**LEVERAGING MACHINE LEARNING APPROACHES FOR DIABETES PREDICTION AND ANALYSIS**

1. **Introduction**
   1. **Background**

Diabetes is a chronic metabolic disorder characterised by high blood glucose levels due to the inability of the pancreas to produce adequate insulin for the body function. This condition, if left unmanaged, can lead to severe health complications, resulting to heart disease, kidney failure and amputations. The disease’s increasing prevalence poses a global health crisis, estimating that diabetes cases will surpass 600 million by 2040 (World Health Organisation, 2023). The alarming projection of diabetes highlights the urgent need for preventive measures, emanating from early detection and effective care management strategies (Houngue & Bigirimana, 2022).

Traditional diagnostic approaches for diabetes, such as fasting plasma glucose (FPG) tests, oral glucose tolerance tests (OGTTs), and glycated haemoglobin (HbA1c) measurements are widely used but often limited in sensitivity or accuracy, particularly for early-stage and prediabetic cases. These methods do not fully capture the complex relationships among multiple risk factors, including genetic predisposition, demographic attributes, lifestyle habits and biochemical markers. Recent advancements in Machine learning (ML) have opened new possibilities for diabetes prediction, allowing researchers to analyse diverse risk factors and identify subtle, complex patterns in data (Khongorzul Dashdondov et al., 2024).

Machine learning approaches are increasingly applied to medical science due to their ability to process vast datasets and uncover valuable insights. By leveraging ML models likes Random Forest, XGBoost and Support Vector (SVM). Researchers and practitioners can develop predictive models that more accurate and reliable than traditional methods. For example, models developed by Kasula (2023) and Rani (2020) achieved high accuracy rate of 85% for Random Forest and 99% for Decision Trees. demonstrating the effectiveness of ML in diabetes prediction. Also, ensemble techniques such as those used by Jain et al. (2024), have proven effective in improving model performance across diverse patient demographics. The innovation and involvement of ML technology for diabetes prediction has led to improved and precise healthcare management interventions.

1.2. **Problem Statement**

Diabetes is an intricate, compound medical condition that is influenced by a wide range of genetic, environmental and lifestyle factors, making it challenging to detect in its early stages using conventional diagnostic techniques alone. While traditional methods provide basis for diabetes test, they often fail to accurately screen individuals at risk before the disease symptoms manifest, thereby exposing the individual to chances of complications. Additionally, the linear nature of these methods limits their capabilities in capturing the non-linear associations among various indicators of diabetes advancement.

The use of machine learning in diabetes prediction aims to address these challenges by effectively capturing complex patterns related with the disease onset. However, there is a significant gap in choosing the model that could accurately perform best with robust and reliable predictions for diabetes diagnosis across varied patient groups. Furthermore, outliers, feature selection and data imbalances often found with healthcare datasets can negatively impact model performance. This study aims to address these limitations by evaluating and comparing the performance of three ML algorithms, namely Random Forest, XGBoost and Support Vector Machines. Moreso, it seeks to identify patterns and compare insights from two datasets (Large and Small) from Exploratory Data Analysis (EDA) for better decision making.

1.3. **Justification of the Study**

The growing rate of diabetes and the related health risks emphasize the need for innovative, precise and proactive procedure to its detection and management. Early and accurate diagnosis is essential for effective action, as it enables healthcare professionals to mitigate complications and improve patient outcomes.

This study aims to evaluate which machine learning models, such as Random Forest, XGBoost and SVM is most suited for diabetes prediction in clinical settings. Through Exploratory Data Analysis, the research attempts to discover trends and relationships which compares understanding from two datasets of different sizes, offering actionable insights for treatment and prevention.

Ultimately, this research contributes to the growing body of knowledge on machine learning in healthcare by supporting the development of predictive tools for early detection, management and personalized care of diabetes. Its findings may not only improve health outcomes but also help to create more effective, data-driven strategies to address the global diabetes epidemic.

1.4. **Research Questions**

i. Model Performance Comparison: How do Random Forest, XGBoost and Support Vector Machines (SVM) compare in their predictive performance for diabetes progression?

ii. Exploratory Data Insights: What insights, trends or patterns can be uncovered through Exploratory Data Analysis (EDA) in relation to diabetes progression?

1.5. **Objectives**

i. To develop machine learning models for predicting early detection of diabetes.

ii. To perform Exploratory Data Analysis for identifying trends and patterns in diabetes progression.

iii. To compare insights from two datasets – large and small

iv. To evaluate and compare the performance of different machine learning models.

v. To provide actionable insights for healthcare providers, to improve health outcomes.

2.0. **Literature Review**

This chapter provides a comprehensive review of existing papers on the use of machine learning approaches in diabetes prediction and analysis. It explores the key studies that have contributed to understanding how machine learning models can enhance the accuracy of diabetes diagnostics and the early identification of risk factors. This literature review plans to put in perspective the research problem, identify gaps in the current knowledge and provide a baseline for the chosen methodologies in this work.

Building upon the growing body of research advocating for machine learning in healthcare, this provides a comparative analysis that highlights the efficacy of advanced machine learning models in diabetes risk prediction (Prasetyo and Izdihar, 2024). Their study addresses the critical global health challenges posed by diabetes and underscores the importance of effective predictive methodologies to mitigate its widespread impact. By analysing three prominent ML models; Gaussian Naïve Bayes, Decision Tree and Artificial Neural Network (ANN). The authors asses each model’s performance in diabetes risk prediction. Utilising the Behavioural Risk Factor Surveillance System (BRFSS) dataset, a comprehensive resource for health-related risk behaviours. The paper reveals that the ANN model achieves the highest accuracy with an impressive 84.73%, thereby outperforming both the Gaussian Naïve Bayes and Decision Tree models. This finding stresses the ANN’s potentials in advancing diabetes risk prediction. The study’s contributions are significant as they pave the way for developing precise diagnostic tools and customised interventions, enhancing diabetes management and addressing its societal burden (Prasetyo, Izdibar & Nabiilah, 2024).

Expanding on the application of machine learning in diabetes prediction, Sivaranjani et al. (2021) focus on using predictive models to prevent diabetes by analysing patterns within a dataset comprising both diabetic and non-diabetic individuals. Leveraging the diabetes 130-US hospitals dataset, which spans from 1999 to 2008. The research employs several machine learning models, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest and Support Vector machine (SVM). Findings reveal that feature selection plays a pivotal role in optimising model performance, with Rando Forest model standing out by impressive 99.8% accuracy using the raw dataset. Logistic Regression, though initially less accurate, showed potential for enhancement when combined with multiple sampling techniques.

In response to the rising occurrence of diabetes and its acute health implications, Kasturi (2024) investigated a range of machine learning and deep learning models for detection of diabetes. Utilising the Pima Indian Diabetes dataset, this study aims to classify individuals as diabetic and non-diabetic by employing various machine learning algorithms, including Logistic Regression (LR), K-Nearest Neighbours (KNN), Random Forest (RF) and Support Vector Machine (SVM). In addition, a Multi-Layered Feed Forward Neural Network (MLFNN) is implemented from a deep learning perspective. Among these approaches, the MLFNN achieved the highest accuracy at 92%, suggesting its capabilities for even greater accuracy if applied to larger datasets.

N Nagarjuna and Lakshmi (2024) explore the application of machine learning techniques in diabetes prediction, with a concentration of leveraging various algorithms to enhance predictive accuracy and disease management. Their study employs multiple machine learning models, such as Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and XGBoost, each of which is evaluated for its effectiveness in classifying diabetic and non-diabetic individuals. Random Forest and XGBoost outshine others with notable performance of 82% and 80% respectively.

Gupta (2024) addresses the challenges of predicting diabetes accurately, especially with issues like outliers and missing data in labelled datasets. To handle these issues, the research introduces a comprehensive prediction framework that incorporates essential data preprocessing techniques, including outlier rejection, missing value imputation, data standardisation, feature selection and K-fold cross-validation. The framework employs various machine learning algorithms, such as K-nearest Neighbour (K-NN), Decision Trees, Random Forest, AdaBoost, Naïve Bayes, XGBoost, and Multi-Layer Perception (MLP). A key innovation in this study is the introduction of a weighted ensembling techniques designed to improve prediction accuracy. To further optimise results, hyperparameter tuning is carried out using grid search. The experiment with the Pima Indian Diabetes dataset, demonstrate the effectiveness of the ensembling classifier, with sensitivity of 0.789, specificity of 0.934, false omission rate of 0.092, a diagnostic odds ratio of 66.234 and an AUC of 0.950, which is a 2% improvement over existing methods.

Gufran et al. (2024) address the growing prevalence of diabetes and the importance of early prediction to prevent its onset. Their research uses lifestyle data from the UCI database and evaluates the effectiveness of six machine learning techniques (MLTs) for diabetes prediction. These techniques include Logistic Regression (LR), Decision Tree Classification (DTC), Random Forest Classification (RFC), Support Vector Classification (SVC), and K-Nearest Classification (KNC). The research shows that Logistic Regression outperforms other algorithms with an accuracy of 93%. By examining patients’ lifestyle data and applying techniques like embedding, filter and hybrid feature selection methods, the research demonstrates the advantages of using refined input characteristics.

Elmenshawy et al. (2024) address the global impact of diabetes, which affects 537 million people worldwide and contributes to serious health issues like heart disease, kidney damage and diabetic retinopathy. This study introduces a diabetes prediction framework built using a private Bangladeshi dataset and various machine learning algorithms, such as Decision Tree, SVM, Random Forest, Logistic Regression, K-Nearest Neighbour (KNN) and XGBoost. The XGBoost classifier combined with the ADASYN technique for handling imbalance data, had an accuracy of 80%. Moreover, a stacked ensemble of three classifiers achieved an outstanding accuracy of 99.3%. The framework’s adaptability is demonstrated by the use of domain adaptation techniques, enhancing its predictive capability across various settings.

3.0 **Methodology**

This section outlines the research techniques I applied, for predicting diabetes risk and analysing patterns related to the disease. It details the approach used for data collection, ethical considerations, Exploratory Data Analysis (EDA) and Preprocessing.

3.1. **Overview**

In this study my objective is to predict diabetes risk and analyse patterns associated with the disease using health-related survey responses. I utilized two different datasets for comparison: one large dataset from the CDC’s Behavioural Risk Factor Surveillance System (BRFSS), and a small dataset sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the Pima Indians Diabetes dataset.

For this analysis, I used three different models: namely Random Forest, XGBoost and Support Vector Machine (SVM). These models were chosen due to their ability to handle complex, high-dimensional data and their effectiveness in classification tasks. Additionally, I performed EDA to discover meaningful insights and patterns within the data, and also, I preprocessed the data for suitability. After building the models, I evaluated their performance using metrics, such as Accuracy, Precision, Recall and F1 Score.

3.2. **Data Collection**

The two datasets used in this study are publicly available on Kaggle, ensuring their accessibility for academic and research purposes. These datasets were chosen for their relevance to diabetes prediction and the variety of features they offer for analysis.

1. **Large dataset:**

This dataset is part of the CDC’s Behavioural Risk Factor Surveillance System (BRFSS), collected in 2015. It contains 253,680 survey responses from the U.S. residents, with 22 features associated with health indicators that are important in understanding diabetes risk. The dataset has a target variable of **Diabetes\_012** with three classes containing **0 for no diabetes**, **1 for prediabetes** and **2 for diabetes.**

See Link: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

1. **Small dataset:**

This dataset was collected in 1990 and was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and focuses on health indicators related to diabetes among the Pima Indians in Arizona, USA. It consists of 768 records with 8 features. The target variables indicates whether an individual has **diabetes** or **not,** making it suitable for binary classification tasks.

See Link: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

3.2.1. **Ethical Considerations**

i. **GDPR Compliance:** The two datasets used in this study are anonymized, ensuring that personal identifiable information (PII) cannot be traced back to individuals, in line with GDPR principles, such as data anonymization, lawful usage and fairness. Since these datasets are publicly available and do not involve active collection of personal data, they do not fall under stricter GDPR requirements related to personal data collection.

ii. **UH Ethical Policies:** As the datasets are publicly accessible and anonymized, University of Hertfordshire (UH) Ethics Committee approval was not required. This study relies on secondary data from Kaggle, which complies with UH ethical policies for responsible data usage in academic research, ensuring participant privacy and confidentiality.

1. **Permission to Use the Data:** The datasets are governed by Kaggle’s term of use, which allows for academic and non-commercial research purposes. There is no need to obtain explicit permission from individual participants, as the data is openly shared for research under these terms.
2. **Ethical Data Collection:** The study does not involve active data collection or surveys. The anonymised data ensures that no personal information is used inappropriately. The findings are intended solely for general knowledge and research purposes, not for diagnosing or treating individual cases, thus minimizing the risk of misusing the results.

3.3. **Exploratory Data Analysis (EDA)**

The exploratory data analysis (EDA) of both diabetes datasets revealed class imbalances in the target variables, which were addressed using Synthetic Minority Over-sampling Technique (SMOTE).

The large dataset offered a comprehensive analysis, identifying key risk factors for diabetes progression, such as high blood pressure, cholesterol and BMI, through count and bar plots. It also highlighted protective factors like physical activities and vegetable consumption, which were more common in non-diabetic individuals. Additionally, mental and physical health declined with diabetes progression. Demographic factors like age and income were linked to higher diabetes prevalence. Correlation analysis confirmed strong relationships between these factors and diabetes.

The small dataset focused on clinical features, including glucose, BMI, insulin levels, pregnancies and genetic predisposition. Bar plots showed that diabetic individuals had higher glucose, insulin and BMI levels, with clear relationships between pregnancies, age and diabetes. Correlation analysis confirmed strong association between glucose, BMI and age with diabetes progression.

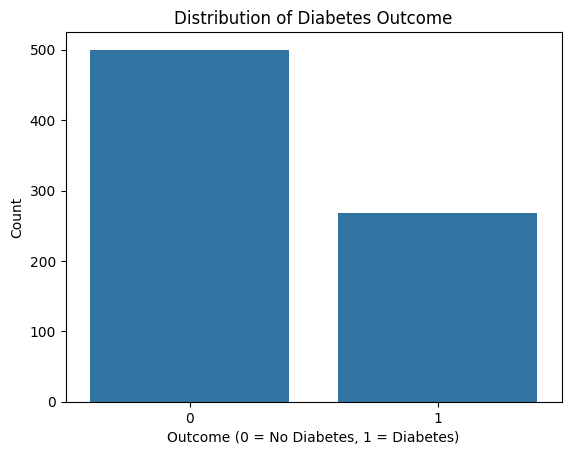
The following plots are particularly valuable, as they effectively highlight key insights relevant to the research objectives vis-à-vis the research questions for both diabetes datasets.

3.3.1. **Comparative Analysis of Key Insights using Visualizations from Large and Small Diabetes Datasets**

**Class Distribution of Target Variables**

Fig 1a. Large dataset Fig 1b. Small dataset

A graph of a diabetes

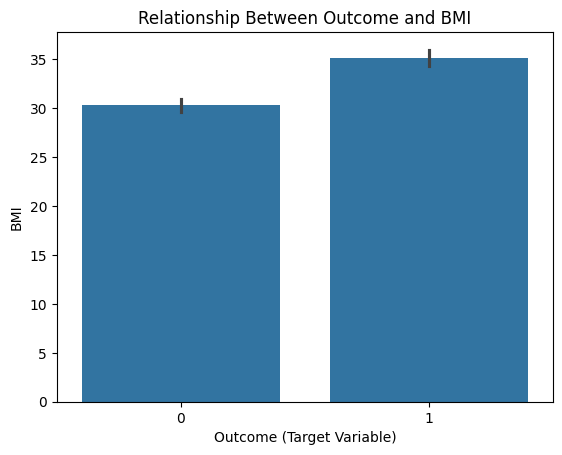
Description automatically generated with medium confidence

The large dataset shows a severe class imbalance, with 0.0 for non-diabetic at about 180,000 observations, 2.0 for diabetic at around 30,000 and 1.0 for prediabetic being negligible. The small dataset has a moderate imbalance with 500 non-diabetic and nearly 300 diabetic cases. To address this issue, SMOTE is crucial for balancing the minority classes for fair model performance.

**BMI Across Diabetes Categories**

Fig 2a. Large dataset Fig 2b. Small dataset

A graph of a couple of colored squares

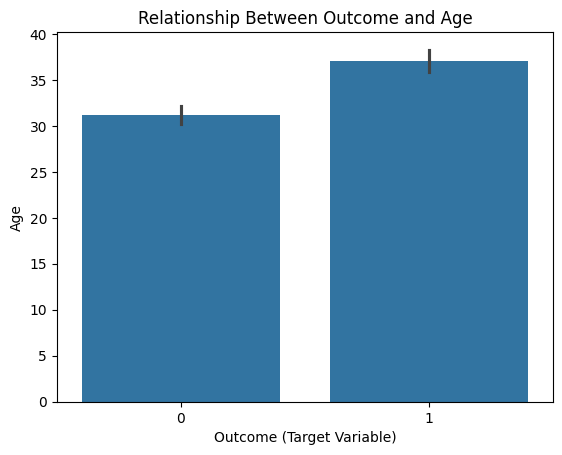
Description automatically generated with medium confidence

In the large dataset, the average BMI for non-diabetic individuals (0.0) is approximately 30, for prediabetics (1.0) is around 31 and for diabetics (2.0) it’s slightly exceeds 31. This indicates a positive correlation between BMI and diabetes progression. In the small dataset, non-diabetics (0) have an average BMI of 30, while diabetics (1) have a higher BMI of 35. Both plots show that higher BMI is strongly associated with diabetes, making it a key risk factor.

**Age vs. Target Variable (Age distribution Across Diabetes)**

Fig 3a. Large dataset Fig 3b. Small dataset

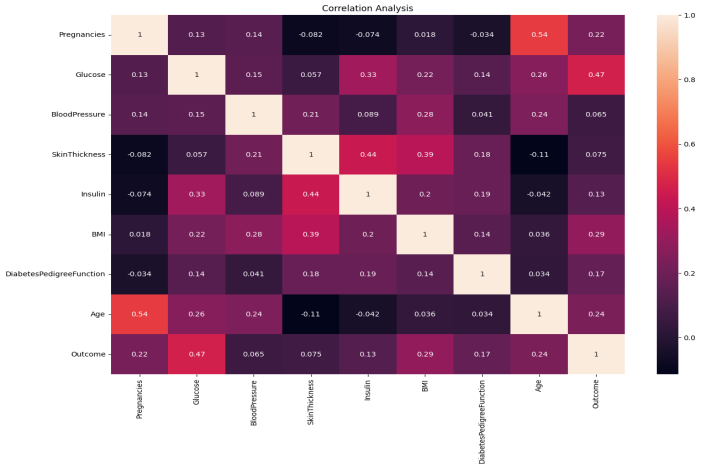
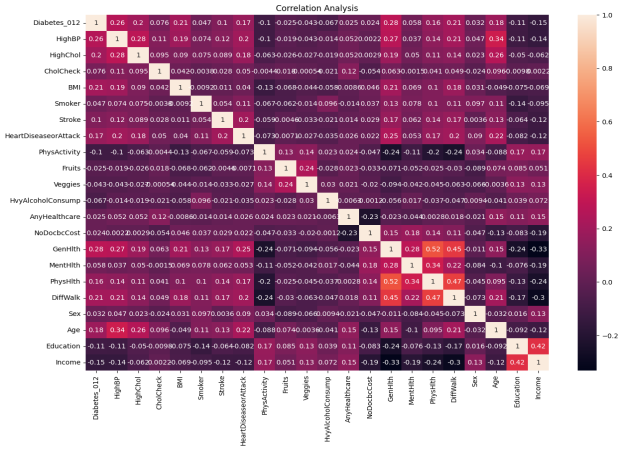
A graph of a number of bars

Description automatically generated with medium confidence

In the large dataset, the average age is 8 for non-diabetics (0.0), around 9 for pre-diabetics (1.0), and slightly above 9 for diabetics (2.0). The average age in the small dataset is approximately 30 for non-diabetics (0) and 35 for diabetics (1). This demonstrates that older individuals are at a higher risk of diabetes. Both plots confirm that age is strongly associated with diabetes progression, making it a critical feature for predictive modelling.

**Correlation Analysis of Features Influencing Diabetes Progression**

Fig 4a. Large dataset Fig 4b. Small dataset



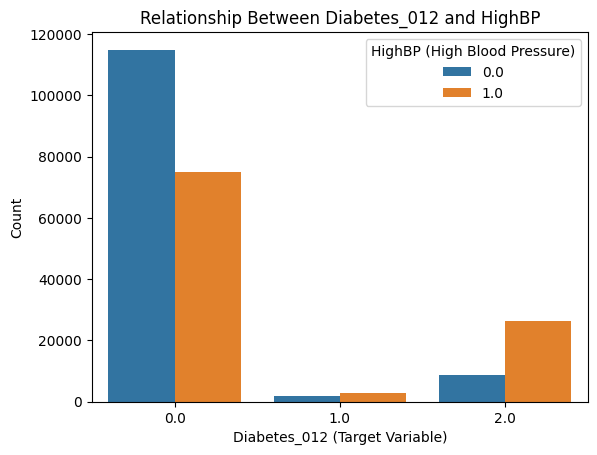
In the large dataset, the target variable Diabetes\_012 is moderately correlated with Age (0.34), High Blood Pressure (0.26) and Cholesterol (0.28), making them key predictors. Socioeconomic factors like Income and Education (0.42) also show a positive correlation with diabetes risk. General health is strongly correlated with physical health (0.52) and Difficulty Walking has strong positive correlations with both general health (0.45) and physical health (0.47). While Difficulty Walking (0.47) is more strongly correlated with diabetes progression than Age (0.34), it is likely a symptom of advanced diabetes rather than a cause. Note that correlation measures the strength of the relationships between two variables but does not imply causation.

In the Small dataset, the target variable Outcome is moderately correlated with Age (0.54), Glucose (0.47) and BMI (0.29), indicating these as the strongest predictors. Other variables show weak correlations and there is no significant multicollinearity. Both datasets emphasized that Age and BMI are consistent predators of diabetes.

3.3.2. **Additional Visualizations with Key Insights from Large and Small Diabetes Datasets**

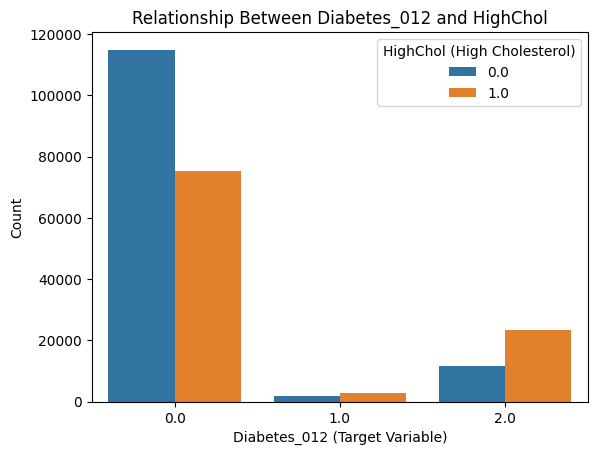
i. **Large dataset** (Blue bar = No, Orange bar = Yes)

**Fig 1.** **Relationship Between Diabetes and High Blood Pressure**



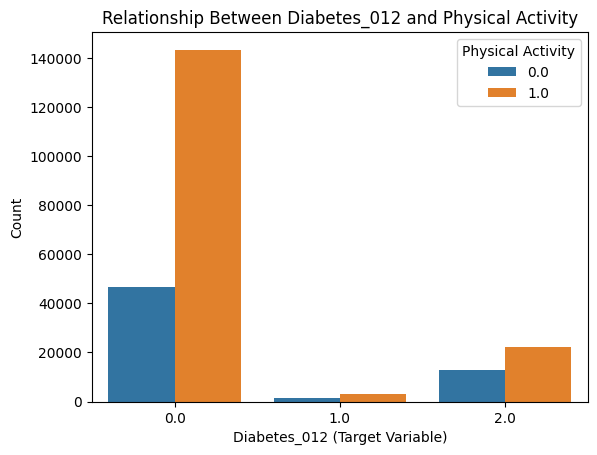
In the above chart, among the non-diabetics, approximately 61% (110,000) do not have high blood pressure, while about 39% (70,000) do. In pre-diabetics, high blood pressure is slightly more common. Among diabetics, nearly 71% (25,000) have high blood pressure and about 29% (1o,000) do not. This trend aligns with a moderate correlation of 0.26 between high blood pressure and diabetes progression, highlighting high blood pressure as a significant risk factor.

**Fig 2.** **Prevalence of High Cholesterol Across Diabetes Progression**



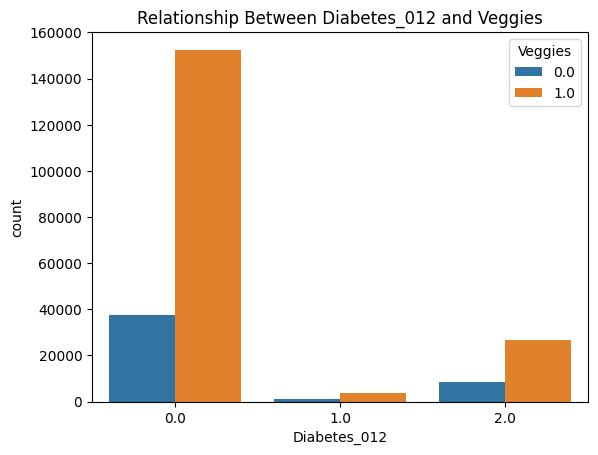
The chart illustrates that, among non-diabetics (0.0), approximately 110,000 individuals (61%) do not have high cholesterol, while about 70,000 (39%) do. In the pre-diabetic group (1.0), high cholesterol affects a slight majority of about 51%. Among diabetics (2.0), nearly 25,000 individuals (71%) have high cholesterol, whereas about 10,000 (29%) do not. This trend indicates high cholesterol as a major risk factor.

**Fig 3.** **Physical Activity and Diabetes Progression**



The chart demonstrates that physical activity declines slightly with diabetes progression. Among non-diabetics (0.0), 74% (140,000) are active and 26% (50,000) are not. For pre-diabetics (1.0), 55% are active and 45% are inactive. Among diabetics (2.0), 63% (25,00) remain active and 37% (15,000) are inactive. The weak negative correlation (-0.10) suggests physical activity offers modest protection against diabetes.

**Fig 4.** **Relationship between Vegetable Consumption Across Diabetes Progression**



In the plot, vegetable consumption slightly decreases with diabetes progression. Among non-diabetics (0.0), approximately 74% consume vegetables, while about 26% do not. Among diabetics (2.0), about 71% consume vegetables and 29% are not. A weak negative correlation (-0.14) suggests a minor protective effect of vegetable consumption against diabetes.

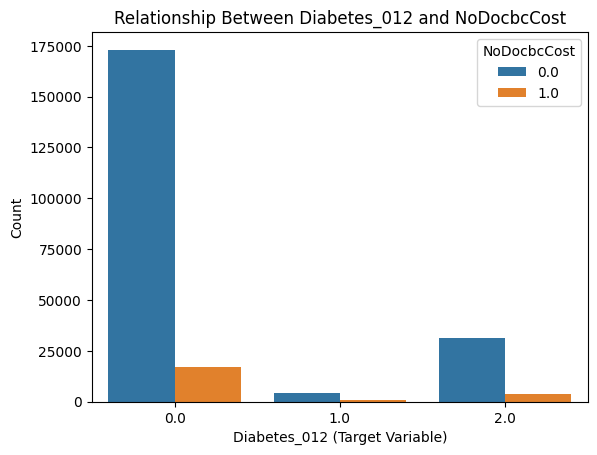
**Fig 5.** **Income Levels Across Diabetes Progression Categories**

A graph of a number of bars

Description automatically generated with medium confidence

The plot shows that income decreases by approximately 8% as diabetes progresses, with non-diabetics averaging about 6 and diabetes about 5.5. A weak negative correlation (-0.15) links lower income to higher diabetes risk, showcasing potential socioeconomic barriers.

**Fig 6.** **Cost-Related Barriers to Healthcare Across Diabetes Progression**



Cost-related barriers slightly increase with diabetes progression. Among non-diabetics, 88% had no barriers, compared to 78% for diabetics. This reflects a weak positive correlation of 0.07.

**Fig 7.** **Impact of Diabetes Progression on Mental and Physical Health**

7a. Mental Health 7b. Physical Health

A graph of a couple of colored bars

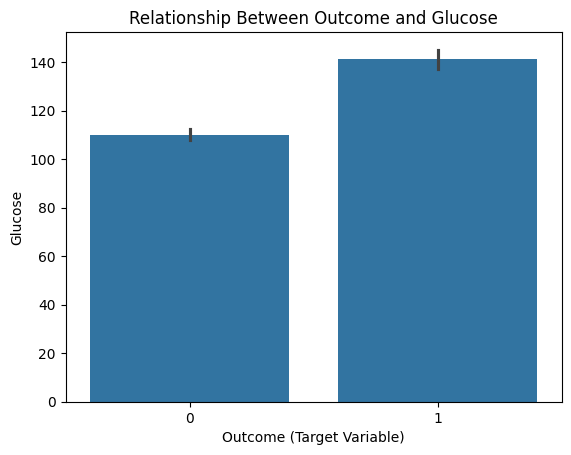
Description automatically generated with medium confidenceA graph of a number of people with diabetes

Description automatically generated with medium confidence

As diabetes progresses, mental health worsens by 50% and physical health by 167%. Physical health shows a stronger decline. Positive correlations stood at 0.16(weak) and 0.27 (moderate) for mental and physical health respectively.

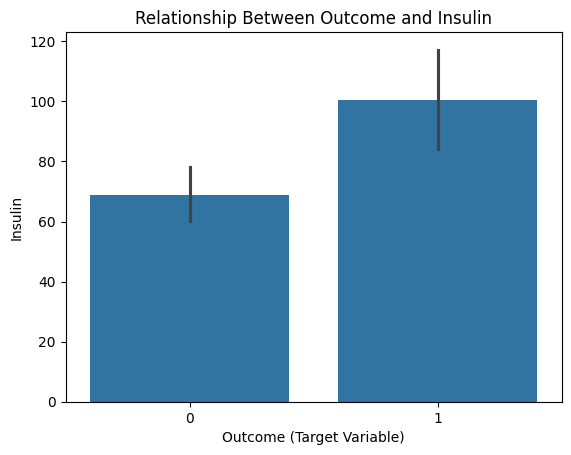
**ii. Small Dataset**

**Fig 1. Relationship between Diabetes Presence and Glucose Levels**



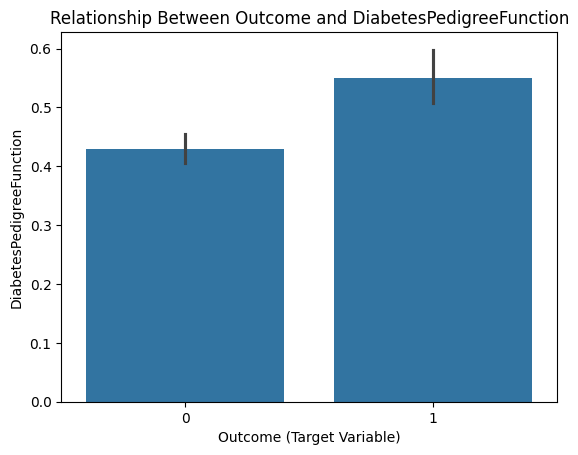
Non-diabetics account for about 44% of average glucose levels, diabetics for 56%, with about 27% (110 mg/dL) glucose increase from non-diabetics to diabetics. Its strong positive correlation of 0.47, confirms glucose levels as a key diabetes predictor.

**Fig 2. Diabetes Presence and Insulin Levels**



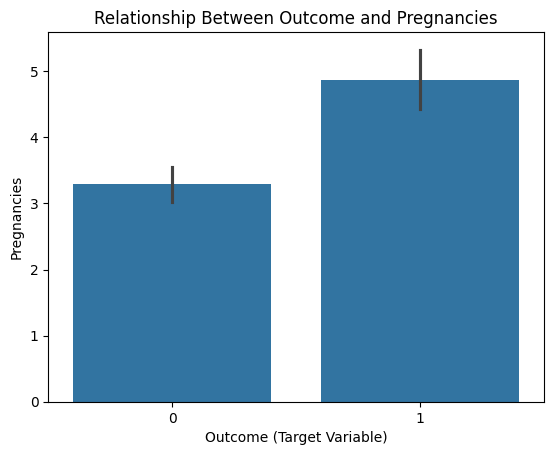
From the plot, diabetic individuals have about 43% higher average insulin level (600 pmol/L) compared to non-diabetics (420 pmol/L). A weak positive correlation (0.13) signifies a modest association between high insulin levels and diabetes.

**Fig 3. Diabetes Presence and Genetic Predisposition**



Diabetic individuals have a 31% higher average Diabetes Pedigree Function (0.55) compared to non-diabetics (0.42). A moderate positive correlation of 0.17 points out the role of genetic predisposition in diabetes risk.

**Fig 4. Diabetes Presence and Number of Pregnancies**



The plot reveals that people with diabetes have 67% more pregnancies on average (5) than non-diabetics individuals (3). A moderate positive correlation (0.22) expresses a link between pregnancies and diabetes risk, potentially due to gestational diabetes.

3.4. **Data Preprocessing**

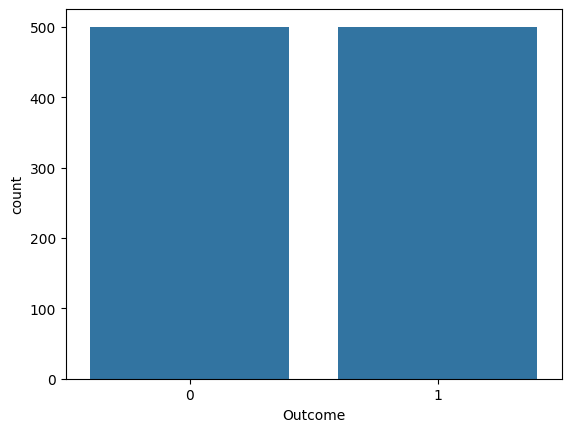
Preprocessing is a vital step that directly impacts the accuracy, efficiency and interpretability of machine learning models, by ensuring that the data is appropriately prepared and suitable. The following preprocessing steps were carried out.

1. **Data Cleaning:** The large dataset contained duplicate values; this was handled by dropping the duplicates. The small dataset does not contain duplicates.
2. **Handling Class Imbalance:** Both datasets showed class imbalances, this was addressed by applying Synthetic Minority Over-Sampling Techniques (SMOTE) to balance the classes and ensure that the models could effectively learn from all classes.
3. **Data Splitting:** I split the dataset into training and testing sets, using an 80-20 split. I used the training data to train the models and the testing data to evaluate their performance.
4. **Feature Scaling:** I applied StandardScaler to normalize the features so they have a mean of 0 and a standard deviation of 1. The scaler is fitted to the training data and applied to the testing data

**Imbalanced Data Handled Using SMOTE**

**Fig 1a. Large Dataset Fig 1b. Small Dataset**

**A graph of a diabetes

Description automatically generated with medium confidence**

The two datasets are now evenly distributed. The application of SMOTE effectively balances the class distributions, ensuring fair training of the models and reducing bias toward majority classes.

3.5. **Machine Learning Algorithms**

In this project, I used three different machine learning algorithms to predict diabetes status and evaluate their effectiveness, as they are fundamental tools in extracting patterns and making predictions from data. They include Random Forest, XGBoost and Support Vector Machine (SMV). These algorithms stand out for their capabilities to handle classification and regression tasks effectively.

1. **Random Forest:** This is an ensemble learning algorithms that build multiple decision trees during training and merges their outputs (averaging for regression or majority vote for classification). It operates on the principle of bagging (Bootstrap Aggregation), where subsets of the data are created with replacement and used to train individual decision trees. During tree construction, features for splits are randomly selected which introduces diversity and reduces correlation among trees.

When applied to diabetes large dataset, Random Forest handles the high dimensionality of features effectively, providing robust predictions while mitigating the risk of overfitting. It can effectively manage missing data and noise, making it suitable for real-world healthcare datasets. For a small dataset, Random Forest’s ensemble approach stabilizes predictions by reducing variance, though care must be taken to avoid overfitting due to limited data.

1. **XGBoost:** The eXtreme Graadient Boosting (XGBoost) is a high-performance implementation of gradient boosting that focuses on both speed and accuracy. It works by building decision trees subsequentially, where each new tree aims to correct the errors of the previous one. This is achieved by minimizing a loss function, such as mean squared error for regression or log loss for classification, through gradient descent.

With large diabetes dataset, XGBoost excels in delivering high predictive accuracy by effectively handling missing values and leveraging parallel processing for faster training. The algorithm’s regularization techniques ensure robust generalization. This, even with complex feature interactions common in large medical datasets. For small diabetes dataset, XGBoost can still perform well but overfitting is a potential concern. Proper hyperparameter tuning, such as adjusting the learning rate and depth becomes key to prevent overfitting and optimize performance.

1. **Support Vector Machine (SVM):** SVM is a supervised learning algorithm designed to solve classification and regression problems by identifying the optimal hyperplane that separates data into classes. The objective is to maximize the margin which is the distance between the hyperplane and the nearest data point. (Support Vector) from each class. This ensures better generalization and robustness to small changes in the dataset.

In the case of a large diabetes dataset, SVM’s performance can be computationally intensive due to the high dimensionality and size of the data. The use of kernel functions, while powerful may significantly increase computation time, making it less practical for very large dataset without dimensionality reduction techniques. For a small diabetes dataset, SVM has the ability to handle high-dimensional spaces and its reliance on a few support vectors makes it a strong candidate. With appropriate kernel selection and parameter tuning, SVM can achieve high accuracy even with limited data, making it ideal for smaller well-defined datasets.

* 1. **Evaluation Metrics**

To evaluate the performance of the models, I used the confusion matrix and the classification report as basic tools to calculate the performance of the models.

1. **Confusion Matrix:** This is a table that displays the performance of a classification model by comparing the predicted and the actual values (true) or actual class labels. It has four components. It visualizes the breakdown of correct and incorrect predictions for each class. This helps to identify where the model is performing or where it is making errors.
2. True Positives (TP): Correctly predicted positive instances.
3. True Negatives (TN): Correctly predicted negative instances
4. False Positives (FP): Incorrectly predicted positive instances (Type I error)
5. False Negative FN): Incorrectly predicted negative instances (Type II)
6. **Classification Report:** This provides metrics for evaluating the performance of a classification model. The key metrics typically include:
7. **Accuracy:** This is the ratio of correctly predicted instances to the total instances.

**Accuracy =**

1. **Precision:** This is the proportion of true positive predictions out of all positive predictions made by the model. It evaluates the correctness of positive prediction.

**Precision =**

1. **Recall (Sensitivity):** This is the proportion of actual positives that the model correctly identified. It is the ratio of true positives to the sum of true positives and false negatives

**Recall =**

1. **F1-Score:** This is the harmonic mean of precision and recall, used to balance the two metrics.

**F1-Score = 2 x**

* **Support:**  In a classification report, support is not a performance metric. It refers to the number of actual occurrences of each class in the dataset. It simply shows the size of the actual dataset for each class. It is the count of instances in the test set for each class and serves as a baseline to measure the performance of the model for each specific class. The larger the support, the more reliable the performance, especially with imbalance datasets.
* **Macro Average:** This calculates the metric (e.g., precision, recall and F1-score) independently for each class and then takes the unweighted mean of these metrics.
* **Weighted Average:** This calculates the metric for each class and then takes the mean, weighted by the support (number of samples) for each class.

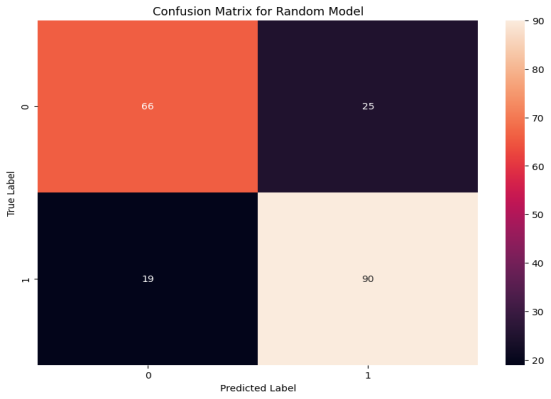
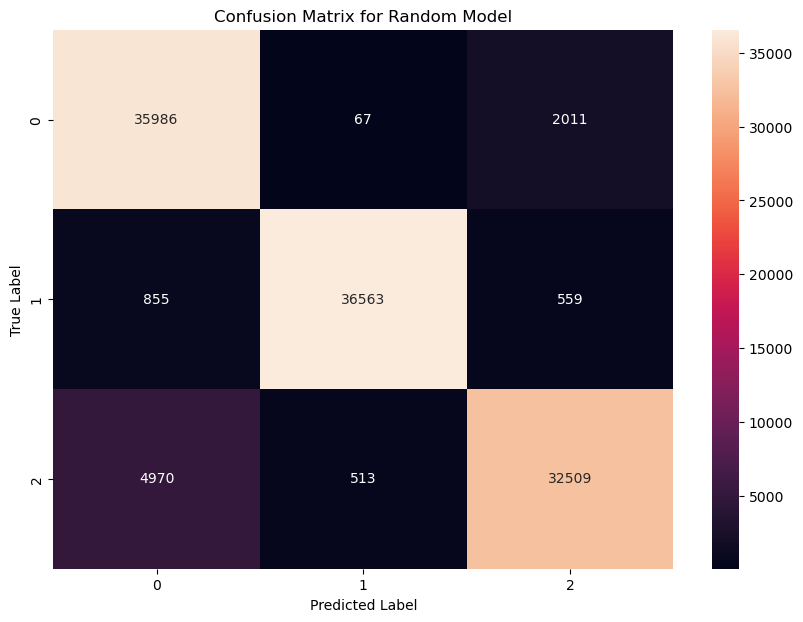
4.0 **Results and Analysis**

In this research, I employed multiple models, namely Random Forest, XGBoost and Support Vector Machine, to predict diabetes progression using two datasets, the BRFSS 2015 Diabetes dataset (large) and the Pima Indians Diabetes dataset (small). These models were evaluated using key performance metrics such as accuracy, precision, recall and F1-score for each class. The target class for BRFSS dataset (0 for non-diabetic, 1 for pre-diabetic and 2 for diabetic) and the Pima Indians dataset (0 for non-diabetic and 1 for diabetic). The metrics provide insight into the model’s ability to correctly classify the target class for both datasets, which is critical for predicting diabetes progression effectively.

4.1. **Analysis of the Models using Confusion Matrix**

1. **Random Forest**

**Fig 1a. Large dataset (BRFSS) Fig 1b. Small dataset (Pima Indians)**



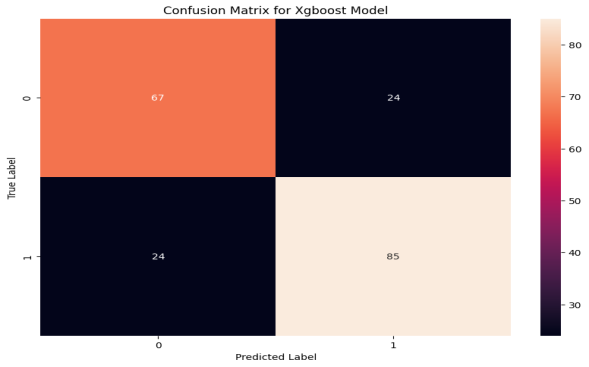
The large dataset being a multi-class classification task, shows strong performance for majority classes (0 and 1), with high true positives (TP) and true negatives (TN). Class 1 achieves 36,563 correctly predicted instances. However, it struggles with the minority class (class 2), leading to significant false negatives FN), such as 4,970 instances of class 2 being misclassified as class 0. Similarly, false positives (FP) are prominent for class 2, where 2,011 instances of class 0 and 559 of class 1, are incorrectly predicted as class 2. These challenges highlight difficulties in handling class imbalance and feature overlap in a large-scale dataset.

In contrast, the small dataset being binary, achieves excellent performance with minimal FN and FP rates. Class 1 is well detected with 90 true positives, while FN (19 instances of class 1 misclassified as class 0) and FP (25 instances of class 0 misclassified as class 1) are low. True negatives (TN) are also high with 66 correctly identified for class 0. The model demonstrates robust accuracy and precision for this simpler classification task, benefiting from the small dataset size and balanced class distribution, unlike the large dataset.

**2. XGBoost**

**Fig 2a. Large dataset Fig 2b. Small Dataset**

**A screenshot of a computer screen

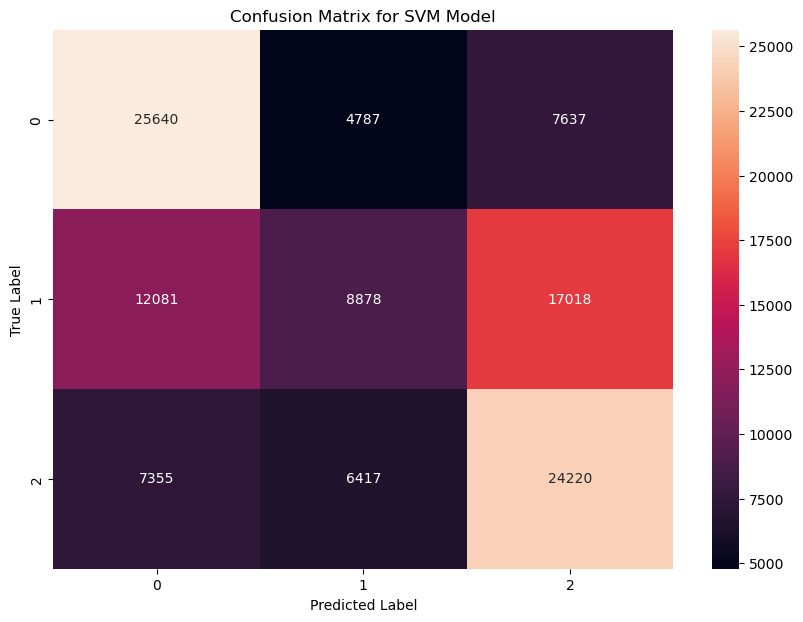
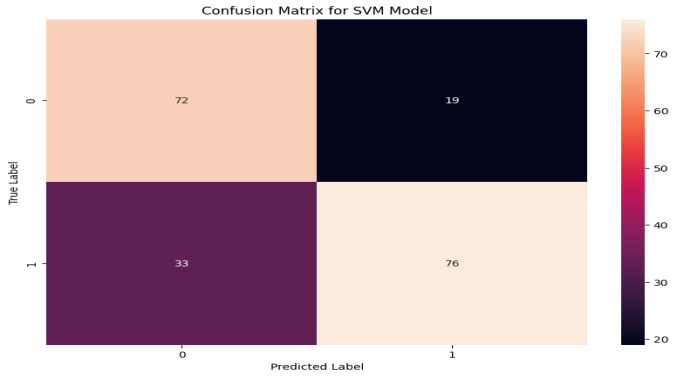
Description automatically generated**

The first confusion matrix represents the performance of an XGBoost model on a large dataset with three classes. (0,1, 2). The model demonstrates good performance with majority of the predictions falling on the diagonal, indicating correct classification. For example, 36,602 instances of class 0, 32,924 of class 1 and 25054 of class 2 were correctly predicted. However, misclassifications are observed, particularly for class 2, which has significant confusion with other classes (7,429 misclassified as class 1). The scale of the dataset with prediction in the thousands highlights the model’s capacity to handle complex dataset, but also emphasizes the need to address class specific misclassification issues.

The second confusion matrix is of a small dataset, involving two classes (0 and 1). The model achieves relatively balanced predictions, with 67 instances of class 0 and 85 of class 1 correctly classified. However, misclassifications are notable with 24 instances of each class being incorrectly predicted as the other. The small scale of the dataset means these misclassifications have a large proportional impact on performance evaluation. Overall, while the model performs well for its size, it reveals challenges such as limited data impacting classification reliability.

1. **SVM**

**Fig 3a. Large dataset Fig 3b. Small dataset**

****

The first confusion matrix measures the SVM model’s performance on a large dataset with three classes (0, 1, 2). The model achieves reasonable accuracy for classes 0 and 2 with 25,640 and 24,220 correct predictions respectively. However, it struggles significantly with class 1, where 12,081 instances are misclassified as class 0 and 17,018 misclassified as class 2. These substantial misclassifications suggest challenges in separating class 1 from the others, possibly due to overlapping features or class imbalance. The large dataset size with predictions in the tens thousands, stresses the need for strategies like better feature selection, class balancing or optimization to improve performance.

The second matrix reflects model’s performance on a small dataset, which achieves balanced predictions with 72 and 76 correct classifications for classes 0 and 1 respectively. It represents a notable proportion of misclassifications. 19 instances of class 0 are misclassified as class 1 and 33 of class 1 as class 0. These misclassifications have a large impact due to the dataset’s small size, emphasizing the importance of careful tuning and potentially augmenting the data to improve model reliability and accuracy.

**4.2 Classification Report Analysis**

**1. Random Forest**

**Table 1a. Large Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.86** | **0.95** | **0.90** | **38064** |
| **1** | **0.98** | **0.96** | **0.97** | **37977** |
| **2** | **0.93** | **0.86** | **0.89** | **37992** |
| **Accuracy** |  |  | **0.92** | **114033** |
| **Macro avg** | **0.92** | **0.92** | **0.92** | **114033** |
| **Weighted avg** | **0.92** | **0.92** | **0.92** | **114033** |

**Table 1b. Small dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.78** | **0.73** | **0.75** | **91** |
| **1** | **0.78** | **0.83** | **0.80** | **109** |
| **Accuracy** |  |  | **0.78** | **200** |
| **Macro avg** | **0.78** | **0.78** | **0.78** | **200** |
| **Weighted avg** | **0.78** | **0.78** | **0.78** | **200** |

The first classification report assesses the performance of a Random Forest model on a large dataset with three classes (0, 1, 2). The model achieves high overall accuracy of 92%, with macro and weighted averages for precision, recall and F1-score all at 0.92. Class 1 performs particularly well, with an F1-score of 0.97, reflecting excellent balance between precision and recall. However, Class 2 has a slight lower recall of 0.86, indicating some challenges in identifying all instances of this class. Overall, the results demonstrate the model’s effectiveness in handling large datasets with strong class-level performance.

In contrast, the second table is for model performance on a small dataset with two classes (0 and 1). The overall performance is lower with an accuracy, macro average and weighted average of 0.78 for precision, recall and F1-score. Class 1 performs slightly better with an F1-score of 0.80 compared to 0.75 for class 0. The lower scores highlight the challenges of using a Random Forest model on small datasets, where limited data can restrict the model’s ability to generalize effectively. This comparison underscores the model’s scalability and the influence of dataset size on its predictive performance with stronger results observed in large datasets.

**2. XGBoost**

**Table 2a. Large dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.85** | **0.96** | **0.90** | **38064** |
| **1** | **0.82** | **0.87** | **0.84** | **37977** |
| **2** | **0.82** | **0.66** | **0.73** | **37992** |
| **Accuracy** |  |  | **0.83** | **114033** |
| **Macro avg** | **0.83** | **0.83** | **0.82** | **114033** |
| **Weighted avg** | **0.83** | **0.83** | **0.82** | **114033** |

**Table 2b. Small dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.74** | **0.74** | **0.74** | **91** |
| **1** | **0.78** | **0.78** | **0.78** | **109** |
| **Accuracy** |  |  | **0.76** | **200** |
| **Macro avg** | **0.76** | **0.76** | **0.76** | **200** |
| **Weighted avg** | **0.76** | **0.76** | **0.76** | **200** |

Table 2a illustrates performance of the XGBoost model on a large dataset with three classes (0, 1, 2). The overall accuracy is 83% with a macro and weighted average F1-score of 0.82, indicating strong but not perfect performance. Class 0 achieves the highest F1-score of 0.90, benefiting from a high recall of 0.96, while class 2 has the lowest F1-score of 0.73, due to a relatively low recall of 0.66. This suggests that the model struggles to correctly classify some instances of class 2, likely due to overlapping features or imbalanced data. These results show the model’s capability to handle large datasets effectively, though improvements in class-specific recall especially for class 2, could enhance overall performance.

The second classification report summarises the model’s performance on the small dataset with two classes (0 and 1). The overall accuracy, macro average and weighted average F1-score are lower at 76%, reflecting the limitations of working with a small dataset. Class 1 performs slightly better with an F1-score of 0.78 compared to 0.74 for class 0. The balanced but relatively modest precision and recall values for both classes suggest that the model’s ability to generalize is constrained by the limited size of the dataset. These results illustrate how dataset size impacts the XGBoost model’s performance, with better generalization in the large dataset compared to the small one.

1. **SVM**

**Table 3a. Large dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.57** | **0.67** | **0.62** | **38064** |
| **1** | **0.44** | **0.23** | **0.31** | **37977** |
| **2** | **0.50** | **0.64** | **0.56** | **37992** |
| **Accuracy** |  |  | **0.52** | **114033** |
| **Macro avg** | **0.50** | **0.51** | **0.49** | **114033** |
| **Weighted avg** | **0.50** | **0.52** | **0.49** | **114033** |

**Table 3b. Small dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.69** | **0.79** | **0.73** | **91** |
| **1** | **0.80** | **0.70** | **0.75** | **109** |
| **Accuracy** |  |  | **0.74** | **200** |
| **Macro avg** | **0.74** | **0.74** | **0.74** | **200** |
| **Weighted avg** | **0.75** | **0.74** | **0.74** | **200** |

Table 3a evaluates the performance of an SVM model on the large dataset, having overall accuracy of 52%, with macro and weighted averages for precision, recall and F-score around 0.50, indicating poor overall performance. Class 0 achieves the highest F1-score of 0.62, due to relatively better recall of 0.67. But classes 1 and 2 perform poorly, particularly class 1, which has a recall of just 0.23. These results suggest significant difficulty in distinguishing between the classes, likely due to the complexity of the dataset or limitations in the SVM model’s ability to handle complex multi-class problems effectively.

The second report explains the model’s performance on the small dataset, having overall accuracy of 74%, with macro and weighted averages for precision, recall and F1-score also at 0.74, showing a moderate improvement compared to the larger dataset. Class 1 performs slightly better, achieving an F1-score of 0.75 compared to 0.73 for class 0. These results express that while the SVM model struggles with larger and more complex datasets, it can perform adequately on small, simpler tasks with fewer classes. This comparison highlights the importance of dataset size and complexity in evaluating SVM performance and suggests the need for alternative approaches or model optimizations for large datasets.

**4.3. Optimization of Models for both Datasets Using Hyperparameter Tuning**

Model optimization refers to the process of improving model’s performance by fun-tuning its parameters, hyperparameters and structure to better fit the underlying data. Hyperparameter tuning is a crucial step in enhancing the performance of models. By systematically optimizing parameters, models can better adapt to the underlying patterns in the data. The table below shows the results summary of the models’ optimization using hyperparameter tuning.

Table 1. **Results Summary of Models’ Optimization for the Two Datasets**

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Datasets** | **Stratified Sampling**  **(Yes or No)** | **Accuracy (%)** |
| Random Forest | Large | Yes (5% subset) | 76.6 |
|  | Small | No | 80.1 |
| XGBoost | Large | Yes (5% subset) | 76.3 |
|  | Small | No | 80.5 |
| SVM | Large | Yes (5% subset) | 47.8 |
|  | Small | No | 76.0 |

In this project, I performed optimization using hyperparameter tuning. Random Forest, XGBoost and SVM were evaluated on large and small diabetes datasets. On the large dataset, hyperparameter tuning could not execute on the full dataset until a 5% stratified subset, due to computational constraints. Random Forest achieved the highest accuracy of 76.6%, followed by XGBoost with 76.3%, while SVM struggled with an accuracy of 47.8%, reflecting its limitations with large-scale data.

On the small dataset, which did not require stratified sampling, XGBoost performed best with 80.5%, followed by Random Forest and SVM with 80.1% and 76.0% respectively. These results highlight XGBoost’s efficiency in small datasets, while Random Forest demonstrated consistent reliability across dataset sizes. SVM performed well on small dataset but struggled significantly with the large one.

**4.4. Model Performance Overview**

The table below shows the results summary before and after mode’s optimization for the two diabetes datasets.

Table 1. **Results Summary before and after Model’s Optimization for the two Datasets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Datasets** | **Stratified Sampling (Yes or No)** | **Accuracy before Optimization (%)** | **Accuracy after Optimization (%)** |
| Random Forest | Large | Yes (5% subset) | 92.0 | 76.6 |
|  | Small | No | 78.0 | 80.1 |
| XGBoost | Large | Yes (5% subset) | 83.0 | 76.3 |
|  | Small | No | 76.0 | 80.5 |
| SVM | Large | Yes (5% subset) | 52.0 | 47.8 |
|  | Small | No | 74.0 | 76.0 |

The performance summary illustrates that Random Forest and XGBoost consistently outperformed SVM. For the large dataset, a 5% stratified subset of the data was evaluated. After optimization, Random Forest and XGBoost’s accuracies dropped from 92.0% to 76.6% and 83.0% to 76.3% respectively, likely due to challenges with tuning on a subset. SVM lagged significantly with accuracy declining from 52.0% to 47.8%, indicating its limitations in handling complex, non-linear relationships even after optimization.

For the small dataset, Random Forest improved from 78.0% to 80.1%, and XGBoost achieved the highest accuracy, increasing from 76.0% to 80.5% after optimization. SVM also improved moderately from 74.0% to 76.0%. The results emphasize that Random Forest and XGBoost are more robust across dataset sizes. Particularly effective in capturing intricate patterns of non-linear relationships, with XGBoost excelling after tuning on small dataset, while SVM remains more suitable for small tasks.

**5. Analysis and Discussion**

5.1. **Discussion**

**What do the Results Mean:** The results from this study provide comprehensive understanding into the performance of three Machine Learning models, such as Random Forest, XGBoost and SVM, on two datasets of varying size and complexity. Random Forest and XGBoost outperformed SVM across both datasets. On the large dataset, performance declined after optimization due to tuning on a 5% stratified subset, with Random Forest dropping from 92.0% to 76.6% and XGBoost from 83.0% to 76.3%. This reflects challenges in capturing the full dataset’s complexity. SVM performed poorly with accuracy falling from 52.0% to 47.8%, showcasing its inability to handle complex, non-linear relationships in large tasks.

On the small dataset, Random Forest improved from 78.0% to 80.1%, and XGBoost achieved the highest accuracy, increasing from 76.0% to 80.5%. SVM also improved moderately from 74.0% to 76.0%, demonstrating better suitability for smaller, simpler tasks.

**What Model Works the best and why:** From the results, Random Forest emerged as the overall best-performing model due to its consistent performance across both datasets. Its ensemble approach, which combines multiple decision trees, allows it to reduce overfitting and improve generalization. Why Random Forest emerged best also include its robustness to noise, ability to handle both categorical and continuous data and resilience to class imbalance when paired with strategies like SMOTE. These characteristics makes it versatile and effective for diverse datasets.

5.2. **Comparison with the Literature**

The results align with prior research emphasizing the superior performance of the ensemble methods like Random Forest and XGBoost for diabetes prediction. Studies by Kasula (2023) and Rani (2020) reported high accuracies of 85% and 99% respectively, for Random Forest and Decision Trees, supporting this study’s findings. However, the challenges with XGBoost’s recall for the diabetic class and SVM’s overall poor performance are consistent with findings from Kasturi (2024), which highlighted the difficulty of handling imbalanced datasets. The use of SMOTE helped mitigate these issues but did not completely resolve them. Compared to the literature, the results for SVM were lower, reflecting the model’s sensitivity to dataset size and complexity.

5.3. **Limitations of the Results**

One major limitation is the persistent class imbalance, particularly with the large dataset. Despite the application of SMOTE, this imbalance negatively affected XGBoost’s recall for the diabetic class. The second challenge is the tuning on a 5% subset for the large dataset, which negatively impacted the optimization process for Random Forest and XGBoost, leading to reduced accuracy. Thirdly, the small size of the Pima Indians Diabetes dataset also restricted the model’s ability to generalize effectively. Moreover, the interpretability of Random Forest and XGBoost remains a challenge as these models provide high accuracy but lack transparency compared to simpler models like Logistics Regression.

5.4. **Relation to Project Objectives and Answering the Research Question**

This investigation met its objectives by evaluating model performance and uncovering key predators like BMI, glucose levels, and age through EDA. Random Forest was identified as the most accurate and reliable model, with XGBoost performing well on small dataset. The Research question was addressed by showcasing Random Forest’s superiority and providing insights into feature relationships with diabetes progression.

5.5. **Relation to Project Application and Practical Use of the Models**

Random Forest and XGBoost demonstrated strong practical applicability in healthcare for diabetes prediction. Random Forest high accuracy, robustness and ability to handle complex data make it ideal for early detection, risk identification and management strategies. XGBoost, with further tuning is also effective particularly for smaller datasets. Insights from EDA, such as BMI’s correlation with diabetes progression, contribute actionable knowledge. However, SMV’s poor performance makes it unsuitable for practical deployment in this context.

5.6. **Conclusion**

**Summary of Key Results:** Random Forest and XGBoost consistently outperformed SVM in predicting diabetes progression. On the large dataset, Random Forest and XGBoost accuracy decreased after optimization on a 5% subset of the data from 92.0% to 76.6% and 83.0% to 76.3% respectively. On the small dataset, Random Forest’s accuracy increased from 78.0% to 80.1%, and XGBoost achieved the highest accuracy of 80.5%. SVM performed poorly on the large dataset (47.8%) but showed moderate improvement on the small dataset (76.0%)

**Justifiable Conclusions:** Random Forest proved to be the most reliable and effective model, with XGBoost excelling on small dataset. SVM struggled with complex and imbalanced data making it less suitable for this task.

**Application and Real-World Use:** Random Forest and XGBoost can be used in healthcare systems for early diabetes detection, risk assessment and tailored interventions. Insights from EDA can guide prevention programs and public health initiatives.

**Future Work:** Future research should focus on improving model interpretability, addressing class imbalance with advanced techniques and exploring Neural Networks or hybrid models for enhanced performance. Expanding datasets to include lifestyle and genetic factors can further refine predictions.