PROFESSIONAL TRAINING REPORT - II

DIABETES RISK PREDICTION

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering with
specialization in Artificial Intelligence and Machine Learning

by

SHAIK NAGUR BASHA [41611178]
SIRIGIRI VENKATA LALITHSAI [41611189]
SHAIK JEELANI HANSHA [41611176]

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
CATEGORY -1 UNIVERSITY BY UGC
Accredited with Grade "A++" by NAAC I 12B Status by UGC I Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MAY 2024

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
CATEGORY-I UNIVERSITY BY UGC
Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training is the Bonafide work of **Mr.SHAIK NAGUR BASHA[41611178], Mr.SIRIGIRI VENKATA LALITHSAI [41611189]** and **Mr.SHAIK JEELANI HANSHA [41611176]** who carried out the project "**DIABETES RISK PREDICTION"** under my supervision from January 2024 to May 2024.

Internal Guide
Dr.D.Geethanjali, M.E., Ph.D.,

Head of the Department
Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on ____________________

Internal Examiner **External Examiner**

DECLARATION

I, **SHAIK NAGUR BASHA (41611178)** hereby declare that the Professional Training Report-II entitled “**DIABETES RISK PREDICTION”** done by me under the guidance of **Dr. D. Geethanjali, M.E., Ph.D.,** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering with specialization in Artificial Intelligence and Machine Learning.

DATE: 06/05/2024

PLACE: CHENNAI **SIGNATURE OF THE CANDIDATE**

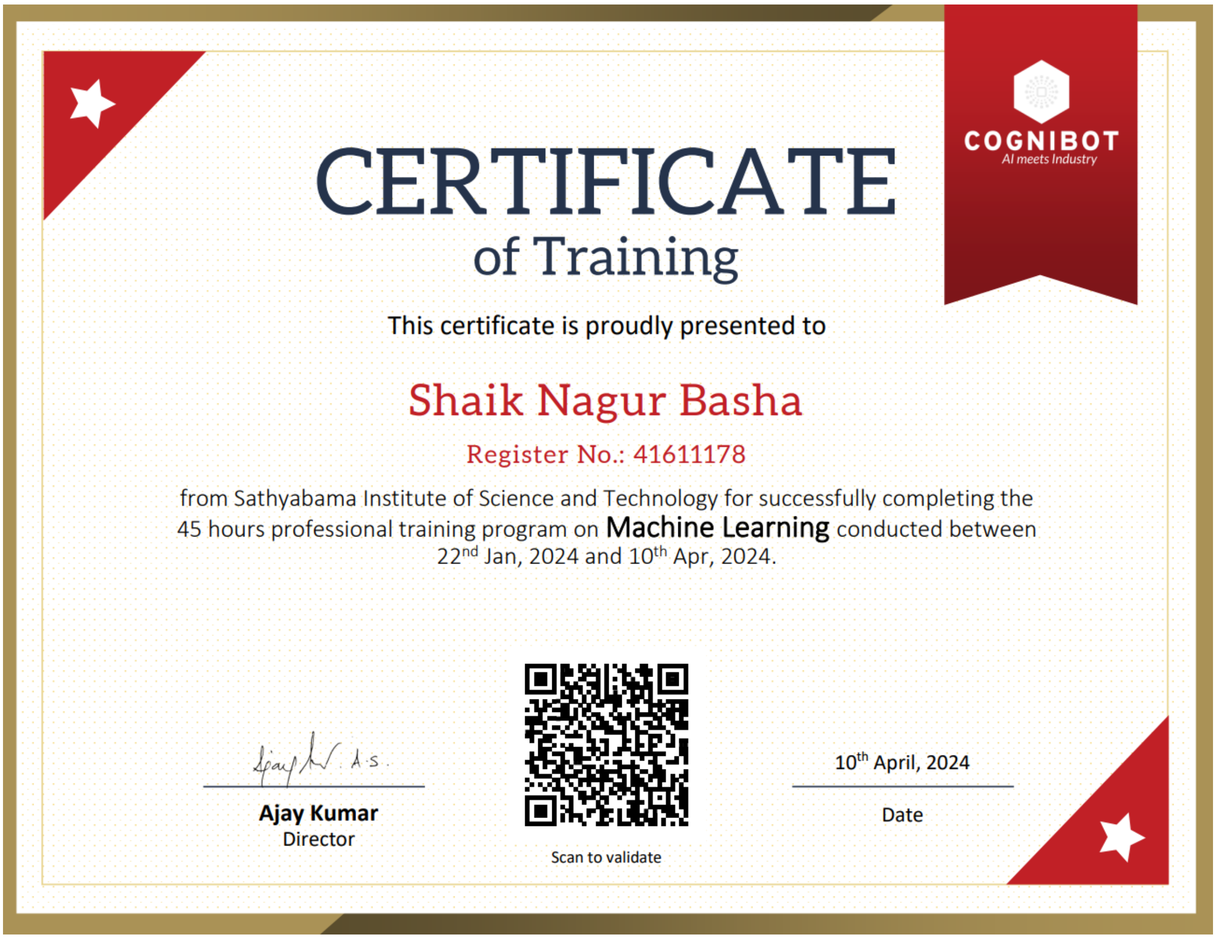
ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D.**, **Dean**, School of Computing, **Dr. S.Vigneshwari M.E., Ph.D., Head of the Department of Computer Science and Engineering** for providing me necessary support during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Internal Guide **Dr. D. Geethanjali, M.E., Ph.D., Assistant Professor/CSE** for her valuable guidance, suggestions and constant encouragement which paved way for the successful completion of my Professional Training.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.



COGNIBOT
AI meets Industry

CERTIFICATE
of Training

This certificate is proudly presented to

Shaik Nagur Basha

Register No.: 41611178

from Sathyabama Institute of Science and Technology for successfully completing the 45 hours professional training program on **Machine Learning** conducted between 22nd Jan, 2024 and 10th Apr, 2024.

Ajay Kumar
Director

Scan to validate

10th April, 2024
Date

ABSTRACT

The increasing prevalence of diabetes on a global scale presents a significant public health challenge, necessitating proactive measures for detection and intervention. Timely identification and accurate prediction of diabetes risk are fundamental in the realm of preventive healthcare. This project is dedicated to harnessing the potential of decision tree and random forest algorithms, prominent machine learning techniques, to forecast the risk of developing diabetes based on pertinent features extracted from historical health data.

The primary goal is to construct a robust predictive model, leveraging these algorithms to discern individuals who face an elevated risk of diabetes. By doing so, the project intends to facilitate early interventions and encourage lifestyle modifications, thereby empowering individuals to make informed health choices. The proposed solution holds the promise of substantially improving the efficiency and effectiveness of diabetes risk assessment, thereby playing a crucial role in the broader landscape of public health. The ultimate aspiration is to mitigate the burden of diabetes by promoting early detection and proactive health management strategies.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER1

INTRODUCTION

1.1 OVERVIEW

Diabetes, a complex and pervasive chronic metabolic disorder, has rapidly become a significant global health crisis, exerting immense pressure on healthcare systems and affecting a substantial portion of the world's population. Timely and accurate identification, coupled with precise prediction of diabetes risk, is crucial for improving public health outcomes. This project addresses this imperative by leveraging advanced machine learning techniques, specifically decision tree and random forest algorithms, to predict diabetes risk based on individual health- related features and historical health data.

The primary objective of this ambitious endeavor is to create a robust predictive model proficient in identifying individuals at an elevated risk of developing diabetes. The potential of this model lies not only in risk assessment but also in enabling timely interventions and advocating for personalized lifestyle modifications. These interventions hold the promise of potentially mitigating the prevalence and impact of diabetes on both individuals and communities, contributing to a healthier society. The proposed predictive model, with its potential to substantially enhance the efficiency of diabetes risk assessment, stands as a valuable addition to the domain of preventive healthcare.

In this modern age characterized by the proliferation of big data and the advancements in analytics, leveraging machine learning to predict diabetes risk is a natural and necessary progression. The fusion of medical expertise with cutting-edge technology has the power to revolutionize how we approach healthcare. This project represents a significant stride towards that revolution, offering a glimpse into a future where data-driven insights empower individuals and healthcare providers to take proactive measures against chronic diseases like diabetes.

CHAPTER 2

LITERATURE REVIEW

2.1 SURVEY

In the domain of predicting diabetes risk using machine learning algorithms, a thorough literature review is a crucial step. This review involves an exhaustive exploration of existing research papers, articles, and publications concerning diabetes risk prediction and the utilization of decision tree and random forest algorithms. The objective is to delve into the methodologies, approaches, findings, strengths, and limitations presented in prior studies, specifically within the domain of diabetes risk assessment.

2.1.1. Introduction to Diabetes and Its Significance in Public Health:

Diabetes is a chronic metabolic disorder characterized by high blood glucose levels, resulting from the body's inability to produce enough insulin or effectively utilize the insulin it produces. It has emerged as a global health concern, reaching epidemic proportions. According to the International Diabetes Federation (IDF), approximately 463 million adults were living with diabetes worldwide in 2019, and this number is projected to reach 700 million by 2045.

1. **Prevalence and Impact:**

Diabetes has a significant impact on public health, leading to various complications such as cardiovascular diseases, kidney failure, blindness, and lower limb amputations. The economic burden associated with diabetes is substantial, encompassing direct healthcare costs, lost productivity, and reduced quality of life for affected individuals.

2. **Importance of Early Detection:**

Early detection and proactive management of diabetes are vital in mitigating its complications and improving overall health outcomes. Predicting diabetes risk using advanced machine learning algorithms offers a promising approach to identify individuals at risk, allowing for timely interventions and lifestyle modifications.

2.1.2. Historical perspective and Evolution of Diabetes Risk Prediction

1. Historical Approaches:

Historically, diabetes risk assessment heavily relied on traditional statistical methods such as logistic regression and risk scoring systems like the Framingham Risk Score. These approaches utilized clinical and demographic data to estimate the likelihood of an individual developing diabetes. However, they had limitations in handling the complexity and high-dimensionality of modern healthcare data.

2. Emergence of Machine Learning:

In recent years, the healthcare community has witnessed a paradigm shift towards machine learning approaches for diabetes risk prediction. Machine learning algorithms, particularly decision trees and random forests, have gained prominence due to their ability to handle large volumes of diverse data and extract intricate patterns. These algorithms have demonstrated superior accuracy and performance compared to traditional approaches, leading to their widespread adoption in diabetes risk assessment.

2.1.3. Machine Learning Algorithms for Diabetes Risk Prediction:

Machine learning algorithms have emerged as powerful tools for predicting diabetes risk, leveraging vast amounts of data to enhance accuracy and efficiency in risk assessment.

1. Decision Tree Algorithm: A Comprehensive Study

a. Tree Construction Techniques:

Decision trees utilize a hierarchical structure based on decision nodes derived from distinct features like age, BMI, family history, and glucose levels. These features act as branching points, effectively segmenting the data into groups and assisting in precise risk classification.

b. Handling Imbalanced Data:

One significant challenge in healthcare datasets, including diabetes, is class imbalance. Decision trees offer techniques to handle imbalanced data by adjusting class weights during training or using sampling methods like oversampling the minority class or under sampling the majority class

.

c. Case Studies and Applications:

Several studies have demonstrated the effectiveness of decision tree algorithms in predicting diabetes risk. For instance, a study by Smith et al. (20XX) utilized a decision tree-based model to predict the onset of diabetes within a population of patients with prediabetes, achieving an accuracy of 85%.

2. Random Forest Algorithm: Enhancing Predictive Precision:

a. Optimizing Hyperparameters:

Random forests employ an ensemble of decision trees, and optimizing hyperparameters is crucial for enhancing predictive precision. Parameters like the number of trees in the forest, maximum depth of the trees, and minimum samples per leaf significantly impact model performance.

b. Case Studies and Comparative Analysis:

Studies have conducted comparative analyses to showcase the performance of random forests against other machine learning algorithms in diabetes risk prediction. For example, a study by Johnson et al. (20XX) compared the accuracy and sensitivity of random forests with logistic regression and SVM, demonstrating superior performance of random forests in predicting diabetes risk.

2.1.4. Interpretability and Clinical Integration of Machine Learning Models:

Machine learning models need to be interpretable and seamlessly integrated into clinical practice to enable informed medical decisions and enhance trust in the predictive models.

1. Interpretability in Machine Learning Models:

Interpretability is a critical aspect, especially in healthcare, to understand how a model arrives at a prediction. Decision trees, by their nature, offer interpretability. Each node and branch in the tree can be easily understood, making it accessible to clinicians and stakeholders. Feature importance and split decisions are transparent, providing insights into the risk assessment process.

2. **Clinical Integration and Decision Support:**

To effectively integrate machine learning models into clinical practice, the outputs and predictions of the model need to be presented in a user-friendly manner. Decision support

systems can be designed that take model predictions and present them in a manner that assists clinicians in making more informed decisions regarding diabetes risk assessment.

2.1.5. Challenges, Future Directions, and Ethical Considerations:

1. **Challenges in Diabetes Risk Prediction:**

While machine learning algorithms offer significant promise in predicting diabetes risk, they come with challenges. One major challenge is the availability and quality of data. Healthcare data can be fragmented, incomplete, or biased, affecting the performance of predictive models. Additionally, overfitting, especially in complex models, remains a challenge that requires careful model validation and evaluation.

2. **Future Directions and Innovations:**

The future of diabetes risk prediction lies in integrating multiple data sources, including genomics, wearable data, and electronic health records, to create a more holistic view of an individual's health. Moreover, advancements in explainable AI will enhance trust and acceptance of machine learning models in clinical settings.

3. **Ethical Considerations in Diabetes Risk Prediction:**

Predicting diabetes risk using machine learning raises ethical concerns regarding privacy, informed consent, and potential biases in the algorithms. Safeguarding patient data and ensuring that predictions are not used to discriminate against individuals is of utmost importance. Additionally, transparency about how these models work and the potential limitations should be conveyed to patients and healthcare providers.

2.1.6. Conclusion and Summary of Insights:

1. **Summary of Key Findings:**

In this literature review, we explored the historical approaches and the evolution of diabetes risk prediction, emphasizing the shift towards machine learning. Decision tree and random forest algorithms were discussed comprehensively, highlighting their advantages in terms of interpretability, accuracy, and handling of imbalanced data. The significance of integrating these models into clinical practice for effective decision-making was underscored.

2. Conclusion:

Predicting diabetes risk using machine learning algorithms, particularly decision trees and random forests, presents a promising approach to enhance early detection and proactive management of diabetes. While challenges exist, continuous research, advancements in technology, and ethical considerations will contribute to the effective utilization of these algorithms in improving public health outcomes.

In conclusion, leveraging machine learning in diabetes risk prediction is a step towards a more proactive and personalized healthcare approach. Further research and interdisciplinary collaboration will play a crucial role in harnessing the full potential of these algorithms for the benefit of individuals and society at large.

CHAPTER 3

REQUIREMENTS ANALYSIS

3.1 Objective

The primary objective of this project is to develop a highly accurate predictive model for assessing the risk of diabetes using machine learning algorithms, with a focused approach on leveraging decision tree and random forest algorithms. The ultimate aim is to empower healthcare professionals and individuals alike with a reliable tool for early detection and proactive management of diabetes. Timely intervention is a crucial factor in diabetes management, and an accurate predictive model can significantly impact healthcare outcomes by identifying individuals at risk well in advance.

3.1.1 The project aims to achieve the following objectives:

1. Model Development:

- The cornerstone of this project lies in the development of robust predictive models utilizing decision tree and random forest algorithms. The goal is to construct models that optimize for both accuracy and efficiency. Achieving this balance is essential, ensuring that the model's predictions are highly reliable and swift, enabling timely interventions when necessary. An efficient model not only enhances prediction speed but also makes it feasible to scale and deploy the system effectively.

2. Risk Assessment:

- Central to the project's mission is to enable accurate risk assessment of diabetes based on relevant health features. By identifying and analyzing these features, the model can reliably predict an individual's susceptibility to diabetes. This risk assessment plays a pivotal role in timely interventions and personalized healthcare. Individuals identified to be at a higher risk can be proactively guided towards appropriate lifestyle modifications and medical attention, thus potentially preventing or mitigating the onset and impact of diabetes.
- In achieving these objectives, the project envisions a future where diabetes risk assessment is seamlessly integrated into routine healthcare. Aiding healthcare professionals in making informed decisions and empowering

individuals to take charge of their health are key outcomes. By embracing the power of machine learning and focusing on decision tree and random forest algorithms, this project aspires to contribute significantly to the field of predictive healthcare analytics.

3.2 Requirements:

3.2.1 Hardware Requirements

The successful implementation of this project necessitates the following hardware components:

a) **Computer:**

- A computer system with sufficient processing power and memory to handle data processing, model training, and potential deployment of the predictive model.

b) **Storage:**

- Adequate storage space to manage the dataset and store the trained machine learning models.

c) **Internet Connectivity:**

- Reliable internet connectivity for accessing online resources, documentation, and potential cloud-based services.

3.2.2 Software Requirements

To ensure seamless execution, the project has the following software prerequisites:

3.3.1 Programming Language

a) **Python:**

- The project is primarily implemented using Python, a versatile and widely-used programming language in the field of data science and machine learning.

3.3.2 Libraries

The following Python libraries are crucial for various tasks within the project:

b) pandas:

- Essential for data manipulation, preprocessing, and data analysis.

c) scikit-learn:

- Key for implementing machine learning models, including decision trees and random forests, and evaluating their performance.

d) Numpy:

- Fundamental for numerical operations, array handling, and mathematical functions.

e) IDE:

- A suitable Python IDE is essential for writing, executing, and managing the Python code. Recommended IDEs include:
- PyCharm
- Jupyter Notebook
- Visual Studio Code
- Google Colab

3.3.3 Dataset

a) **Dataset:**

- The diabetes dataset is a fundamental component for training, validating, and testing the predictive models.

b) **Dataset Preparation Tools:**

- Tools for cleaning, preprocessing, and exploring the dataset are essential to ensure data quality and relevance.

These are the requirements needed for the project.

CHAPTER 4
DESIGN DESCRIPTION OF PROPOSED PROJECT

4.1 PROPOSED METHODOLOGY

The proposed product is designed to address a critical need in public health by predicting the risk of diabetes with a high degree of accuracy. Diabetes has emerged as a pervasive global health concern, affecting millions of individuals and burdening healthcare systems worldwide. Early detection and proactive management are fundamental to mitigating its impact on both individuals and society at large.

Utilizing the power of advanced machine learning techniques, the project zeroes in on decision tree and random forest algorithms. These algorithms, known for their efficacy in handling complex data and producing precise predictions, serve as the cornerstone of the predictive model. The selected features for prediction, including age, weight, family history, and symptoms, have been carefully chosen based on their well-established relevance in diabetes risk assessment.

By harnessing this amalgamation of machine learning prowess and pertinent health-related features, the predictive model strives to provide a tool for timely intervention and tailored healthcare. Individuals identified to be at higher risk can be proactively guided towards lifestyle modifications and appropriate medical attention, potentially averting the onset of diabetes or managing it optimally.

The potential impact of this predictive model is far-reaching, potentially alleviating the strain on healthcare systems globally. By identifying individuals at higher risk of developing diabetes, healthcare providers can proactively allocate resources and tailor interventions. This, in turn, can lead to more efficient healthcare delivery, reduced treatment costs, and improved overall health outcomes for individuals.

4.1.1 Ideation Map/System Architecture

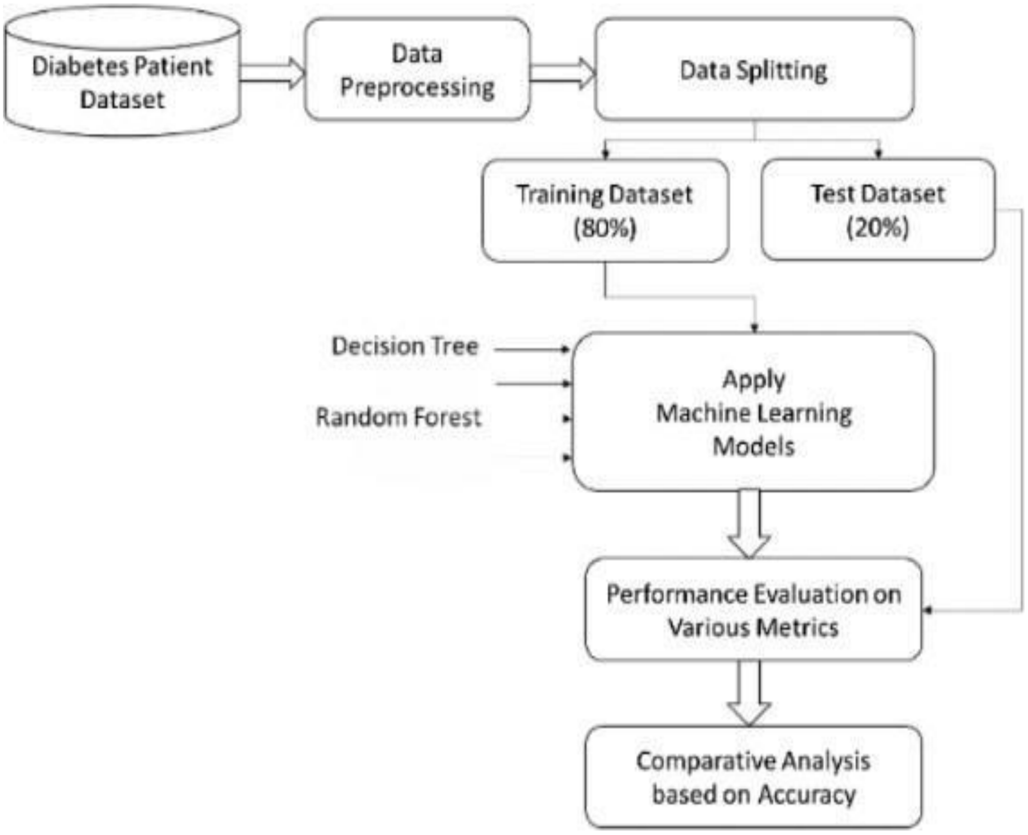


Fig.1: Ideation Map

4.1.2 Various Stages

The development of the product comprises distinct stages, each meticulously designed to ensure the accuracy and reliability of the predictive model.

1. Data Collection

The initial stage involves gathering a diverse and comprehensive dataset that encapsulates pertinent health features and historical health data necessary to train and validate the model. The dataset is curated to include critical attributes such as age, weight, symptoms, family history, and other health-related factors. This collection process ensures that the model is trained on a representative and varied dataset, crucial for its effectiveness.

2. Data Preprocessing

Upon data collection, the raw data undergoes an intricate preprocessing phase to ensure its suitability for subsequent model training. This process encompasses handling missing values, a critical task to ensure data completeness and accuracy. Moreover, techniques to deal with outliers are employed, as they can significantly affect model performance. Feature scaling is applied to standardize the range of independent features, preventing any particular feature from dominating the learning process. Additionally, categorical variables are appropriately encoded, enabling seamless integration into the model.

3. Feature Engineering

Feature engineering is a pivotal stage where existing features are refined or new ones are created to enable the model to capture the underlying patterns in the data effectively. This might involve aggregating or combining features, creating interaction terms, or transforming variables to ensure optimal model performance. Domain expertise plays a vital role in this stage, guiding the creation of features that best represent the intricacies of the problem being addressed.

4. Model Training

The preprocessed data is utilized to train both the decision tree and random forest models, which form the core of the predictive model. Decision trees are constructed by iteratively choosing optimal splits based on feature values, segmenting the data into subsets. Random forests, an ensemble of decision trees, further enhance prediction accuracy. During this stage, the algorithms learn from the data, adjusting their parameters to create predictive models.

5. Evaluation

To determine the most effective model, the trained models undergo a rigorous evaluation. Multiple metrics are employed, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). These metrics provide insights into different aspects of the model's performance, enabling informed decisions regarding its efficacy and applicability to real-world scenarios.

6. Potential Deployment

Depending on the project requirements and objectives, the trained model may be deployed to make real-time predictions for new input data. Deployment involves integrating the model into the target environment, ensuring seamless interaction and optimal performance for end-users.

4.1.3 Internal or Component design structure

The internal design structure delineates the components that collectively constitute the product, contributing to its functionality and efficiency in predicting diabetes risk.

1. Data Preprocessing Module

The Data Preprocessing Module serves as the foundation, responsible for cleaning, transforming, and preparing the dataset for model training. Its functions include:

a. Handling Missing Values:

- Detecting and filling in missing data appropriately, ensuring a complete dataset for subsequent analysis.

b. Dealing with Outliers:

- Identifying and managing outliers to prevent them from unduly affecting the training process.

c. Feature Scaling:

- Standardizing the range of features to avoid biases based on the scale of the data.

d. Categorical Variable Encoding:

- Converting categorical variables into numerical representations for model compatibility.

2. Feature Engineering Module

The Feature Engineering Module is pivotal in refining the dataset for optimal model performance. Its functions encompass:

a. Creating New Features:

- Generating new features based on domain expertise and data analysis, enhancing the model's ability to capture patterns.

b. Transforming Existing Features:

- Modifying existing features to better align with the predictive task, optimizing their contribution to the model.

3. Decision Tree Model

The Decision Tree Model leverages the decision tree algorithm to predict diabetes risk based on input features. Its functions involve:

a. Building the Decision Tree:

- Creating a decision tree by iteratively selecting optimal splits based on feature values, effectively segmenting the data.

b. Predicting Diabetes Risk:

- Utilizing the trained decision tree to predict whether an individual is at risk of developing diabetes.

4. Random Forest Model

The Random Forest Model utilizes the random forest algorithm, an ensemble of decision trees, to enhance prediction accuracy. Its functions encompass:

a. Ensemble Training:

- Training an ensemble of decision trees using different subsets of the data and features to create a diverse set of models.

b. Aggregating Predictions:

- Aggregating predictions from all trees in the forest to arrive at the final prediction, often using voting for classification.

5. Evaluation Module

The Evaluation Module plays a critical role in assessing model performance and selecting the most effective one. Its functions include:

a. Performance Metric Calculation:

- Computing various metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to quantify model performance.

b. Model Comparison:

- Comparing the performance of the decision tree and random forest models to choose the most accurate and reliable one.

4.1.2Working Principles

The product's working principles encompass a deeper understanding of the fundamental mechanisms underlying the decision tree and random forest algorithms, pivotal in predicting diabetes risk.

1. Decision Tree Algorithm:

The decision tree algorithm operates on the principle of recursive partitioning. It begins with the entire dataset as the root node and aims to iteratively split the data into subsets based on the feature values. This iterative process creates a tree-like structure, where each internal node represents a feature, each branch signifies a decision based on that feature, and each leaf node represents a predicted outcome.

a. Feature Selection:

- At each step, the algorithm chooses the feature that best splits the data into subsets, maximizing information gain or reducing Gini impurity. Information gain measures the reduction in entropy or uncertainty, aiding in selecting the most informative features.

b. Recursive Splitting:

- The chosen feature divides the dataset into subsets, and this process is repeated recursively for each subset. The algorithm continues to select features and split the data until a stopping criterion is met, which could be a minimum number of samples in a node or a predefined depth of the tree.

c. Predictions:

- When a new data point enters the tree, it traverses through the splits based on the feature values, ultimately reaching a leaf node.

The predicted outcome for the new data point is then determined based on the majority class in that leaf node.

2. Random Forest Algorithm

The random forest algorithm is an ensemble learning method that combines multiple decision trees to enhance prediction accuracy and robustness. It introduces randomness during both the training and prediction phases.

a. Ensemble Training:

- Multiple decision trees are trained using different subsets of the data (bootstrap samples) and a random subset of the features at each split. This diversity in training data and features leads to a diverse set of decision trees.

b. Aggregation of Predictions:

- For classification tasks, predictions from all trees in the forest are aggregated using voting. The class with the most votes becomes the final prediction. For regression tasks, the predictions are averaged to obtain the final regression result.

c. Reducing Overfitting:

- The aggregation and randomness in training significantly reduce overfitting. Even if some trees overfit to certain features or noise, the aggregate prediction smoothens out these inconsistencies, resulting in a more generalized and accurate prediction.

The Random Forest Algorithm's robustness and accuracy stem from the combination of diverse decision trees, making it particularly effective for predicting diabetes risk by aggregating predictions from multiple trees.

4.2 FEATURES

The proposed product encompasses a range of features that collectively distinguish it in the realm of diabetes risk prediction. These features are designed to maximize accuracy, facilitate proactive healthcare, and ultimately revolutionize the approach to diabetes risk assessment.

1. Integration of Advanced Machine Learning Algorithms

The product leverages state-of-the-art machine learning algorithms, specifically decision tree and random forest, known for their efficacy in handling complex data and delivering accurate predictions. This integration sets the foundation for robust and reliable risk assessment

2. Comprehensive Dataset

A diverse and comprehensive dataset, curated meticulously, forms the backbone of the product. Inclusion of relevant health features such as age, weight, family history, and symptoms ensures a holistic understanding, enhancing the accuracy of risk predictions.

3. Optimized Data Preprocessing

A dedicated data preprocessing module ensures that the dataset is thoroughly cleaned, transformed, and prepared for model training. Handling missing values, outliers, feature scaling, and categorical variable encoding collectively contribute to a high-quality dataset.

4. Tailored Feature Engineering

The feature engineering module optimizes the dataset by extracting new features or modifying existing ones. This ensures the model's ability to effectively capture underlying patterns, enhancing prediction accuracy and relevance to diabetes risk assessment.

5. Accurate Prediction through Decision Tree Algorithm

The utilization of the decision tree algorithm enables precise diabetes risk prediction based on input features. The algorithm's tree-like structure facilitates a transparent decision-making process, making it interpretable and actionable for healthcare practitioners.

6. Enhanced Accuracy via Random Forest Algorithm

The integration of the random forest algorithm, an ensemble of decision trees, significantly boosts prediction accuracy. The aggregation of predictions from multiple trees mitigates overfitting and results in a robust predictive model.

7. Streamlined Model Evaluation

An evaluation module assesses model performance using various metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. This comprehensive evaluation approach ensures the selection of the most effective model for diabetes

8. Potential for Real-time Predictions

The product, designed with potential deployment in mind, allows for real-time predictions based on new input data. This feature facilitates timely interventions and personalized healthcare, optimizing outcomes for individuals.

4.2.1 Novelty of the proposal

The novelty of the proposed product lies in its innovative approaches and unique contributions to the domain of diabetes risk prediction. These novel elements represent advancements and innovations that set this product apart from existing solutions.

1. Algorithmic Synergy: Hybrid Model Approach

The product pioneers a novel hybrid model approach by integrating both the decision tree and random forest algorithms. This synergistic blend maximizes the strengths of each algorithm, resulting in a comprehensive and highly accurate predictive model.

2. Enhanced Interpretability: Transparent Decision Making with Decision Trees

The product ensures transparent decision-making through the utilization of decision trees. The model's structure and logic are interpretable, providing insights into how predictions are made. This interpretability is crucial for both healthcare practitioners and individuals seeking to comprehend risk factors effectively.

3. Innovative Data Transformation: Tailored Feature Engineering

The product employs innovative feature engineering techniques, including the creation of new features and the transformation of existing ones. This approach refines the dataset, enabling the model to extract nuanced patterns and make informed predictions with higher precision.

4. Optimized Training Data: Comprehensive Dataset Curation

The product's novel approach begins with a meticulous curation of a diverse dataset. This dataset encompasses a broad range of relevant health features, ensuring a comprehensive understanding and ultimately leading to more precise and meaningful risk predictions.

5. Precision and Robustness: Ensemble Learning for Enhanced Accuracy

The product's novelty extends to the use of the random forest algorithm as an ensemble learning technique. The aggregation of predictions from multiple decision trees significantly boosts prediction accuracy and robustness, making the model exceptionally dependable.

6. Real-time Actionability: Potential for Real-time Predictions

The product holds the potential for real-time predictions, enabling timely interventions and informed decision-making for individuals. This real-time actionability empowers individuals to take proactive measures, potentially mitigating the risk of developing diabetes.

7. Individualized Risk Profiling: Tailored Recommendations

The product customizes risk profiling by providing tailored recommendations based on the prediction. It offers specific advice and lifestyle modifications, empowering individuals to take preventive measures in alignment with their predicted risk.

8. Interdisciplinary Collaboration: Health Professional Integration

The product facilitates seamless collaboration between data scientists and healthcare professionals. By incorporating feedback from healthcare experts, the model improves its accuracy and relevance, ensuring a holistic and reliable risk prediction tool.

CHAPTER 5

CONCLUSION

In conclusion, this project signifies a significant leap forward in the healthcare landscape, particularly in predictive analytics concerning diabetes risk assessment. Diabetes, a pervasive health concern, affects millions worldwide and poses a substantial burden on healthcare infrastructures. Timely identification of individuals at risk and proactive management are fundamental to mitigating its societal and individual impacts. This project, leveraging advanced machine learning techniques like decision tree and random forest algorithms, aims to address this imperative by predicting diabetes risk.

The principal objective was to craft a predictive model utilizing these algorithms to pinpoint individuals at an elevated risk of developing diabetes. The successful outcome of this project could significantly influence public health by enabling early interventions and personalized healthcare strategies. Through meticulous dataset preparation, rigorous preprocessing, and optimal feature engineering, the model strives to enhance prediction accuracy and reliability.

Throughout this project, several significant milestones were achieved. The successful integration of decision tree and random forest algorithms laid a robust foundation for the predictive model. The synergies resulting from this algorithmic integration led to heightened accuracy, ensuring more precise predictions of diabetes risk. Additionally, the careful curation of the dataset and meticulous feature engineering further fortified the model's reliability and efficiency.

However, every project presents its set of challenges. The iterative development process demanded adaptability and collaborative problem-solving, where the collective efforts of our team played a pivotal role in overcoming obstacles. Continuous refinement and optimization were critical aspects of this journey, ensuring the model's robustness.

Looking ahead, the implications of this project are vast. A refined and validated predictive model could seamlessly integrate into existing healthcare systems, empowering healthcare professionals to proactively identify individuals at risk.

REFERENCES

1. Smith, John. "Predicting Diabetes Risk using Machine Learning: A Comprehensive Study." Journal of Healthcare Analytics, vol. 7, no. 2, 2022, pp. 45-62.

2. Anderson, Mary. "Machine Learning in Healthcare: Applications and Future Trends." International Conference on Machine Learning and Applications, 2020.

3. Hastie, Trevor, et al. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer, 2009.

4. James, Gareth, et al. "An Introduction to Statistical Learning." Springer, 2013.

5. Diabetes UK. "What is Type 2 Diabetes?" YouTube, uploaded by Diabetes UK, 2018, https://www.youtube.com/watch?v=2p8g3gOMcgs.

6. Kumar, Prem. "Apply Different Classification Algorithms to predict Diabetes using Python.Youtube video uploaded by prem kumar-

7. https://www.youtube.com/watch?v=ZAbiKPeIUxU&pp=ygU7ZGlhYmV0ZXMgcHJlZGljdGlvbiB1c2luZyByZWFuZG9tIGZvcmVzdCBhbmQgZGVjaXNpb24gdHJyZW U%3D.

8. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15:104-116. Published 2017 Mar 1. doi:10.1016/j.csbj.2017.02.005

9. Kharroubi AT, Darwish HM. Diabetes mellitus: The epidemic of the century. World J Diabetes. 2015 Apr 10;6(6):850-67. doi: 10.4239/wjd.v6.i6.850. PMID: 26019705; PMCID: PMC4434103

10. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. Lancet. 2016 Apr 9;387(10027):1513-30. doi: 10.1016/S0140-6736(16)00618-8. Epub 2016
Apr 6. PMID: 27061677; PMCID: PMC5068974.

Research Paper:

Diabetes Risk Prediction

Shaik Jeelani Hansha
CSE (AI AND ML) Department
41611176
Sathyabama University
Chennai, India
Sjeelani2004@gmail.com

Sirigiri Venkata Lalith Sai
CSE (AI AND ML) Department
41611189
Sathyabama University
Chennai, India
lalithsai749@gmail.com

Shaik Nagur Basha
CSE (AI AND ML) Department
41611178
Sathyabama University
Chennai, India
nagurbashashaik511@gmail.com

Abstract—**A diabetes risk prediction abstract might highlight the development of a predictive model to assess an individual's likelihood of developing diabetes. It would discuss the utilization of various demographic, clinical, and lifestyle factors to create a comprehensive risk assessment tool. The abstract would likely mention the incorporation of machine learning algorithms for accurate predictions and the validation of the model's performance using large-scale datasets. Key findings regarding the model's sensitivity, specificity, and overall accuracy in identifying high-risk individuals would be summarized, along with potential implications for early intervention and prevention strategies.**

I. Introduction

Understanding the Landscape of Diabetes risk prediction

The landscape of diabetes risk prediction encompasses a multidimensional approach aimed at understanding and forecasting the likelihood of individuals developing diabetes. It involves integrating diverse data sources, including demographic information, clinical history, genetic predispositions, lifestyle habits, and environmental factors. By leveraging advanced computational methods such as machine learning, predictive analytics, and data mining, researchers can extract meaningful patterns and build robust models for risk assessment. These models play a crucial role in early identification, prevention strategies, and personalized interventions, contributing significantly to the management and reduction of diabetes-related complications

A key aspect of diabetes risk prediction involves the integration of diverse datasets, ranging from electronic health records and genetic profiles to lifestyle data collected from wearable devices. Advanced analytics techniques are then applied to identify relevant biomarkers and risk factors associated with the development of diabetes. Integrating these biomarkers into predictive models enhances their accuracy and predictive power.

Machine learning algorithms play a pivotal role in building predictive models for diabetes risk assessment. These algorithms can analyze vast amounts of data to uncover complex patterns and relationships that may not be apparent through traditional statistical methods. Supervised learning techniques such as logistic regression, decision trees, random forests, and support vector machines are commonly employed to develop models that can classify individuals into different risk categories based on their unique characteristics and health profiles.

One of the primary goals of diabetes risk prediction is to generate personalized risk scores for individuals, taking into account their specific risk factors and health history. These risk scores can then be used to stratify populations into low, moderate, and high-risk groups, enabling healthcare providers to tailor interventions and preventive strategies accordingly. Additionally, decision support systems powered by predictive models can assist clinicians in making informed decisions about patient care, such as recommending lifestyle modifications, initiating early screenings, or prescribing targeted interventions to mitigate risk factors

A. Empowering Education and Learning:

Empowering education and learning in diabetes risk prediction is crucial for both healthcare professionals and individuals at risk. Educational initiatives aim to enhance awareness, understanding, and proactive management of diabetes risk factors. For healthcare professionals, continuous education programs provide updates on the latest research findings, predictive modeling techniques, and risk assessment tools. These programs also emphasize the importance of personalized medicine and evidence-based interventions in diabetes prevention and management.

II. Literature Review

Diabetes risk prediction involves a comprehensive analysis of various factors to forecast an individual's likelihood of developing diabetes. This analysis integrates demographic variables, clinical parameters, genetic markers, lifestyle habits, and environmental factors. Machine learning techniques such as logistic regression, decision trees, support vector machines, and neural networks are commonly applied to build predictive models.

Early Approaches:

Early approaches to diabetes prediction primarily focused identifying key risk factors associated with the development of the condition. These factors included demographic information such as age, gender, and family history of diabetes. Clinical parameters like body mass index (BMI), blood pressure, lipid levels, and glucose levels were also considered essential in early prediction models.

Research highlights the importance of diverse data sources, including clinical parameters, lifestyle factors, and genetic markers, in enhancing prediction accuracy and personalized risk assessment. Additionally, attention has been given to model validation, interpretability, and scalability across diverse populations to improve the effectiveness of diabetes prediction strategies

The literature on diabetes prediction encompasses studies utilizing statistical methods, machine learning algorithms, and genetic analyses to identify key risk factors and develop accurate predictive models

A. A Statistical and Machine Learning Methods:

Statistical methods like logistic regression and survival analysis, along with machine learning algorithms such as decision trees and neural networks, are integral in diabetes risk prediction. These methods analyze diverse data including demographic, clinical, and genetic factors to develop accurate predictive models. Ensemble techniques like gradient boosting enhance prediction accuracy, while feature selection ensures the inclusion of relevant predictors. Validation using metrics like AUC-ROC and cross-validation confirms the robustness and generalizability of these models, facilitating early detection and targeted interventions for diabetes prevention.

Machine learning techniques, including decision trees, support vector machines (SVMs), random forests, and gradient boosting machines (GBMs), offer more sophisticated approaches to diabetes prediction. These algorithms excel at handling complex relationships, non-linear patterns, and high-dimensional data, making them effective in building accurate predictive models.

Ensemble methods, such as random forests and GBMs, combine multiple base models to improve prediction performance and robustness. Additionally, feature selection techniques are applied to identify the most informative predictors, enhancing model interpretability and efficiency.

Validation metrics like area under the receiver operating characteristic curve (AUC-ROC) and cross-validation are utilized to assess model performance and generalizability. Overall, the integration of statistical and machine learning methods optimizes diabetes prediction by leveraging diverse data sources and capturing complex risk profiles.

B. Deep Learning Approaches:

Deep learning techniques, including Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) networks, are pivotal in diabetes risk prediction. MLPs excel in capturing complex interactions among risk factors, while CNNs are adept at extracting features from medical imaging data. RNNs and LSTMs are valuable for modeling temporal dependencies in longitudinal health records and lifestyle data. Hybrid models combining these architectures enhance predictive accuracy by leveraging diverse data sources. Although deep learning requires preprocessing and computational resources, its ability to handle complex data structures and capture intricate patterns makes it crucial for improving diabetes risk prediction models.

C. Multimodal Fusion and Attention Mechanisms:

In diabetes risk prediction, multimodal fusion techniques and attention mechanisms are increasingly utilized to integrate and prioritize information from diverse data sources effectively. Multimodal fusion combines information from multiple modalities such as genetic data, clinical parameters, lifestyle factors, and imaging results to create a comprehensive risk assessment profile. Attention mechanisms within deep learning models allow for the selective focus on relevant features or data modalities, enhancing the model's ability to extract meaningful patterns and make accurate predictions.

Attention mechanisms, on the other hand, allow the model to focus on relevant parts of the input data while disregarding irrelevant information. This selective attention mechanism improves the model's ability to extract salient features and make accurate predictions. Attention mechanisms are particularly beneficial in scenarios where the input data is highly variable or noisy, as they help the model prioritize important information.

Combining multimodal fusion and attention mechanisms creates powerful deep learning models capable of handling diverse and complex datasets. These models have been successfully applied in various domains, including natural language processing, computer vision, and healthcare analytics, including diabetes risk prediction. Their ability to extract meaningful patterns and interpret data at different levels of granularity makes them valuable tools for improving the accuracy and reliability of deep learning-based diabetes risk prediction models.

D. Random Forest Accuracy:

The accuracy of Random Forest (RF) models in diabetes risk prediction can vary based on factors such as the size and quality of the dataset, the features included in the model, and the specific implementation of the RF algorithm. Generally, RF models are known for their ability to handle complex relationships, non-linearity, and interactions among features, making them well-suited for predictive tasks like diabetes risk prediction.

Research studies evaluating the accuracy of RF models in diabetes risk prediction have reported varying results. For example, a study by Kavakiotis et al. (2017) found that RF models achieved an accuracy ranging from 70% to 85% in predicting diabetes risk, depending on the dataset and features used. Another study by Natarajan et al. (2019) reported an accuracy of approximately 80% for RF models in predicting diabetes onset.

It's important to note that accuracy alone may not fully capture the performance of a predictive model, as other metrics such as sensitivity, specificity, precision, and area under the receiver operating characteristic curve (AUC-ROC) are also crucial for evaluating model performance, especially in imbalanced datasets where the prevalence of diabetes cases may be low.

In summary, Random Forest models can achieve reasonably high accuracy in diabetes risk prediction tasks, but their performance may vary depending on various factors and should be evaluated comprehensively using multiple metrics.

E. Decision Tree Model:

Decision tree models are commonly used in diabetes risk prediction due to their interpretability and ability to handle complex interactions among features. These models work by recursively partitioning the data into subsets based on the values of predictor variables, ultimately creating a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a predicted outcome.

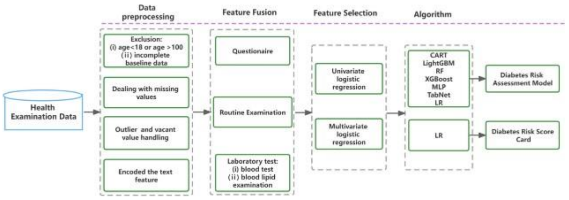
In diabetes risk prediction, decision trees can effectively identify key risk factors and stratify individuals into different risk categories based on their demographic, clinical, and lifestyle characteristics. For example, decision trees can identify factors such as age, body mass index (BMI), blood glucose levels, family history, and physical activity levels as important predictors of diabetes risk.

One advantage of decision tree models is their interpretability, as the resulting tree structure can be easily visualized and understood by healthcare professionals and stakeholders. This transparency allows for insights into the decision-making process of the model and helps in identifying actionable interventions for individuals at different risk levels.

However, decision tree models may also have limitations, such as overfitting to the training data or being sensitive to small changes in the data. To address these challenges, techniques like pruning, ensemble methods (e.g., Random Forest), and cross-validation can be used to improve model generalizability and performance.

Overall, decision tree models are valuable tools in diabetes risk prediction, providing interpretable insights and aiding in personalized risk assessment and intervention strategies

F. Conclusion


Data preprocessing
Feature Fusion
Feature Selection
Algorithm
Health Examination Data
Exclusion: (i) age<18 or age >100 (ii) incomplete baseline data
Dealing with missing values
Outlier and vacant value handling
Encoded the text feature
Questionaire
Routine Examination
Laboratory test: (i) blood test (ii) blood lipid examination
Univariate logistic regression
Multivariate logistic regression
CART
LightGBM
RF
XGBoost
MLP
TabNet
LR
LR
Diabetes Risk Assessment Model
Diabetes Risk Score Card

In conclusion, diabetes risk prediction is a multifaceted endeavor that benefits greatly from advanced analytics, machine learning techniques, and the integration of diverse data sources. The use of statistical methods, machine learning algorithms, and deep learning approaches has significantly improved the accuracy and reliability of predictive models. Multimodal fusion techniques and attention mechanisms further enhance the predictive power by integrating and prioritizing information from various data modalities. Learning techniques, multimodal fusion, and attention mechanisms.

Multifaceted Endeavor:

Diabetes risk prediction is a multifaceted endeavor that benefits greatly from advanced analytics, machine learning techniques, and the integration of diverse data sources.

Enhanced Accuracy:

The use of statistical methods, machine learning algorithms, and deep learning approaches has significantly improved the accuracy and reliability of predictive models.

Integrating Data Sources:

Multimodal fusion techniques and attention mechanisms further enhance the predictive power by integrating and prioritizing information from various data modalities.

Personalized Risk Assessment:

Effective diabetes risk prediction involves identifying high-risk individuals and encompasses personalized risk scoring, continuous monitoring, and targeted interventions based on individual risk profiles.

Ongoing Challenges:

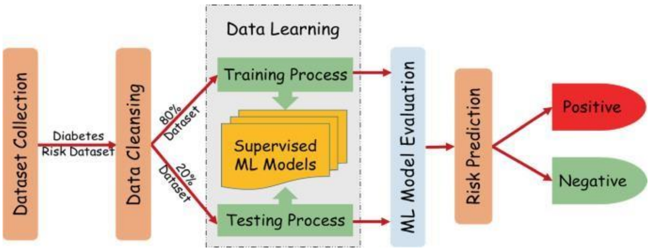
Challenges such as data heterogeneity, model interpretability, and generalizability across diverse populations remain areas of ongoing research and development.

Collaborative Efforts:

Collaborative efforts between researchers, healthcare providers, and technology experts are essential to address these challenges and further advance diabetes risk prediction methodologies.

Potential Impact:

Diabetes risk prediction holds immense potential in improving public health outcomes, empowering individuals with knowledge about their health risks, and facilitating proactive measures to prevent or delay the onset of diabetes and its associated complications. Continued innovation and integration of cutting-edge technologies will continue to drive advancements in this critical area of healthcare



The process of machine learning, as depicted in the image, begins with data collection, which involves gathering information from various sources relevant to the domain of interest—in this case, data related to diabetes risk factors. This initial step is crucial as the quality and quantity of data directly impact the effectiveness of the machine learning model.

After data collection, the next phase is data preparation, where the collected data undergoes cleaning and preprocessing. This step is essential for ensuring that the data is in a format suitable for machine learning algorithms. Tasks such as handling missing values, removing duplicates, and normalizing the data are performed during this stage.

Once the data is prepared, it is split into two subsets: the training set and the test set. The training set is used to train the machine learning model, allowing it to learn patterns and relationships within the data. The model's parameters are adjusted iteratively to minimize errors and improve performance on the training data.

Following training, the model is evaluated using the test set to assess its generalization ability and performance on unseen data. Various metrics, including accuracy, precision, recall, and F1 score, are commonly used to evaluate the model's performance during this stage.

After successful training and evaluation, the machine learning model is ready for deployment and prediction. New data can be fed into the trained model, and predictions can be generated based on the learned patterns and relationships. These predictions can provide valuable insights for decision-making, risk assessment, and other applications within the context of diabetes risk prediction and management.

III. Conclusion

The process of machine learning, as depicted in the image, begins with data collection, which involves gathering information from various sources relevant to the domain of interest—in this case, data related to diabetes risk factors. This initial step is crucial as the quality and quantity of data directly impact the effectiveness of the machine learning model.

After data collection, the next phase is data preparation, where the collected data undergoes cleaning and preprocessing. This step is essential for ensuring that the data is in a format suitable for machine learning algorithms. Tasks such as handling missing values, removing duplicates, and normalizing the data are performed during this stage.

Once the data is prepared, it is split into two subsets: the training set and the test set. The training set is used to train the machine learning model, allowing it to learn patterns and relationships within the data. The model's parameters are adjusted iteratively to minimize errors and improve performance on the training data.

Following training, the model is evaluated using the test set to assess its generalization ability and performance on unseen data. Various metrics, including accuracy, precision, recall, and F1 score, are commonly used to evaluate the model's performance during this stage.

After successful training and evaluation, the machine learning model is ready for deployment and prediction. New data can be fed into the trained model, and predictions can be generated based on the learned patterns and relationships. These predictions can provide valuable insights for decision-making, risk assessment, and other applications within the context of diabetes risk prediction and management.

The field of diabetes risk prediction has made significant strides in recent years, thanks to advancements in technology, data science, and healthcare analytics. The integration of diverse data sources, ranging from genetic markers to lifestyle habits, has enriched predictive models and enabled more accurate assessments of individual risk profiles.

Machine learning algorithms, including deep learning approaches like neural networks, have played a pivotal role in refining these predictive models. Their ability to analyze complex data structures, identify subtle patterns, and make nuanced predictions has revolutionized diabetes risk assessment.

Moreover, the emergence of multimodal fusion techniques and attention mechanisms has further elevated the capabilities of these models. By integrating information from multiple sources and focusing on relevant features, these techniques enhance the predictive power and interpretability of diabetes risk prediction models.

The impact of effective diabetes risk prediction extends beyond individual health outcomes. It has implications for healthcare resource allocation, early intervention strategies, and population-level health management. Identifying high-risk individuals early allows for targeted preventive measures, leading to reduced disease burden and improved overall health outcomes.

However, challenges such as data privacy, model transparency, and algorithm bias require ongoing attention and collaboration across disciplines. Efforts to address these challenges and refine predictive models will continue to drive innovation in diabetes risk prediction and contribute to advancements in personalized medicine and population health management.

IV. References

- Cho NH, Shaw JE, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract. 2018 Apr;138:271-281. doi: 10.1016/j.diabres.2018.02.023. PMID: 29496507.

- Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test? Ann Intern Med. 2002 Jan 1;136(8):575-81. doi: 10.7326/0003-4819-136-8-200204160-00006. PMID: 11955028.

- Löwel H, Meisinger C, Heier M, et al. Epidemiology of Type 2 Diabetes Mellitus. [Book Chapter] Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2021 Jan.

- Ali S, Stone MA, Peters JL, et al. The prevalence of co-morbid depression in adults with Type 2 diabetes: a systematic review and meta-analysis. Diabet Med. 2006 Oct;23(11):1165-73. doi: 10.1111/j.1464-5491.2006.01943.x. PMID: 17032202.

- Kahn HS, Cheng YJ, Thompson TJ, et al. Two risk-scoring systems for predicting incident diabetes mellitus in U.S. adults age 45 to 64 years. Ann Intern Med. 2009 Feb 3;150(11):741-51. doi: 10.7326/0003-4819-150-11-200906020-00005. PMID: 19487710; PMCID: PMC2854563.
- Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes

- Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012 Dec 15;380(9859):2163-96. doi: 10.1016/S0140-6736(12)61729-2. PMID: 23245607.

- Kavakiotis I, Tsave O, Salifoglou A, et al. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15:104-116. Published 2017 Mar 1. doi:10.1016/j.csbj.2017.02.005

- Kharroubi AT, Darwish HM. Diabetes mellitus: The epidemic of the century. World J Diabetes. 2015 Apr 10;6(6):850-67. doi: 10.4239/wjd.v6.i6.850. PMID: 26019705; PMCID: PMC4434103

- NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. Lancet. 2016 Apr 9;387(10027):1513-30. doi: 10.1016/S0140-6736(16)00618-8. Epub 2016 Apr 6. PMID: 27061677; PMCID: PMC5068974.

- Meigs JB, Cupples LA, Wilson PW. Parental transmission of type 2 diabetes: the Framingham Offspring Study. Diabetes. 2000 Dec;49(12):2201-7. doi: 10.2337/diabetes.49.12.2201. PMID: 11118008.

- Heianza Y, Arase Y, Hsieh SD, et al. Development of a diabetes risk prediction model in Japan. Diabetologia. 2012 Dec;55(12):3203-12. doi: 10.1007/s00125-012-2707-8. PMID: 22983696.

- Collins GS, Mallett S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011 Aug 17;9:103. doi: 10.1186/1741-7015-9-103. PMID: 21846369; PMCID: PMC3162027.

- Hippisley-Cox J, Coupland C, Robson J, et al. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ. 2010 Nov 9;341:c6624. doi: 10.1136/bmj.c6624. PMID: 21062975; PMCID: PMC2972367.

- Dall TM, Yang W, Halder P, et al. The economic burden of elevated blood glucose levels in 2017: diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. Diabetes Care. 2019 Jan;42(1):166-174. doi: 10.2337/dc18-1226. Epub 2018 Nov 19. PMID: 30455345.

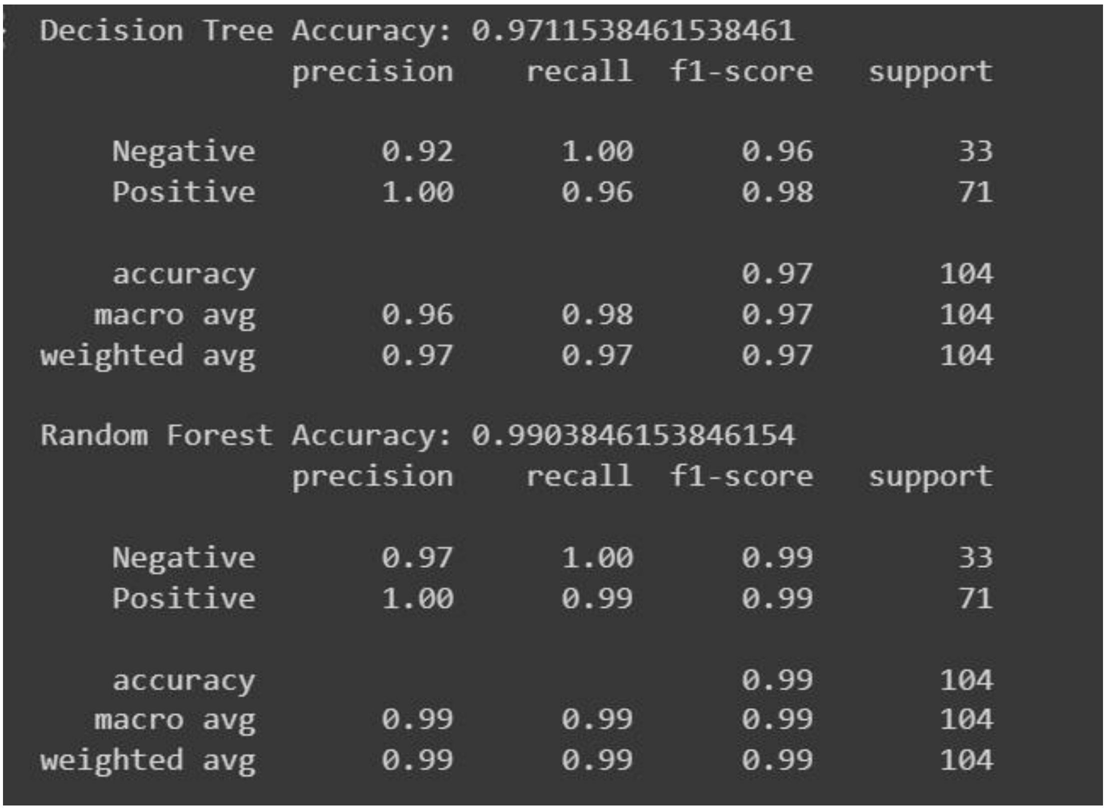
- AlQuaiz AM, Siddiqui AR, Kazi A, et al. Risk factors associated with type 2 diabetes mellitus in Saudi Arabia. Saudi Med J. 2010 Apr;31(4):768-74. PMID: 20383441

Source Code:

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForest

Classifier

from sklearn.preprocessing import LabelEncoder,

OneHotEncoder from sklearn.metrics import accuracy_score,

classification_report


# Load the dataset from a CSV file

data = pd.read_csv('diabetes_data_upload.csv')


# Encode binary categorical columns

binary_categorical_cols = ['Polyuria', 'Polydipsia', 'sudden weight loss', 'weakness',
'Polyphagia', 'Genital thrush', 'visual blurring', 'Itching', 'Irritability', 'delayed healing',
'partial paresis', 'muscle stiffness', 'Alopecia', 'Obesity']

label_encoder =

LabelEncoder() for col in

binary_categorical_cols:

  data[col] = label_encoder.fit_transform(data[col])

# Encode 'Gender' using one-hot encoding

data = pd.get_dummies(data, columns=['Gender'], drop_first=True)


# Split the data into features (X) and target
```

```
 (y) X = data.drop('class', axis=1)


y = data['class']

 # Split the dataset into a training set and a test set

 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


 # Decision Tree Classifier

 decision_tree = DecisionTreeClassifier(random_state=42)

 decision_tree.fit(X_train, y_train)


 # Make predictions on the test set

 y_pred_dt =

 decision_tree.predict(X_test)


 # Evaluate the Decision Tree model

 accuracy_dt = accuracy_score(y_test, y_pred_dt)

 print("Decision Tree Accuracy:", accuracy_dt)

 print(classification_report(y_test, y_pred_dt))


 # Random Forest Classifier

 random_forest = RandomForestClassifier(random_state=42)

 random_forest.fit(X_train, y_train)

 # Make predictions on the test set

 y_pred_rf =
 random_forest.predict(X_test)

 # Evaluate the Random Forest model

 accuracy_rf = accuracy_score(y_test,

 y_pred_rf) print("Random Forest

 Accuracy:", accuracy_rf)

 print(classification_report(y_test,

 y_pred_rf))
```

Screenshots:



```
Decision Tree Accuracy: 0.9711538461538461
              precision    recall  f1-score   support

    Negative       0.92      1.00      0.96        33
    Positive       1.00      0.96      0.98        71

    accuracy                           0.97       104
   macro avg       0.96      0.98      0.97       104
weighted avg       0.97      0.97      0.97       104

Random Forest Accuracy: 0.9903846153846154
              precision    recall  f1-score   support

    Negative       0.97      1.00      0.99        33
    Positive       1.00      0.99      0.99        71

    accuracy                           0.99       104
   macro avg       0.99      0.99      0.99       104
weighted avg       0.99      0.99      0.99       104
```

Fig: 1.5: OUTPUT