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

**1.0. Introduction**

**1.1. Background**

Diabetes mellitus (DM) also termed Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels due to inadequate insulin production or ineffective insulin use by the body. This imbalance, if left unmanaged, can lead to hyperglycemia and severe health complications, including cardiovascular disease, neuropathy, retinopathy, kidney failure, and amputations. The disease’s increasing prevalence presents a global health crisis, estimating that diabetes cases will surpass 600 million by 2040 (World Health Organization, 2023). The alarming projection of DM highlights the urgent need for preventive measures, early detection, and effective management strategies (Hounguè & Bigirimana, 2022).

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

Machine learning approaches are increasingly applied to medical fields due to their ability to process vast datasets and uncover valuable insights. By leveraging ML models like Random Forest, XGBoost, and Support Vector Machines (SVM), researchers and practitioners can develop predictive models that are more accurate and reliable than traditional methods. For example, models developed by Kasula (2023) and Rani (2020) achieved high accuracy rates—85% for Random Forest and 99% for Decision Trees, demonstrating the effectiveness of ML in diabetes prediction. Additionally, ensemble techniques, such as those used by Jain et al. (2024), have proven effective in improving model performance across diverse patient demographics. The integration of ML in diabetes prediction thus represents a promising step towards more precise, individualized healthcare interventions.

**1.2. Problem Statement**

Diabetes is a complex, multifactorial disease that is influenced by an array of genetic, environmental, and lifestyle factors, making it challenging to detect in its early stages using conventional diagnostic methods alone. While traditional methods provide a foundation for diabetes diagnosis, they often fail to identify individuals at-risk before the disease progresses, thereby delaying treatment and increasing the likelihood of complications. Furthermore, the linear nature of these methods limits their effectiveness in capturing the non-linear and complex relationships among various predictors of diabetes progression.

The use of machine learning in diabetes prediction aims to address these limitations by providing tools that can analyze diverse data sources and capture intricate patterns associated with disease onset and progression. Despite its potential, there remains a significant gap in understanding which machine learning models offer the most accurate, robust, and reliable predictions for diabetes progression across varied patient groups. Additionally, outliers, feature selection and data imbalances commonly found in medical datasets can impact model performance, further complicating the development of effective ML-driven prediction systems. This study aims to address these challenges by evaluating and comparing the performance of three machine learning algorithms—Random Forest, XGBoost, and SVM—in predicting diabetes progression. Furthermore, it seeks to identify patterns and insights through Exploratory Data Analysis (EDA) that could guide preventive and personalized interventions.

**1.3. Justification of the Study**

The growing prevalence of diabetes and the associated health risks underscore the necessity for innovative, precise, and proactive approaches to its detection and management. Early and accurate diagnosis is critical for effective intervention, as it enables healthcare providers to mitigate complications and improve patient outcomes. Machine learning offers a transformative potential in this domain by providing the means to identify individuals at-risk, even before clinical symptoms fully manifest, through the analysis of multi-dimensional data and complex correlations.

This study is justified by the need to evaluate which machine learning models are best suited for diabetes prediction, especially in a clinical context where decision-making can have life-altering consequences. By examining the performance of Random Forest, XGBoost, and SVM, this research seeks to identify the model that best balances accuracy, sensitivity, and generalizability across different patient demographics and data environments. Moreover, through Exploratory Data Analysis, the study aims to uncover trends, patterns, and associations related to diabetes progression, potentially offering healthcare practitioners actionable insights that can inform treatment and prevention strategies.

By contributing to the growing body of knowledge on machine learning in healthcare, this research has the potential to support the development of predictive tools that can aid healthcare professionals in early detection, proactive management, and personalized care for diabetes patients. The insights gained from this study may not only improve individual health outcomes but also provide a valuable resource for designing more effective, data-driven healthcare 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 diabetes progression.

ii. To perform Exploratory Data Analysis to identify trends and patterns in diabetes progression.

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

iv. To provide actionable insights for healthcare professionals to improve health outcomes.

**2.0. Literature Review**

This chapter provides a comprehensive review of existing literature 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. By examining the methodologies, findings, and limitations of previous research, this chapter establishes a foundation for the current study’s focus on comparing the predictive performance of models such as Random Forest, XGBoost, and Support Vector Machines (SVM). This literature review aims to contextualize the research problem, identify gaps in the current knowledge, and provide a rationale for the chosen methodologies in this study.

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 analyzing three prominent machine learning models—Gaussian Naive Bayes, Decision Tree, and Artificial Neural Network (ANN)—the authors assess each model's performance in diabetes risk prediction. Utilizing the Behavioral Risk Factor Surveillance System (BRFSS) dataset, a comprehensive resource for health-related risk behaviors, the study reveals that the ANN model achieves the highest accuracy, with an impressive 84.73%, thereby outperforming both the Gaussian Naive Bayes and Decision Tree models. This finding underscores ANN’s potential in advancing diabetes risk prediction and offers a compelling case for the application of machine learning techniques in healthcare. The study’s contributions are significant, as they pave the way for developing precise diagnostic tools and customized interventions, enhancing diabetes management and addressing its societal burden (Prasetyo, Izdihar & 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 analyzing patterns within a dataset comprising both diabetic and non-diabetic individuals. Leveraging the Diabetes 130-US hospitals dataset, which spans from 1999 to 2008, this study examines critical features through preprocessing and feature selection, while addressing data imbalance with various sampling techniques to improve model accuracy. 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 optimizing model performance, with the Random Forest model standing out by achieving an impressive 99.8% accuracy using the raw dataset. Logistic Regression, though initially less accurate, demonstrated potential for enhancement when combined with multiple sampling techniques, suggesting its viability for future improvements in diabetes prediction. This study emphasizes the value of advanced machine learning techniques and data handling strategies in refining predictive capabilities and enhancing early diabetes detection.

In response to the rising prevalence of diabetes and its severe health implications, Kasturi (2024) investigates a range of machine learning and deep learning models for the early detection of diabetes. Utilizing the PIMA Indian Diabetes dataset, this study aims to classify individuals as diabetic or non-diabetic by employing various machine learning algorithms, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM). Additionally, a Multi-Layered Feed Forward Neural Network (MLFNN) is implemented from a deep learning perspective, offering a robust comparison with traditional machine learning models. Each model’s accuracy and execution time are recorded to facilitate a comprehensive evaluation, with various activation functions, learning algorithms, and methods for handling missing data considered to enhance model performance. Among these approaches, the MLFNN achieved the highest classification accuracy at 92%, suggesting its potential for even greater accuracy if applied to larger datasets. This study contributes to the field by improving diabetes prediction from clinical data, enabling the identification of risk factors associated with pre-diabetes, and ultimately supporting proactive diabetes management to elevate patient care standards.  
  
N Nagarjuna and Lakshmi (2024) explore the application of machine learning techniques in diabetes prediction, with a concentration on leveraging various algorithms to enhance predictive accuracy and disease management. Their study employs multiple machines learning models, including Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, each of which is evaluated for its effectiveness in classifying diabetic and non-diabetic individuals. Emphasis is placed on feature engineering and addressing class imbalance through the Synthetic Minority Oversampling Technique (SMOTE) to improve model performance. Data preprocessing steps, including handling missing values and extracting relevant features, further enhance model accuracy and efficiency. Among the models, Random Forest and XGBoost demonstrate notable performance, achieving accuracy rates of 82% and 80%, respectively, in predicting diabetes. This research highlights the critical role of feature engineering and data balancing in machine learning models for diabetes prediction, offering valuable insights for improving disease management strategies

Gupta (2024) addresses the challenge of predicting diabetes accurately, especially with issues like outliers and missing data in labeled datasets. To address these issues, the research introduces a comprehensive prediction framework that incorporates essential data preprocessing techniques, including outlier rejection, missing value imputation, data standardization, 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, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP).A key innovation in this study is the introduction of a weighted ensembling technique designed to improve prediction accuracy, where each model’s weight in the ensemble is determined by its Area Under the ROC Curve (AUC) performance. To further optimize results, hyperparameter tuning is carried out using grid search. The experiments, conducted on the Pima Indian Diabetes Dataset under consistent conditions, demonstrate the effectiveness of the ensembling classifier. This classifier achieves a 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 proposed framework emphasizes early detection through lifestyle-based predictors, using tenfold cross-validation for accuracy verification. The research finds that Logistic Regression outperforms other algorithms with an accuracy of 93%, precision of 92%, recall score of 94%, F1 score of 93%, and a weighted average of 90%. This framework incorporates Canonical-Correlation Analysis (CCA) for feature selection, improving the model’s accuracy and reducing training time. 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. Their study introduces a diabetes prediction framework built using a private Bangladeshi dataset and various machine learning algorithms, including Decision Tree, SVM, Random Forest, Logistic Regression, K-Nearest Neighbour (KNN), and XGBoost. The framework uses a semi-supervised approach with high-inclination support to predict insulin levels.The study highlights that the XGBoost classifier, combined with the ADASYN technique for handling imbalanced data, achieved an accuracy of 80%, an F1 score of 0.81, and an AUC of 0.84. Moreover, a stacked ensemble of three classifiers achieved a remarkable 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 associated with the disease. It details the approach used for data collection, ethical considerations. I employed various statistical techniques to process and explore the data, develop predictive models, and I evaluated their effectiveness in predicting diabetes risk.

**3.1. Overview**

In this study, my objective was 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), collected in 2015, and another small dataset, the Pima Indians Diabetes Dataset.

The large dataset contains 253,680 survey responses from U.S. residents, with 22 features related to health indicators that are important in understanding diabetes risk.

The small dataset, **Pima Indians Diabetes Dataset** contains **768 records** and **8 features**, such as **age, BMI, glucose levels**, and **insulin levels**. This small dataset was included in the study for model comparison and evaluation in a less complex environment, providing insights into how well the models perform on both large and small datasets.

For this analysis, I applied three different predictive models: Random Forest, XGBoost, and Support Vector Machines (SVM). These models were chosen due to their ability to handle complex, high-dimensional data and their effectiveness in classification tasks. Additionally, I performed Exploratory Data Analysis (EDA) to uncover 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, F1Score, and AUC-ROC.

### 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: Diabetes Health Indicators 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 U.S. residents, with **22 health-related features**, both categorical (e.g., smoking status, high blood pressure) and continuous (e.g., age, BMI). The dataset includes a target variable (**Diabetes\_012**) indicating diabetes status (**no diabetes, prediabetes, or diabetes**). This dataset is valuable for analyzing a wide range of health indicators to predict diabetes risk and explore patterns related to lifestyle factors, healthcare access, and other relevant variables.

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

1. **Small Dataset: Pima Indians Diabetes Database**

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**, such as age, BMI, glucose levels, and insulin levels. The target variable indicates whether an individual has **diabetes or not**, making it suitable for binary classification tasks.

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

**3.2.1. Ethical Considerations**

### ****GDPR Compliance:**** The two datasets used in this study are **anonymised**, ensuring that **personally identifiable information (PII)** cannot be traced back to individuals, in line with **GDPR** principles such as **data anonymisation**, **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. ****UH Ethical Policies:**** As the datasets are publicly accessible and anonymised, **University of Hertfordshire (UH) Ethics Committee** approval was not required. This study relies on **secondary data** from publicly available datasets on **Kaggle**, which complies with **UH ethical policies** for responsible data usage in academic research, ensuring participant privacy and confidentiality. ****Permission to Use the Data:**** The datasets are governed by **Kaggle’s terms of use**, which allow 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. ****Ethical Data Collection:**** The study does not involve **active data collection** or **surveys**, and 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)**

For EDA, I started by calculating basic descriptive statistics for the continuous variables (e.g., **BMI**, **Age**) to understand their distributions. I then created visualizations such as histograms and box plots to observe the range and spread of these features. These visualizations helped me identify any outliers or unusual data points.

Next, I used scatter plots to explore the relationships between the continuous features and between the continuous and categorical variables. I also created a correlation matrix to see how strongly the continuous variables were related to each other. This allowed me to identify any potential multicollinearity between features, which could affect the performance of the models.

**3.4. Data Preprocessing**

To ensure the dataset was ready for model training and to optimize performance, I performed the following preprocessing steps:

1. **Data Cleaning**:
   1. I checked for missing values and there was none, ensuring the data remained consistent and reliable for model training.
2. **Feature Engineering**:
3. Since one-hot encoding was already completed, I did not need to perform any additional transformations on categorical features. However, I verified that the features were correctly encoded.
4. I created new features where necessary based on insights gathered from the exploratory analysis, which helped improve the models' performance.
5. **Normalization/Scaling**:
   1. I normalized the continuous variables, such as **BMI** and **Age**, to ensure that they were on a similar scale. This is especially important for models like **SVM**, which are sensitive to the scale of input data.
6. **Handling Class Imbalance**:
   1. As the target variable **Diabetes\_012** showed class imbalance, I applied techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the classes and ensure that the models could effectively learn from all classes.
7. **Data Splitting**:
   1. 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.
   2. I applied **k-fold cross-validation** to further assess the model's reliability and generalizeability across different subsets of the data.

**3.5 Machine Learning Algorithms**

I used three different machine learning models to predict **diabetes status** and evaluate their effectiveness:

* + 1. **Random Forest**: **Random Forest** is an ensemble learning method based on decision trees. It works by constructing multiple decision trees during training and outputting the most frequent class from the trees. I selected Random Forest because it handles both categorical and continuous data well and works effectively with large datasets.
    2. **XGBoost**: **XGBoost** is a boosting algorithm that builds models sequentially to correct the errors made by previous models. I included XGBoost because of its high predictive power and ability to handle imbalanced datasets efficiently.
    3. **Support Vector Machine (SVM)**: **SVM** is a classification algorithm that finds the optimal hyperplane that separates the data into different classes. I chose **SVM** because it works well with clear margins of separation between classes and performs well on binary classification tasks.

**3.6 Evaluation Metrics**

To evaluate the performance of the models, I used the following metrics:

1. **Accuracy**: This measures the percentage of correct predictions from the total predictions. While it is an important metric, I considered other metrics due to the class imbalance in the dataset.
2. **Precision**: Precision calculates the proportion of correctly predicted diabetes cases among all the predicted positive cases, which is useful for understanding how reliable the model is in identifying diabetes.
3. **Recall**: Recall measures how many of the actual diabetes cases were correctly identified by the model. It is important in medical diagnostics to ensure that as many positive cases as possible are detected.
4. **F1 Score**: The **F1 score** is the harmonic mean of precision and recall, offering a balanced evaluation metric, especially in imbalanced datasets.
5. **AUC-ROC**: The **AUC-ROC** curve evaluates the trade-off between the true positive rate and false positive rate. A higher AUC score indicates better model performance in distinguishing between the classes, making it a crucial metric for assessing the models in this study.

By using these metrics, I was able to comprehensively evaluate the models' effectiveness in predicting diabetes risk.

**References**

Elmenshawy, K., Wael, N., Ahmed, R. and El-Douh, A.A. (2024). Diabetes Prediction using Machine Learning and Explainable Artificial Intelligence Techniques. *SciNexuses*, 1, pp.28–43. doi:https://doi.org/10.61356/j.scin.2024.1306.

Gufran Ahmad Ansari, Salliah Shafi Bhat and Mohd Dilshad Ansari (2024). Machine Learning Techniques for Diabetes Mellitus Based on Lifestyle Predictors. *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, 17. doi:https://doi.org/10.2174/0123520965291435240508111712.

Gupta, N.P. (2024). Diabetes Prediction Using Machine Learning. *Deleted Journal*, 20(7s), pp.2244–2257. doi:https://doi.org/10.52783/jes.3960.

‌Hounguè, P. and Bigirimana, A.G. (2022) Leveraging pima dataset to diabetes prediction: Case study of deep neural network’, *Journal of Computer and Communications*, 10(11), pp. 15–28. doi:10.4236/jcc.2022.1011002.

Jain, R., Nitin Kumar Tripathi, Pant, M., Chutiporn Anutariya and Chaklam Silpasuwanchai (2024). Investigating Gender and Age Variability in Diabetes Prediction: A Multi-Model Ensemble Learning Approach. IEEE Access, [online] pp.1-1. doi:https://doi:.org/10.1109/access.2024.3402350.

Khongorzul Dashdondov, Lee, S. and Munkh-Uchral Erdenebat (2024). Enhancing Diabetes Prediction and Prevention through Mahalanobis Distance and Machine Learning Integration. *Applied Sciences*, 14(17), pp.7480–7480. doi:https://doi.org/10.3390/app14177480.

Kasturi, K. (2024). Comparison of Machine Learning Models for Diabetes Prediction. *International Journal of Advanced Research in Science Communication and Technology*, pp.531–536. doi:https://doi.org/10.48175/ijarsct-19072.

Kasula, B.Y. (2023). Machine Learning Applications in Diabetic Healthcare: A Comprehensive Analysis and Predictive Modeling. *International Numeric Journal of Machine Learning and Robots,* [online] 7(7). Available at: <https://injmr.com/index.php/fewfewf/article/view/19>.

Mishra, H.V.K., Singh, A.P., Sharma, V. and Khanna, E. (2024). Exploring the Effectiveness of Various Machine Learning Algorithms in Detecting Alzheimer’s Disease: A Comparative Analysis. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.4776479.

‌N Nagarjuna and Dr. Lakshmi HN (2024). Predictive Modeling of Diabetes Mellitus Utilizing Machine Learning Techniques. *CVR Journal of Science & Technology*, 26(1), pp.112–117. doi:https://doi.org/10.32377/cvrjst2618.

Rani, K.J. (2020). Diabetes Prediction Using Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 6(4), pp.294-305. doi:https://doi.org/10.32628/cseit206463.

Simeon Yuda Prasetyo and Zahra Nabila Izdihar (2024). Multi-layer Perceptron Approach for Diabetes Risk Prediction using BRFSS Data. pp.303–308. doi:https://doi.org/10.1109/icsima62563.2024.10675535.

Simeon Yuda Prasetyo, Zahra Nabila Izdihar and Ghinaa Zain Nabiilah (2024). Analyzing Machine Learning Approaches for Diabetes Risk Prediction: Comparative Performance Assessment Using BRFSS Data. pp.324–329. doi:https://doi.org/10.1109/icicos62600.2024.10636871.

Sivaranjani, S., Ananya, S., Aravinth, J. and Karthika, R. (2021). *Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICACCS51430.2021.9441935.

World Health Organization (2023). *DIABETES*. [online] World Health Organisation. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes.

‌

‌

‌

‌

‌