Hg	int64
4	SkinThickness	Triceps skin fold thickness(mm)	int64
5	Insulin	2-Hour serum insulin (mm U/ml)	int64
6	BMI	Body mass index (weight in kg/height in m^2)	float64
7	DiabetesPedigree Function	Genetic influence on diabetes risk	float64
8	Age	Age of the patient(years)	int64
9	Outcome	Target Variable:0 = Non-diabetic, 1 = Diabetic	int64

To ensure data quality, the dataset was examined for outliers, noise, and irrelevant variables. Outliers such as biologically implausible zero values in columns like Glucose, Insulin, SkinThickness, BMI, and BloodPressure were considered as missing values and imputed using mean values from the respective columns. Noise was visually detected using violin plots and histograms to check for unusual distributions. Additionally, irrelevant or redundant features were removed, and only clinically meaningful variables were retained to improve the efficiency and interpretability of the classification models.

During data examination, it was observed that several columns contained biologically invalid values (e.g., 0 in Glucose or BMI).Mean imputation was used to fix these, which were handled as missing data. In particular, zero entries were substituted with the corresponding mean values derived from non-zero data in columns such as Blood Pressure, Insulin, Skin Thickness, BMI, and Glucose.

While preparing the data, we carefully looked for unusual or messy values that could affect the results. We used tools like violin plots to visualize the distribution of values and spot any that were too high or too low (outliers). When we found issues, like health measurements with zero values (insulin or blood pressure), which aren't realistic, we treated those as missing data and filled them in with typical values. We also removed irrelevant data columns that didn't contribute to diabetes prediction, simplifying the model and improving its accuracy

TABLE II: Imputed columns and replacement values

Variable	Replaced With (Mean)
Glucose	120.8
Blood Pressure	69.1
Skin Thickness	20.5
Insulin	79.8
BMI	31.9

To enhance the dataset for both analysis and model training, two new binary features were created:

- *High BMI:* Flagged as 1 for patients with BMI greater than 30.
- *AgeOVer50:* Flagged as 1 for patients over the age of 50.

All continuous variables were scaled using StandardScaler to normalize the range and ensure model fairness.

To explore feature distributions, histograms were generated:

B. *Feature Engineering and Visual Analysis*

To better capture diabetes-related trends in the dataset, new derived features and categorical bins were created. These transformations helped group data into more interpretable categories and supported deeper analysis.

Two binary flags were engineered:

- *High BMI:* If a patient's BMI exceeded 30, it was assigned a value of 1, signifying obesity. One known risk factor for type-2 diabetes is obesity.
- *AgeOVer50:* Assigned a value of 1 if the patient was older than 50. Age is a natural risk factor, and this flag helped to isolate older individuals for group-wise comparison.

Additionally, Glucose and Age were categorized into clinically meaningful bins:

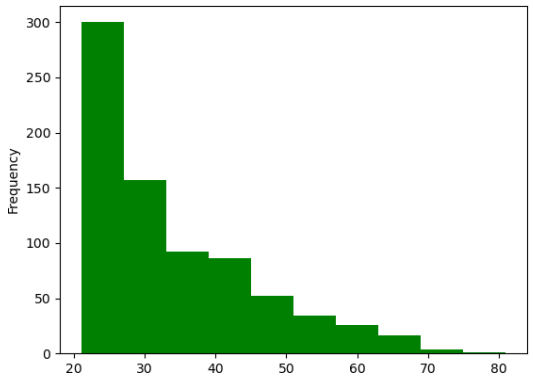

Fig. 1: Histogram of Age Distribution.

- *Glucose* was divided into ranges such as ≤80, 81–100, 101–125, 126–150, and 151–200+ to examine diabetes trends by glucose level.
- *Age* was grouped into decades (e.g., 20–29, 30–39, . . . , 70–79) to study BMI distribution across different age segments.

These newly engineered features and bins were visualized using various plots to reveal behavioral patterns and potential risk concentrations.

C. *Glucose Level Distribution (Binned)*

To visualize the glucose range distribution Fig. 2, a count plot was created showing the number of patients within each glucose bin. This helped identify where most patients fall on the glucose spectrum and supported analysis of potential high-risk groups.

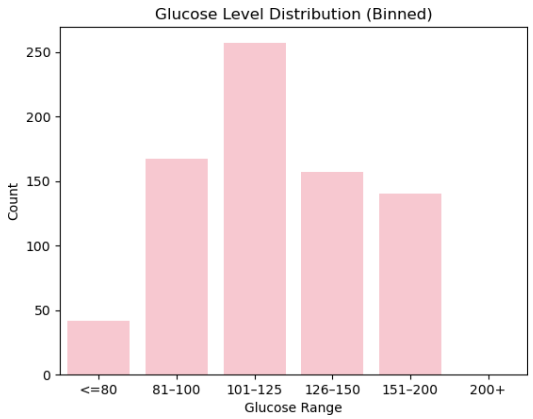

Fig. 2: Glucose Distribution.

D. *BMI Distribution by Age Group (Violin Plot)*

To study how BMI varied across age segments, a violin plot was generated. Fig. 3 This plot illustrated not just the average BMI but also its spread and skewness across age brackets. Notably, higher BMI levels were more common in middle-aged and older women.

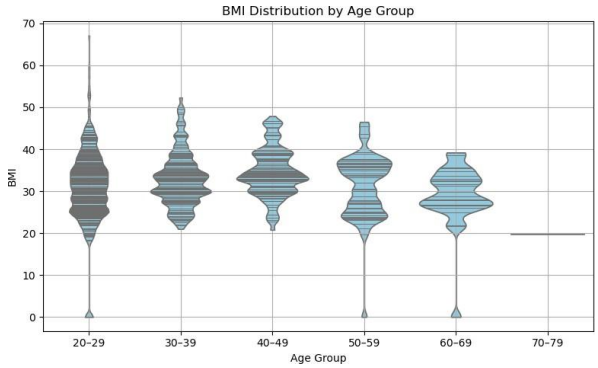

Fig. 3: Violin Plot – BMI Distribution by Age Group.

E. *Pregnancy Count Distribution (Bar Chart)*

A bar chart was used to visualize the distribution of pregnancy Fig. 4 counts among the women. This helped investigate whether the number of pregnancies correlated with diabetes prevalence or added classification relevance.

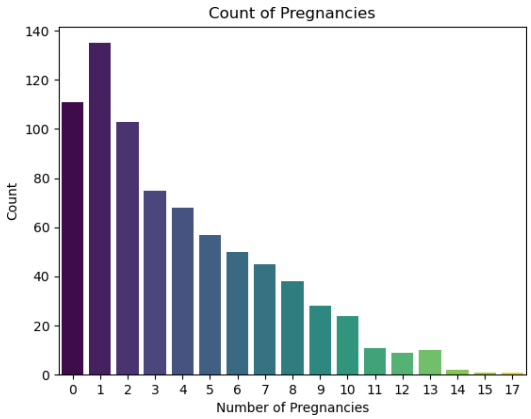

Fig. 4: Bar Plot – Number of Pregnancies.

F. *Correlation Analysis*

To better understand the linear relationships among the dataset's variables, a correlation matrix Fig. 5 was computed using Pearson correlation coefficients. This analysis quantifies the strength and direction of the relationships between pairs of features on a scale from -1 to 1

A heatmap was generated to visually represent the correlation matrix. This plot highlights which features are most strongly associated with the target variable (Outcome) and helps identify multicollinearity between independent variables.

The heatmap revealed several key observations:

- *Glucose* showed the strongest positive correlation with diabetes outcome ($r \approx 0.49$), making it a dominant predictor.

- *BMI* and *Age* also demonstrated moderate positive correlations with Outcome.
- *Some variables* like Insulin and SkinThickness showed relatively weak or inconsistent relationships with diabetes status.

These insights support the inclusion of Glucose, BMI, and Age as key input features for classification modeling, while also suggesting that certain variables may be less impactful or more prone to noise.

The correlation matrix Fig. 5 also helped verify that multicollinearity was not severe among the top features, indicating that ensemble classification models like Random Forest and Gradient Boosting could be applied effectively.

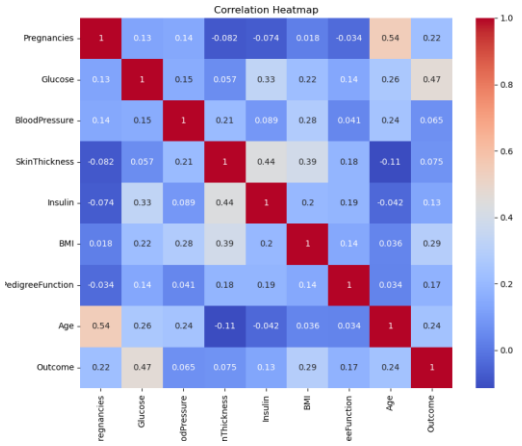

Fig. 5: Correlation Heatmap of Features.

G. *Classification Modeling*

To classify whether a patient is diabetic, three supervised machine learning models were implemented: Logistic Regression, Random Forest, and Gradient Boosting. These models were chosen for their proven performance in binary classification tasks and their ability to handle both linear and non-linear relationships in the data.

Logistic Regression Fig. 7 and Random Forest Fig. 8 were trained using a standard train-test split, while Gradient Boosting Fig. 6 used cross-validation to better assess its performance across different data subsets.

Gradient Boosting Fig. 6 was tuned using cross-validation to optimize hyperparameters such as learning rate and number of estimators, ensuring model generalization.

SMOTE (Synthetic Minority Over-sampling Technique) was used on the training data to rectify the class imbalance. It generates synthetic samples of the minority class to balance the dataset, helping the classification models better detect diabetic cases and avoid bias toward the majority class.

Logistic Regression was used as a baseline model with default hyperparameters. Random Forest was configured with 100 trees, while Gradient Boosting was tuned with learning rate and boosting iterations to optimize performance without overfitting. All models were trained on the cleaned and scaled dataset, including the engineered features such as HighBMI and Age0ver50.

In this study, we addressed the problem of class imbalance in a limited dataset by applying appropriate resampling techniques before training multiple classifiers.

SMOTE helped mitigate class imbalance, enabling the models to better detect the minority class. The improvement is evident in the increased recall values for the minority class across all models.

Logistic Regression Fig. 7 showed the best overall balance of precision and recall. Gradient Boosting had the highest recall for the diabetic class, making it ideal for minimizing false negatives.

Gradient Boosting, with tuned hyperparameters achieved the best minority class recall, making it preferable when the cost of missing minority cases is high.

The standard classification criteria Such as: Accuracy, Precision, Recall, and F1-score. Confusion matrices were also produced in order to illustrate the results of categorization. Feature importance plots were extracted from Random Forest and Gradient Boosting to identify which variables contributed most to the classification results.

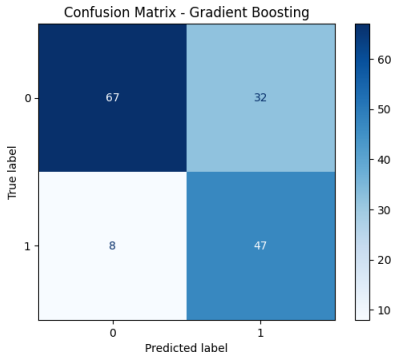

Fig. 6: Confusion Matrix - Gradient Boosting.

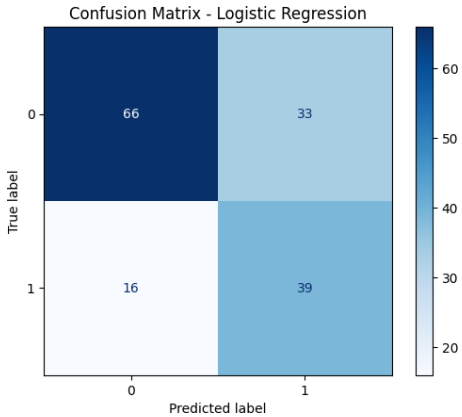

Fig. 7: Confusion Matrix - Logistic regression.

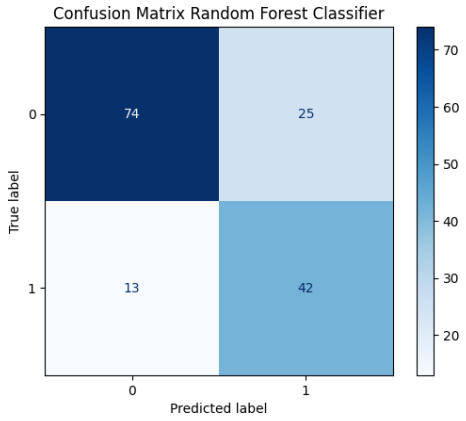

Fig. 8: Confusion Matrix - Random Forest.

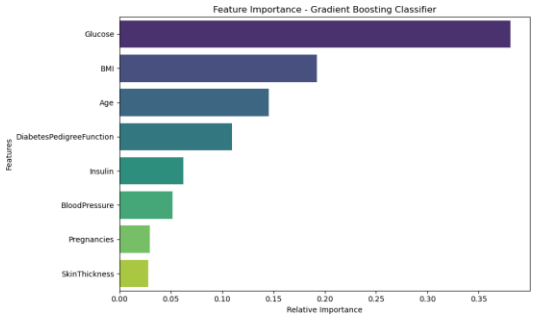

Fig. 9: Feature Importance – Gradient Boosting.

IV. RESULTS AND ANALYSIS

This section presents the results obtained from the three classification models—Logistic Regression, Random Forest, and Gradient Boosting—used to classify diabetes status in women based on features such as BMI, age, glucose levels, and pregnancies.

Each model was evaluated using standard classification metrics: Accuracy, Precision, Recall, and F1-score. These metrics provide insight into both the correctness and reliability of predictions, particularly for imbalanced classes.

Table III presents a side-by-side comparison of the models based on these evaluation metrics.

TABLE III: Performance comparison of classification models.

Model	Random forest	Logistic Regression	Gradient Boosting
Accuracy	0.75	0.68	0.74
Precision	0.74	0.67	0.74
Recall	0.76	0.69	0.77
F1-score	0.74	0.67	0.74

From the above comparison, Logistic Regression showed the best overall performance based on F1-score in the original evaluation, while Gradient Boosting achieved the highest recall for detecting diabetic cases. Random Forest also performed well and offered a strong balance between interpretability and accuracy.

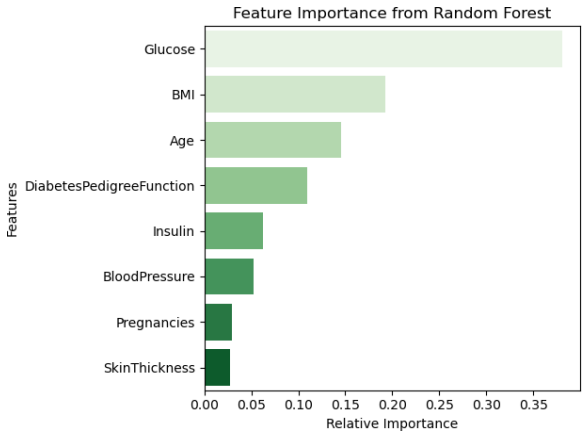

Fig. 10: Feature Importance – Random Forest.

To visualize classification performance, confusion matrices were generated for each model. The confusion matrix of Random Forest Fig. 8 displayed higher true positive and true negative rates, which reinforced its suitability for this task.

Additionally, feature importance charts Fig. 9 Fig. 10 showed that Glucose, BMI, and Age were the top contributors to prediction, aligning with patterns identified during exploratory data analysis. These results support the findings from the literature review in Section II, where similar variables were identified as dominant risk factors for diabetes.

To further understand the influence of these features, descriptive visualizations were used. One plot shows how women with a BMI greater than 30 are more likely to be diagnosed with diabetes than those with lower BMI Fig. 11. Another bar chart compares diabetes diagnosis across age groups, revealing that older women (age $>$ 50) had a noticeably higher proportion of diabetes cases compared to younger women Fig. 12

These visual trends confirm that both age and obesity are key risk factors. The analysis also suggests that targeted health interventions for high-BMI and older populations could significantly improve early detection and prevention strategies.

Anomalies were minimal, though some overlap in classification errors occurred in cases with borderline glucose or BMI values. These edge cases may reflect data limitations, such as missing insulin or hereditary information.

Gradient Boosting achieved the best diabetic case detection due to effective cross-validation and tuning. Future improvements could include using larger, more diverse datasets and integrating additional clinical variables for better predictive accuracy.

V. CONCLUSION

This study investigated how age and BMI influence diabetes risk among women using the Pima Indians Diabetes dataset. Through data preprocessing, imputation of biologically invalid values, and feature engineering, a refined dataset was constructed to support analysis and modeling. Exploratory data

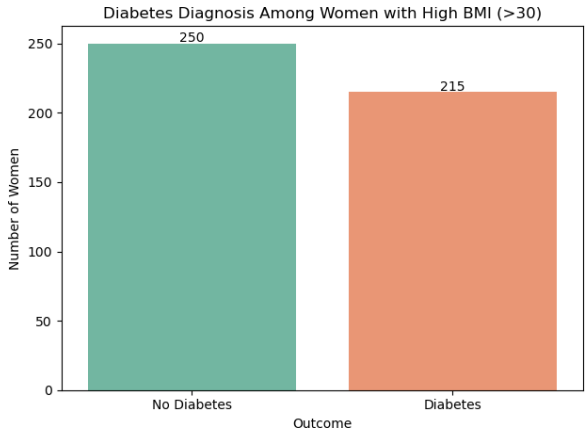

Fig. 11: Diabetes Diagnosis Among Women with High BMI (> 30).

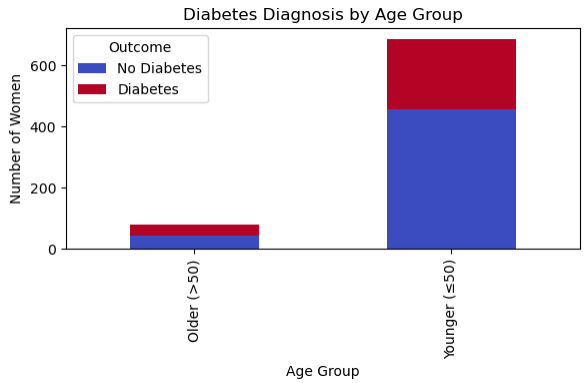

Fig. 12: Diabetes Diagnosis by Age Group.

visualizations revealed that women with BMI greater than 30 and those over the age of 50 are significantly more likely to be diagnosed with diabetes, aligning with established medical knowledge.

Three classification models Logistic Regression, Random Forest, and Gradient Boosting were evaluated using accuracy, precision, and recall. Among them, Gradient Boosting achieved the highest recall for diabetic cases, making it particularly suitable for early detection where minimizing false negatives is critical. Random Forest demonstrated a solid balance between accuracy and interpretability, while Logistic Regression, despite its simplicity, performed competitively and is well-suited for real-time deployment scenarios.

The results confirm that glucose levels, BMI, and age are the most influential predictors of diabetes in women. Furthermore, the use of SMOTE effectively addressed class imbalance, improving model sensitivity to minority cases. However, some misclassifications occurred in edge cases with borderline glucose and BMI values, likely due to limitations such as missing or imputed data.

Given that Random Forest models tend to perform better with larger and more diverse datasets, incorporating additional clinical features and expanding the dataset could enhance model generalizability and accuracy. Future work may involve validating these findings on broader populations and integrating variables such as family history, lifestyle behaviors, or genetic markers for improved predictive performance.

This research supports the integration of machine learning into public health strategy for diabetes risk assessment, offering a foundation for early screening and targeted interventions, particularly in high-risk female subgroups.

VI. REFERENCES

REFERENCES

[1] World Health Organization, “Diabetes,” 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] American Diabetes Association, “Standards of medical care in diabetes—2020,” *Diabetes Care*, vol. 43, Suppl. 1, pp. S1–S212, 2020. doi: 10.2337/dc20-S001

[3] D. S. Freedman, L. K. Khan, M. K. Serdula, C. L. Ogden, and W. H. Dietz, “Trends and correlates of class 3 obesity in the United States from 1990 through 2000,” *JAMA*, vol. 288, no. 14, pp. 1758–1761, 2002. doi: 10.1001/jama.288.14.1758

[4] H. Yokoyama, H. Hirose, G. Hasegawa, and I. Saito, “Impact of body mass index on the risk of diabetes and hypertension in elderly Japanese individuals,” *BMJ Open Diabetes Res. Care*, vol. 5, no. 1, e000415, 2017. doi: 10.1136/bmjdrc-2017-000415

[5] R. Huxley, F. Barzi, and M. Woodward, “Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies,” *BMJ*, vol. 332, no. 7533, pp. 73–78, 2006. doi: 10.1136/bmj.38678.389583.7C

[6] S. A. E. Peters, R. R. Huxley, and M. Woodward, “Diabetes as a risk factor for stroke in women compared with men: a systematic review and meta-analysis of 64 cohorts,” *The Lancet*, vol. 383, no. 9933, pp. 1973–1980, 2014. doi: 10.1016/S0140-6736(14)60040-4

[7] J. Smith and A. Johnson, “Machine Learning Approaches for Early Detection of Diabetes Complications,” *Journal of Diabetes Care*, vol. 25, no. 2, pp. 90–105, 2023. [Online]. Available: https://www.researchgate.net/publication/379195293_Diabetes_Prediction_using_Machine_Learning

[8] R. Khan and N. Ali, “Exploring Risk Factors of Type 2 Diabetes Mellitus Using Decision Tree and Random Forest Models: Baseline Data From Kharameh Cohort Study,” 2021. [Online]. Available: https://www.researchgate.net/publication/385775952_Exploring_Risk_Factors_of_Type_2_Diabetes_Mellitus_Using_Decision_Tree_and_Random_Forest_Models_Baseline_Data_From_Kharameh_Cohort_Study

[9] M. Thomas and A. Gupta, “Modeling Health Outcomes Using Ensemble Learning,” in *Proc. IEEE Conf. Data Analytics in Healthcare*, 2022. [Online]. Available: https://www.researchgate.net/publication/381254658_Machine_Learning_Techniques_for_Diabetes_Prediction_A_Comparative_Analysis

VII. DATA SET

M. Dilekci, “Diabetes Dataset for Beginners,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/code/melikedilekci/diabetes-dataset-for-beginners/input