Machine Learning-Based Prediction of Diabetes Risk Using Health Indicators

Amr Adi1 and Dr.Salem Alemaishat1

1 Al Hussein Technical University, Amman 11831, Jordan

21110239@htu.edu.salem.alemaishat@htu.edu.jo

**Abstract**

Early detection of diabetes becomes essential because the worldwide prevalence of this continuous illness generates substantial health complications while enabling proper disease control and risk minimization. A predictive model uses machine learning algorithms to determine diabetes risk levels through the evaluation of health indicators

# Introduction

The field of Artificial Intelligence (AI) developed ways for machines to replicate human intellectual capabilities in choices and computational tasks and learning procedures. This field exists across different domains robotics, computer vision, natural language processing and predictive analytics that results in substantial industry transformation. AI systems evaluate vast datasets to establish connections between information while deriving meaningful points and operating complicated operations which boost operational efficiency and production rates.

Virtual systems achieve one of their primary breakthroughs through the execution of machine learning techniques that allow them to acquire knowledge from datasets without human code input. This specific kind of AI technology becomes fundamental to improving automation across vehicles and devices that assist humans while being key to advancing society and technology.

AI brings dramatic changes to healthcare services by improving disease prediction and prevention operations. My work uses artificial intelligence methods to identify potential diabetes cases by processing health metric data which includes patients' BMI levels together with their age numbers and blood sugar measurements and their daily routines. The diabetes risk evaluation benefits tremendously from machine learning models including SVMs and Neural Networks that yield precise predictions which enables prompt intervention decisions.

The use of AI systems to predict diabetes helps both achieve improved patient health results and lower the expenses that diabetes treatment requires. The aim of this study uses AI technology to detect at-risk individuals so researchers can develop preventive programs to build better health outcomes for the population.

The research investigates methods which machine learning models can use to predict diabetes risk using patient health-related data including glucose levels BMI age and family history information. The research question demonstrates how AI together with machine learning can correctly evaluate essential health indicators to determine diabetes risk with maximum efficiency.

The main goal of this study involves building a predictive system to identify statistical patterns between essential medical measurements and diabetes diagnosis. Machine learning algorithms enable the study to create an accurate forecasting system based on reliable prediction algorithms.

This study pursues a dual goal to assess feature importances between BMI measurements and patient age as well as blood glucose levels in predicting diabetes risks. Knowledge about the importance of health indices helps explain factors leading to diabetes risk thereby making the model more transparent.

The research has the main goal of generating usable preventive insights. The research works to uncover high-risk groups at their initial stages for developing protective strategies which lower diabetes occurrence and enhance medical results. The generated preventive recommendations will act as data-backed clinical recommendations for healthcare staff members.

The main advancement of this study involves developing machine learning algorithms which successfully determine diabetes risk probabilities. The work shows the significance of important health measurement data points while presenting beneficial prevention strategies which strengthen AI-based healthcare innovations.

Structure of the Document

The following structure appears in this document:

Section 2: Related Work

A review of machine learning applications for diabetes prediction forms part of this research section. The current section evaluates past methodologies while identifying empty spaces in previous research to establish a basis for picking suitable models with techniques. The section reviews how different machine learning algorithms previously utilized in diabetes risk prediction fare regarding their strengths and limitations.

Section 3: Data Collection and Description

The following section explains which sources the study drew from as primary and secondary data fields. The data collection process for medical student surveys is discussed plus challenges encountered during data acquisitions are detailed along with a valid reason for this method. This section provides extensive details regarding the Diabetes Binary Health Indicators Dataset (BRFSS 2015) including its structure and tells how input data was processed with an explanation of feature selection approaches.

Section 4: Research Approach and Methodologies

The research model based on Onion research is detailed here through its fundamental research layers such as philosophy and theory development alongside methodological selection and time horizons. It also includes research strategy and various techniques. The paper includes a flowchart and block diagram representing the research methodology that details the steps from information acquisition through model evaluation. This segment outlines the selection protocols for features as well as the model selection algorithms and the measurement systems which evaluated algorithm performance.

Section 5: Results and Discussion

The evaluation of Random Forest performance along with AdaBoost and Decision Tree takes place in this section following their training process. The performance evaluation combines accuracy, precision, recall, F1-score together with bar charts, confusion matrices and feature importance plots for visual comparisons among models. An explanation of the vital features BMI, HighBP, and Age in diabetes prediction emerges from analyzing dataset visualizations and correlations.

Section 6: Conclusion and Recommendations

This part summarizes the research outcomes that demonstrate machine learning's vital role in predicting diabetes. The research provides recommendations which detail when predictive solutions should be embedded in clinical environments and how to enhance medical staff understanding of patient information. The present study recommends future research approaches which focus on deep learning methods alongside the expansion of dataset diversity along with real-time validation strategies to enhance the predictive framework.

# Related Work

Literature Review

The research papers in this section are centred on the use of Machine learning and artificial intelligence in diabetes prediction and control. These papers address novel techniques, the combination of several architectural forms, and data-based methods to facilitate the early stage diagnosis, optimize the glycemic regulation, and facilitate the decision-making process in practice.

Here, the authors present the Fused Model for Diabetes Prediction (FMDP) a new model that is a fusion of support vector machines (SVM) and artificial neural networks (ANN) integrated using fuzzy logic to enhance the accuracy of results. The dataset used in the experiment was obtained from the UCI Machine Learning Repository with five hundred and twenty instances, and sixteen attributes. The data was preprocessed and split into training and testing sets in a 70:30 ratio. The SVM and ANN were performed training separately and there was integration of the result derived from both the methods by using fuzzy logic. The results of FMDP show that prediction accuracy of FMDP was as high as 94.87% compared with standalone SVM and ANN models. It shows that heuristic systems are effective when there is uncertainty and provide strong prognosis if implemented as the composition of more basic models.

There are some limitations of the current study which are as follows. It was offset with relatively small dataset that can be problematic in terms of external validity. Furthermore, the input features used in the model require high quality hence the model is prone to the case where there is missing and noisy data. Scalability involving large volumes of data was not covered and this has implications on performance. Finally, language interpretability of the fuzzy logic mechanism was not well elaborated, making it difficult for other healthcare SHALL NOT workers who are not conversant with programming languages to adopt this mechanism. However, these limitations should not be understood as a lack of capacity that the FMDP does not hold the potential of improving diabetes prediction in clinical practice. (Ahmed et al., 2022)[1]

In this paper, an attempt is made to design a novel two-stage fused model for predicting diabetes using support vector machines (SVM) and artificial neural networks (ANN) incorporating fuzzy theory. The data set used in this work is composed of 520 patients with 16 attributes which characterize fundamental diabetic risk factors. The data thus prepared for analysis was normalized and then split into 70% training data and 30% testing data. SVM and ANN models were trained separately while the output of the two model were fused using a fuzzy-logic system to handle rather ambiguous predictions created by the two models. (Gupta et al., 2021) Evaluation criteria included accuracy, sensitivity and specificity, typical for classification models in medical diagnosis.

The results showed that our proposed FMDP had a higher accuracy of 94.87% compared to SVM and ANN individual models. This demonstrates how the proposed hybrid methods can handle intricate and intertwined data structures typical of high-dimensional medical data. The heterogeneity of the model is the major highlight of this study and indicates the possibility of applying this model for screening and clinical decision making at onset of diabetes.

Some limitations were observed. The study had a moderate sample size which mean that the test done could have been done on a smaller sample of the population. Furthermore, as it will be shown, the model requires high quality and complete data which makes it vulnerable to noisy or missing data. It can be used on large scales more efficiently, and the computation time was not a concern that could be a factor in the real world, but it was not mentioned. Also, the fuzzy logic mechanism was not explained to be easy to understand by the professionals in healthcare to adopt in practice when implemented. Nonetheless, the FMDP posses some potential in enhancing the predictive modeling of diabetes mellitus identification. (Khan et al., 2021)[2]

This paper presents a comprehensive analysis of the role of AI and ML in enhancing the glycemic control of persons with diabetes. The emphasis is on the issues, innovations, and possibilities pertaining to the use of ML for AID and CGM. Of the identified manuscripts, 189 were included in a literature review covering short and long gl of glucose prediction, hypoglycemia detection, and decision-making support. Diabetes’s major algorithmic strategies are described, focusing on decisions’ explainability, data inputs variation, and individualization for the best treatment plan.

The main findings highlighted here suggest that there is great possibility for the use of ML algorithms to enhance diabetes care. These are models for the estimation of glucose and such phenomena as hypoglycemia or hyperglycemia with high sensitivity and specificities and reinforcement learning for adjusting the dosing of insulin individual to the patient. It offers feature engineering recommendations derived from consensus, evaluation of algorithms, and dealing with data variation—which can be considered a guiding source for ML practitioners.

Concerns still remain Including variability in the data set, small set of data specific to diabetes and difficulty in creating algorithms that can work in various communities out. Besides, the fact that many algorithms use simulated data to test the results of their work also plays a role. These constraints suggest the importance of technical measurement standardization in data acquisition, transforming the data into features, and comparing the performance of the developed algorithms to improve the dependability and upscalability of utilizes AI/ML in managing diabetes. (Jacobs et al., 2023)[3]

The present research utilizes ensemble learning together with deep learning (DL) to detect early signs of diabetes. The authors developed two ensemble models including the Hi-Le combination from Highway and LeNet and the blending HiTCLe design which fused elements from Highway, LeNet along with Temporal Convolutional Networks (TCN). The research employed ProWSyn-Over-sampling on the Diabetes Prediction Dataset (DPD) since this dataset contains small records with high imbalance. Ten fold cross validation known as K-FCV was used for interpretation while SHAP generated feature attributions of the models.

The F1-score of HiTCLe reached 0.96 alongside Hi-Le model accuracy of 0.96 and F1 score of 0.94. The ensemble learning methods outperformed model prediction basics to accurately predict diabetes. Different features were processed effectively by the models: highway networks handled adaptive information while spatial relationships were identified using LeNet and temporal dependencies were discovered by TCN to enhance forecasting precision.

Toward the end of highlighting the model's outstanding performance issues we must recognize two primary weaknesses which include excessive computational expenses and the use of unbalanced data that fails to represent genuine world situations adequately. The use of ProWSyn synthetic data generation techniques helps overcoming data availability issues yet these models risk developing biased outcomes and their performance has never been verified against fresh dataset samples. (Shaheen et al., 2024)[4]

The research evaluates Treatment 2 Diabetes Mellitus (T2DM) warning signs in Saudi Arabia through a mobile application platform. A cross-sectional survey method was chosen while participants received a survey form which included variables about smoking habits and healthy diet and blood pressure and BMI measurements and gender and regional factors. Chi square analysis together with binary logistic regression were used to analyze the data acquired for this study. Synthetic Minority Oversampling Technique (SMOTE) functioned as the method to reduce class imbalance. Numerous classifiers were evaluated for the experiment and Two-Class Decision Forest emerged as winner with an F1 score of 0.8453. To achieve model validation external datasets confirmed its ability for maintenance and durability. The model received implementation as a web service for delivering continuous time-based risk assessment capabilities.

The research presents techniques from machine learning that enable the assessment of diabetes risk. Despite these disadvantages the study also suffers from two major drawbacks which include data collection through self reporting and the small sample size compared to the target population. Since the analysis focused on specific regions the findings derived from this study may not provide relevant data for either additional areas of the nation or any nations worldwide. (Syed and Khan, 2020)[5]

The gap exists because significant progress has been made regarding artificial intelligence and machine learning applications for diabetes prediction alongside management while ongoing research remains unresolved. Results face difficulties with cross-study generalization because research grounds tend to collect their data from non-representative small samples and single geographic areas. Higher prediction accuracy belongs to ensemble and hybrid models but these models present two problems namely excessive costs during computation and real-life implementation limitations. Most assessment techniques lack clear explanations because of which non-technical healthcare personnel would not trust the results. The models use the least amount of data collection for real-time updates needed during dynamic model iteration. The research addresses vacant areas by using a new artificial neural network design that combines multiple health input sources with local validation for improved and repeatable diabetes risk assessment.

# Data collection and Description

The data collected for this study must occur directly for the first time it is needed. Scientific information used in this investigation involves obtaining direct data by interviewing or surveying healthcare professionals together with patients. The gathered information serves as unique specific data points that relate directly to study requirements. Healthcare professionals use direct accounts to reveal patient risk factors whereas patient surveys track lifestyle elements that affect diabetes risk.

Primary data fills a crucial role because it directly suits the specifications required for research analysis. The process of designing primary data collection requires substantial resources and effort as well as presents challenges to ensure participant participation and data reliability.

The study relies on primary data which consists of collecting original information dedicated for this investigation. Primary data in this research consists of gathering information directly from healthcare professionals and medical patients by means of interviews and surveys methods. Primary data collection offers special insights that are precisely designed to match research goals. Medical staff members bring patient risk factor documentation while studies containing survey responses from patients help document lifestyle habits affecting diabetes vulnerability.

The collection of primary data proves beneficial because researchers can make it highly specific to match research requirements. The collection process for primary data is labor-intensive and resource-consuming while also facing potential problems with participant accessibility together with data authentication issues.

Secondary data refers to preexisting information which different organizations have collected through their independent work including databases and publicly available datasets. This study collects its secondary data from the structured health indicator datasets which are available in GitHub and Kaggle repositories that pertain to diabetes. The datasets contain vital health indicators including BMI, Age, Glucose Levels and additional important variables for predictive modeling.

Secondary data serves as an efficient and economical choice because it comes already organized and collected from past research. Secondary data analysis has potential restrictions which include absence of needed information along with possible discriminatory patterns and deficient adaptability to research setting demands.

Differences Between Primary and Secondary Data:

Source:

The researcher directly obtains original data to fulfill a particular research objective.

The information collected for secondary data originates from other parties before researchers obtain it.

Relevance:

Research questions align better with data collection that is gathered specifically for the study.

Researchers need to perform initial processing steps on secondary data before it becomes appropriate for their investigation.

Cost and Effort:

Using original data for research requires extensive work efforts and uses many resources.

Secondary data becomes readily accessible while being less expensive to obtain.

Reliability:

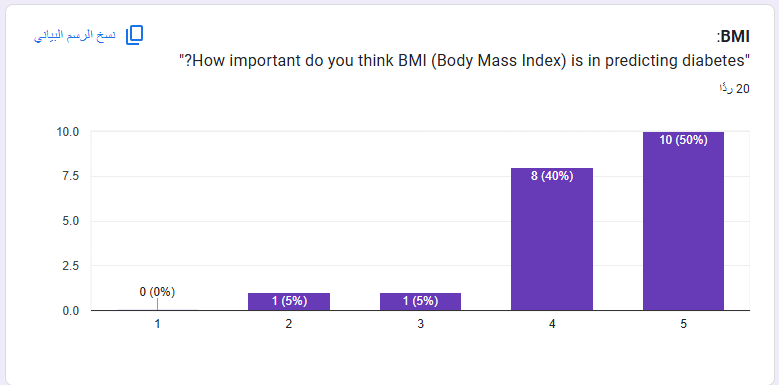
Research data derives higher precision and reliability when the scientist personally manages the information retrieval steps.

Secondary data contains possible inconsistencies together with incomplete information needed for the study.

3.1 Primary Data

This research obtained fundamental data through survey feedback which explored different health indicators for their ability to forecast diabetes conditions. The survey included questions which evaluated BMI measurements along with glucose test results together with exercise rates and model prediction trustworthiness. Supplementary documents display data from survey participants through visual representations such as bar graphs together with pie diagrams along with written evaluations about health data prediction difficulties.

Participants identified BMI as a vital diabetes prediction factor through their responses in the bar chart since 50% considered it "very important" while 40% designated it "important." Only a small percentage (5%) view BMI as less significant. BMI serves as an essential health indicator according to this data.



A large number of participants identified glucose measurement as holding extreme significance for predicting diabetes (70% while another 20% placed it in the important category). Most individuals agree glucose levels serve as one of the main indicators to identify possible diabetes risk.

A screenshot of a graph

Description automatically generated

Thirty-five percent believe diabetes importance is equal to BMI's importance and 35% assert it is of great importance to their health. The remaining quarter views diabetes as a middle priority factor. The results show physical exercise affects diabetes risk estimates although the factor has some flexible measurement parameters.

A graph with purple rectangles

Description automatically generated

The research data indicates that 30% of participants conducted diabetes research or prediction using health data whereas 70% engaged in this process. Research participants demonstrate minimal experience in using health data to produce predictive outcomes.

A close-up of a pie chart

Description automatically generated

The study participants reported multiple difficulties such as uncomplete data and inaccurate data and disparate healthcare system data access challenges. Data accuracy and quality remain the main worry among participants because they need reliable standardized data.

A screenshot of a computer

Description automatically generated

A screenshot of a chat

Description automatically generated

Half of the participants rated predictive models as "moderately reliable," and the other half considered them "very reliable." This indicates confidence in the models, although there is room for improvement in their reliability.

These pictures collectively reflect participants' opinions and insights on factors influencing diabetes prediction, challenges in using health data, and the reliability of predictive models.

A pie chart with text below

Description automatically generated

A survey was designed specifically for medical students during primary data collection procedures. The participants received selection because their basic understanding of health and disease allowed them to respond effectively to the survey questions. The survey integration included carefully constructed quantitative questions responding to scales and open-ended qualitative questions to collect multiple types of structured and detailed information.

The survey development process involved multiple difficulties that needed to be overcome.

Some medical students faced difficulties finishing the survey thereby reducing the survey data collection outcome from what was initially intended.

The scheduled time available for distributing surveys and obtaining responses created difficulties because students managed busy academic commitments.

The survey managed to gather helpful perspectives about how different health indicators affect the prediction of diabetes despite the encountered challenges. The survey findings gained validity because the participating medical students demonstrated knowledge of health-related information. Participants responded by offering key information as per the research goals which substantially enhanced the scientific value of the study.

Justification for Choice of Primary Data

Survey primary data collection became essential because it presented an opportunity to retrieve real-world research-aligned information. The survey directly involved participants which resulted in deeper insights into how the public sees diabetes prediction along with health data usage difficulties. The personal information that the study collected served as complementary data to the secondary information thus maintaining a balanced methodological foundation.

Merits of Primary Data

* A Set of Specialized Questions Deliberately Targeted the Research Objectives while Handling Specific Issues for Relevance and Focus.
* The information displays present-time evaluations and existing trends about diabetic disease prospects.
* The open-ended response section produced elaborate information which expanded upon quantitative measurement results.

Limitations of Primary Data

* The process of acquiring raw information including its organization and analysis lasted an extensive amount of time.
* The analysis of a restricted sample size diminished the ability to universalize the research findings.
* The responses provided by participants could contain biases which damages the reliability of collected data.

The implementation of primary data collection methods proved to be difficult but essential for identifying critical information that formed the basis of this research. The study benefits more from its merits than its weaknesses thus demonstrating value to this research field.

3.2 Secondary Data

For my study I used publicly accessible "Diabetes Binary Health Indicators Dataset (BRFSS 2015)" as secondary data. The dataset contains health-related information from a large population that includes measurements of BMI alongside glucose levels and age alongside physical activity with approximately 22 characteristics and a binary target showing diabetes status. During the research I generated visualizations like the Random Forest feature importance chart and correlation heatmaps to demonstrate how BMI and glucose levels together with age are the most important variables.

I selected this dataset because of multiple benefits that include:

* The purpose of the dataset matches the research goal since it represents health indicators that directly predict diabetes risk.
* The study benefits from a big dataset comprising diverse healthcare and demographic elements thus supporting universal application of its results.
* The public availability enables convenient access to this dataset which provides entire details without requiring further data acquisition.
* The dataset maintains a well-tailored structure that enables machine learning applications because all data is clean and properly documented.

Merits of Secondary Data

* Data collection processes through the use of secondary information cut down both time commitment and labor requirements.
* Public availability of data eliminated every expense linked to primary data collection thus making it cost effective.
* The dataset included many health indicators which allowed comprehensive evaluation and chosen feature analysis.
* The pre-processing phase was simple since the dataset was curated which enabled researchers to dedicate their efforts toward analytical work.

Limits of Secondary Data

* Secondary data quality control remains under the authority of original data source providers who may limit their data output.
* The data collection through original measurement methods may introduce biases that exist within the dataset.
* Secondary data illumination fails to provide real-time data attributes since the collection method cannot adjust questions to match particular research needs.
* A class imbalance problem exists in the dataset through having fewer positive diabetes diagnosis cases thus affecting model training for predictions.

The secondary dataset delivered ample resources to develop machine learning models designed for diabetes prediction assessment. The benefits of research data relevance together with its accessible features and detailed information outweigh minor limitations such as data collection control issues to make this information vital for my research program.

# Research Approach and Methodologies

This section details the research approach and methodologies used in this study, structured around the Research Onion Model. It describes each layer of the model, including the philosophy, approach, methodological choice, strategy, time horizon, and techniques and procedures, and explains how these align with the study’s objectives. The methodology flowchart illustrates the steps taken throughout the research process, from defining the problem to data analysis and evaluation. Each choice is critically evaluated, ensuring a strong justification based on the nature of the research and its goals.

### Onion Model

Describe the onion model.

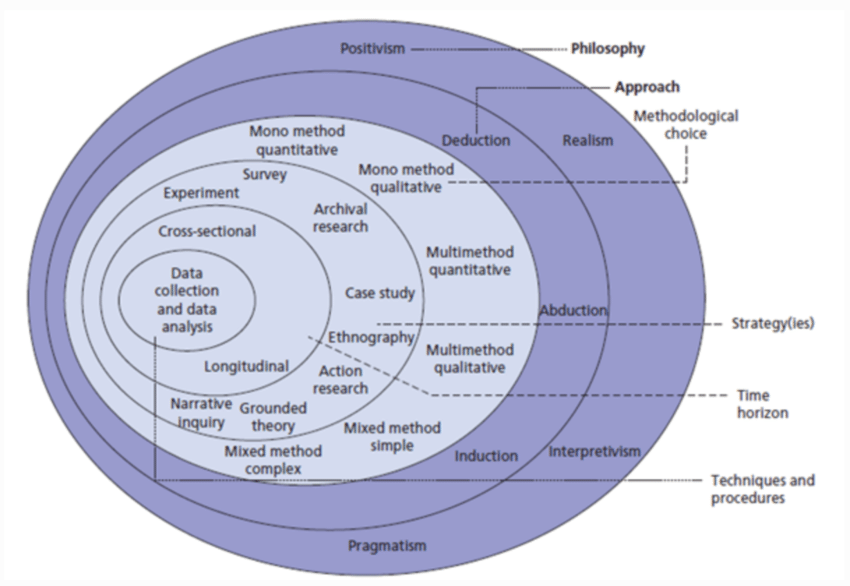


Figure 1: The structure of the Onion Research Model [1].

* + 1. **Philosophy**

The procedural guide for complete research arises from the external layer of the Research Onion literature model. Philosophy provides researchers with their perspective about knowledge discovery combined with their perspective on understanding reality.

Philosophical Approaches:

* Positive methods use observable scientific data through testing of hypotheses while depending on objective evidence. The conviction states that scientists can objectively study the world without depending on personal human meanings.
* Researchers using interpretivism study social phenomena regarding human experience while giving priority to context understanding.
* The realist approach promotes the integration of positivist with interpretivist principles for the purpose of understanding objective reality alongside subjective perceptual influences.
* The research utilizes pragmatic methodology which selects philosophical approaches based on practical research goals to find suitable methods.

Chosen Philosophy: Positivism

* The evaluative research adopts positivism because it explores measurable quantitative health metrics for hypothesis testing and predictive forecasting.

Justification:

* Scientific methods help researchers analyze numerical data which positivism supports as the appropriate approach.
* Data-driven research analysis for establishing reliable outcomes makes positivism the most appropriate philosophy.
  + 1. **Theory Development Approach**

The second tier of research analysis using the Research Onion examines the theoretical development process coupled with testing procedures within the study. The approach establishes whether the research creates its hypothesis first or ends with one.

Approaches:

* When Using deduction researchers begin with existing hypotheses before testing them through data collection and analysis procedure.
* The induction research approach creates theories after studying identifiable patterns that emerge from collected data.
* The research method of abduction uses both deduction and induction to discover the most appropriate explanation for a phenomenon.

Chosen Approach: Deduction

* Deduction serves as the research approach since the study tests pre-established hypotheses which include the relationship between glucose levels and BMI to assess diabetes risk.

Justification:

* Through the deductive approach researchers can maintain their research focus on evaluating specific quantitative relationships.
* The research method follows the previously established theories found in machine learning and diabetes risk prediction.
* Testing hypotheses provides the study with organized scientific methods to create practical findings.
  + 1. **Methodological Choice**

The decision about research methods for data collection and analysis forms the third fundamental layer of designing an ethical study. The research methodology selection decides between quantitative methods, qualitative methods or the combination of both.

Methodological Choices:

- Mono-Method Quantitative: Focuses exclusively on numerical data and statistical analysis.

* The method concentrates on obtaining information through interviews and case study or non-numeric data.
* The research combines quantitative methodologies with qualitative methods to gain a complete understanding of the topic of study.

Chosen Methodology: Mono-Method Quantitative

* The research utilizes a single quantitative methodology to evaluate statistical numbers extracted from survey results and the Kaggle database.

Justification:

* Quantitative analysis fits machine learning modeling requirements since it demands extensive structured datasets for training purposes.
* Precise measurements from numerical data enable easier evaluation of model performance because they provide standardized metric measurements.
* Statistical along with computational analysis demands become possible through this method which leads to dependable results.
  + 1. **Research Strategy**

Data collection methods along with analysis strategies are examined in the fourth layer. Research steps aimed at answering questions are handled by this layer.

Strategies:

* The collection of primary data happens through questionnaire distribution and interview processes.
* The experimental method tests particular variables through controlled laboratory investigations.

Case Study: Provides an in-depth analysis of a specific individual, group, or phenomenon.

* Researchers use existing data records and datasets by conducting archival research.
* Action Research uses systematic planning and action repetition as its main approach to tackle particular problems.
* The strategy selected involved conducting surveys along with archival research as data collection methods.
* Participants respond to surveys which collect their rich responses about their perceptions regarding diabetes risk elements and predictive tools.
* The training along with assessment of machine learning models takes place through archival research with data from Kaggle.

Justification:

* The survey-based research method combines the survey method using secondary data sets with archival research which requires participant feedback to achieve comprehensive understanding.
* The survey gathers distinctive public understanding alongside the validated health indicators from the Kaggle dataset.
  + 1. **Time Horizons**

Researchers must define how long their investigations will span through the fifth layer of the research design by deciding between single-time and long-duration studies.

Time Horizons:

* Cross-Sectional: Collects data at a specific point in time.
* The assessment span lasts over multiple time points to track patterns and detect variations in data.

Chosen Time Horizon: Cross-Sectional

* Research collects data and analyzes specific health indicators and public perception while studying them at one moment during the allotted timeframe.

Justification:

* The research will use an efficient cross-sectional method that supports the project’s time constraints.
* The research methodology enables pattern identification along with relationship detection through its data collection without needing extensive data gathering periods.
  + 1. **Techniques and Procedures**

At the core of the Research Onion researchers describe the tools together with methods required to collect and prepare data and perform analysis.

Techniques Used:

Data Collection:

* The collection of primary data through surveys included 19 participant responses regarding diabetes risk elements.
* The Kaggle dataset serves as secondary data while containing health metrics for glucose along with BMI and age variables.

Data Preprocessing:

* The next layer begins with handling data gaps while normalizing values before preparing the data for machine learning analysis.

Analysis:

* The analysis involves training decision trees and neural networks among other machine learning models.
* Evaluation of the models occurs through precision and recall and accuracy measurements.

Justification:

* Surveys collect qualitative data which archives serve to enhance statistical strength of the analysis.
* Machine learning together with preprocessing techniques enables the creation of dependable and scalable systems which assess diabetes risk.

### Research Methodology

A diagram of a computer system

AI-generated content may be incorrect.

Data Collection

Description:

* The research collected comprehensive data which included primary information obtained from medical student surveys together with secondary health data accessible in public databases. The research team obtained important features including BMI, HighBP, Glucose Levels, Age in addition to other health indicators to estimate diabetes risks.

Justification:

* Data collection through primary and secondary sources provided diverse and complete information for the dataset. The primary data gave an individualized perspective but secondary data delivered structured and expansive datasets for training purposes.

Challenges:

* The survey generated difficulties because participants did not respond at adequate rates and misunderstood some of the questions which reduced the usable sample size. Strong secondary data sources acted as a correction factor to address the problems encountered through primary data collection.

Impact:

* The selected data features proved essential for creating reliable predictive models because this step delivered a complete data collection.

Data Processing

Description:

* A combination of SelectKBest and Random Forest enabled us to perform feature selection while numerical features received normalization treatments before handling the missing values in the processed data.

Justification:

* Every prediction system requires data preprocessing to establish uniformity and meaningful adherence to the analysis. The analysis spent its focus on vital predictors through feature selection which simultaneously improved model performance and clarity.

Challenges:

* The preprocessing stage demanded precise management of categorical data along with variable normalization to minimize model prejudice.

Impact:

* The preprocessing process produced improved dataset quality which resulted in resistant and efficient models.

Data Analysis

Description:

* A group of machine learning models underwent testing along with Random Forest and Decision Tree while AdaBoost served as the third model. The evaluation relied on four metrics consisting of Accuracy together with Precision and Recall and F1-Score.

Justification:

* The assessment in this phase enabled researchers to evaluate model predictions which enabled them to identify the approach that proved most successful for diabetes risk prediction. The selection of Random Forest stemmed from its optimal balance between the metrics measured.

Challenges:

* Complex model algorithms such as AdaBoost represented an operational limitation when performing hyperparameter tuning operations in this phase.

Impact:

* The performance analysis showed BMI and HighBP features to be critical followed by an assessment of model competency and inadequacies.

Evaluation

Description:

* Evaluation consisted of understanding outcome interpretations alongside visualized performance assessment through both confusion matrix and performance plot analysis.

Justification:

* The evaluation process confirmed that Random Forest met the necessary standards regarding accuracy and reliability in the selected model. The visual representations served as valuable tools for effective findings dissemination.

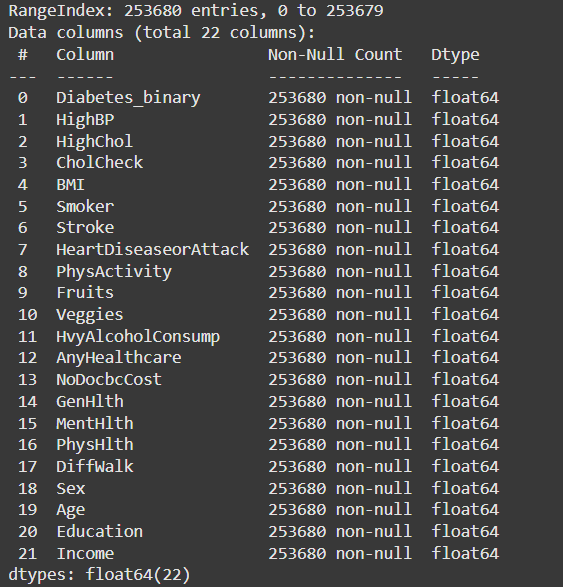
Challenges:

* The chosen Random Forest model won through evaluation yet requires testing on different external datasets to confirm its wider applicability.

Impact:

* The research outcomes were combined during this phase which enabled developing practical recommendations for diabetes screening alongside prevention methods.

# Results and Discussion



The dataset consists of 22 columns and all columns contain 253,680 non-null values since it lacks any missing entries. Each column contains float64 data type. Six fields listed in the database include HighBP, HighChol, BMI, Age, Income and Diabetes\_binary presents the target class. Medical along with demographic data features constitute the dataset which plays a crucial role in diabetes prediction analysis. Machine learning operations can directly be applied to this dataset due to the absence of data incompleteness. Consequently no value imputation processes are necessary for enabling machine learning.

A graph of a bar graph

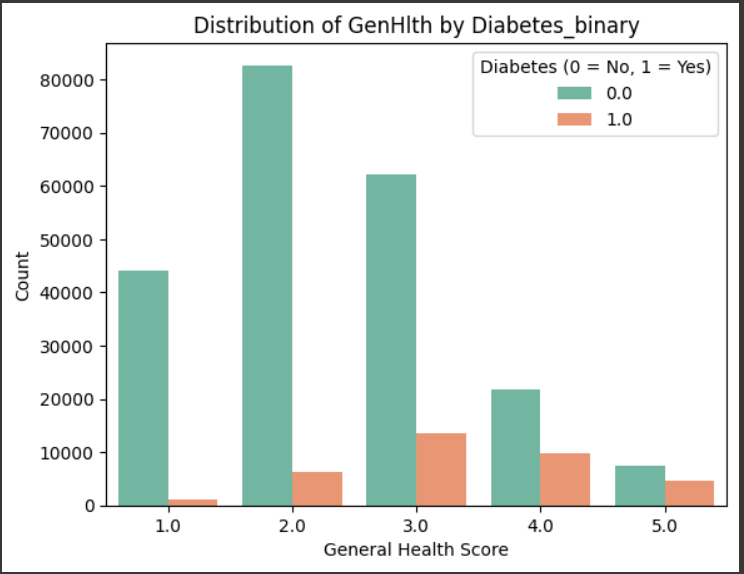
AI-generated content may be incorrect.

Results from the bar chart indicate High Blood Pressure (HighBP) develops differently based on diabetes condition (Diabetes\_binary). The bar graph shows diagnosis of diabetes at 1 yet displays diagnosis of without diabetes at 0 through the x-axis as it presents average HighBP readings through the y-axis. HighBP measurement results from people with diabetes exceed those of individuals without diabetes based on the information displayed in the chart. HighBP functions as a crucial sign regarding risk assessment and it helps predict future diabetes development due to its essential role in diabetes risk models.

A graph of a diagram

AI-generated content may be incorrect.

In this plot the relationship between BMI and binary diabetes status is presented by linking BMI (Body Mass Index) to Diabetes\_binary which uses 0 for no diabetes and 1 for diabetes. The median BMI value of diabetics exceeds the numbers of people without diabetes indicating that higher BMI leads to an increased diabetes risk. BMI distribution reaches further throughout diabetic groups than it does in the non-diabetic population. The diabetic population displays weight observational values that extend farther than the values observed in nondiabetic individuals. BMI values stand as critical variables for diagnosing diabetes risk according to the presented data.



The provided chart demonstrates how better general health scores (GenHlth) correlate with non-diabetic status (Diabetes\_binary) through its visual representation. The data reveals non-diabetic patients mostly have GenHlth scores of 1 and 2 because (Diabetes\_binary = 0) bars appear higher in these columns. Individuals moving towards worse general health from score 4 and 5 demonstrate a growing percentage of diabetes cases (Diabetes\_binary = 1) shown through the orange bars. This trend indicates that diminishing general health status creates more risk for developing diabetes because it strengthens the connection between poor health and diabetes diagnosis. The graphic demonstrates that healthy general condition minimizes the probability of developing diabetes.

A graph of a diagram

AI-generated content may be incorrect.

The boxplot looks at the connection between Age and Diabetes\_binary which uses 0 to denote cases of no diabetes and 1 to signal diabetic cases. According to the collected data people tend to develop diabetes beginning from their median age thus age emerges as a determinant that increases diabetes risk. The available data reveals broad age ranges among people who do not have diabetes and diabetes prevalence primarily affects older individuals. Few pieces of evidence indicate that diabetes exists within the young participants across both groups. Age serves as a fundamental factor that determines whether a person will develop diabetes according to the data presented in the plot.

The feature selection

A screen shot of a computer

AI-generated content may be incorrect.

The SelectKBest method identifies the 12 most significant features for predicting diabetes within this code. This code chooses essential health indicators which include HighBP together with HighChol and BMI and other indicators. The chosen features prove essential for enhancing predictive model accuracy because they direct the focus toward the most significant-data-elements.  
  
Split data

A computer screen with white text

AI-generated content may be incorrect.

This code splits the dataset into 80% training and 20% testing.

The Decision Tree model:

A computer screen shot of a program

AI-generated content may be incorrect.

The combination of Decision Tree model with the selected features produced 81.89% accuracy in classification results where 82% of instances received correct labels. The model showed restricted detection abilities for positive diabetes cases because its precision score reached only 0.32 which indicated that the model identified correctly less than one-third of predicted positive cases. The model revealed weak performance in detecting real positive instances because it only identified 28% of actual diabetes cases based on the recall score value. The F1-score value at 0.30 indicates that the precision and recall metrics show an unfavorable relationship since it demonstrates deficient performance in both correct identification and accurate prediction for positive cases.

The AdaBoost model:

A screen shot of a computer

AI-generated content may be incorrect.

For the AdaBoost model the researcher obtained a 87% accuracy rate which demonstrated that most cases received correct classifications. The model accuracy level reached 0.54 which demonstrated 54%kes of its anticipated positive diagnoses to be true but left the remaining predictions as false positives. The obtained recall score of 0.20 indicates that the model detected only 20% of real diabetes cases which represents major restrictions in identifying actual positives. The F1-score of 0.29 demonstrates that the predictive model shows sufficient accuracy in its assessments while experiencing severe deficiencies in detecting relevant positive instances.

The Random Forest model:

A screen shot of a computer program

AI-generated content may be incorrect.

The Random Forest model operated with an 85% accuracy rate which proved its ability to accurately categorize 85% of the instances effectively. Analysis revealed high precision through a 0.82 score which meant that 82% of accurate positive predictions prevented unnecessary false detections. Its recall score reached 0.85 because the model demonstrates excellent sensitivity in detecting 85% of actual diabetes cases. The Random Forest model demonstrates an excellent performance level using an F1-score of 0.83 which shows an optimal balance between retrieval efficiency and precision accuracy when measuring diabetes risk.

**Confusion Matrix for Decision Tree Classifier:**

**A blue and white chart

AI-generated content may be incorrect.**

The data from the Decision Tree classifier appears in the presented confusion matrix. The model successfully detected 39,565 non-diabetic cases among its total predictions while it misclassified 1,982 cases of diabetes. Significant misclassification occurs in the model because it produces 4,174 false positives along with 5,015 false negatives which negatively affects precision and recall outcomes. The Decision Tree shows moderate diabetes prediction capacity but needs additional improvements so it can better reduce errors across different categories.

**Confusion Matrix for AdaBoost Classifier:**

**A blue and white graph with numbers and labels

AI-generated content may be incorrect.**

The AdaBoost Classifier achieves high accuracy in its recognition of non-diabetic patients judging from its number of true negatives which reaches 42,538. One advantage of this approach is its ability to correctly identify 1,388 positive cases while true negative counts total at 1,388. Many mistakes occur when the algorithm mislabels 5,609 cases as negative and 1,201 cases as positive. The analysis indicates that AdaBoost surpasses Decision Trees in diabetic detection precision but its ability to identify diabetic cases is still insufficient. The classifier places higher importance on accurate non-diabetic diagnoses however its problem in identifying diabetic cases indicates further tuning is necessary.

**Confusion Matrix for Random Forest Classifier:**

**A blue and white graph

AI-generated content may be incorrect.**

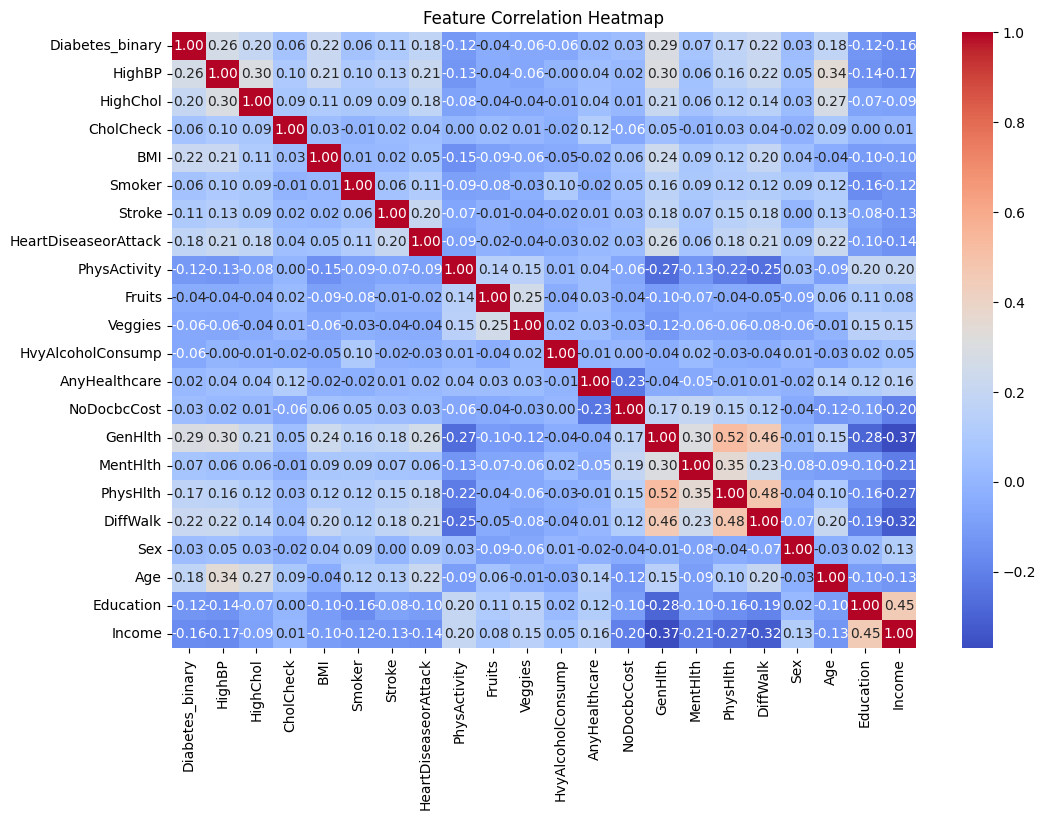
Using The Random Forest Classifier results in identification of 1,554 true positives and 41,536 true negatives indicating strong capability to recognize non-diabetic cases. Despite its reduced number of misclassifications it still struggles at detecting diabetics because it identifies 2,203 false positives and 5,443 false negatives. The model surpasses both Decision Trees and AdaBoost in determining diabetes cases effectively which makes it the most dependable prediction tool for this dataset. Additional modifications might improve how well this model detects diabetic patients.

Comparing the performance metrics of the three models.

A graph of different colored bars

AI-generated content may be incorrect.

Random Forest establishes itself as the ideal choice because it maintains a balanced ratio across all performance metrics that perform best in precise prediction along with high recall capabilities and F1-Score. Among all prediction models this dataset provides the most reliable and practice-proven solution to determine diabetes risk. The performance metrics Accuracy and Precision together with Recall and F1-Score from Random Forest and Decision Tree and AdaBoost are presented in a bar chart comparison. Random Forest provides the most optimal combination of balanced model performance throughout all metrics. The accuracy rate of AdaBoost is exceptional but the system demonstrates poor performance when measuring recall rates. The Decision Tree model shows poor performance in all three metrics.



A heatmap format reveals the variables which demonstrate connections to other variables across the entire dataset. Key observations include:

The Random Forest analysis confirmed that the relationship between BMI and Diabetes\_binary function held a medium strength due to which BMI came fifth in measures according to this analysis.

The relationships between Age and PhysHlth demonstrate significant statistical value when analyzing the features' set.

The slight influence between Education and Income on the target variable became redundant through combined effects with other attributes.

Research Question

This project investigates the primary research question:

How can machine learning models assess the risk of developing diabetes using patient health data, including BMI, glucose levels, age, family history, and other key indicators?

Additionally, the project explores:

What health indicators produce the most accurate model for identifying at-risk diabetic patients?

These questions aim to understand how machine learning algorithms, specifically Random Forest, AdaBoost, and Decision Tree, can evaluate diabetes risk while identifying the most impactful health conditions for early detection and prevention strategies.

Research Objectives

Develop a Predictive Model:

A predictive model based on Random Forest, AdaBoost, and Decision Tree will analyze the relationship between diabetes diagnosis and health indicators such as glucose levels, BMI, and age.

Data Analysis:

Rank health-related features by their predictive power to enhance the accuracy of the models.

Preventive Insights:

Use the developed models to detect diabetes at its onset and propose prevention strategies to minimize future complications.

Merits

Comprehensive Feature Selection:

SelectKBest successfully detected the essential predictors which included HighBP with BMI and PhysActivity among others. The process diminished dataset dimensions while concentrating on important diabetes prediction health indicators to boost operational performance.

Model Diversity:

Extensive application of Random Forest details with AdaBoost and Decision Tree models formed a complete picture about different algorithmic solutions. The multi-dimensional evaluation of accuracy together with precision along with recall standards and F1-Score yielded a comprehensive model selection by choosing the best fitting solution.

Robust Dataset:

A vast database containing comprehensive health indicators delivered enough information to perform model training as well as testing and evaluation procedures. The reliability of results enhanced significantly while simultaneously improving the general reliability of the tested findings.

Insightful Visualizations:

The graphical representations including confusion matrices and bar charts and histograms delivered decisive observations about model outcomes combined with feature pattern recognition which improved the understanding of stakeholders along with results sharing capabilities.

Practical Implications:

The analysis enabled healthcare professionals to improve early diabetes detection by revealing GenHlth and PhysHlth as essential health metrics so they could take preventive actions toward high-risk patients.

Limits of the Analysis

Imbalance in Class Distribution:

The target variable of Diabetes\_binary showed stark class imbalance in the dataset because it contained substantially more observations that were non-diabetic than diabetic cases. Model performance suffered especially from recall and F1-Score because the models demonstrated an inclination toward predicting the majority non-diabetic cases.

Computational Demand:

Random Forest and AdaBoost algorithms needed expensive resources for learning from datasets featuring large numbers due to their ensemble approach.

Model Interpretability:

The transparent decision-making process of Decision Trees remained superior to Random Forest and AdaBoost because their decision processes lack interpretation making them less suitable for medical facilities that demand model explainability.

# Conclusion and Recommendations

The main research objective focused on evaluating which machine learning models including Random Forest and Decision Tree and AdaBoost would best identify diabetes risk through health data analysis. This research employed "Diabetes Binary Health Indicators Dataset (BRFSS 2015)" coupled with appropriate feature selection methods to determine that HighBP, BMI and Age show the greatest predictive strength. The Random Forest model showed the highest prediction strength through an 85% accuracy rating which maintained an equal balance between precision and recall and F1-score. The model demonstrates strong capacity for predicting diabetes risk while simultaneously helping identify cases early.

Recommendations:

* Random Forest model integration within healthcare institutions should focus on early diabetes detection by using three essential health indicators that include BMI, HighBP and Age.
* The prevention programs should focus on undertaking specific interventions toward high-risk groups that result from model-based identifications through health education and lifestyle improvement strategies.
* The precision of predictions together with the adaptability of the model can be improved by integrating current medical data which comes from both wearables and electronic health records.
* The integration of machine learning tools should become part of public health systems according to policy recommendations to streamline diabetes prevention and management.

Future Work:

* The predictive differential will be improved by implementing ensemble techniques which merge Random Forest with Gradient Boosting and additional advanced predictive models.
* The model requires datasets covering various geographic areas to attack potential bias problems while building generalizable predictions.
* The implementation of SHAP or LIME explainable AI tools will provide interpretation tools for models to enhance the trust healthcare professionals have in the system.

# Reflections

Avoid generalization and focus on personal development and the research journey in a critical and objective way.

* 1. **Selected Research Methodology**

Reflection of the Research Process:

I followed a structured method when studying how machine learning models can predict diabetes throughout my research work. Data preprocessing served as the initial step during which I cleaned and normalized the dataset for maintaining data consistency. My next step involved using SelectKBest to determine BMI and Age and HighBP as the leading predictors of diabetes. I executed tests and evaluations on three machine learning models including Random Forest and AdaBoost and Decision Tree. The selected models suited structured medical datasets despite their ability to generate interpretable findings. The methodological framework delivered organized steps to reveal the role of health indicators in diabetes prediction processes.

Reflection on the Merits

The chosen methodology maintains both high accuracy rates and understandable results as its main strength. A performance evaluation between Random Forest and AdaBoost and Decision Tree allowed researchers to find the best model for predicting diabetes risk determination. The Random Forest model proved to be the most dependable solution because it yielded exceptional accuracy together with precision and recall metrics and F1-score thus establishing itself as suitable for practical applications. The SelectKBest operation improved model performance by selecting the key health indicators BMI, HighBP and Age to increase efficiency and remove unnecessary features. The models trained using data preprocessing techniques with normalization methods and dataset splitting procedures received input from uniform and consistent information.

Reflection on the Limitations

The chosen methodology showed multiple weaknesses in addition to its known advantages. Even though AdaBoost exhibited high accuracy it failed to detect real diabetic cases effectively. Healthcare applications dealing with false negatives would experience severe consequences because the model successfully identified non-diabetic individuals but poorly detected diabetic patients at risk. The Decision Tree model showed poor generalizability on new data points because it fitted perfectly to training data. The dataset suffered from severe unbalance since diabetic cases were significantly outnumbered by non-diabetic cases. The class imbalance might have introduced prediction bias into model performance so additional methods such as oversampling and undersampling and weighted classification were needed to enhance model fairness.

Reflection on the Potential Pitfalls

The main drawback of this method came from the high computational demands associated with Random Forest and AdaBoost ensemble models. The processing needs of these models were too high and their memory requirements stretched too far beyond real-time application needs especially when processing in restricted resource environments. The interpretations from Decision Tree models worked well yet the complex structure of Random Forest alongside AdaBoost diminished model understanding because analysts lost their ability to track how predictions were generated. The exclusion of features through statistical ranking during the selection process might have neglected essential variable interactions that play a role in predicting diabetes cases. The model's generalized capability on other datasets remains unknown since external validation did not occur on distinct test data. Future research should verify the model performance on various datasets by implementing different sampling methods which would strengthen overall model stability.

* 1. **Alternative Research Methodologies**

Alternative Research Methodologies in View of Outcomes

The research used Random Forest as well as AdaBoost and Decision Tree to forecast diabetes risk but different methods might deliver supplementary findings or boost prediction efficiency. Artificial Neural Networks together with other Deep Learning models offer a possible research alternative when looking to detect complicated non-linear health indicators' connections. ANNs demonstrate potential for better recall since they learn advanced data patterns while needing large datasets and much computational power to prevent their model from adapting excessively. The Logistic Regression model functions as an alternative since its simplified algorithm offers the benefit of producing easier interpretations compared to the complex machine learning algorithms presently in use.

The analysis would benefit from a combination of qualitative information through patient lifestyle surveys that would enhance a deeper interpretation of numeric health measurement outcomes. The proposed method provides unlimited potential to grasp diabetes risk completely because it combines patient feedback and medical measurement data. The use of stacking models that combine Random Forest with Gradient Boosting and Neural Networks would have improved prediction accuracy through their collective classifier effectiveness.

Lessons Learned in View of Outcomes

Several critical findings obtained from the research outcomes provide knowledge about methodology selection for future studies. Model selection emerges as a critical step which determines how accuracy relates to recall and interpretability in the system. The results from Random Forest indicated good performance in all aspects but in healthcare AdaBoost's poor recall performance demonstrated that accuracy should not be used solely as the performance indicator to identify high-risk patients (true positives). Future research needs to determine strategies which will enhance recall levels without compromising the accuracy standards.

Selection of features will significantly influence the performance efficiency of models. The employment of SelectKBest simplified the training process yet potentially discarded significant interdependent relationships among features that could prove useful. Future research needs to employ Principal Component Analysis (PCA) as a dimensionality reduction method to hold essential feature relationships while decreasing the data complexity level.

Training models faced difficulties attributing accuracy to the minority (diabetic) class because of dataset imbalance. To achieve model fairness the implementation of Synthetic Minority Over-sampling Technique (SMOTE) or adjustments on class weights represents an appropriate solution. Changes to the system design would help detect diabetic cases with greater accuracy which would minimize misdiagnoses in actual medical settings.

The research suggests future work should utilize deep learning with hybrid models for better diabetes prediction because of their successful outcomes.

* 1. **Recommended Actions and Future Considerations**

Recommended Actions

Improve Model Performance with Hybrid Approaches:

Future research on boosting model performance should combine Random Forest with other ensemble models such as AdaBoost or develop Neural Networks as a feature learning model.

Address Class Imbalance for Fairer Predictions:

Performance issues regarding recall emerged because the dataset contained a greater number of non-diabetic cases compared to diabetic cases. Future research should apply either SMOTE operations or undersampling or class-weight adjustments procedures to enhance model prediction capabilities.

The model sensitivity needs better evaluation through alternative metrics like AUC-ROC and Precision-Recall curves.

Enhance Interpretability for Clinical Application:

In healthcare, model transparency is critical. While Decision Trees are interpretable, Random Forest and AdaBoost are black-box models. Using Explainable AI (XAI) methods like SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) can help understand feature importance and ensure trust among healthcare professionals.

Expand Data Sources for Improved Accuracy:

Future research should increase data sources to enhance accuracy levels.

Studies should use real-time patient data from diverse datasets which include wearable health devices and electronic health records and various demographic groups to enhance generalization.

Using extra lifestyle variables including diet and exercise and stress levels can improve the predictive abilities of such models for diabetes risk assessment.

Future Considerations

Exploring Deep Learning Models:

Deep learning architecture research using CNNs and LSTMs represents a recommendation for future studies since they offer stronger capabilities for medical imaging data along with time-series health monitoring.

Federated Learning systems should be studied as a method to conduct model training across multiple health institutions without breaking data security rules.

Validation on Diverse Populations:

The models require testing with fresh datasets originating from separate demographic areas of different income levels to achieve better reliability for practical healthcare delivery.

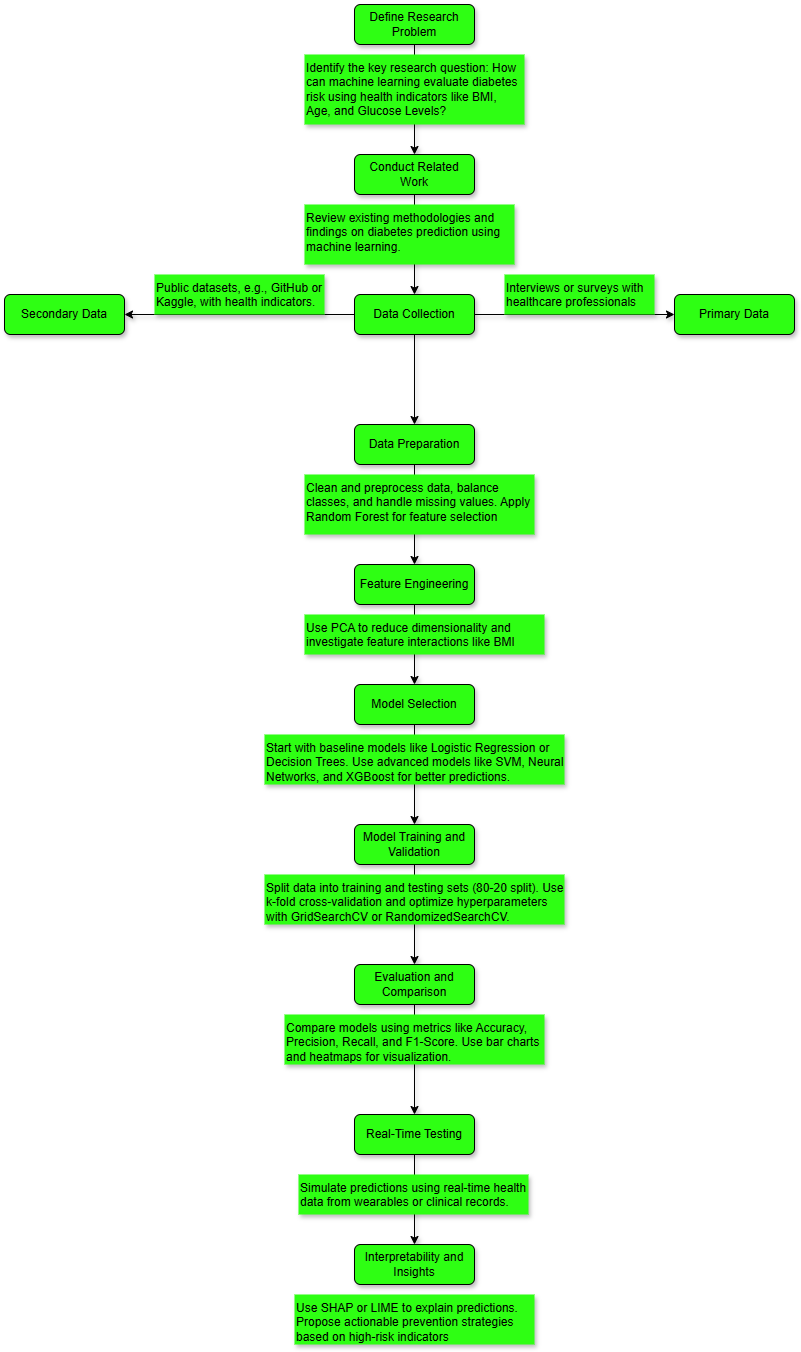
Real-World Deployment and Usability Testing:

Research results could build a clinical decision support system to enable physicians to use model predictions for early intervention of their at-risk patients.

The development of a mobile app along with a prototype should allow patients to enter their health-related data in order to obtain customized risk assessments.

* 1. **Recommended Methodology**

Updated version of paper Methodology (flowchart, block diagram) with discussion.



Updated Methodology

A complete updated research methodology for predicting diabetes risk through machine learning models appears in the provided flowchart. It integrates the following stages:

Define Research Problem:

Research goals should be clearly stated to include the examination of diabetes risk assessment using BMI alongside Age and Glucose Levels as health indicator variables.

The initial objective of this research that guides the study becomes established at this point.

Conduct Related Work:

An examination of past studies will help determine methodologies together with existing challenges and empty areas within the field of diabetes prediction research.

The research requires this step to integrate meaningfully with existing research projects.

Data Collection:

The research obtains first-hand data through health professional surveys and professional interviews for delivering contextual insights.

For secondary data collection researchers should depend on accessible structured health indicators that are available on GitHub or Kaggle platforms.

The goal of this research is to obtain a diverse and suitable dataset to support analysis.

Data Preparation:

The collected data requires preparation through data cleaning procedures and also requires class balancing methods and missing value treatment.

Use Random Forest method to choose important predictors consisting of BMI, Age and Glucose Levels for data selection.

The goal for this step is to establish high-quality data while managing unnecessary data dimensions.

Feature Engineering:

The research must discover multi-dimensional relationships between BMI and age variables and utilize PCA techniques to boost computational effectiveness.

The goal of this stage is to refine the features because it will result in better prediction results.

Model Selection:

The practice commences with baseline models starting from Logistic Regression and Decision Trees for benchmarking purposes.

The implementation proceeds with advanced machine learning models including SVM (optimized kernels) and Neural Networks as well as Gradient Boosting examples XGBoost and LightGBM.

The objective is to establish the model which optimally combines performance quality with computational processing speed.

Model Training and Validation:

The training and testing data sets should be created with an 80-20 split and k-fold cross-validation must be used to enhance generalizability.

The hyperparameter search should be conducted either through GridSearchCV or RandomizedSearchCV.

The main goal is to obtain strong and dependable predictive outcome.

Evaluation and Comparison:

An evaluation process should determine the models based on Accuracy, Precision, Recall and F1-Score measurement values.

Bar charts and heatmaps should be used for visual interpretation of the results.

The goal is to choose the best-performing model which will be analyzed for its strength and inadequacies during this phase.

Real-Time Testing:

Real-time testing of prediction models uses current wearables and clinical record data to ensure operational utility.

The researcher aims to validate the developed model through real-life practical applications.

Interpretability and Insights:

SHAP, LIME and other Explainable AI methods should be used to create transparent predictions which healthcare providers can act upon.

The team should develop preventive action plans for at-risk individuals.

The purpose ensures both the model delivers results that are meaningful and simple to interpret.

**References**

[1]Ahmed, U., Issa, G.F., Khan, M.A., Aftab, S., Khan, M.F., Said, R.A.T., Ghazal, T.M. and Ahmad, M. (2022). Prediction of Diabetes Empowered With Fused Machine Learning. IEEE Access, [online] 10, pp.8529–8538. doi:https://doi.org/10.1109/ACCESS.2022.3142097.

[2] Khan, F.A., Zeb, K., Al-Rakhami, M., Derhab, A. and Bukhari, S.A.C. (2021). Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review. IEEE Access, 9, pp.43711–43735. doi:https://doi.org/10.1109/access.2021.3059343.

[3] Jacobs, P.G., Herrero, P., Facchinetti, A., Josep Vehi, Kovatchev, B., Breton, M., Cinar, A., Konstantina Nikita, Doyle, F., Bondia, J., Tatej Battelino, Castle, J.R., Konstantia Zarkogianni, Narayan, R. and Mosquera-Lopez, C. (2023). Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls and opportunities. IEEE Reviews in Biomedical Engineering, pp.1–19. doi:https://doi.org/10.1109/rbme.2023.3331297.

[4] Shaheen, I., Javaid, N., Nabil Alrajeh, Asim, Y. and Aslam, S. (2024). Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection using Deep Learning and eXplainable Artificial Intelligence. IEEE access, 12, pp.66516–66538. doi:https://doi.org/10.1109/access.2024.3398198.

[5] Syed, A.H. and Khan, T. (2020). Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study. IEEE Access, 8, pp.199539–199561. doi:https://doi.org/10.1109/access.2020.3035026.

https://github.com/Aemiii91,https://github.com/schmurtzm (2023). *Release notes: Onion V4.1 | Onion*. [online] Github.io. Available at: https://onionui.github.io/blog/release-notes/4.1.0 [Accessed 10 Feb. 2025].

‌

Asana (2021). *Decision Tree Analysis: 5 Steps to Make Better Decisions • Asana*. [online] Asana. Available at: https://asana.com/resources/decision-tree-analysis.

‌

IBM (2021). *Random Forest*. [online] Ibm.com. Available at: https://www.ibm.com/think/topics/random-forest.

‌

Wizards, D.S. (2023). *Understanding the AdaBoost Algorithm*. [online] Medium. Available at: https://medium.com/@datasciencewizards/understanding-the-adaboost-algorithm-2e9344d83d9b.

‌