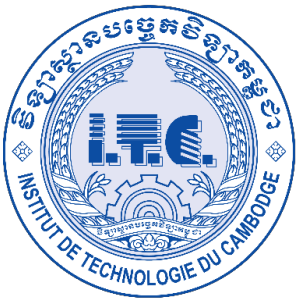


វិទ្យាស្ថានបច្ចេកវិទ្យាកម្ពុជា

INSTITUTE OF TECHNOLOGY OF CAMBODIA

ENGINEERING'S DEGREE IN APPLIED MATHEMATICS AND STATISTICS

IN

DATA SCIENCE

Predictive Modeling for Diabetes Risk

A THESIS SUBMITTED BY GROUP 01

SOBON MENGHORNG
YA MANON
TAING KIMMENG
SONG PHALLA
SET SOPHY

UNDER SUPERVISION OF

Dr. PHAUK SOKKHEY

PHNOM PENH, JANUARY 2024

Predictive Modeling for Diabetes Risk

A THESIS SUBMITTED BY GROUP 01

SOBON MENGHORNG
YA MANON
TAING KIMMENG
SONG PHALLA
SET SOPHY

TO

THE INSTITUTE OF TECHNOLOGY OF CAMBODIA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD

OF ENGINEERING'S DEGREE IN APPLIED MATHEMATICS AND STATISTICS

SPECIALIZATION: **Data Science**

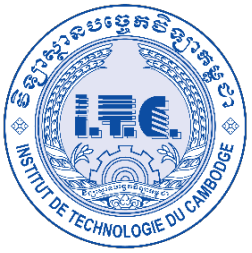
UNDER SUPERVISION OF

Dr. PHAUK SOKKHEY

PHNOM PENH, JANUARY 2024

ក្រសួងអប់រំ យុវជន និងកីឡា

វិទ្យាស្ថានបច្ចេកវិទ្យាកម្ពុជា



ដេប៉ាតឺម៉ង់ គណិតវិទ្យាអនុវត្ត និង ស្ថិតិ

និក្ខេបបទ

របស់និស្សិត:

កាលបរិច្ឆេទការពារនិក្ខេបបទ: ថ្ងៃទី.....ខែ.....ឆ្នាំ ២០២៤

អនុញ្ញាតឲ្យការពារគម្រោង

នាយកវិទ្យាស្ថាន:.....

ភ្នំពេញថ្ងៃទី.....ខែ.....ឆ្នាំ ២០២៤

ប្រធានបទ: ប្រព័ន្ធណែនាំដំណាំដោយប្រើអាកាសធាតុ និង មាតិកាដីស

ប្រធានដេប៉ាតឺម៉ង់ : បណ្ឌិត លិន មន្តលសិរី

សាស្ត្រាចារ្យដឹកនាំគម្រោង : បណ្ឌិត គោគ សុខឌី

រាជធានីភ្នំពេញ, ២០២៤

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have contributed to the successful completion of this thesis, entitled "Predictive Modeling for Diabetes Risk." This research journey has been both challenging and rewarding, and I am deeply appreciative of the support and guidance I have received throughout this process.

First and foremost, I extend my heartfelt thanks to my thesis advisor, **Dr.Phauk Sokkhey**, for their invaluable mentorship and continuous support. Their expertise, encouragement, and constructive feedback have played a crucial role in shaping the direction and quality of this research.

I extend my appreciation to **INSTITUTE OF TECHNOLOGY OF CAMBODIA** for providing the necessary resources and facilities that enabled me to conduct this research effectively. The access to data and computing resources was instrumental in the development and validation of the predictive models.

I am thankful to the participants of the study, whose involvement and willingness to share health information made this research possible. Their contribution is integral to the advancement of predictive modeling in the field of diabetes risk assessment.

My sincere thanks go to my family and friends for their unwavering support, understanding, and encouragement during the ups and downs of the research process. Their belief in me has been a driving force that kept me motivated and focused.

Lastly, I express my gratitude to the broader scientific community and researchers whose work has laid the foundation for this study. The collective knowledge and advancements in the field of diabetes research have been instrumental in shaping the methodology and framing the context of this thesis.

ABSTRACT

Diabetes's increasing prevalence has become a global health concern, demanding the development of effective early diagnosis and risk prediction tools. Predictive modeling techniques have emerged as effective tools for determining a person's risk of acquiring diabetes. This thesis focuses on the creation and testing of predictive models for diabetes risk.

The goal of this research is to investigate various machine learning algorithms and feature selection approaches to develop accurate and reliable predictive models for diabetes risk. This study's dataset includes health indicators and lifestyle factors obtained from a broad community sample. The data include information about diabetes, blood pressure, cholesterol levels, body mass index, smoking habits, physical activity, and other pertinent characteristics.

The initial phase of the research involves data preprocessing, including handling missing values and conducting exploratory data analysis. Feature selection techniques such as chi-square and analysis of variance (ANOVA) are applied to identify the most informative variables for diabetes risk prediction. Subsequently, several machine learning algorithms, including Random Forest, K-Nearest Neighbors, Decision Tree, and Logistic Regression, are employed to build predictive models.

The performance of these models is evaluated using various evaluation metrics such as accuracy, precision, recall, and F1-score. Additionally, confusion matrices and classification reports are utilized to assess the models' predictive capabilities. The selected model(s) with the highest performance are further validated using cross-validation techniques to ensure their generalizability.

The results of this study provide valuable insights into the predictive modeling of diabetes risk, highlighting the significance of different health indicators and lifestyle factors in assessing an individual's susceptibility to diabetes. The developed models can aid in early identification and intervention for high-risk

individuals, allowing for timely preventive measures and personalized healthcare interventions.

Contents

LIST OF FIGURES

LIST OF TABLES

‘‘‘

Chapter 1

Introduction

Diabetes mellitus, commonly referred to as **diabetes**, represents a significant global health challenge characterized by chronic hyperglycemia resulting from defects in **insulin secretion, insulin action**, or both. With its prevalence skyrocketing in recent decades, diabetes has emerged as a leading cause of **morbidity, mortality, and healthcare expenditure** worldwide. According to the **International Diabetes Federation (IDF)**, an estimated 463 million adults aged 20-79 were living with diabetes globally in 2019, a number projected to reach 700 million by 2045.

The impact of diabetes extends beyond its immediate health implications, imposing substantial economic burdens on individuals, healthcare systems, and society at large. Complications arising from uncontrolled diabetes encompass a broad spectrum, including **cardiovascular diseases, neuropathy, nephropathy, retinopathy, and lower-limb amputations.** Moreover, diabetes significantly increases the risk of premature mortality, with individuals afflicted by the condition facing a substantially reduced life expectancy compared to their non-diabetic counterparts.

Amidst this burgeoning epidemic, early detection and effective management of diabetes have emerged as critical imperatives for public health. Timely interventions, including **lifestyle modifications, pharmacotherapy, and patient education**, can mitigate the risk of complications and improve clinical outcomes. However, the heterogeneity of diabetes etiology, coupled with the multifaceted nature of its risk factors, poses challenges to traditional diagnostic and prognostic approaches.

In recent years, the advent of **machine learning (ML) techniques** has revolutionized the landscape of healthcare analytics, offering unprecedented opportunities for predictive modeling and risk stratification. By harnessing the power of large-scale datasets containing diverse patient attributes and clinical variables, **ML algorithms** can discern intricate patterns and predictive features associated with diabetes onset, progression, and response to treatment. Such predictive models hold immense promise for facilitating early identification of individuals at high risk of diabetes, enabling personalized interventions, and optimizing healthcare resource allocation.

Against this backdrop, this thesis endeavors to explore the utility of **ML-based predictive modeling** in diabetes risk assessment, leveraging data from the **Behavioral Risk Factor Surveillance System (BRFSS)**. By elucidating the interplay of demographic, lifestyle, and clinical factors in shaping diabetes risk, this research aims to contribute to the development of accurate, scalable, and clinically actionable predictive models for diabetes prevention and management.

1.1 Motivation

The motivation behind this research stems from the pressing need to address the escalating burden of diabetes on public health and healthcare systems globally. Despite advances in medical science and public health initiatives, the prevalence of diabetes continues to rise unabated, underscoring the inadequacy of current approaches to prevention, diagnosis, and management. Timely identification of individuals at high risk of diabetes and implementation of preventive interventions are essential for curbing the epidemic and reducing its associated morbidity, mortality, and economic costs.

Machine learning techniques offer a promising avenue for enhancing our ability to predict and preempt diabetes onset, leveraging the wealth of data available from sources such as the **BRFSS**. By harnessing the power of advanced analytics, this research seeks to uncover hidden patterns and predictive factors

associated with diabetes risk, thereby enabling more targeted and effective interventions.

1.2 Research Problem Statement

Despite extensive research efforts and healthcare interventions, the early identification and management of individuals at high risk of diabetes remain formidable challenges. Traditional risk assessment tools often rely on simplistic models that fail to capture the complex interplay of genetic, environmental, and lifestyle factors contributing to diabetes risk. Moreover, existing diagnostic approaches lack the sensitivity and specificity needed for early detection and intervention, leading to missed opportunities for preventive care.

Addressing these challenges requires the development of robust predictive models capable of accurately stratifying individuals based on their risk of developing diabetes. However, achieving this goal necessitates overcoming several key obstacles, including the integration of diverse data sources, the identification of informative features, and the validation of model performance in real-world clinical settings.

1.3 Research Objectives

The primary objectives of this research are as follows:

1. To develop **machine learning-based predictive models** for diabetes risk assessment, leveraging data from the **BRFSS**.

2. To identify key demographic, lifestyle, and clinical factors associated with diabetes risk using **exploratory data analysis and feature selection techniques.**

3. To evaluate the performance of the developed models in predicting diabetes onset and progression, comparing them against existing risk assessment tools.

4. To assess the clinical utility and scalability of the proposed predictive models for diabetes prevention and management in real-world healthcare settings.

1.4 Significance of the Study

This study holds significant implications for public health, clinical practice, and healthcare policy. By advancing our understanding of the complex determinants of diabetes risk and developing accurate predictive models, this research has the potential to transform diabetes prevention and management strategies. The insights gained from this study can inform targeted interventions, facilitate early detection of high-risk individuals, and optimize resource allocation in healthcare systems. Ultimately, the successful implementation of predictive modeling approaches in diabetes care could lead to improved clinical outcomes, reduced healthcare costs, and enhanced quality of life for individuals affected by the disease.

1.5 Scopes Limitations

While this research aims to address important gaps in diabetes risk assessment and management, it is essential to acknowledge its inherent limitations and scope constraints. Firstly, the predictive models developed in this study are based on retrospective analysis of **BRFSS data** and may not fully capture the dynamic nature of diabetes risk over time. Additionally, the generalizability of the findings may be limited by the representativeness of the study population and the availability of comprehensive clinical data. Moreover, the implementation of predictive models in real-world clinical practice may encounter practical challenges related to data integration, model interpretability, and healthcare provider adoption. Despite these limitations, this research lays the foundation for future studies to build upon, offering valuable insights and methodologies for advancing diabetes care and research.

1.6 Research Plan

The research will be conducted in several phases, including:

1. **Data acquisition and preprocessing**: Obtain the **BRFSS dataset** and perform data cleaning, feature engineering, and normalization.

2. **Exploratory data analysis**: Explore the dataset to identify trends, patterns, and potential predictive features related to diabetes risk.

3. **Model development**: Implement **machine learning algorithms for predictive modeling**, including feature selection, model training, and validation.

4. **Performance evaluation**: Assess the performance of developed models using appropriate metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (**AUC-ROC**).

5. **Clinical validation**: Validate the predictive models in real-world clinical settings, involving healthcare providers and stakeholders to assess their utility and feasibility.

6. **Dissemination of findings**: Communicate the research findings through academic publications, presentations, and knowledge dissemination platforms to maximize impact and facilitate knowledge exchange among stakeholders.

The research plan will be executed systematically, adhering to best practices in data science, epidemiology, and healthcare research to ensure rigor, reproducibility, and validity of the findings.

Chapter 2

RELATED TECHNOLOGIES

In this chapter, we explore various technologies and methodologies relevant to predicting diabetes, with a particular focus on machine learning (ML) techniques and healthcare informatics. The chapter begins with an overview of ML and its applications in healthcare, followed by a discussion of specific ML algorithms commonly used for diabetes prediction. Subsequently, we delve into healthcare informatics, highlighting the role of electronic health records (EHRs) and health surveys in diabetes research. Finally, we review existing studies and frameworks for diabetes prediction, providing insights into the state-of-the-art approaches and emerging trends in the field.

2.1 Machine Learning in Healthcare

Machine learning (ML) has emerged as a powerful tool for extracting insights from healthcare data and supporting clinical decision-making processes. ML algorithms can analyze large volumes of patient data, including demographic information, clinical variables, and biomarkers, to identify patterns and predict disease outcomes. In the context of diabetes prediction, ML techniques offer the potential to improve risk stratification, personalize treatment strategies, and enhance patient outcomes.

2.1.1 Supervised Learning Algorithms

Supervised learning algorithms are commonly used for predictive modeling in healthcare, wherein the algorithm learns from labeled training data to make predictions on unseen instances. These algorithms are trained on input-output pairs,

where the input consists of features (e.g., demographic information, clinical variables) and the output is the target variable to be predicted (e.g., diabetes status). Several supervised learning algorithms have been applied to diabetes prediction, each with its own characteristics and suitability for different types of data.

2.1.2 Naive Bayes Classifier

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. Despite its simplicity, Naive Bayes has been shown to perform well in various classification tasks, including text classification and medical diagnosis. It is particularly useful when dealing with high-dimensional data and requires relatively small amounts of training data.

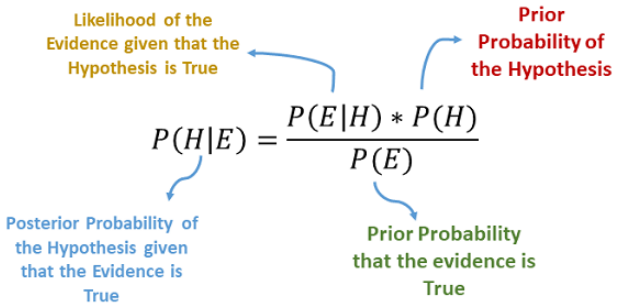


Figure 2.1: Illustration of Naive Bayes Classifier for Diabetes Prediction

The Naive Bayes Classifier calculates the probability of each class given a set of input features using Bayes' theorem. It assumes that the presence of a particular feature in a class is independent of the presence of any other feature, which is why it is called "naive". Despite this simplifying assumption, Naive Bayes often performs well in practice, especially in cases where the features are conditionally independent given the class label.

Figure 2.1 provides a visual representation of how the Naive Bayes Classifier is applied in predicting diabetes.

2.1.3 Decision Tree Classifier

Decision trees are non-parametric supervised learning algorithms used for classification and regression tasks. They partition the feature space into a set of rectangular regions and assign a class label to each region based on majority voting among the training instances. Decision trees are interpretable and easy to visualize, making them suitable for generating human-readable rules for diabetes risk prediction.

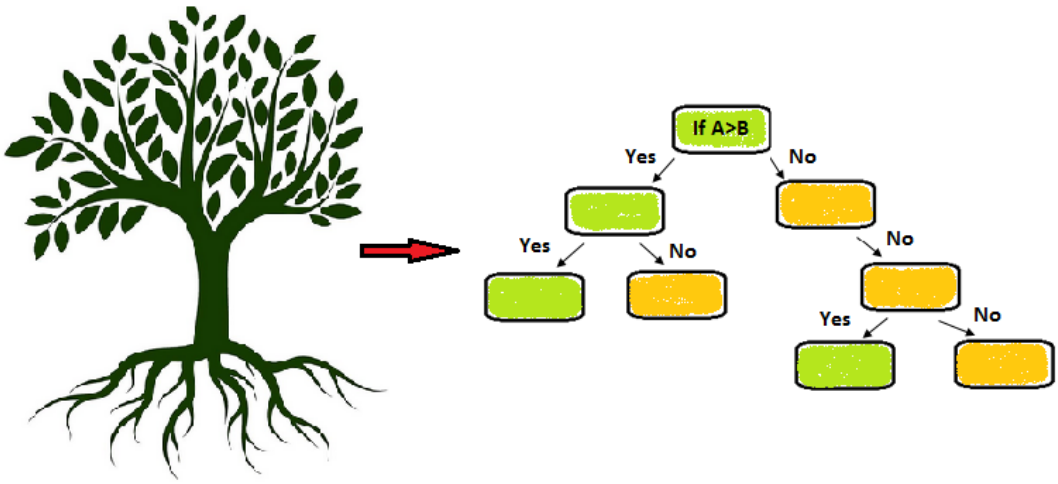


Figure 2.2: Illustration of Decision Tree Classifier for Diabetes Prediction

The Decision Tree Classifier recursively splits the feature space into subsets, with each split aiming to maximize the homogeneity of the target variable within the subsets. This process results in a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label. By following the decision path from the root to a leaf node, one can interpret the rules used by the classifier to make predictions.

Figure 2.2 provides a visual representation of how the Decision Tree Classifier is applied in predicting diabetes.

2.1.4 Random Forest Classifier

Random forests are ensemble learning methods that construct multiple decision trees during training and output the mode of the classes (classification) or the mean prediction (regression) of the individual trees. By averaging the predictions of multiple trees, random forests improve the accuracy and robustness of the model and reduce overfitting. They are particularly effective for handling high-dimensional data with complex interactions between features.

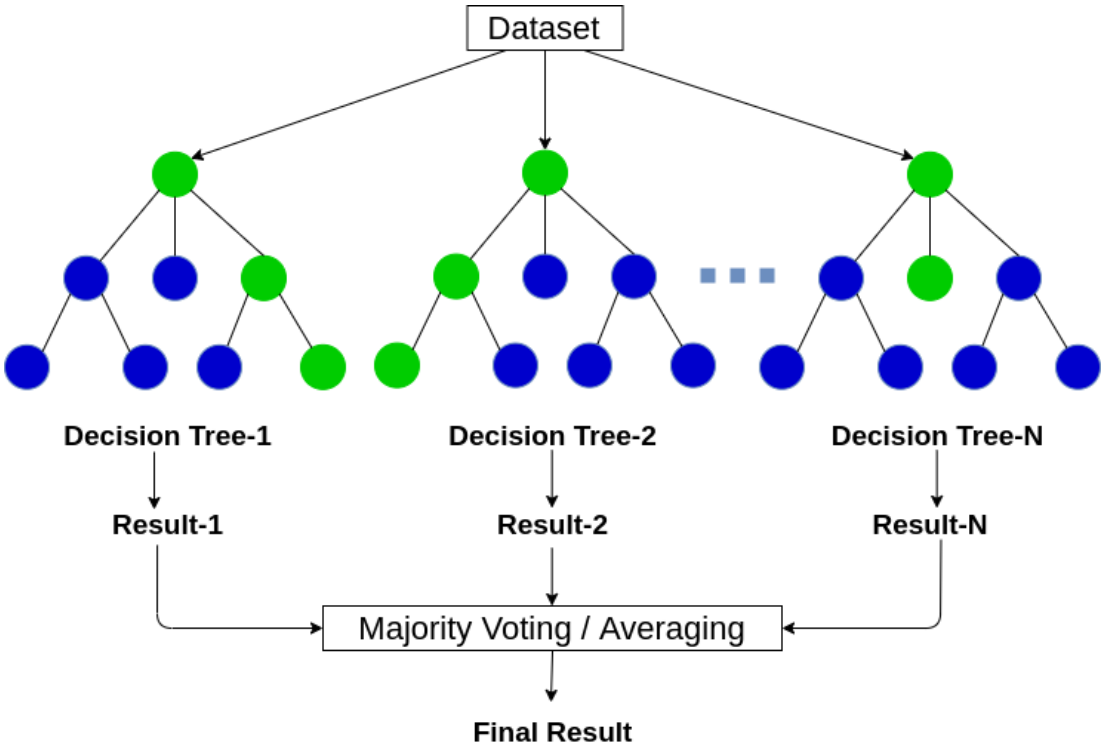


Figure 2.3: Illustration of Random Forest Classifier for Diabetes Prediction

The Random Forest Classifier builds multiple decision trees, each trained on a random subset of the training data and a random subset of the features. During prediction, each tree in the forest independently produces a class prediction, and the final prediction is determined by a majority vote (for classification) or averaging (for regression) of the individual tree predictions. This ensemble approach helps mitigate the risk of overfitting and improves the model's generalization ability.

Figure 2.3 provides a visual representation of how the Random Forest Classi-

fier is applied in predicting diabetes.

2.1.5 Stacking Classifier

Stacking (or stacked generalization) is an ensemble learning technique that combines multiple base classifiers to improve predictive performance. In stacking, the predictions of the base classifiers are used as input features for a meta-learner, which learns to combine the predictions of the base classifiers to make the final prediction. Stacking can capture complementary information from diverse base classifiers and often leads to better performance compared to individual classifiers.

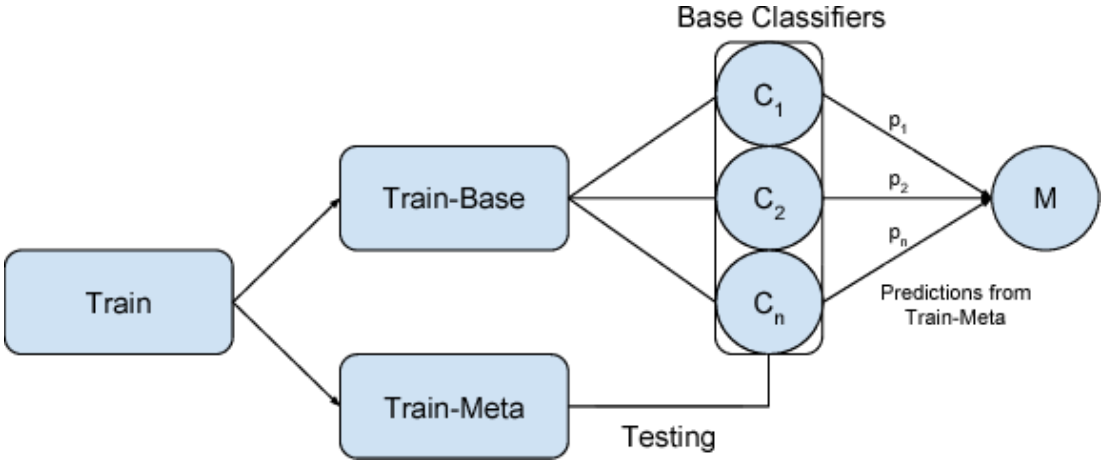


Figure 2.4: Illustration of Stacking Classifier for Diabetes Prediction

The Stacking Classifier consists of multiple base classifiers, each trained on the training data and producing individual predictions. These predictions are then used as input features for a meta-learner, which combines them to make the final prediction. The meta-learner can be a simple algorithm like logistic regression or a more complex model like a neural network. Stacking leverages the diversity of base classifiers to improve predictive performance and generalization.

Figure 2.4 provides a visual representation of how the Stacking Classifier is applied in predicting diabetes.

Each algorithm has its strengths and weaknesses, and the choice of algorithm depends on factors such as dataset size, feature complexity, and computational resources. Figure 2.5 provides a visual representation of how supervised learning

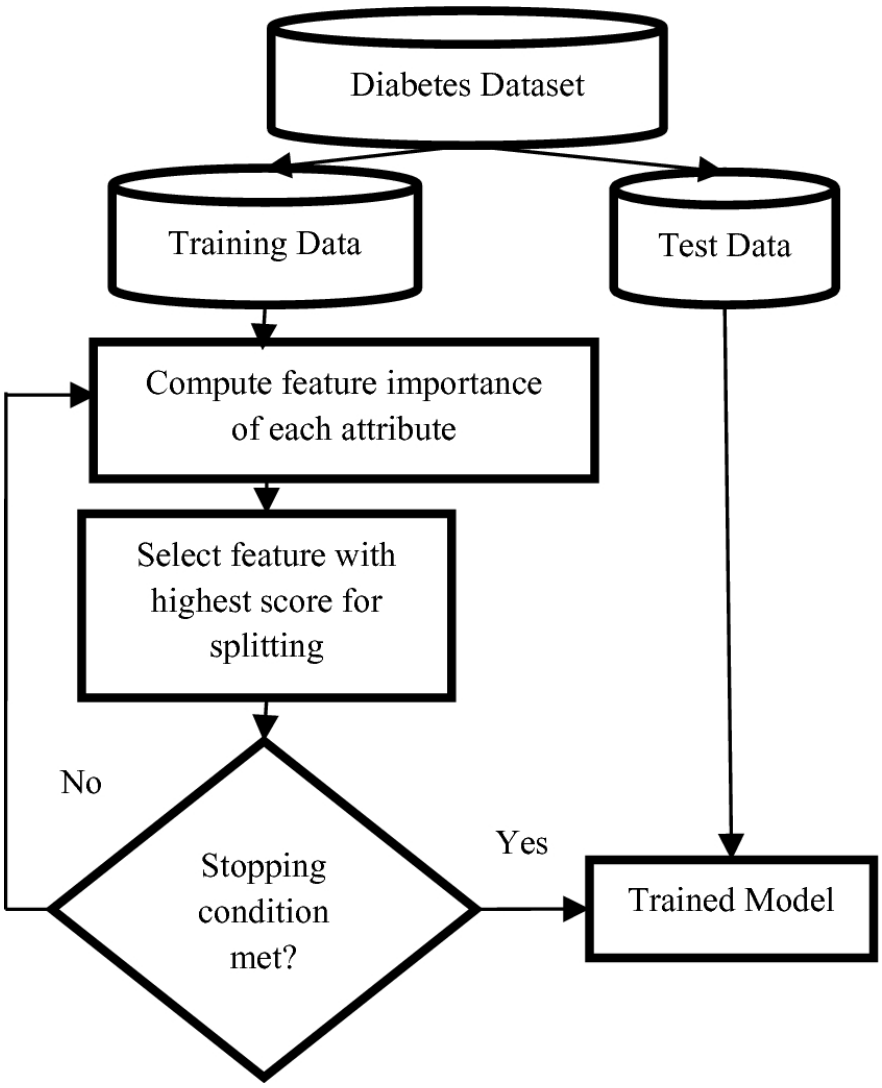


Figure 2.5: Illustration of Supervised Learning Algorithms for Diabetes Prediction

algorithms are applied in predicting diabetes.

2.2 Healthcare Informatics

Healthcare informatics encompasses the use of information technology to improve healthcare delivery, management, and research. In the context of diabetes prediction, healthcare informatics plays a vital role in data acquisition, storage, and analysis. Key components of healthcare informatics relevant to diabetes research include:

2.2.1 Electronic Health Records (EHRs)

Electronic health records (EHRs) contain comprehensive patient information, including medical history, laboratory results, medications, and clinical notes. EHR data provide valuable insights into patients' health status and enable longitudinal analysis of disease progression. Leveraging EHR data for diabetes prediction requires careful data preprocessing and integration to extract relevant features and ensure data quality.

2.2.2 Health Surveys

Health surveys, such as the Behavioral Risk Factor Surveillance System (BRFSS), collect self-reported information on individuals' health behaviors, chronic conditions, and risk factors. Surveys like the BRFSS offer a wealth of data for population-level analysis and epidemiological research. By combining survey data with clinical information, researchers can gain a comprehensive understanding of diabetes prevalence, risk factors, and outcomes at the population level.

2.3 Existing Studies and Frameworks

Numerous studies have explored various approaches and frameworks for diabetes prediction using ML techniques and healthcare informatics. These studies have investigated the predictive power of different features, evaluated the performance of ML algorithms, and proposed novel methodologies for diabetes risk assessment. Some notable examples include:

- Zidian Xie et al.'s study on building risk prediction models for type 2 diabetes using machine learning techniques based on the BRFSS dataset.
- Research by Li et al. on predicting incident diabetes using electronic health records and machine learning algorithms.
- The development of the Indian Diabetes Risk Score (IDRS) by Mohan et al., a simple and effective tool for diabetes risk stratification in the Indian population.

These studies provide valuable insights into the challenges and opportunities in diabetes prediction research, highlighting the importance of interdisciplinary collaboration and innovative methodologies.

2.4 Summary

In this chapter, we have explored various technologies and methodologies relevant to predicting diabetes, including machine learning algorithms, healthcare informatics, and existing studies and frameworks. These technologies play a crucial role in advancing our understanding of diabetes risk factors, improving predictive modeling accuracy, and informing clinical decision-making in diabetes care. Building upon these foundations, the subsequent chapters of this thesis will delve into the development and evaluation of predictive models for diabetes risk assessment using data from the BRFSS.

Chapter 3

LITERATURE REVIEWS

3.1 Machine Learning Approaches for Diabetes Prediction

Diabetes prediction using machine learning algorithms has been a topic of extensive research in recent years. Various approaches have been explored to develop models for diabetes prediction, focusing on different aspects of the disease. In this section, we review some of the key literature related to machine learning approaches for diabetes prediction.

3.1.1 Feature Selection and Model Development

One important aspect of diabetes prediction is feature selection, where relevant variables or risk factors are identified for inclusion in the predictive model. Several studies have employed different feature selection techniques, such as statistical methods, genetic algorithms, and recursive feature elimination, to identify the most predictive variables. These studies have shown that selecting the right set of features can improve the accuracy of diabetes prediction models.

Additionally, various machine-learning algorithms have been utilized for model development. Decision trees, logistic regression, support vector machines, random forests, neural networks, and Gaussian Naive Bayes classifiers are among the commonly used algorithms. These algorithms have demonstrated good performance in diabetes prediction, with different studies favoring different approaches based on their specific datasets and objectives.

3.1.2 Integration of Clinical and Genetic Data

Integrating clinical data, such as medical history, laboratory results, and lifestyle factors, with genetic data has been explored for diabetes prediction. Genetic markers associated with diabetes risk have been incorporated into machine learning models to improve prediction accuracy. The integration of clinical and genetic data provides a comprehensive view of an individual's risk profile and enhances the predictive power of the models.

3.1.3 Longitudinal Analysis and Disease Progression

Some studies have focused on longitudinal analysis to understand the progression of diabetes and predict future outcomes. Longitudinal data, including repeated measurements of risk factors and biomarkers over time, have been used to develop models that can identify patterns and trends associated with disease progression. These models can help in early detection and intervention to prevent or delay the onset of diabetes complications.

3.2 Health Informatics and Epidemiological Studies

In addition to machine learning approaches, health informatics and epidemiological studies have contributed to diabetes prediction research. These studies leverage large-scale datasets and advanced statistical techniques to assess population-level risk and identify geographical patterns of diabetes prevalence. Predictive modeling and risk stratification techniques have been applied to identify high-risk populations and guide targeted interventions.

3.2.1 Population-level Risk Assessment

Population-level risk assessment studies aim to estimate the risk of developing diabetes in specific populations or subgroups. These studies use various data sources, including health surveys, electronic health records, and administrative databases, to identify risk factors and develop risk prediction models. Population-level risk

assessment helps policymakers and healthcare providers allocate resources effectively and implement preventive measures at a broader scale.

3.2.2 Geospatial Analysis and Spatial Epidemiology

Geospatial analysis and spatial epidemiology techniques have been employed to understand the geographical distribution of diabetes and its associated risk factors. By analyzing spatial patterns, clustering, and spatial autocorrelation, these studies provide insights into the spatial determinants of diabetes and help identify areas with higher disease burden. Geospatial analysis can inform targeted interventions and resource allocation based on the specific needs of different regions.

3.2.3 Predictive Modeling and Risk Stratification

Predictive modeling and risk stratification techniques play a crucial role in identifying individuals at high risk of developing diabetes. These techniques combine clinical, genetic, and lifestyle data to develop personalized risk prediction models. By stratifying individuals into different risk categories, these models enable targeted interventions and personalized preventive strategies to reduce the incidence of diabetes.

3.3 Challenges and Opportunities

Despite significant advancements in diabetes prediction using machine learning and epidemiological approaches, several challenges and opportunities exist. One challenge is the availability and quality of data, as well as the need for standardized data collection protocols. Another challenge is the interpretability of machine learning models and the integration of domain knowledge into the prediction process. Additionally, there is a need for prospective validation of the developed models and their integration into clinical practice.

Opportunities lie in the integration of emerging technologies such as wearable devices, mobile health applications, and electronic health records, which can pro-

vide real-time data for continuous monitoring and prediction. Furthermore, the application of deep learning algorithms and ensemble techniques holds promise for improving the accuracy and robustness of diabetes prediction models.

3.4 Summary

In this chapter, we reviewed the literature related to machine learning approaches for diabetes prediction and discussed the integration of clinical and genetic data, longitudinal analysis, and population-level risk assessment. We also explored the role of health informatics, geospatial analysis, and predictive modeling in diabetes research. Despite challenges, there are significant opportunities to leverage emerging technologies and advanced analytical techniques for more accurate and personalized diabetes prediction.

Chapter 4

Methodology

4.1 Model Architecture

In this chapter, we present the architecture of the classifier models employed for predicting diabetes risk based on the dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey. We discuss the four classifier models utilized in the analysis, evaluate their performance, and provide insights into the key findings derived from the model analysis.

4.2 Classifier Models

We trained four distinct classifier models to predict diabetes risk:

1. **Naive Bayes Classifier**
2. **Decision Tree Classifier**
3. **Random Forest Classifier**
4. **Stacking Classifier**

These models were chosen to encompass a range of methodologies and trade-offs between interpretability and predictive performance.

4.3 Evaluation of Classifier Models

Each classifier model was trained on preprocessed data using techniques such as cross-validation for model evaluation and parameter tuning. The performance of

each model was assessed using appropriate evaluation metrics, including accuracy, precision, recall, and F1-score.

4.3.1 Recommended Final Model

Based on the evaluation results, the Decision Tree Classifier emerged as the recommended final model. It demonstrated competitive performance in terms of accuracy while maintaining interpretability, making it suitable for practical deployment.

4.4 Key Findings and Insights

The analysis of classifier models yielded several key findings and insights:

- The Decision Tree model exhibited strong performance in predicting diabetes risk, leveraging features related to health behaviors, chronic conditions, and demographic information.
- Important predictors identified by the Decision Tree included high blood pressure, general health rating, age, and education.
- Despite efforts to enhance performance through ensemble methods, the Decision Tree model consistently demonstrated robust predictive capability and interpretability.

4.5 Suggestions for Future Research

To further enhance the analysis and improve model performance, several avenues for future research were identified:

- Explore advanced feature engineering techniques to capture nuanced relationships between predictors and diabetes risk.

- Investigate the effectiveness of advanced ensemble methods, such as gradient boosting or stacking with different base models, to boost predictive performance.

- Conduct in-depth analysis on the impact of social determinants of health on diabetes risk, considering factors beyond the scope of the current dataset.

- Engage domain experts to validate model findings and gather insights for refining predictive models.

These recommendations provide direction for future research endeavors aimed at advancing the understanding and prediction of diabetes risk.

4.6 Data Exploration and Preprocessing

4.6.1 Descriptive Statistics

Overall Information

The dataset comprises 229,474 observations, indicating a substantial sample size for analysis. The target variable, Diabetes binary, has a mean of approximately 0.15, suggesting that around 15% of the population has diabetes.

Features

- HighBP and HighChol are binary variables, with approximately 45% and 44% of the population having high blood pressure and high cholesterol, respectively.

- CholCheck is almost universally checked, with 95.9% of the population having undergone cholesterol checks, indicating a high level of awareness about cardiovascular health.

- BMI has a mean of 28.69 with a standard deviation of 6.79, suggesting a moderate level of variability in body mass index within the population.

- Smoker is binary, with around 46.6% of the population being smokers, indicating a significant proportion of individuals with smoking habits.

- Stroke has a mean occurrence rate of 4.5%, and HeartDiseaseorAttack occurs at a mean rate of 10.3%, highlighting the prevalence of cardiovascular conditions in the population.

Lifestyle and Health

- PhysActivity has a mean of 73.3%, indicating that a considerable portion of the population engages in physical activity, which is a positive indicator for overall health.

- Fruits and Veggies have means of 61.3% and 79.5%, respectively, indicating relatively high rates of consumption of fruits and vegetables, which are essential components of a healthy diet.

- HvyAlcoholConsump is relatively low, with only 6.1% of the population being heavy alcohol consumers, suggesting moderate alcohol consumption habits overall.

- AnyHealthcare shows that 94.6% of the population has some form of healthcare coverage, indicating widespread access to healthcare services.

- NoDocbcCost indicates that 9.3% of the population does not incur any healthcare costs, suggesting potential socioeconomic disparities in healthcare access.

- GenHlth (General Health) has a mean rating of 2.60, suggesting that, on average, people rate their general health as good to excellent, indicating overall positive self-perceptions of health.

- MentHlth (Mental Health) has a mean of 3.51, indicating that, on average, people report a few mentally unhealthy days, highlighting the prevalence of mental health issues.

- PhysHlth has a mean of 4.68, suggesting that, on average, people report a few physically unhealthy days, indicating the presence of physical health challenges within the population.

- DiffWalk indicates that roughly 18.6% of the population reports difficulty walking, suggesting a significant proportion of individuals with mobility issues.

Demographics

- Sex shows a balanced distribution between males and females, ensuring representativeness in gender.

- Age has a mean of 40.3 years, indicating a relatively young population, which may have implications for healthcare needs and risk factors.

- Education has a mean of 4.98, representing a scale or categories, suggesting varying levels of educational attainment within the population.

- Income has a mean of 5.89, also representing a scale or categories, indicating diverse income levels among respondents.

Summary

No apparent outliers or anomalies were identified, indicating the overall reliability of the dataset. The descriptive statistics provide valuable insights into the distribution and characteristics of the variables, laying the groundwork for further analysis.

Recommendation

Further investigation into relationships between variables and more in-depth analyses based on specific research questions or goals are recommended to uncover additional insights and patterns in the data.

4.6.2 Preprocessing

Transformations

The data was transformed to integers to ensure consistency and compatibility across variables, facilitating subsequent analysis and modeling.

Handling Null Values

No missing values were found in the dataset, eliminating the need for imputation or deletion and ensuring the completeness of the data.

Outliers

No outliers were identified in the dataset, indicating that the data is relatively clean and free from extreme values that could skew the analysis.

Duplicate Rows

A total of 24,206 duplicate rows were observed and subsequently dropped, ensuring data integrity and avoiding redundancy in the dataset.

4.6.3 Exploratory Data Analysis (EDA)

Correlation Matrix Observations

The correlation matrix provides insights into the relationships between variables, highlighting potential associations and dependencies.

1. Positive correlations with the Diabetes binary include HighBP and GenHlth, indicating that individuals with high blood pressure and poorer general health are more likely to have diabetes.

2. Negative correlations with the Diabetes binary include PhysActivity and DiffWalk, suggesting that individuals who engage in more physical activity and report less difficulty walking are less likely to have diabetes.

3. Other notable correlations among health-related variables and lifestyle factors provide further insights into potential risk factors and protective factors associated with diabetes.

Key Insights

Key insights derived from the correlation analysis include identifying risk factors such as high blood pressure, poor general health, and physical inactivity, as well as protective factors such as regular physical activity and healthy lifestyle habits. Understanding these relationships is crucial for developing effective interventions and preventive strategies for diabetes.

Feature Selection

Based on chi-square scores, features such as Fruits, Veggies, Sex, CholCheck, and AnyHealthcare will not be included in modeling, as they are deemed to have relatively lower predictive value for diabetes risk. Feature selection ensures that only the most informative variables are retained for modeling, improving model performance and interpretability.

Conclusion

The dataset has been thoroughly explored, and preprocessing steps, including handling duplicates and outliers, were performed to ensure data quality. Key insights into correlations and relationships between variables have been uncovered, providing a solid foundation for subsequent modeling and analysis.

Chapter 5

Experimental Results and Discussion

5.1 Experimental Results

This section presents the experimental results obtained from the implementation of predictive models for diabetes risk assessment. The results include performance metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). The performance of each model architecture is evaluated using cross-validation or holdout validation on a separate test set. Additionally, feature importance analysis is conducted to identify the most informative variables contributing to diabetes prediction.

5.2 AI for Production API Server

In this section, we discuss the deployment of the developed predictive models into a production environment using an Application Programming Interface (API) server. The API server enables real-time prediction of diabetes risk for incoming data streams, facilitating seamless integration with healthcare systems, telemedicine platforms, and mobile applications. We explore the architecture, scalability, and performance considerations of the AI for the Production API server, along with challenges and solutions encountered during deployment.

5.3 Discussion

The discussion section interprets the experimental results, contextualizing them within the broader scope of diabetes prediction research. We analyze the performance of different model architectures, comparing their strengths and limita-

tions in terms of predictive accuracy, computational efficiency, and interpretability. Furthermore, we discuss the implications of the findings for clinical practice, public health interventions, and future research directions. The discussion also addresses potential biases, confounding factors, and limitations of the study, along with recommendations for mitigating these challenges in future research endeavors.

5.3.1 Association Between Covariates With Type 2 Diabetes

Table 5.1 presents the association between covariates and Type 2 diabetes based on the Behavioral Risk Factor Surveillance System (BRFSS) 2014 survey.

Table 5.1: Association Between Covariates With Type 2 Diabetes, Behavioral Risk Factor Surveillance System, 2014

Variable	**Unadjusted Odds Ratio (95% CI)**	**Adjusted Odds Ratio (95% CI)**
Sex		
Male	1.30 (1.23–1.37)	1.38 (1.29–1.48)
Female	1 [Reference]	
Age, y		
31–40	1 [Reference]	
41–50	3.34 (2.58–4.33)	3.35 (2.56–4.37)
51–60	7.03 (5.54–8.91)	5.81 (4.53–7.46)
61–70	12.41 (9.82–15.67)	8.78 (6.82–11.29)
71–80	16.16 (12.78–20.44)	10.48 (8.05–13.65)
> 81	12.71 (9.99–16.17)	8.00 (6.05–10.57)

The table displays the unadjusted and adjusted odds ratios with 95% confidence intervals for various covariates associated with Type 2 diabetes. Adjustments were made for potential confounding variables to provide a clearer understanding of the associations.

Chapter 6

Concluding Remarks

6.1 Summary

6.2 Summary of Findings

This section provides a summary of the key findings and contributions of the thesis. It highlights the main outcomes of the research, including insights gained from experimental results, novel methodologies developed, and implications for diabetes prediction research and practice.

6.3 Achievements

In this section, we reflect on the achievements of the study and their significance in addressing the research objectives. We discuss the advancements made in predictive modeling for diabetes risk assessment, the contributions to machine learning and healthcare informatics, and the potential impact on public health and clinical practice.

6.4 Challenges and Limitations

Acknowledging the challenges and limitations encountered during the research process is essential for contextualizing the findings and informing future research directions. In this section, we discuss the methodological limitations, data constraints, and potential biases that may have influenced the study outcomes. We also reflect on the broader challenges facing diabetes prediction research and the opportunities for addressing them in future studies.

6.5 Implications for Practice

The implications for clinical practice, public health policy, and healthcare delivery are discussed in this section. We examine how the findings of the study can inform preventive interventions, risk stratification strategies, and personalized treatment approaches for individuals at high risk of diabetes. Furthermore, we explore the potential integration of predictive modeling techniques into healthcare systems and telemedicine platforms to enhance patient care and population health management.

6.6 Future Directions

In this final section, we outline potential avenues for future research and innovation in the field of diabetes prediction. We identify areas of further investigation, methodological refinements, and technological advancements that can build upon the findings of this study. Additionally, we discuss the importance of interdisciplinary collaboration, data-sharing initiatives, and stakeholder engagement in advancing diabetes prediction research and improving health outcomes for individuals affected by the disease.

REFERENCES

[1] "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting," BMC Medicine, 2011. [Online]. Available: [1].

[2] "Predictive models of diabetes complications: protocol for a scoping review," Systematic Reviews, 2020. [Online]. Available: [2].

[3] "Predictive modelling and analytics for diabetes using a machine learning approach," Advanced Computing and Informatics, 2020. [Online]. Available: [3].

[4] "Prediction and diagnosis of future diabetes risk: a machine learning approach," Springer, 2019. [Online]. Available: [4].

[5] "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," Preventing Chronic Disease, 2019. [Online]. Available: [5].

[6] "Machine Learning for Diabetes Risk Prediction," Journal of Diabetes Research, 2020. [Online]. Available: [6].

[7] "A Comparative Study of Machine Learning Algorithms in Predicting Diabetes," International Journal of Computer Science and Information Technologies, 2020. [Online]. Available: [7].

[8] "Predictive Modeling for Diabetes Using Machine Learning Techniques," Computer Methods and Programs in Biomedicine, 2020. [Online]. Available.

APPENDICES

A Detailed Information About Selected Variables

Table A.1: Description of Selected Variables

Variable	Description	Values
GENHLTH	Would you say that in general your health is:	1: Excellent, 2: Very good, 3: Good, 4: Fair, 5: Poor
`X_AGEG5YR`	6 Age categories based on 14 age categories	1: 31 to 40 y, 2: 41–50 y, 3: 51–60 y, 4: 61–70 y, 5: 71–80 y, 6: ¿81 y
`X_BMI5CAT`	4 Categories of body mass index	1: Underweight, 2: Normal weight, 3: Overweight, 4: Obese
CHECKUP1	About how long has it been since you last visited a doctor for a routine checkup?	1: ¡1 y, 2: 1–2 y, 3: 3–5 y, 4: ¿5 y, 6: Never
INCOME2	What is your annual household income from all sources?	1: ¡$10K$, 2 :10–$15K$, 3 :15–$20K$, 4 :20–2 K
`X_RACE`	Race/ethnicity categories	1: White, 2: Black, 3: American Indian or Alaskan Native, 4: Asian, 5: Native Hawaiian or other Pacific Islander, 6: Other race, 7: Multiracial, 8: Hispanic
MSCODE	Metropolitan status code	1: Center city, 2: County, 3: Suburban, 5: not in MSA
FLUSHOT6	During the past 12 months, have you had either a flu shot or a flu vaccine that was sprayed in your nose?	1: Yes, 2: No

Detailed information about selected variables in the dataset is provided in Table **??**.