School of Computing, Engineering, and Digital Technologies
Teesside University | Middlesbrough TS1 3BA.

An Investigation into the Significance Machine Learning Models for Diabetes Risk Prediction

Analysis, Design, and Report

10th January 2024.
Babatunde Olusegun | B1605208
Submitted as partial requirements for the degree of MSc in Data Science
Supervisor: Zahid Iqbal

ACKNOWLEDGMENT

The success of this postgraduate research project is attributed chiefly to the altruistic supervision of Mr Zahid Iqbal, who provided invaluable direction and timely recommendations. Similarly, I would like to express gratitude to my project's second reader, Guo Qiang, and the several module instructors and lecturers who generously contributed their time and resources to impart their extensive knowledge throughout my postgraduate studies.

Furthermore, I would like to express my gratitude to my colleagues, specifically Deji Ladipo, Olutayo Oladeinbo, Fabian Umeh, and others, whose motivation, encouragement, and emotional support propelled me to successfully finish this project during a challenging period when my family experienced a loss of pregnancy.

I must acknowledge my wife, Abimbola, and our children, Eriife and Tisefunmi who are the primary motivations behind my relentless pursuit of success. I appreciate your affection and generous assistance.

DECLARATION

I, Babatunde Abiodun OLUSEGUN, at this moment, certify that the project titled "An Investigation into the Pros and Cons of Machine Learning Models for Diabetes Risk Prediction" has been totally done by me and has not been partially or entirely plagiarised from any other source, except for adequately cited sources.

This dissertation was my labour; no other people were involved in its production. I further declare that this dissertation was created exclusively to meet the criteria of this course and that it had never been utilised for another endeavour before its submission.

An Investigation into the Pros and Cons of Machine Learning Models for Diabetes Risk Prediction

Abstract

In response to the growing global burden of diabetes, this project addresses the dynamic convergence of data science and healthcare. It looks at the possible benefits and drawbacks of using machine learning models to forecast the risk of diabetes. Effective illness management depends on early risk detection, and machine learning presents the possibility of tailored interventions and better patient outcomes. However, this study also recognises the difficulties with model interpretability, data privacy, and ethical issues.

To educate stakeholders and encourage appropriate healthcare applications, it provides insights into the complex field of machine learning's role in diabetes risk prediction in a succinct exploration. The project research explores research questions that seek to determine how machine learning models for Diabetes risk prediction should be adopted for clinical practices. A research review carried out in the project studies previous work on diagnosing, treating, and predicting people with diabetes.

The study, despite limitations, shows significant benefits and challenges that come with using machine learning models for diabetes' risk prediction. The research gap and limitations are equally highlighted to promote further advancement in the application of machine learning to predict diabetes' risk.

Keywords: Continuous Glucose Monitors (C.G.M.s), Random Forest (R.F.), Logistic regression (L.R.), Gradient boosting (G.B.), k-nearest neighbour (k-NN)

Contents

TABLE OF FIGURE

Chapter Two

Introduction

1 INTRODUCTION

1.1 Background and Context

The global impact of diabetes has led to research that seek to identify improved ways to detect the disease at early stage. Based on the data from I.D.F. Atlas, it is projected that in 2021, there were approximately 536.6 million adults aged 20-79 with diabetes across 215 nations and territories. By 2045, it is projected that 783.2 million people will have Diabetes (Sun H. et al., 2022). There are primarily two classes of diabetes, namely, Type-1, known as Insulin-Dependent Diabetes Mellitus (IDDM) and Type-2 is also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). The former is caused by the inability of the human body to generate sufficient insulin while the latter is caused by seen when body cells cannot use insulin effectively. Type-3 gestational Diabetes on the other hand can be found in pregnant women due to increased blood sugar not detected early.

D.M. has long-term complications associated with it. Also, there are high risks of various health problems for a diabetes person (Bhat, S.S. et al., 2023). Reducing the burden of the disease, enabling personalised interventions, and revolutionising healthcare are all possible with machine learning for risk prediction (Verma, V. et al., 2023). Machine learning algorithms must be employed for improved diabetes prediction to prevent diabetes illnesses and take preventative action beforehand (Dhande, B. et al., 2022). However, it requires a thorough analysis of the trade-offs, considering issues with data privacy, model interpretability, and ethics. By bridging the gap between technology and healthcare, this research seeks to shed light on the way towards the merit and demerits of machine learning in Diabetes risk prediction. More also, the research project investigates the impact of identified lifestyles and risk factors on prevalence of diabetes in people. Furthermore, emphasis is made to research use of machine learning models in diabetes risk prediction among varying ethically acceptable, efficient, and responsible applications.

1.2 Research Aims & Objectives

This research project explores the benefits and challenges of using machine learning models in diabetes risk prediction. The models' performances are compared using evaluation metrics such as Model Score, Recall Score, Precision Score, and F1 score.

Past literature work in the diagnosis, identification, treatment, and prediction of people with diabetes using machine learning models is studied for understanding the growth and scientific development that have happened in the last few years. Evaluating different machine learning models helps to identify the most suitable model for clinical use.

This project also seeks to investigate robustly the lifestyle, risk factors and significant legal and ethical considerations that often hinder the adaptability of the research work for clinical purposes.

Chapter Three

Literature Review

2 LITERATURE REVIEW

2.1 Overview of Diabetes and Its Significance

Elevated blood sugar levels are the hallmark of diabetes, a metabolic condition that is prevalent and chronic Schumacher-Petersen C. et al., 2019). A thorough literature review on diabetes highlights its complexity and examines its two main variants, Type 1 and Type 2.

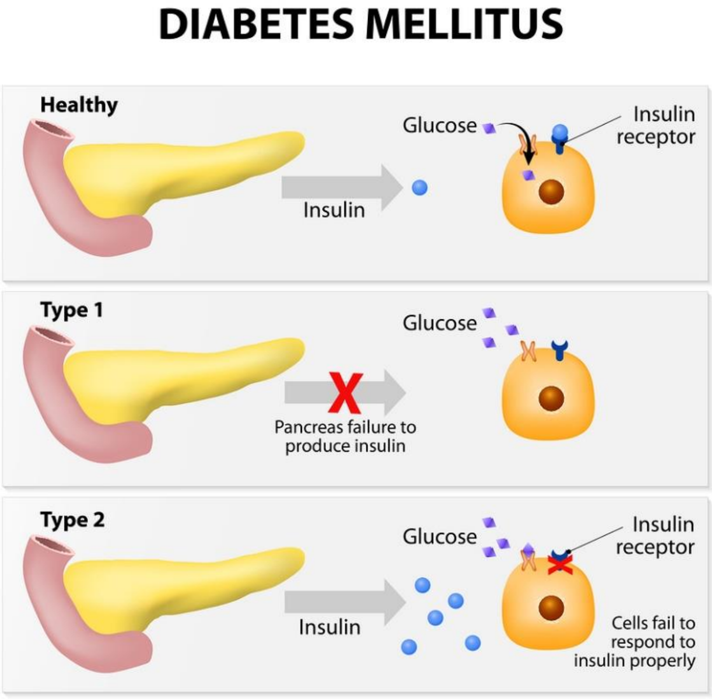
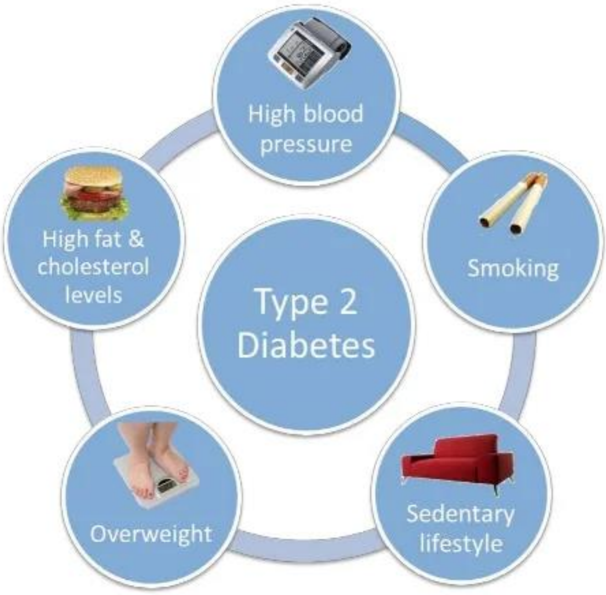


Figure 1: Diabetes Type 1 and 2

Research has shown that Type-1 diabetes often referred to as autoimmune disease, occurs when the immune system unintentionally destroys the beta cells in the insulin producing pancreas. As a result, the body produces little or no insulin. Type-1 diabetes are usually common in infants or adolescence and a lifetime insulin prescription is recommended for disease management. The type-2 diabetes however, results to insulin resistance because of the body's poor response to insulin. The patients' lifestyles such as smoking, inactivity, overweight etc. are linked to type-2 diabetes in adults.



http://www.metacure.com/about-diabetes-2/type-2-diabetes-mellitus-t2dm/

Figure 2: Diabetes Risk Factors

Genetics, risk factors such as smoking, BMI, Cholesterol Level etc. and environmental variables are commonly linked to risk of diabetes. The global menace has called for more concerns due to its link to obesity and sedentary lifestyles common among many people across all demographics. Prompt and early diagnosis coupled with effective risk assessment have been highlighted in much research as viable means to avert the deadly disease. Furthermore, new avenues for diabetes control and individualised care have been made possible by developments in healthcare, particularly the use of machine learning models (Fukushima et al., 2023).

Initially, uncontrolled diabetes has significant health risks as it can lead to many complications, such as peripheral neuropathy, kidney failure, heart disease, and vision issues (Creatore et al., 2010). Creatore [posited that these problems lead to high medical costs and a reduced quality of life for those afflicted. Second, the financial cost of diabetes is debilitating; each year, billions of dollars are spent on direct and indirect care (2010).

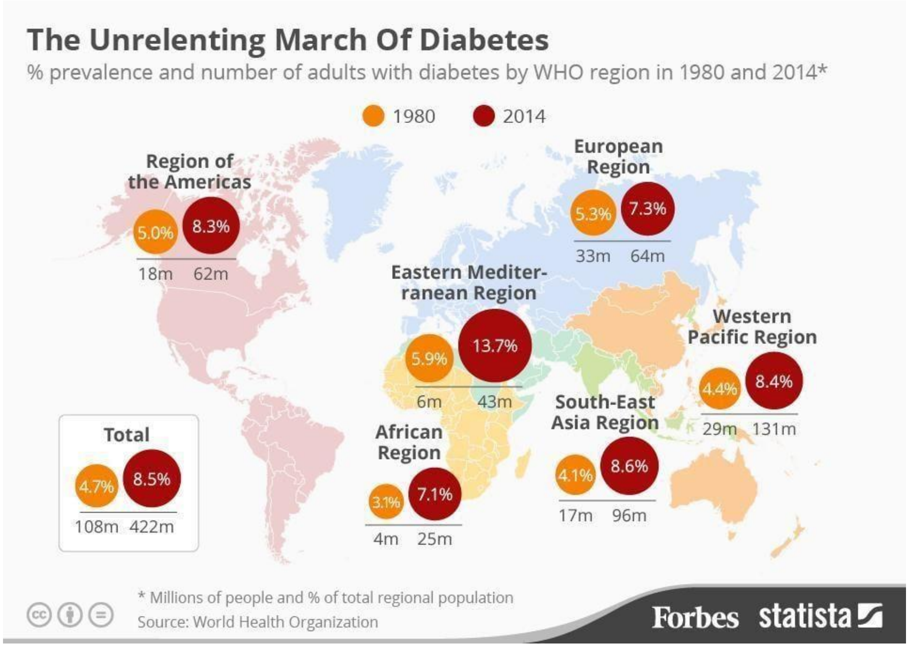
In addition to its clinical and financial ramifications, diabetes poses a public health and socioeconomic burden that demands comprehensive preventative and management initiatives. Two significant lifestyle factors contributing to disease development are diet and exercise (Schumacher-Petersen, C. et al., 2019).

2.2 Treatment and Awareness

Figure 3: Diabetes Awareness

Diabetes is a chronic disease that affects millions of people worldwide. It requires good treatment and widespread sensitisation to reduce its impact on individuals and society (Azadnajafabad, S. et al., 2023). With treatment and public awareness campaigns, the collective effort toward controlling and preventing diabetes has progressed, but it still demands sustained dedication.

Diabetes treatment primarily consists of dietary changes, medication, and, in some instances, insulin therapy. Diet modification is critical to maintaining blood sugar levels, stressing balanced meals, quantity control, and focusing on low glycemic index foods. Regular exercise improves insulin sensitivity and overall health, which complements dietary adjustments.



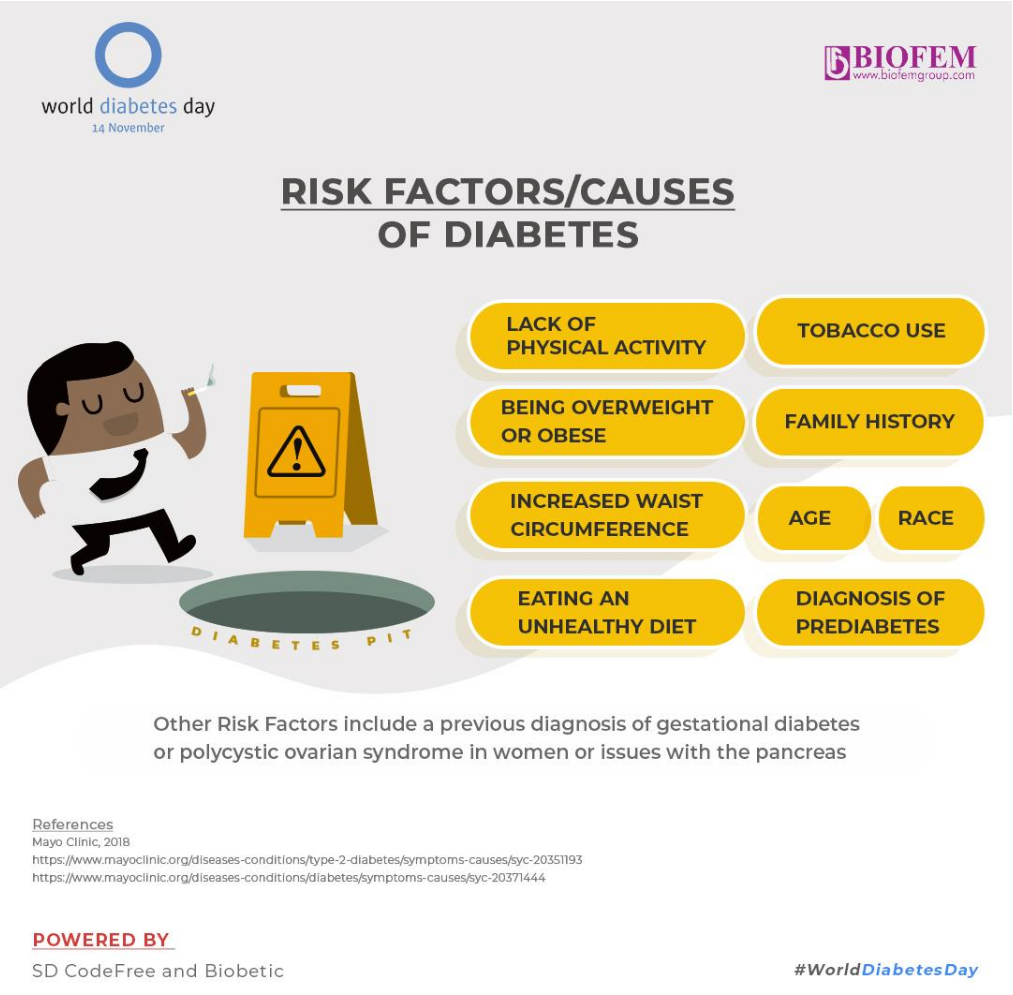
* Millions of people and % of total regional population
Source: World Health Organization

Figure 4: Number of Adults with Diabetes

Medications such as metformin, sulfonylureas, and insulin injections have been identified as remarkable choice of treatment when lifestyle changes have been explored exhaustively. New technological advancement in insulin administration have provided ease and accuracy in dosing which in turn improved the quality of life of patients relying on insulin. Furthermore, developing technology like continuous glucose monitors (C.G.M.s) and insulin pumps provide more exact monitoring and management of blood sugar levels, giving patients and healthcare practitioners valuable data. (Yvan, Y. et al., 2022). Beyond administration of pills and insulin administration technology, education and proper awareness of risk factors and symptoms have proved to be effective diabetes preventive strategies.

2.3 Diabetes risk factors

These are wide range of conditions that have potentials to result to the development of this long-term illness. Although, lifestyles can be altered, certain risk factors are uncontrollable for an individual.



world diabetes day
14 November

BIOFEM
www.biofemgroup.com

RISK FACTORS/CAUSES OF DIABETES

- LACK OF PHYSICAL ACTIVITY
- TOBACCO USE
- BEING OVERWEIGHT OR OBESE
- FAMILY HISTORY
- INCREASED WAIST CIRCUMFERENCE
- AGE
- RACE
- EATING AN UNHEALTHY DIET
- DIAGNOSIS OF PREDIABETES

Other Risk Factors include a previous diagnosis of gestational diabetes or polycystic ovarian syndrome in women or issues with the pancreas

References
Mayo Clinic, 2018
https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193
https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

POWERED BY
SD CodeFree and Biobetic

#WorldDiabetesDay

Figure 5: Diabetes Risk Factors

Family history or genetics is believed to determine individual susceptibility to Diabetes. The probability of having the disease increases with the record of the disease in the family history. Ethnicity is also identified as risk factor. Some racial groups have a higher chance of being diabetes than the other. Lifestyle factors such as poor dietary choices such as excessive sugar intake increase the risk of diabetes in people. More also, age is a significant risk determinant. Aged people are characterized by the decline in healthy physical activity and weight gain. As a result, their chances of Type-2 diabetes are increased. To mitigate this occurrence of this disease, attention must be paid to adjustable risk factors, weight control, regular exercise and healthy diet.

2.4 Critical Concepts of Machine Learning and Diabetes Risk Prediction

Machine learning has shown itself to be a revolutionary force in healthcare, changing how we think about patient care, diagnosis, and illness prediction. It significantly impacts the healthcare sector, especially regarding the prediction of illness (Kumar S. et al., 2023). The diverse application of machine learning includes analysis of patient data such as genetic information, medical imaging and electronic personal health records, to detect early signs of diabetes. Yan defined four categories for Machine learning as follows: supervised, semi-supervised, unsupervised, and reinforcement learning (2022).

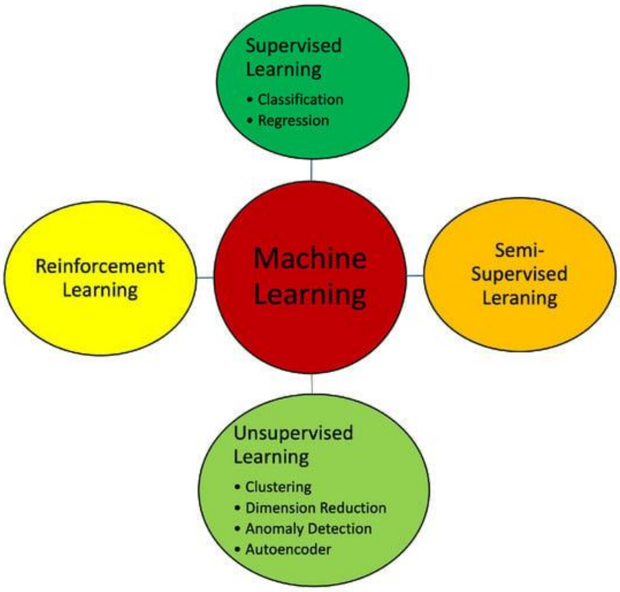


Figure 6: Machine Learning Types

Supervised Learning: When using supervised learning, the algorithm gains knowledge using labelled training data that contains both the input features and the goal outputs that go along with them (Nhangumbe, M. et al.,2023). The model extrapolates patterns in the existing data to forecast new, unknown data. It is extensively employed in tasks where the model is trained to map inputs to known outputs, such as regression and classification (Yan, Y. et al., 2022). Diabetes risk prediction is greatly aided by supervised learning, a key machine learning paradigm. In order to teach the model, the relationship between input features (such as age, BMI, and glucose levels) and the outcome, this strategy requires training algorithms on a labelled dataset where the outcome (the presence or absence of diabetes) is identified (Géron, 2017).

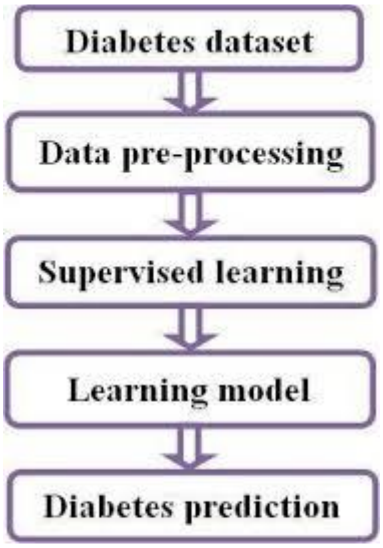


Figure 7: *Supervised Learning*

In the context of diabetes risk prediction, patient data from previous visits where the patient's diabetes status is known is fed into supervised learning algorithms including logistic regression, decision trees, and neural networks. Numerous variables are included in this data, ranging from lifestyle factors (diet, physical activity) to physiological measurements (blood sugar, cholesterol). The algorithm learns by finding patterns and correlations between these features and the likelihood of diabetes (Hastie, & Friedman, 2009). After training, the model is used to estimate a new patient's probability of developing diabetes, offering insightful information for tailored treatment regimens and early intervention. According to James et al. (2013), the algorithm's selection and tuning, together with the quantity and quality of the training data, all have a substantial impact on how accurate these predictions are. Supervised learning is a significant tool in healthcare, particularly for chronic conditions like diabetes where early detection can lead to better management and results. It does this by relying on existing data to produce predictions.

Semi-Supervised Learning: The components of both supervised and unsupervised learning are combined in semi-supervised learning. It uses a dataset in which most of the data is unlabeled, and only a tiny piece is labelled (Nhangumbe, M. et al., 2023). The model generalises from unlabeled data by using the labelled data to direct its learning process (Mohebbi, A. and Shelan, V. 2017). This method works well when getting labelled data is costly or time-consuming. In the context of diabetes risk prediction studied in report, semi-supervised learning—a machine learning technique that makes use of both labelled and unlabeled data—becomes more and more pertinent, especially in situations where obtaining large, comprehensive labelled datasets is difficult or costly. The model presents a useful approach to overcome the difficulty of obtaining substantial collections of completely labelled medical information.

This method usually combines a smaller pool of unlabeled data (e.g., records without diabetes diagnosis) with a larger pool of labelled data (e.g., patient records with known diabetes status). To improve learning accuracy, semi-supervised learning algorithms take advantage of the distribution and structure of both labelled and unlabeled data. For example, they may repeatedly refine the model by using the labelled data to understand the early patterns and then using this learning to categorize the unlabeled data (Zhu et al., 2009).
Semi-supervised learning is quite beneficial in predicting the risk of diabetes. It enables the utilization of extensive electronic health records, even if they are not always completely labelled, to uncover trends and risk factors linked to diabetes. Incorporating a larger range of patient data in this strategy can result in more robust and generalized models. This allows for a more comprehensive understanding of potential risk variables (Chapelle et al., 2006). Semi-supervised learning is a potential method in medical predictive analytics that can improve the efficiency and effectiveness of forecasting diabetes risk in larger patient populations.

Unsupervised Learning Models: Outputs are trained using unsupervised learning techniques without labelled targets. Patterns, structures, or relationships within the data are what the algorithm looks for (Nhangumbe, M. et al., 2023). Typical activities include clustering, which involves assembling similar data points to simplify complex data while preserving crucial information. The learning model functions without labelled outcomes, making it particularly well-suited for uncovering concealed patterns and structures in data. This method is especially beneficial in the context of predicting the risk of diabetes when the goal is to discover previously unknown connections or groupings of patients without pre-established categories. Unsupervised learning can be employed to analyze extensive datasets of patient health records. These records may encompass several variables such as glucose levels, BMI, and lifestyle behaviors**.** The presence or absence of diabetes are not inherently indicated by these datasets. Patients are organized into different groups based on similarities in health data categorized by clustering techniques such as K-means and hierarchical clustering. Patients' vulnerability to diabetes are indicated trends identified in these clusters.
Another utilization of unsupervised learning in this domain is dimensionality reduction, such as Principal Component Analysis. This technique simplifies intricate data while retaining crucial information. Utilizing this technique can facilitate the visualization and comprehension of complex medical datasets, which often contain numerous dimensions. This, in turn, can uncover previously unidentified risk factors or indicators, potentially leading to significant discoveries. Unsupervised learning does not directly forecast the risk of diabetes, but it assists in the exploratory research, offering a more profound comprehension of the data and directing later predictive modelling. This technique is highly effective for generating hypotheses and discovering new insights in the field of diabetes research.

Reinforcement Learning: Through interactions with the environment, a reinforcement learning system can make successive decisions. Based on its activities, the system is rewarded or punished to provide feedback. The system is supposed to figure out the best course of action that maximises cumulative rewards over time. The model is extensively employed in robotics, autonomous systems, and gaming applications. Given its capacity to provide individualised and data-driven insights, machine learning is essential for illness prediction (Zhou, et al., 2020). By mining large datasets, machine learning algorithms can find risk factors and subtle correlations that may escape human notice. Real-time forecasts and ongoing monitoring are also made possible by these models' ability to change and grow in response to new data (Tate, G. et al., 2020)

2.5 Previous Studies on Diabetes Risk Prediction

Various research has concentrated on forecasting the likelihood of developing diabetes, employing diverse methodologies and datasets. The Framingham Heart Study, a groundbreaking longitudinal study, led the way in this field by identifying risk factors such as obesity, hypertension, and familial predisposition. Further research studied other risk factors such as genetics and lifestyle factors. Diabetes Prevention Program (DPP) is an example of research carried out to study the effectiveness of lifestyle change in mitigating the risks of diabetes in people at high risk. Machine learning and artificial intelligence have gained prominence in recent years for diabetes risk prediction. Studies like those by Kavakiotis et al. (2017), Gargeya and Leng (2017) have leveraged machine learning algorithms to analyze electronic health records and clinical data, achieving high accuracy in predicting diabetes onset. In addition, genetic research has discovered multiple susceptibility genes linked to diabetes, with genome-wide association studies (GWAS) uncovering new genetic variations. The UK Biobank project has made substantial contributions to the comprehension of the genetic foundation of diabetes susceptibility.

Furthermore, research on diabetes risk prediction has been essential in expanding our knowledge of this intricate and pervasive illness. The accuracy and application of diabetes risk prediction using machine learning models have increased dramatically in recent years due to several notable academic research and discoveries. Among the significant developments are: Deep Learning and Neural Networks for capturing intricate relationships within the data, leading to more accurate risk assessments. Mohebbi et al. presented a unique deep-learning method for diagnosing type 2 diabetes, demonstrating that Continuous Glucose Monitoring (C.G.M.) signals could identify Type 2 Diabetes patients (Mohebbi, A. et al., 2017). Feature Engineering techniques are some of the other notable research developments that allow more informative features to be produced. By capturing intricate connections and patterns in the data, these features increase the precision of predictions. Advancements in data mining and machine learning have enabled biomedical research to enhance the quality of primary healthcare (Creatore, M.I. et al. 2010). It is necessary to execute a better classification further to increase the prediction rate of the medical datasets, as incorrect classification can result in dire predictions.

The past studies in diabetes risk prediction gave birth to machine learning models that can now offer personalised risk assessments. They consider individual characteristics, lifestyle factors, and genetics to provide tailored predictions, enhancing preventive healthcare. Huang J. et al. studied the relationship and connection between Depression with Physical Activity and Obesity in Older Diabetes Patients (Leah et al., 2022). Integration of Wearable and IoT (Internet of Things) Devices has also been made a reality to provide real-time monitoring and diabetes risk prediction (Tate et al., 2020). Continuous data streams enable early detection of anomalies and risk factors. These breakthroughs and many more such as Interpretable Models, Handling Imbalanced Data, and Handling Imbalanced Data contribute to developing more accurate, interpretable, and personalised machine-learning models for diabetes risk prediction (Huang, J., Li, R. and Tsai, L. 2022). As technology and research in this field continue to advance, the outlook for early diagnosis, preventive interventions, and effective management of diabetes is becoming increasingly promising (Shin, J. et al., 2022). The investigation of diabetes risk factors, the creation of predictive algorithms, and the evaluation of model accuracy have all been the focus of several research projects.

2.6 Pros and Cons of Using Machine Learning Models

Machine learning models, which provide a few advantages and specific difficulties and restrictions, are now essential to many businesses. Determining the appropriate deployment strategy for machine learning models requires understanding the technology's benefits and drawbacks (Kumar S. et al., 2023).

Pros of Using Machine Learning Models

Data-driven decision-making: Large-scale dataset analysis, pattern recognition, and data-driven insight generation are among the many strengths of machine learning models (Balasubramanian et al., 2022). They allow organisations to make well-informed judgments and forecasts supported by data rather than gut feeling. Artificial neural networks can predict the system's behaviour accurately in each scenario (Gomez et al., 2017). Its ability to provide technical data can help decision-makers act more rapidly, identify safety issues, or provide an intelligent system with the potential to use pattern recognition for reactor accident identification and classification (Gomez et al. et al., 2017)

Predictive Capabilities: With great precision, machine learning models can predict future trends, behaviours, and events. Independent validation is an essential step before the clinical implementation of a predictive model since it ensures the assessment of its performance in populations that are different from the one that was involved in its calibration (Mennickent D. *et al.,* 2022) In the management of diabetes, machine learning models have proven to have significant predictive power. They can predict, for example, a person's glycemic reaction to foods or drugs. Treatment plans that are tailored benefit from these forecasts. One example is the development of "closed loop" systems. The "closed loop" systems use continuous glucose monitoring (C.G.M.) devices in conjunction with machine learning algorithms to forecast blood sugar patterns and initiate alerts for modifications in insulin dosage (Gomez et al., 2017). Furthermore, early intervention is made possible by machine learning models that can detect individuals who are at risk of developing diabetes complications like neuropathy or retinopathy (Dhande B. et al., 2022). By customising medicines to each patient's needs, these predictive skills improve not only diabetes control but also patient outcomes.

Automation and Efficiency: Repetitive tasks such as collection, preprocessing and analysis of personal health records can be automated using Machine Learning models, which decreases the demand for manual inputs, decreases errors and increases process efficiency. Predictive models, for instance, can anticipate the risk of contracting diabetes (Moshawrab M. et al., 2023). The continuous monitoring of blood glucose levels is made possible by integrating wearable technology and Internet of Things sensors, allowing for real-time data processing. Predictive algorithms, for example, have been developed using data from continuous glucose monitors (C.G.M.s) to anticipate glycemic trends and notify individuals of impending hypo- or hyperglycemic episodes so that appropriate action can be taken from machine learning (ML) offers individuals and healthcare professionals (Tate and G.H.R., 2020)

Personalization*:*

By customising care to each patient's needs, machine learning models in diabetes therapy offer a revolutionary leap in customisation and user experience (Zou, X. *et al.,* 2023). Predictive algorithms, for instance, can evaluate a patient's lifestyle choices and past glucose data to provide personalised insulin dose recommendations. In addition to improving blood sugar regulation, this lessens the cognitive strain associated with controlling diabetes and improves user experience.
Moreover, ML-powered chatbots and virtual assistants aim to inform and engage patients. With the ability to adjust their answers to each patient's particular circumstances, these AI-driven systems can provide real-time assistance with meal planning, insulin delivery, and glucose monitoring (Tate and G.H.R., 2020). This high degree of customisation encourages improved adherence to treatment plans and enhances the user experience.

Challenges With Using Machine Learning Models in Diabetes Risk Predictions

Data Quality and Bias*:*

The reliability and fairness of predictive models are significantly impacted by data quality and bias in diabetes risk prediction with machine learning (ML). Data Quality Issues include missing data and data imbalance, while examples of bias are data bias, algorithm bias, and feature bias.

Interpretability*:*

The problem of interpretability in Diabetes risk prediction with machine learning (ML) is a critical challenge (Bhat, S.S. et al., 2023). While ML models have demonstrated impressive accuracy, their inherent complexity often leads to "black box" models, making it difficult to understand and trust the decision-making process (Xu, H. and Shuttleworth, K.M.J., 2023). This lack of interpretability raises several concerns, such as compliance issues, failure in clinical adoption, and the need for patient understanding.

Resource-Intensive:

Diabetes prediction presents several resource-intensive issues related to large-scale computing demands and data handling needs. Using machine learning models for diabetes prediction frequently requires substantial processing power and data resources, which presents challenges for practical implementation (Mohebbi A. et al., 2017). Robust hardware is required for complex algorithms with heavy computing demands, such as ensemble approaches or deep learning. For example, significant computational resources are required to train deep neural networks to predict diabetes risk from genetic data.

Overfitting:

One major issue with machine learning (ML) predictive models in managing diabetes is overfitting. Overly sophisticated machine learning models during training may fit the training data too closely, resulting in the capture of noise or random changes instead of actual patterns. Overfit models can be harmful when it comes to the management of diabetes. Consider, for instance, a predictive algorithm that uses a patient's past glucose data to suggest individualised insulin dosages. An overfit model may detect transient variations or inaccurate measurements, leading to unpredictable insulin recommendations that are outside the line with the patient's proper course of care. Such inconsistent advice puts the patient's health at risk, causing blood sugar levels to drop or rise dangerously. Regularisation, cross-validation, and careful feature selection are critical techniques in diabetes therapy models to minimise overfitting. Achieving a balance between model complexity and generalisation is essential to ensure suggestions. Eventually, this will enhance patient outcomes and safety when managing diabetes.

Privacy Concerns:

Applications that use personal data raise concerns about security and privacy. Protecting private information while gaining insightful knowledge is a complex problem. This raises important issues for machine learning (ML) predictive models for diabetes management. These models present serious privacy problems because they frequently rely on sensitive patient data, such as genetic information, medical records, and continuous glucose monitoring (C.G.M.) data. For example, predictive models provide personalized treatments for diabetes using patients' personal health records which may be sensitive medical details or lifestyle choices. Patients concerns range from data breaches, indiscriminate access, and misuse of private information.

Ethical Considerations:

Machine Learning (ML) predictive models for treating diabetes raise serious ethical questions. Although these models have the potential to provide more efficient and customised treatment, they may unintentionally bring up certain moral conundrums: If ML models are trained on biased datasets, they can perpetuate biases in diabetes treatment. For example, if historical data exhibits racial or socioeconomic bias, predictive models may inadvertently discriminate against certain demographic groups in treatment recommendations. Patients often need informed consent to use their data in predictive models. Ensuring that patients fully understand how their data will be utilised and what the potential outcomes may be crucial for ethical data collection. "Black box" models also raised ethical questions.

2.7 Research gaps and Future Work

Data Diversity and Representativeness

The study discovered numerous significant research deficiencies in the field of diabetes risk prediction concerning the diversity and representativeness of data. The gaps encompass the inadequate representation of some populations in datasets, the restricted incorporation of uncommon risk factors, the lack of longitudinal data, potential biases in data collecting, and the absence of detailed contextual information.

It is essential to focus on these areas of study in order to enhance the precision and impartiality of diabetes risk prediction models. In order to achieve this, researchers must give priority to gathering a greater variety of data that spans throughout time and is rich in contextual information. This will ensure that these models can offer fair and efficient evaluations of risks for a wider range of people and factors that contribute to risk.

Interpretability:

The study revealed a notable deficiency in studies within the field of diabetes risk prediction, namely in terms of comprehensibility. This gap is defined by the predominance of black-box machine learning models that lack explicit justifications for their predictions. The absence of interpretability gives rise to difficulties regarding the acceptance and confidence of clinicians, the presence of bias and impartiality, the empowerment of patients, and the fulfilment of regulatory and ethical obligations.

Enhancing black-box models' interpretability is essential to winning over end users' and healthcare professionals' trust (Xu and K.M.J, 2023). Researchers should investigate several approaches to make ML models more visible and comprehensible.

It is crucial to fill this research gap in order to responsibly and efficiently apply machine learning models in predicting the risk of diabetes. Researchers should prioritize the development of strategies that enhance the interpretability of these models, hence increasing the understandability and reliability of their predictions for healthcare professionals and patients. This endeavor should additionally guarantee equity and adherence to ethical standards in the process of making healthcare decisions.

Feature Engineering and Selection

Like every ML application in healthcare, there are social and ethical ramifications to consider. Studies ought to investigate possible model biases, tackle fairness concerns, and consider the social implications of using diabetes risk prediction tools. Feature engineering and selection are crucial in improving the interpretability of models when it comes to predicting the risk of diabetes. Contemporary studies frequently depend on conventional clinical and demographic characteristics, disregarding developing biomarkers and omics data that have the potential to offer useful insights into the risk of diabetes. There is a lack of dynamic feature selection techniques that adapt to changing risk profiles, hindering the modeling of evolving risk factors over time. The current state of customizing feature selection based on individual patient profiles is lacking in development, hence failing to capitalize on the potential for personalized risk prediction. The difficulty of integrating disparate data sources, such as electronic health records and wearable devices, persists and necessitates enhanced techniques for data harmonization.

Through meticulous feature engineering and selection, machine learning models can be enhanced to provide greater transparency, allowing healthcare professionals and patients to fully grasp the underlying principles behind the predictions. This strategy not only preserves the correctness of the model but also guarantees that the decision-making process is more easily understandable and reliable. Within the realm of predicting diabetes risk, these strategies serve as a connection between intricate machine learning models and their comprehensibility, therefore promoting enhanced assurance in and comprehension of the predictions produced by these models.

Incorporation of Behavioral and Environmental Factors

Diabetes risk is mainly determined by behavioural and environmental variables (Agyemang et al.; L., 2021). A more comprehensive knowledge of risk variables requires incorporating information on nutrition, physical activity, lifestyle, and environmental exposures into predictive models. The study underlines notable deficiencies in the integration of behavioral and environmental components in predicting the risk of diabetes. The absence of extensive data on human behaviors and environmental elements frequently obstructs the advancement of comprehensive risk prediction models. Comprehending and simulating the complex interplay of these factors and their influence on the risk of diabetes continue to be difficult. Current models frequently neglect the dynamic aspects of behaviors and settings, hence lacking the ability to adjust to temporal changes.

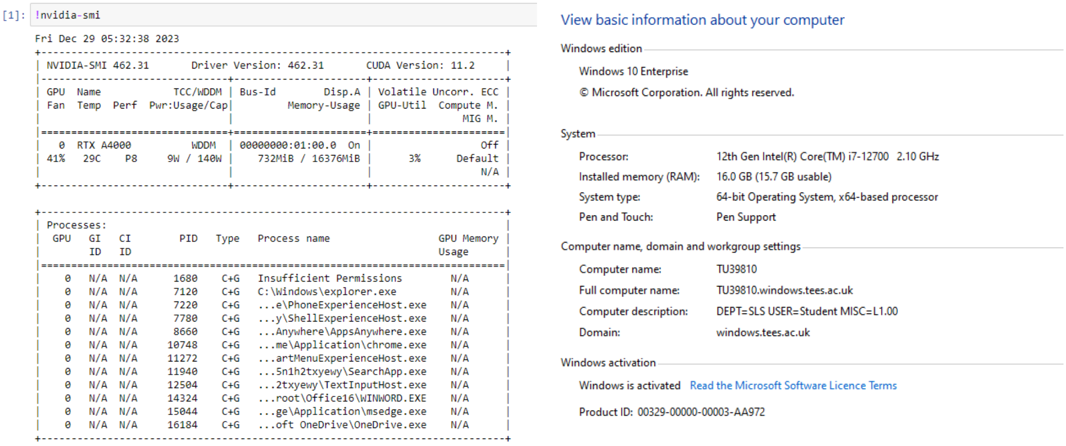
The absence of standardized data and optimal methodologies for gathering and merging behavioral and environmental data undermines the dependability of models. Models must incorporate the variability in these characteristics among diverse ethnic and regional groups. It is crucial to address these areas of study that are currently lacking in order to enhance the precision and significance of diabetes risk prediction models. This will enable the development of more personalized and efficient methods for diabetes prevention and management, considering individual behaviors and environmental factors.

Chapter Four

Methodology

3 METHODOLOGY

3.1 System properties.



[1]: `!nvidia-smi`

```
Fri Dec 29 05:32:38 2023
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 462.31       Driver Version: 462.31       CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name            TCC/WDDM | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  RTX A4000           WDDM | 00000000:01:00.0  On |                  Off |
| 41%   29C    P8     9W / 140W |    732MiB / 16376MiB |      3%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A      1680    C+G   Insufficient Permissions        N/A      |
|    0   N/A  N/A      7120    C+G   C:\Windows\explorer.exe         N/A      |
|    0   N/A  N/A      7220    C+G   ...e\PhoneExperienceHost.exe    N/A      |
|    0   N/A  N/A      7780    C+G   ...y\ShellExperienceHost.exe    N/A      |
|    0   N/A  N/A      8660    C+G   ...Anywhere\AppsAnywhere.exe    N/A      |
|    0   N/A  N/A     10748    C+G   ...me\Application\chrome.exe    N/A      |
|    0   N/A  N/A     11272    C+G   ...artMenuExperienceHost.exe    N/A      |
|    0   N/A  N/A     11940    C+G   ...5n1h2txyewy\SearchApp.exe    N/A      |
|    0   N/A  N/A     12504    C+G   ...2txyewy\TextInputHost.exe    N/A      |
|    0   N/A  N/A     14324    C+G   ...root\Office16\WINWORD.EXE    N/A      |
|    0   N/A  N/A     15044    C+G   ...ge\Application\msedge.exe    N/A      |
|    0   N/A  N/A     16184    C+G   ...oft OneDrive\OneDrive.exe    N/A      |
+-----------------------------------------------------------------------------+
```

View basic information about your computer

Windows edition

Windows 10 Enterprise

© Microsoft Corporation. All rights reserved.

System

Processor:	12th Gen Intel(R) Core(TM) i7-12700 2.10 GHz
Installed memory (RAM):	16.0 GB (15.7 GB usable)
System type:	64-bit Operating System, x64-based processor
Pen and Touch:	Pen Support

Computer name, domain and workgroup settings

Computer name:	TU39810
Full computer name:	TU39810.windows.tees.ac.uk
Computer description:	DEPT=SLS USER=Student MISC=L1.00
Domain:	windows.tees.ac.uk

Windows activation

Windows is activated   Read the Microsoft Software Licence Terms

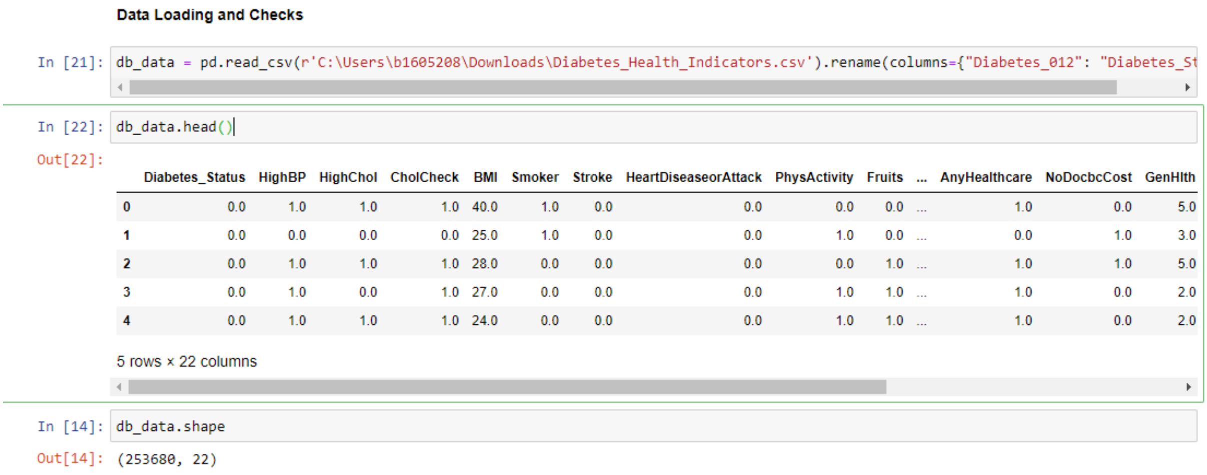
Product ID: 00329-00000-00003-AA972

Figure 8: System properties

3.2 Data Collection and Preprocessing

The dataset used for the purpose of this research is sourced from the Behavioral Risk Factor Surveillance System (BRFSS). The health-related telephone survey which is organized once a year is conducted by The United States Centers for Disease Control and Prevention (C.D.C.). Over 400,000 Americans participate in the annual survey, which gathers information on health-related risk behaviors, chronic illnesses, and the use of preventative services. Since 1984, it has been held annually.

Diabetes _ 012 _ health _ indicators _ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the C.D.C.'s BRFSS2015. The target variable Diabetes_012 has three classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is a class imbalance in this dataset. This dataset has 21 feature variables.
Source; https://www.kaggle.com/alexteboul/diabetes-health-indicators-datasett



Data Loading and Checks

In [21]:
```
db_data = pd.read_csv(r'C:\Users\b1605208\Downloads\Diabetes_Health_Indicators.csv').rename(columns={"Diabetes_012": "Diabetes_St
```

In [22]:
```
db_data.head()
```

Out[22]:

	Diabetes_Status	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0

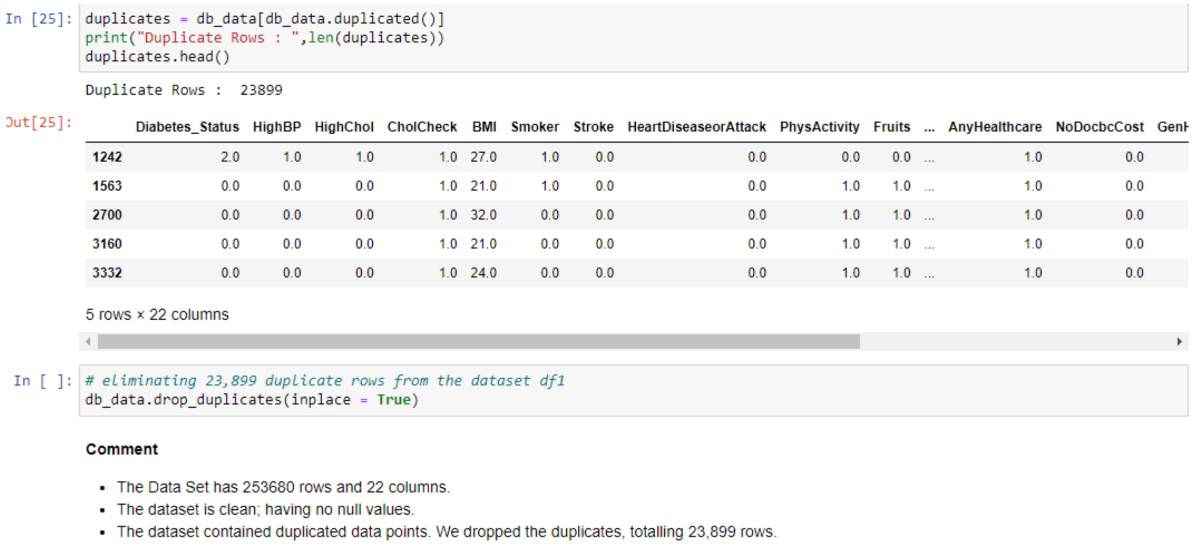
5 rows × 22 columns

In [14]:
```
db_data.shape
```

Out[14]:
```
(253680, 22)
```

Figure 9: Data Loading and Checks

3.3 Data Cleaning.



```
In [25]: duplicates = db_data[db_data.duplicated()]
         print("Duplicate Rows : ",len(duplicates))
         duplicates.head()
```

```
Duplicate Rows :  23899
```

Out[25]:

	Diabetes_Status	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenH
1242	2.0	1.0	1.0	1.0	27.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	
1563	0.0	0.0	0.0	1.0	21.0	1.0	0.0	0.0	1.0	1.0	...	1.0	0.0	
2700	0.0	0.0	0.0	1.0	32.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	
3160	0.0	0.0	0.0	1.0	21.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	
3332	0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	

5 rows × 22 columns

```
In [ ]: # eliminating 23,899 duplicate rows from the dataset df1
        db_data.drop_duplicates(inplace = True)
```

Comment

- The Data Set has 253680 rows and 22 columns.
- The dataset is clean; having no null values.
- The dataset contained duplicated data points. We dropped the duplicates, totalling 23,899 rows.

Figure 10: Data Cleaning

3.4 Heatmap Correlation Matrix

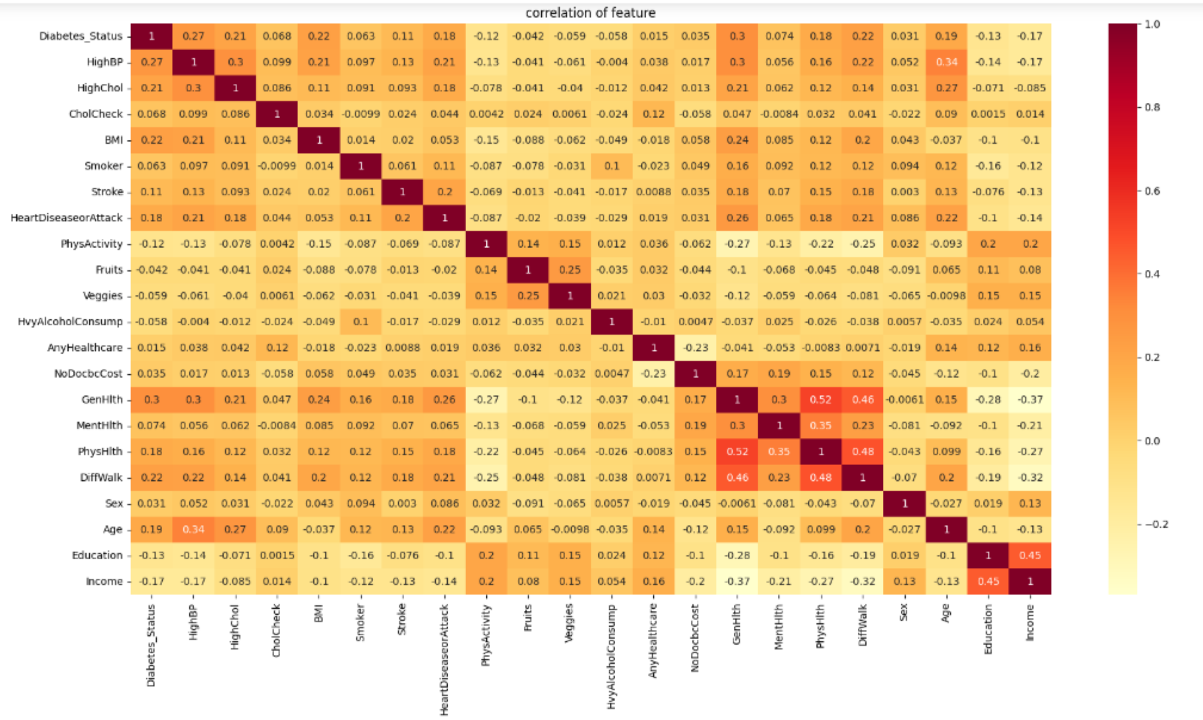


Figure 11: Heatmap Correlation Matrix

Observation

Features that have a positive correlation with diabetes.

- High Blood Pressure
- High Cholesterol
- Cholesterol Check
- B.M.I., Smoker, Age, Sex
- Stroke
- Heart Disease Attack
- Having any healthcare
- Physical activity
- General Health
- Mental health status
- Physical health status
- Difficulty walking

Features that have a Negative correlation with diabetes.

- Fruits (Slight Correlation)
- Veggies (Slight Correlation)
- Heavy alcohol consumption
- Education
- Income

A positive correlation suggests that individuals with features such as High B.P., B.M.I., Smoke etc., may have a higher likelihood of diabetes. Conversely, a negative correlation suggests that according to the dataset, individuals with a high education level and income may have a lower likelihood of diabetes. Further visualisation of the features that are highly correlated with our target column, Diabetes Binary

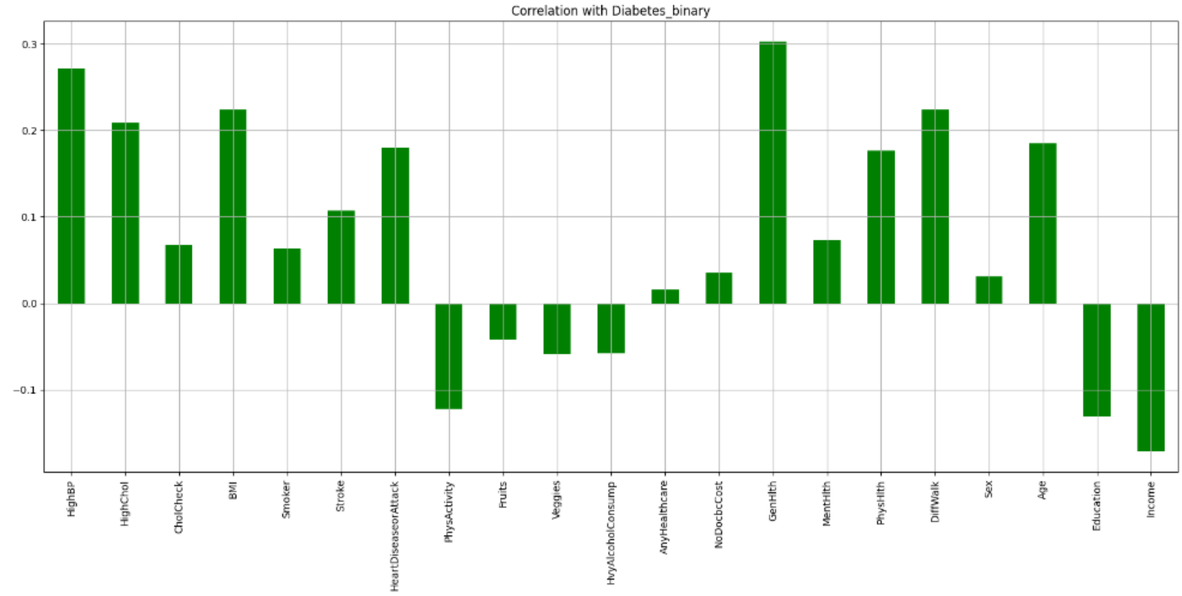
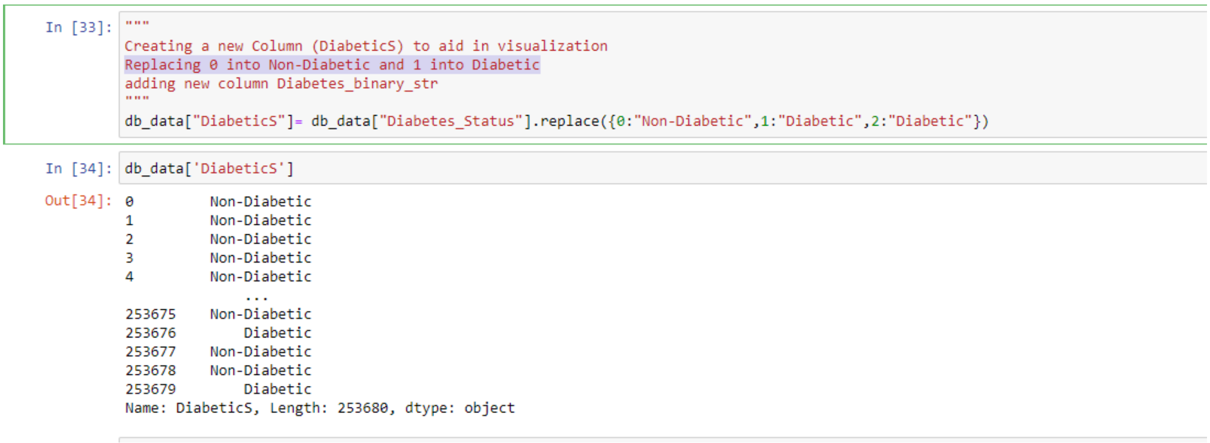


Figure 12: Correlation with Diabetes _binary

3.5 Feature Selection and Engineering

The target feature, the diabetes binary column, is reconstructed into diabetes and non-diabetes categories for better visualisation and analysis.

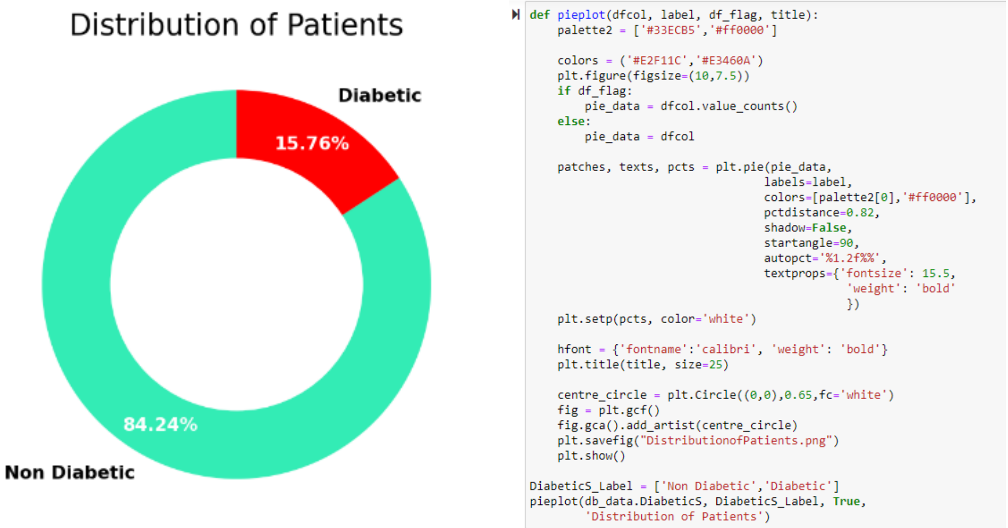


```
In [33]: """
         Creating a new Column (DiabeticS) to aid in visualization
         Replacing 0 into Non-Diabetic and 1 into Diabetic
         adding new column Diabetes_binary_str
         """
         db_data["DiabeticS"]= db_data["Diabetes_Status"].replace({0:"Non-Diabetic",1:"Diabetic",2:"Diabetic"})

In [34]: db_data['DiabeticS']

Out[34]: 0         Non-Diabetic
         1         Non-Diabetic
         2         Non-Diabetic
         3         Non-Diabetic
         4         Non-Diabetic
                       ...
         253675    Non-Diabetic
         253676        Diabetic
         253677    Non-Diabetic
         253678    Non-Diabetic
         253679        Diabetic
         Name: DiabeticS, Length: 253680, dtype: object
```

Figure 13: Feature Engineering

3.6 Data Analysis & Visualization

- **Distribution of patients by health status**



```
def pieplot(dfcol, label, df_flag, title):
    palette2 = ['#33ECB5','#ff0000']

    colors = ('#E2F11C','#E3460A')
    plt.figure(figsize=(10,7.5))
    if df_flag:
        pie_data = dfcol.value_counts()
    else:
        pie_data = dfcol

    patches, texts, pcts = plt.pie(pie_data,
                                   labels=label,
                                   colors=[palette2[0],'#ff0000'],
                                   pctdistance=0.82,
                                   shadow=False,
                                   startangle=90,
                                   autopct='%1.2f%%',
                                   textprops={'fontsize': 15.5,
                                              'weight': 'bold'
                                              })
    plt.setp(pcts, color='white')

    hfont = {'fontname':'calibri', 'weight': 'bold'}
    plt.title(title, size=25)

    centre_circle = plt.Circle((0,0),0.65,fc='white')
    fig = plt.gcf()
    fig.gca().add_artist(centre_circle)
    plt.savefig("DistributionofPatients.png")
    plt.show()

DiabeticS_Label = ['Non Diabetic','Diabetic']
pieplot(db_data.DiabeticS, DiabeticS_Label, True,
        'Distribution of Patients')
```

Figure 14: *Distribution of patients by health status*

Note

The above pie chart has two segments: one represents the non-diabetes patients (in teal) and the other represents the diabetes patients (in red). Their respective values are expressed in percentages.

The pie chart with a donut-like appearance displays that 84.24% of the patients are non-diabetes and 15.76% are diabetes.

- **Gender distribution of patients with stroke but NO diabetes**

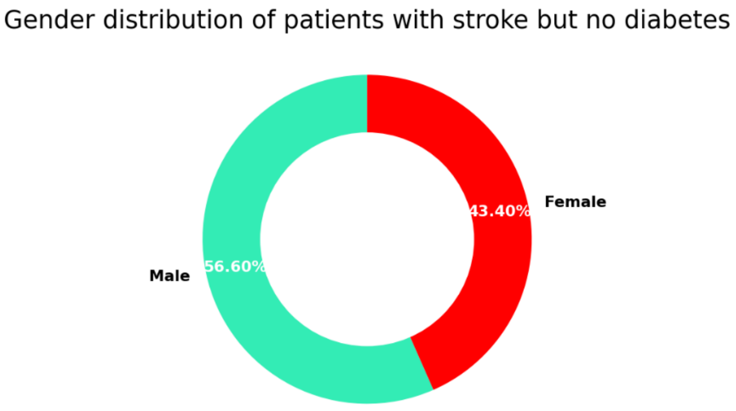
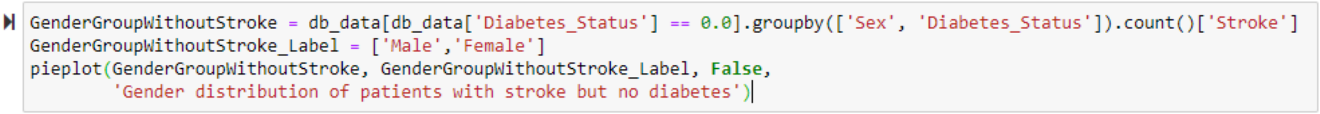


Figure 15: Gender distribution with stroke but no diabetes



```
GenderGroupWithoutStroke = db_data[db_data['Diabetes_Status'] == 0.0].groupby(['Sex', 'Diabetes_Status']).count()['Stroke']
GenderGroupWithoutStroke_Label = ['Male','Female']
pieplot(GenderGroupWithoutStroke, GenderGroupWithoutStroke_Label, False,
        'Gender distribution of patients with stroke but no diabetes')
```

Note

The resulting pie chart with two segments: one representing male patients and the other representing female patients. According to the chart, 56.60% of patients with stroke but no diabetes is male, while 43.40% are female.

- **Gender distribution of patients with stroke and diabetes**

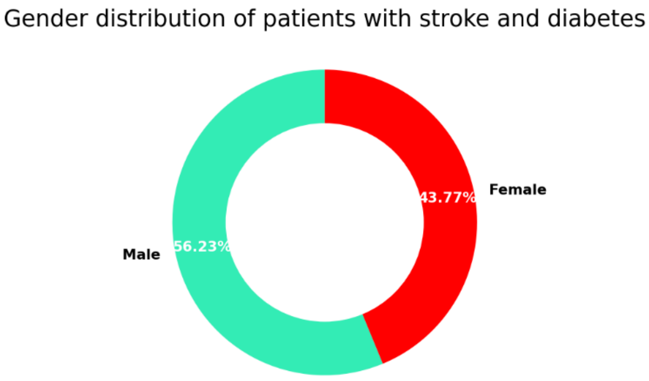
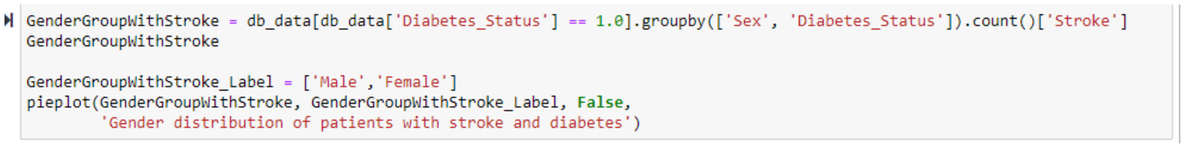


Figure 16: Gender distribution with stroke and diabetes



```
GenderGroupWithStroke = db_data[db_data['Diabetes_Status'] == 1.0].groupby(['Sex', 'Diabetes_Status']).count()['Stroke']
GenderGroupWithStroke

GenderGroupWithStroke_Label = ['Male','Female']
pieplot(GenderGroupWithStroke, GenderGroupWithStroke_Label, False,
        'Gender distribution of patients with stroke and diabetes')
```

Note

The resulting pie chart with two segments: one representing male patients and the other representing female patients. According to the chart, 56.23% of patients with stroke and diabetes is male, while 43.77% are female.

- **Distribution of patients by Age**

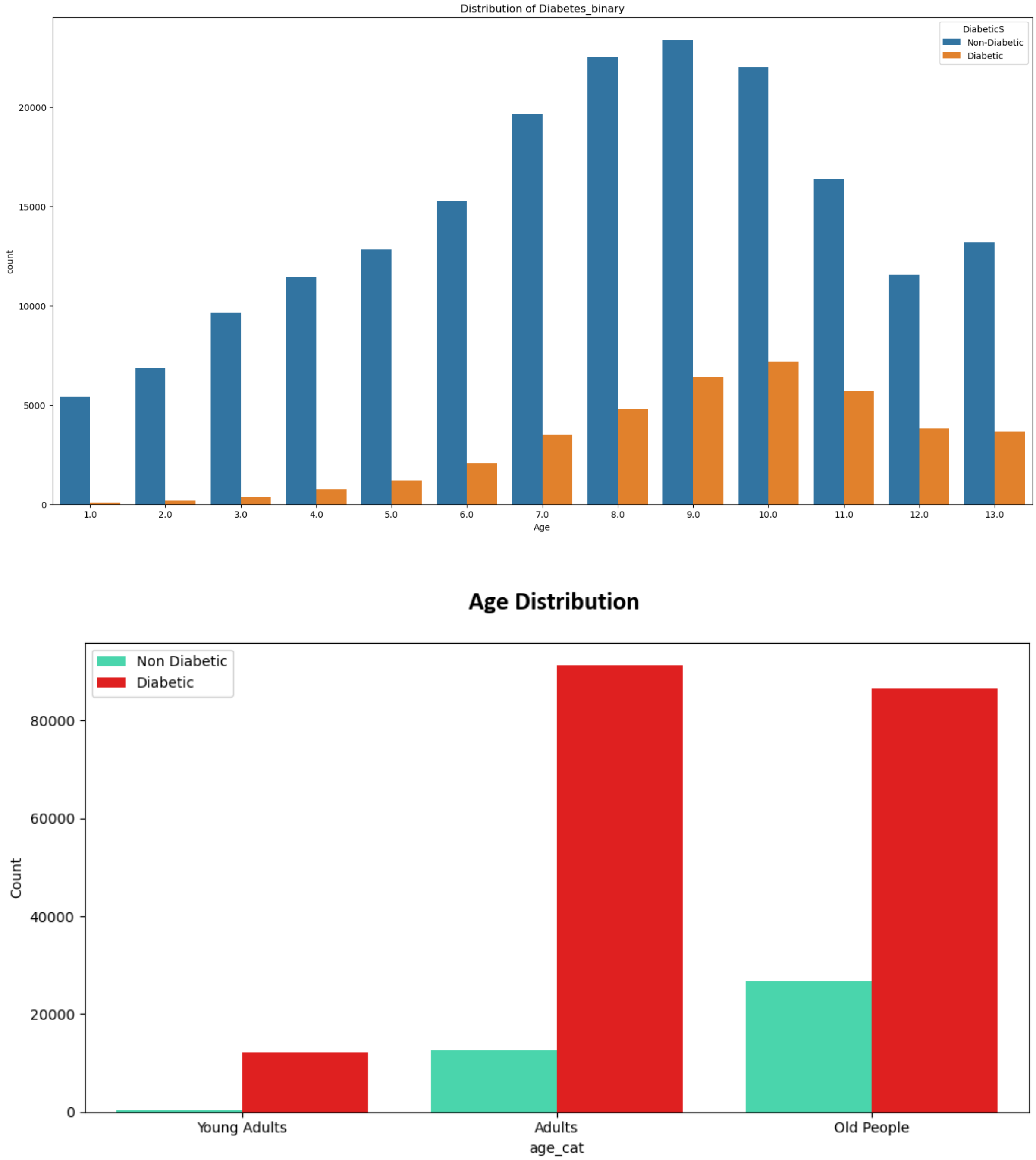


Figure 17: Distribution by Age

Note Given Age: 1-2: Young Adults, Age 3-8: Adults, and Age 9-15: Old,
It is observed that.

- For most age groups, the number of non-diabetes individuals is higher than that of diabetes individuals.
- Diabetes is most common among old individuals, followed by adults.

- **Distribution of Diabetes patients by BMI**

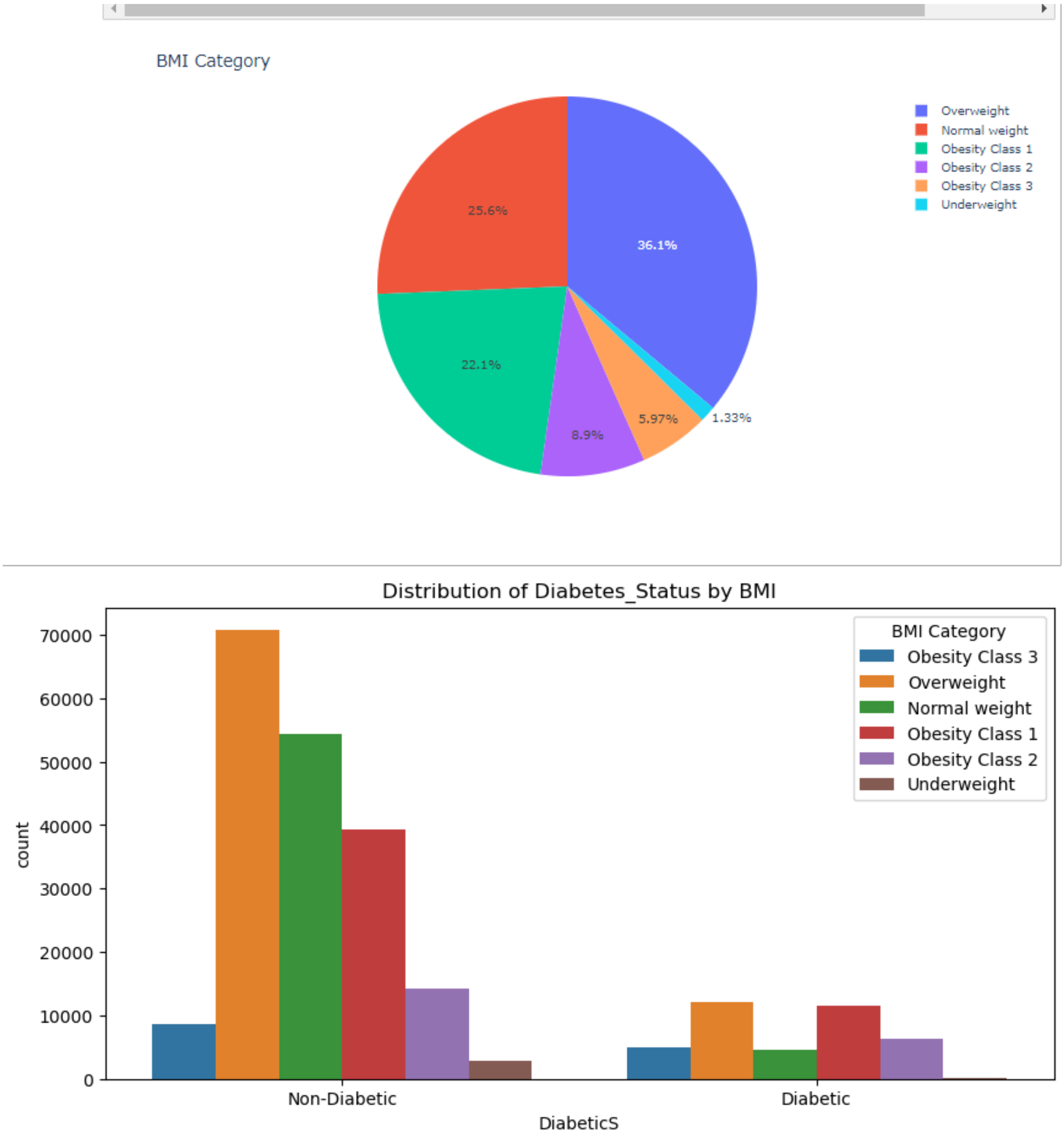


Figure 18: Distribution by BMI

Note
Given the above BMI categories,
It is observed that overweight and obese people have a higher risk of type 2 diabetes.

- **Investigation of Income Risk Factor**

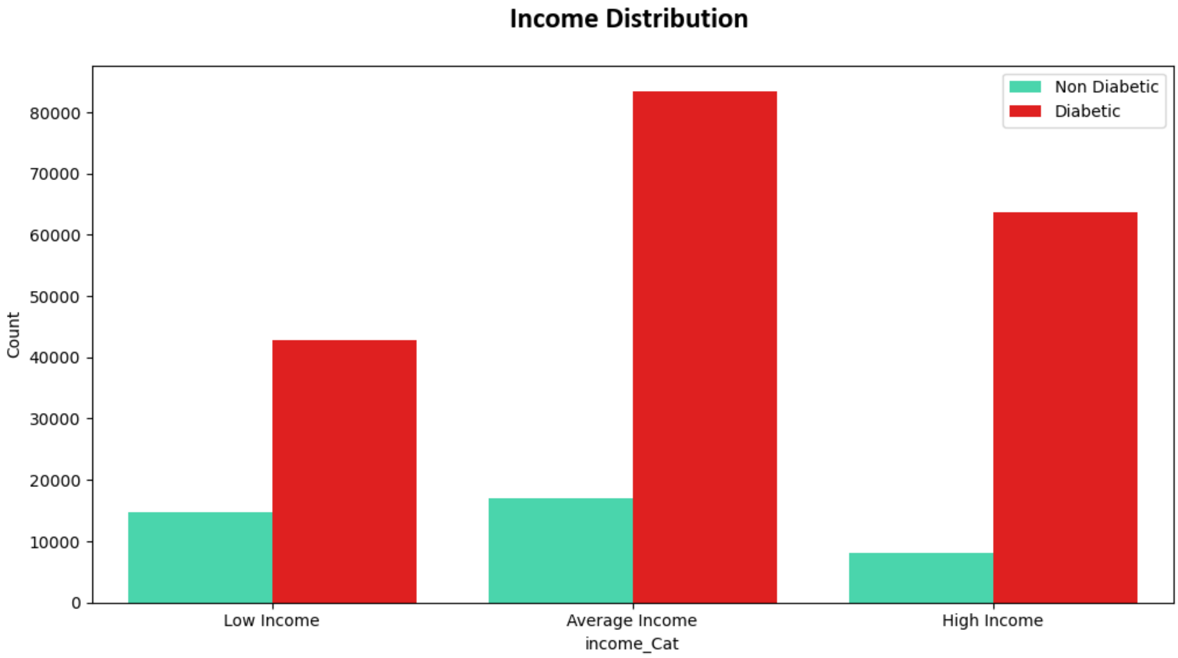


Figure 19: Income Distribution

Observation
Given the income categories above, It is observed that Diabetes is most prevalent among average income earners followed by high income earners

- **Investigation of Smoking as Diabetes Risk Factor¶**

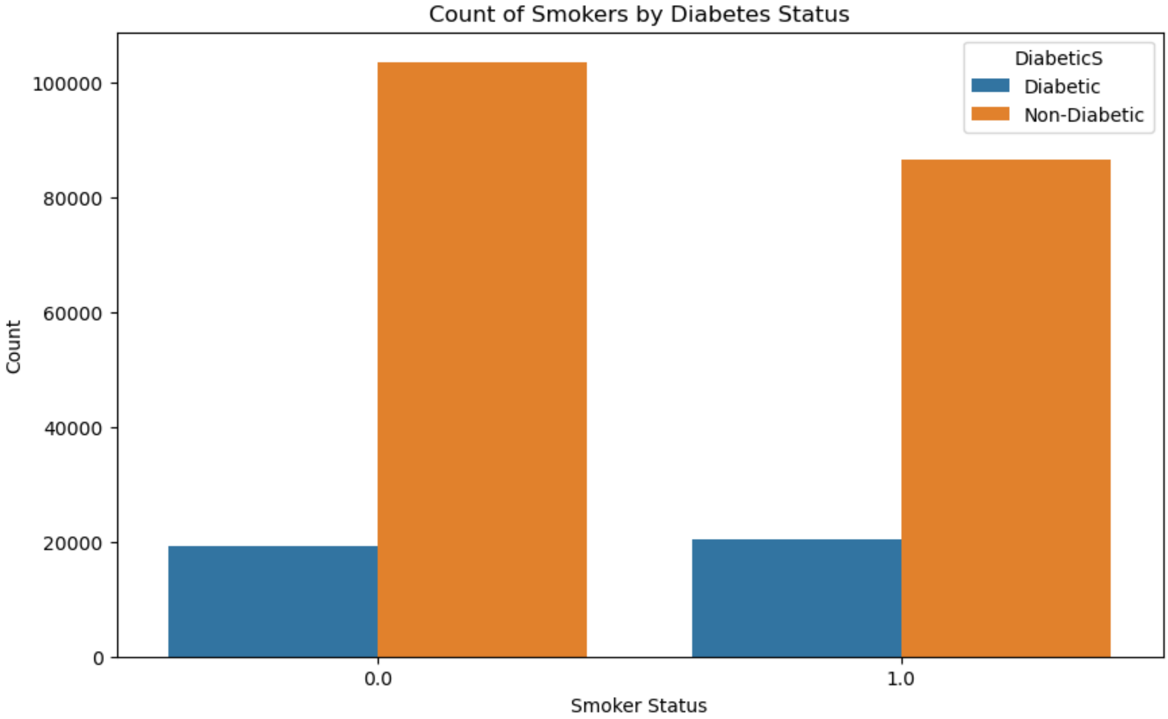


Figure 20: Count of smokers by diabetes status

Observation

Given the bar chat above, It is observed that Diabetes is more prevalent among smokers and less frequent among non-smokers

- **Investigation of High BP as Diabetes Risk Factor**

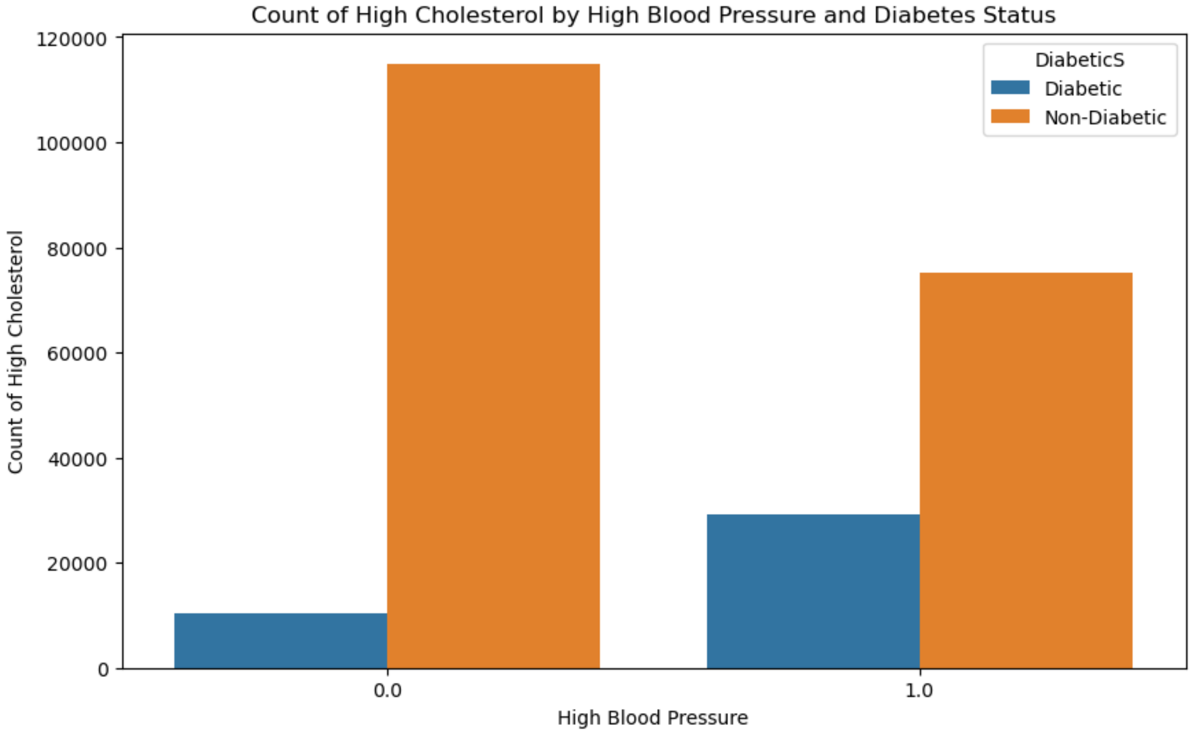


Figure 21: Relationship between High BP and Diabetes

Observation

From the chart above, we observe that.

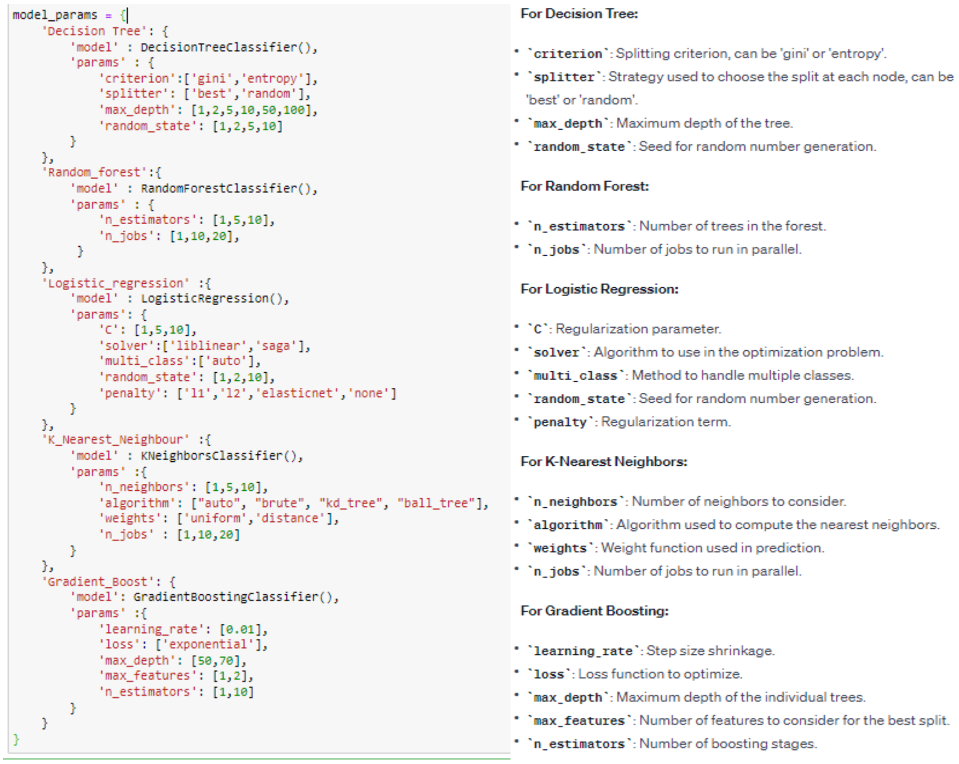
- Nondiabetic patients are mostly individuals with Low Blood Pressure and Low Cholesterol.
- Conversely, Diabetes is higher among individuals with high Blood Pressure and high cholesterol.

3.7 Machine Learning Models

The supervised models explored in this research include The Decision Tree, The Random Forest Algorithm, Logistic Regression, The K Nearest Neighbor Model, and The Gradient Boost.

Hyperparameter Tuning

Optimising hyperparameters is essential for improving the precision and dependability of the prediction model. Diabetes prediction entails using diverse machine-learning techniques that depend on distinct settings or hyperparameters. In this project, we have methodically modified parameters such as the learning rate, depth of the decision trees, or the number of estimators in a random forest classifier. The predictive models can, therefore, greatly enhance their capacity to detect individuals susceptible to diabetes.



```
model_params = {
    'Decision Tree': {
        'model' : DecisionTreeClassifier(),
        'params' : {
            'criterion':['gini','entropy'],
            'splitter': ['best','random'],
            'max_depth': [1,2,5,10,50,100],
            'random_state': [1,2,5,10]
        }
    },
    'Random_forest':{
        'model' : RandomForestClassifier(),
        'params' : {
            'n_estimators': [1,5,10],
            'n_jobs': [1,10,20],
         }
    },
    'Logistic_regression' :{
        'model' : LogisticRegression(),
        'params': {
            'C': [1,5,10],
            'solver':['liblinear','saga'],
            'multi_class':['auto'],
            'random_state': [1,2,10],
            'penalty': ['l1','l2','elasticnet','none']
        }
    },
    'K_Nearest_Neighbour' :{
        'model' : KNeighborsClassifier(),
        'params' :{
            'n_neighbors': [1,5,10],
            'algorithm': ["auto", "brute", "kd_tree", "ball_tree"],
            'weights': ['uniform','distance'],
            'n_jobs' : [1,10,20]
        }
    },
    'Gradient_Boost': {
        'model': GradientBoostingClassifier(),
        'params' :{
            'learning_rate': [0.01],
            'loss': ['exponential'],
            'max_depth': [50,70],
            'max_features': [1,2],
            'n_estimators': [1,10]
        }
    }
}
```

For Decision Tree:

- `criterion`: Splitting criterion, can be 'gini' or 'entropy'.
- `splitter`: Strategy used to choose the split at each node, can be 'best' or 'random'.
- `max_depth`: Maximum depth of the tree.
- `random_state`: Seed for random number generation.

For Random Forest:

- `n_estimators`: Number of trees in the forest.
- `n_jobs`: Number of jobs to run in parallel.

For Logistic Regression:

- `C`: Regularization parameter.
- `solver`: Algorithm to use in the optimization problem.
- `multi_class`: Method to handle multiple classes.
- `random_state`: Seed for random number generation.
- `penalty`: Regularization term.

For K-Nearest Neighbors:

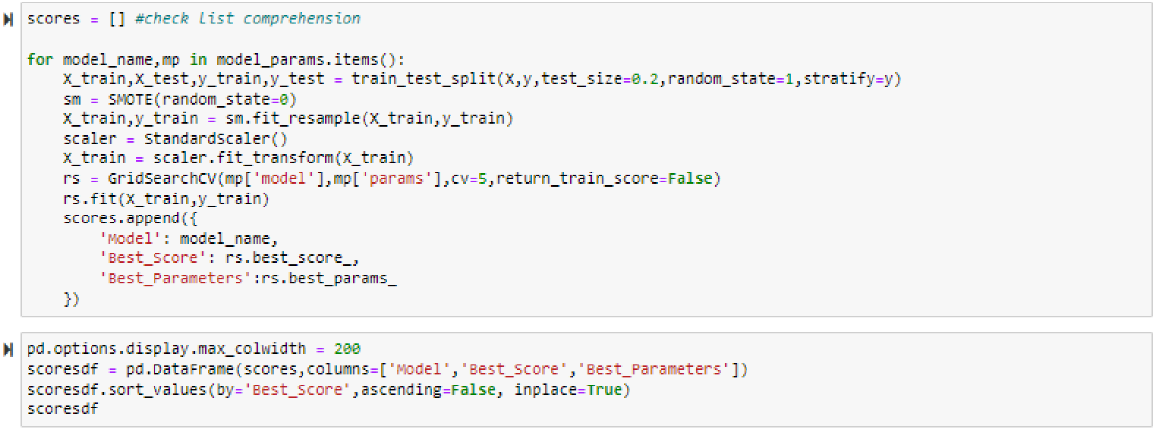
- `n_neighbors`: Number of neighbors to consider.
- `algorithm`: Algorithm used to compute the nearest neighbors.
- `weights`: Weight function used in prediction.
- `n_jobs`: Number of jobs to run in parallel.

For Gradient Boosting:

- `learning_rate`: Step size shrinkage.
- `loss`: Loss function to optimize.
- `max_depth`: Maximum depth of the individual trees.
- `max_features`: Number of features to consider for the best split.
- `n_estimators`: Number of boosting stages.

Figure 22: Hyperparameter Tuning

3.8 Model Training

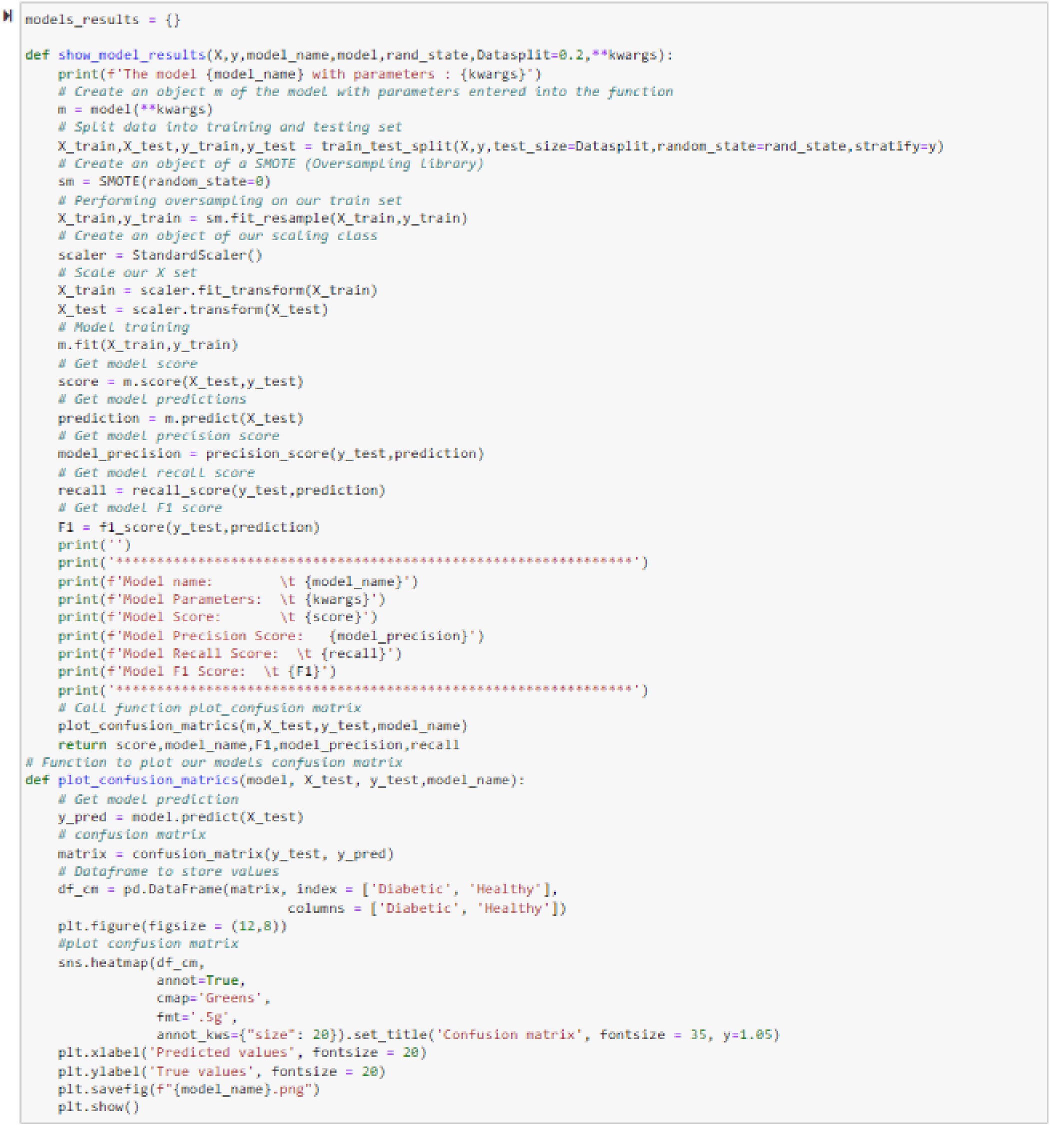


```
scores = [] #check list comprehension

for model_name,mp in model_params.items():
    X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=1,stratify=y)
    sm = SMOTE(random_state=0)
    X_train,y_train = sm.fit_resample(X_train,y_train)
    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    rs = GridSearchCV(mp['model'],mp['params'],cv=5,return_train_score=False)
    rs.fit(X_train,y_train)
    scores.append({
        'Model': model_name,
        'Best_Score': rs.best_score_,
        'Best_Parameters':rs.best_params_
    })
```

```
pd.options.display.max_colwidth = 200
scoresdf = pd.DataFrame(scores,columns=['Model','Best_Score','Best_Parameters'])
scoresdf.sort_values(by='Best_Score',ascending=False, inplace=True)
scoresdf
```

Figure 23: Model Training

3.9 Model Development



```
models_results = {}

def show_model_results(X,y,model_name,model,rand_state,Datasplit=0.2,**kwargs):
    print(f'The model {model_name} with parameters : {kwargs}')
    # Create an object m of the model with parameters entered into the function
    m = model(**kwargs)
    # Split data into training and testing set
    X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=Datasplit,random_state=rand_state,stratify=y)
    # Create an object of a SMOTE (Oversampling library)
    sm = SMOTE(random_state=0)
    # Performing oversampling on our train set
    X_train,y_train = sm.fit_resample(X_train,y_train)
    # Create an object of our scaling class
    scaler = StandardScaler()
    # Scale our X set
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)
    # Model training
    m.fit(X_train,y_train)
    # Get model score
    score = m.score(X_test,y_test)
    # Get model predictions
    prediction = m.predict(X_test)
    # Get model precision score
    model_precision = precision_score(y_test,prediction)
    # Get model recall score
    recall = recall_score(y_test,prediction)
    # Get model F1 score
    F1 = f1_score(y_test,prediction)
    print('')
    print('****************************************************************')
    print(f'Model name:        \t {model_name}')
    print(f'Model Parameters:  \t {kwargs}')
    print(f'Model Score:       \t {score}')
    print(f'Model Precision Score:   {model_precision}')
    print(f'Model Recall Score:  \t {recall}')
    print(f'Model F1 Score:  \t {F1}')
    print('****************************************************************')
    # Call function plot_confusion matrix
    plot_confusion_matrics(m,X_test,y_test,model_name)
    return score,model_name,F1,model_precision,recall
# Function to plot our models confusion matrix
def plot_confusion_matrics(model, X_test, y_test,model_name):
    # Get model prediction
    y_pred = model.predict(X_test)
    # confusion matrix
    matrix = confusion_matrix(y_test, y_pred)
    # Dataframe to store values
    df_cm = pd.DataFrame(matrix, index = ['Diabetic', 'Healthy'],
                            columns = ['Diabetic', 'Healthy'])
    plt.figure(figsize = (12,8))
    #plot confusion matrix
    sns.heatmap(df_cm,
                annot=True,
                cmap='Greens',
                fmt='.5g',
                annot_kws={"size": 20}).set_title('Confusion matrix', fontsize = 35, y=1.05)
    plt.xlabel('Predicted values', fontsize = 20)
    plt.ylabel('True values', fontsize = 20)
    plt.savefig(f"{model_name}.png")
    plt.show()
```

Figure 24: Model development

Here, we evaluated the performance of various machine learning models in a classification job, specifically in the context of biomedical applications such as predicting diabetes. The system performs data preprocessing tasks such as oversampling and scaling, trains the model, assesses its performance using many metrics, and presents the findings using a confusion matrix visualization.

The two functions, **show_model_results** and **plot_confusion_matrics**, which are used for training a machine learning model, evaluating its performance, and visualizing the results are defined as follows.

- **Function Definition**: It takes a dataset (X, y), a model name (model_name), a machine learning model class (model), a random state (rand_state), a data split ratio (Datasplit), and additional model parameters (**kwargs).

- **Model Initialization**: Creates an instance of the model (m) with the specified parameters.

- **Data Splitting**: "Splits the data into training and testing sets using train_test_split, with a test size defined by Datasplit and stratification on y".

- **SMOTE Oversampling**: "Applies SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance by oversampling the minority class in the training data".
- **Data Scaling**: Scales the features using StandardScaler to standardize the feature values.
- **Model Training**: Trains the model on the processed training data.
- **Model Evaluation**: Evaluates the model on the test data, calculating various metrics like model score, precision, recall, and F1 score.
- **Results Display**: Prints out the model's performance metrics.
- **Confusion Matrix Visualization**: Calls plot_confusion_matrics function to display the confusion matrix for the model.
- **Return Statement**: Returns the model's score, name, F1 score, precision, and recall plot_confusion_matrics Function.
- **Function Definition**: Takes a trained model, test data (X_test, y_test), and the model's name.
- **Model Prediction**: Uses the model to make predictions on the test data.
- **Confusion Matrix Creation:** Generates a confusion matrix from the true labels and predictions.
- **Matrix Visualization**: Visualizes the confusion matrix as a heatmap using seaborn, with labels for 'Diabetes' and 'Healthy' classes.
- **Saving the Plot**: Saves the confusion matrix plot as an image file named after the model.

Developing a model to predict the risk of diabetes entails utilising machine learning algorithms to analyse several data sets that include clinical, genetic, and lifestyle aspects. Data analysis methods such as **Logistic Regression, Decision Tree, Gradient Boost, KNN, Random Forest and Artificial Neural Networks** are utilized to develop predictive models. In order to improve the accuracy of the model, we preprocess the data, choose the relevant features, fine-tune the parameters, apply SMOTE to handle the class imbalance, and Data Scaling feature such as StandardScaler to standardize the feature values. The system's robustness is ensured by continuous refinement and validation against various populations. The objective is to develop trustworthy technologies that can detect individuals who are at risk for developing diabetes, hence facilitating early intervention and designing individualized healthcare solutions.

- **The Random Forest Classifier**
  The Random Forest Classifier, a form of ensemble learning, has been this report analytics, specifically for forecasting the likelihood of developing diabetes. It stands out for its accuracy, robustness, and ability to handle large datasets with numerous variables, making it a highly suitable tool for diabetes risk prediction (Breiman, 2001). The Random Forest method utilizes multiple health indicators, such as age, BMI, blood pressure, and cholesterol levels, from the "Diabetes_Health_Indicators.csv" dataset to forecast the probability of an individual acquiring diabetes. This forecast is derived from an ensemble of decision trees, known as a 'forest', where each tree is constructed using a random subset of the data. This approach mitigates the likelihood of overfitting and enhances the precision of predictions (Breiman, 2001).

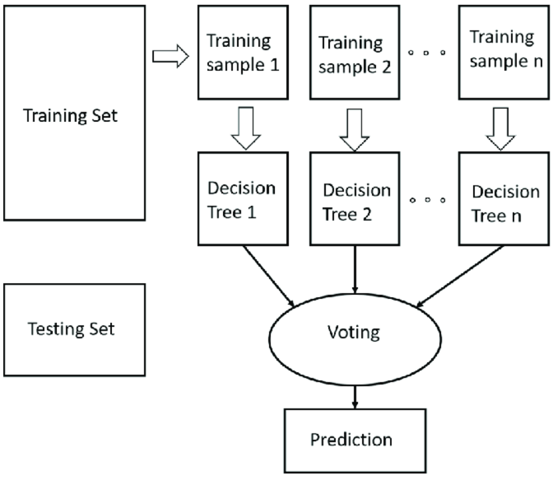
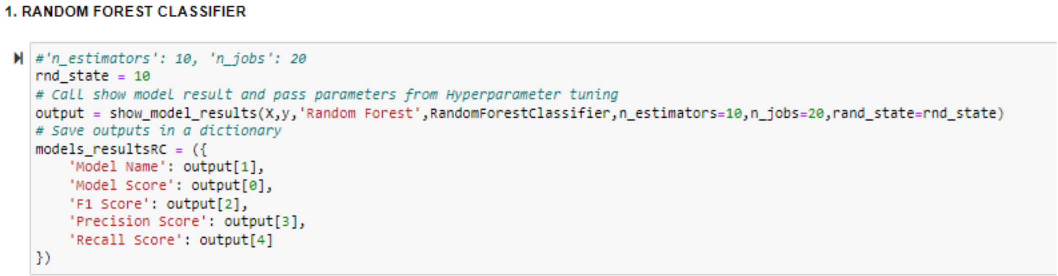


Figure 25: The Architecture of Random Forest Classifier

Random Forest has numerous benefits in this circumstance. The underlying technique of averaging or merging the outcomes of various decision trees aids in mitigating errors. If some trees are incorrect, others are likely to be correct, leading to a high overall accuracy (Breiman, 2001). Moreover, it has the capability to process both quantitative and qualitative data and can simulate intricate relationships among variables. However, Random Forests can become complex and less interpretable when dealing with many trees. This "black box" nature can make it challenging to discern how specific variables influence the overall prediction (Liaw & Wiener, 2002).



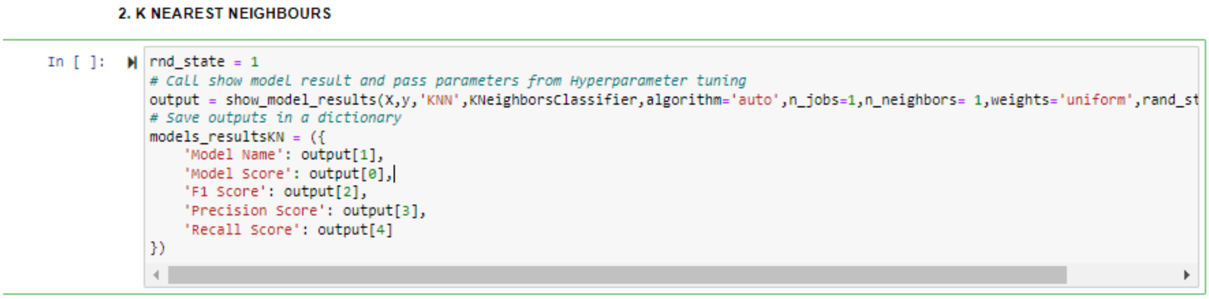
1. RANDOM FOREST CLASSIFIER

```
#'n_estimators': 10, 'n_jobs': 20
rnd_state = 10
# Call show model result and pass parameters from Hyperparameter tuning
output = show_model_results(X,y,'Random Forest',RandomForestClassifier,n_estimators=10,n_jobs=20,rand_state=rnd_state)
# Save outputs in a dictionary
models_resultsRC = ({
    'Model Name': output[1],
    'Model Score': output[0],
    'F1 Score': output[2],
    'Precision Score': output[3],
    'Recall Score': output[4]
})
```

Figure 26 - Random Forest Classifier

- **KNN**

When using the KNN algorithm on the dataset to predict the risk of diabetes, which includes health markers like BMI, age, and blood pressure, the algorithm determines the 'k' individuals who are closest to a certain patient based on these factors. The patient's diabetes status is subsequently predicted by determining the most prevalent status (diabetic or non-diabetics) among these neighboring individuals. The selection of 'k', the quantity of neighbors to contemplate, is crucial and might substantially impact the accuracy of the model. The K-Nearest Neighbors (KNN) algorithm, a non-parametric method used for classification and regression, finds widespread application in medical diagnostics, including predicting diabetes risk. KNN operates on the principle that similar things exist in proximity (Altman, 1992).



2. K NEAREST NEIGHBOURS

```
In [ ]:  rnd_state = 1
         # Call show model result and pass parameters from Hyperparameter tuning
         output = show_model_results(X,y,'KNN',KNeighborsClassifier,algorithm='auto',n_jobs=1,n_neighbors= 1,weights='uniform',rand_st
         # Save outputs in a dictionary
         models_resultsKN = ({
             'Model Name': output[1],
             'Model Score': output[0],
             'F1 Score': output[2],
             'Precision Score': output[3],
             'Recall Score': output[4]
         })
```

Figure 27 - K Nearest Neighbors

The KNN algorithm uses the health indicators of the people who are closest to them in the dataset to determine if a person has diabetes or not.

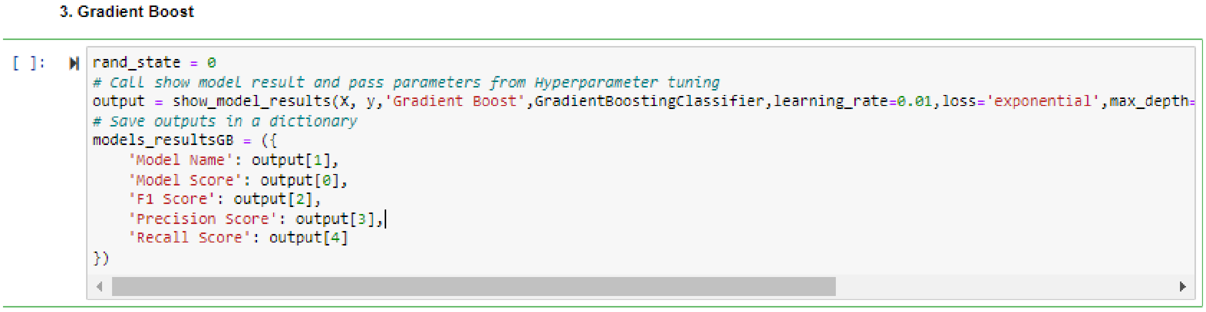
KNN offers a significant benefit in terms of its simplicity and effectiveness, particularly in situations when the decision boundary is non-uniform. The algorithm requires no explicit training phase, making it unique among other standard algorithms (Altman, 1992). However, KNN can suffer from high computational cost as the dataset grows, and its performance can be negatively impacted by the presence of irrelevant or redundant features (Dasarathy, 1991).

Although KNN has several limits, its capacity to adjust to new data renders it a significant tool in medical predictive analytics. It is especially advantageous in situations where prompt and dependable judgements are required, such as in the timely identification of the danger of diabetes.

- **GRADIENT BOOST**

When used for predicting the risk of diabetes, Gradient Boosting begins with a basic model and gradually enhances it by specifically addressing the cases that were categorized incorrectly throughout the training phase. Each subsequent model attempts to correct the errors of its predecessor, with the learning process continuing until no significant improvements can be made or a predetermined number of trees are added. (Natekin., et al., 2013).

An important advantage of the algorithm is its capacity to effectively process different types of data and distributions, rendering it very suitable for a wide range of intricate medical datasets. Furthermore, Gradient Boosting can capture intricate non-linear associations between variables and the target variable, which is essential for precise prediction of diabetes, a disease that is influenced by diverse biology and lifestyle factors.



3. Gradient Boost

```
[ ]:  rand_state = 0
      # Call show model result and pass parameters from Hyperparameter tuning
      output = show_model_results(X, y,'Gradient Boost',GradientBoostingClassifier,learning_rate=0.01,loss='exponential',max_depth=
      # Save outputs in a dictionary
      models_resultsGB = ({
          'Model Name': output[1],
          'Model Score': output[0],
          'F1 Score': output[2],
          'Precision Score': output[3],
          'Recall Score': output[4]
      })
```

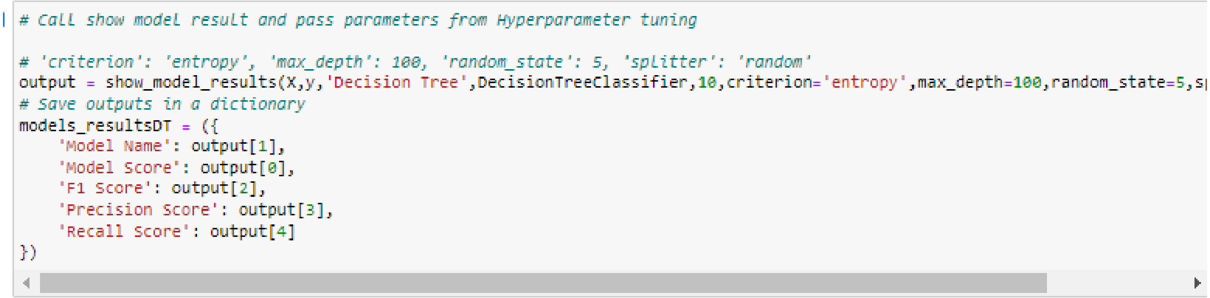
Figure 28 - Gradient Boost

Nevertheless, the approach necessitates meticulous adjustment of parameters such as the quantity and depth of trees, as well as the learning rate, in order to prevent overfitting. In addition, compared to approaches such as logistic regression, this technique is more advanced and requires more processing resources. However, it may be less interpretable and is frequently referred to as a 'black box' (Natekin and Knoll, 2013). Although there are difficulties, the predictive capability of Gradient Boosting renders it a valuable resource in medical predictive analytics. It can offer healthcare professionals valuable information about the risk factors that contribute to diabetes, assisting in early detection and tailored treatment approaches.

- **DECISION TREE**

The Decision Tree algorithm is a crucial tool in medical predictive analytics, specifically in predicting the risk of diabetes, due to its inherent simplicity and interpretability. Decision trees are a modelling technique that divides a dataset into subgroups based on the values of predictive variables. They provide a clear understanding of how various factors contribute to the risk of diabetes.

When utilized on the dataset with indicators such as BMI, age, and blood pressure, a decision tree model functions by generating binary divisions based on these variables. Every individual node within the tree symbolizes a decision rule, and the depth of the tree can be modified to manage the complexity of the model. Decision trees are particularly valuable for deciphering intricate patterns in data that would otherwise be challenging to analyze (Quinlan, 1986).



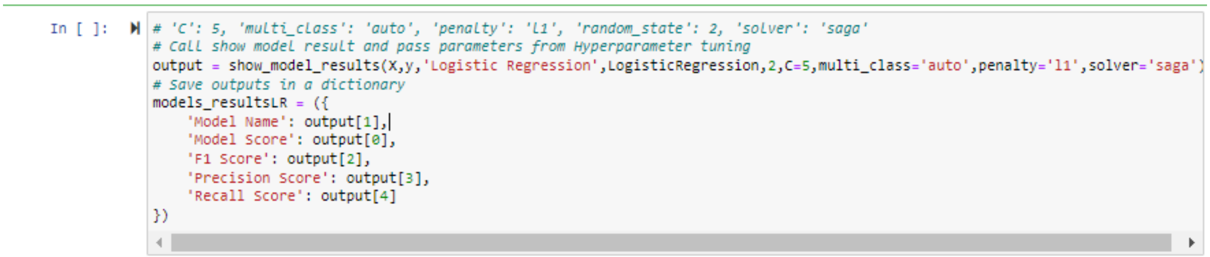
```
# Call show model result and pass parameters from Hyperparameter tuning

# 'criterion': 'entropy', 'max_depth': 100, 'random_state': 5, 'splitter': 'random'
output = show_model_results(X,y,'Decision Tree',DecisionTreeClassifier,10,criterion='entropy',max_depth=100,random_state=5,s
# Save outputs in a dictionary
models_resultsDT = ({
    'Model Name': output[1],
    'Model Score': output[0],
    'F1 Score': output[2],
    'Precision Score': output[3],
    'Recall Score': output[4]
})
```

Figure 29 - Decision Trees

The primary benefit of the decision trees in forecasting the risk of diabetes rests in their capacity to be easily understood and interpreted. Decision trees, unlike more intricate models, may be visually depicted and readily comprehended by clinicians. This facilitates the translation of data-driven insights into effective medical decisions (Breiman et al., 1984). Furthermore, they are non-parametric, meaning that they do not require data to follow a predefined distribution.

Decision trees have a drawback in that they are susceptible to overfitting, especially when dealing with many attributes or missing proper pruning. Overfitting results in a model that exhibits high performance on the training data but performs badly on data that it has not been exposed to (Breiman et al., 1984). Methods such as pruning, which involves lowering the size of the tree, and ensemble techniques like Random Forests are commonly employed to tackle this issue.

- **LOGISTIC REGRESSION**
  Logistic regression is favored due to its simplicity, interpretability, and efficacy in situations when there are only two possible outcomes. This study utilizes an algorithm to evaluate the likelihood of an individual having diabetes by considering these criteria. It expresses the relationship between the binary dependent variable (diabetes or not) and one or more independent variables by estimating probabilities using a logistic function (Hosmer Jr et al., 2013).

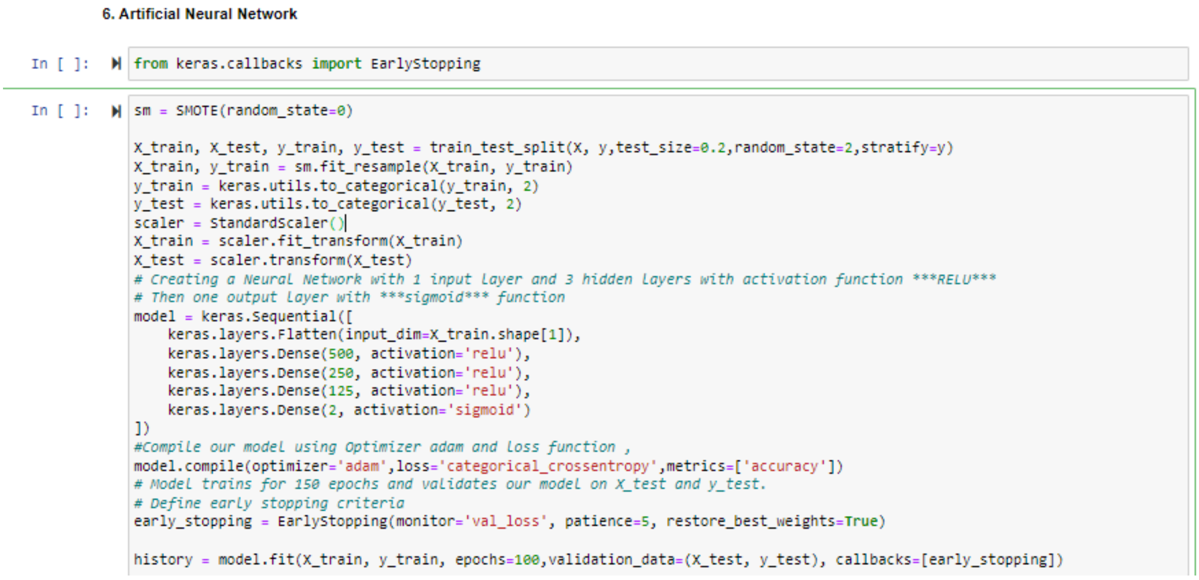


```
In [ ]:  # 'C': 5, 'multi_class': 'auto', 'penalty': 'l1', 'random_state': 2, 'solver': 'saga'
         # Call show model result and pass parameters from Hyperparameter tuning
         output = show_model_results(X,y,'Logistic Regression',LogisticRegression,2,C=5,multi_class='auto',penalty='l1',solver='saga')
         # Save outputs in a dictionary
         models_resultsLR = ({
             'Model Name': output[1],
             'Model Score': output[0],
             'F1 Score': output[2],
             'Precision Score': output[3],
             'Recall Score': output[4]
         })
```

Figure 30 - Logistic Regression

Logistic Regression has a notable advantage in this context as it can generate probability scores that accurately indicate the likelihood of the development of diabetes. Additionally, it permits the assessment of the impact of each predictor variable, helping healthcare providers to comprehend which elements contribute most significantly to the risk of diabetes. The ability to understand and explain the meaning of anything is crucial for making healthcare decisions and creating specific tactics to prevent certain outcomes.
Remarkable to note that a linear relationship between the independent factors and the result may not always hold true, contrary to the assumption made by logistic regression. The model can also be prone to underperforming if there are nonlinear relationships in the data (Peng et al., 2002).

- **ARTIFICIAL NEURAL NETWORK**
  When compared to other models utilized in this study, Artificial Neural Networks (ANNs) have shown to produce the most accurate predictive model. The dataset "Diabetes_Health_Indicators.csv," which contains variables like income levels, blood pressure, BMI, and other lifestyle characteristics, was analyzed by ANN, and it performed exceptionally well at finding complex patterns and correlations between the variables. This is particularly advantageous in diabetes prediction, where the relationship between risk factors and the disease is often non-linear and multifaceted (Rumelhart et al., 1986). Artificial neural networks (ANNs) acquire knowledge by iteratively modifying the synaptic weights that connect neurons, employing algorithms such as backpropagation, throughout a training procedure. Once trained, the network can make predictions about new data, classifying individuals as at risk or not for diabetes based on their health indicators (LeCun et al., 2015).



6. Artificial Neural Network

```
In [ ]:  from keras.callbacks import EarlyStopping
```

```
In [ ]:  sm = SMOTE(random_state=0)

         X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.2,random_state=2,stratify=y)
         X_train, y_train = sm.fit_resample(X_train, y_train)
         y_train = keras.utils.to_categorical(y_train, 2)
         y_test = keras.utils.to_categorical(y_test, 2)
         scaler = StandardScaler()
         X_train = scaler.fit_transform(X_train)
         X_test = scaler.transform(X_test)
         # Creating a Neural Network with 1 input Layer and 3 hidden Layers with activation function ***RELU***
         # Then one output Layer with ***sigmoid*** function
         model = keras.Sequential([
             keras.layers.Flatten(input_dim=X_train.shape[1]),
             keras.layers.Dense(500, activation='relu'),
             keras.layers.Dense(250, activation='relu'),
             keras.layers.Dense(125, activation='relu'),
             keras.layers.Dense(2, activation='sigmoid')
         ])
         #Compile our model using Optimizer adam and Loss function ,
         model.compile(optimizer='adam',loss='categorical_crossentropy',metrics=['accuracy'])
         # Model trains for 150 epochs and validates our model on X_test and y_test.
         # Define early stopping criteria
         early_stopping = EarlyStopping(monitor='val_loss', patience=5, restore_best_weights=True)

         history = model.fit(X_train, y_train, epochs=100,validation_data=(X_test, y_test), callbacks=[early_stopping])
```

Figure 31 - Artificial Neural Network

An important advantage of artificial neural networks (ANNs) is their inherent flexibility and plasticity, enabling them to accurately represent intricate non-linear connections. Nevertheless, this intricacy is accompanied by disadvantages: Artificial neural networks (ANNs) necessitate substantial quantities of data for efficient training and are frequently characterized as 'black boxes' owing to their limited interpretability in contrast to more straightforward models such as logistic regression. The opaque aspect of the 'black box' can provide difficulties in therapeutic environments where comprehending the underlying reasoning behind predictions is essential (LeCun et al., 2015).

Chapter Four

Evaluation And Discussion

4 EVALUATION AND DISCUSSION

4.1 Models Results and research Findings.

The six machine models explored in this research measure in comparison with the following metrics namely **Model Score**, commonly known as accuracy (measured in percentage), measures the proportion of adequately predicted occurrences out of the total instances in the dataset. According to the table below, the Artificial Neural Network has the highest model score, followed by the random Forest model. The high score signifies that a substantial proportion of the model's predictions of people with diabetes align with the actual outcomes in the diabetes dataset used for testing or validation. **The model precision** score quantifies the proportion of accurately predicted positive observations among all the expected positives. The high precision score recorded by ANN signifies that when the model detects individuals as susceptible to diabetes, a significant percentage of them are indeed at risk based on the ground truth or factual data. **The model's recall score,** also known as sensitivity, is recorded highest for ANN, meaning the model has sufficient sensitivity in detecting a significant proportion of the individuals who are at risk for diabetes in comparison to the overall number of actual positive cases. **The Model's F1** score indicates the model's proficiency in accurately detecting positive cases of persons at risk for diabetes and reducing the occurrence of false positives. The ANN model recorded the highest score compared to other models.

Results

Model name	Model Score	Model Precision Score	Model Recall Score	Model F1 Score
Random Forest	81.6%	43.5%	21.3%	28.6%
KNN	76.0%	31.4%	32.8%	32.1%
Gradient Boost	79.9%	43.3%	52.2%	47.3%
Decision Tree	76.1%	31.9%	33.9%	32.9%
Logistic Regression	71.5%	34.7%	74.0%	47.3%
ANN	82.4%	63.9%	67.6%	62.2%

Figure 32 - Model Results

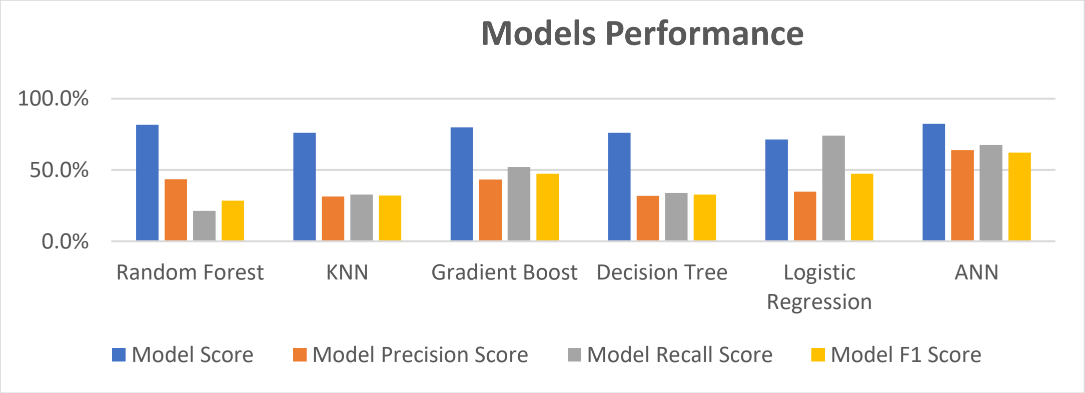
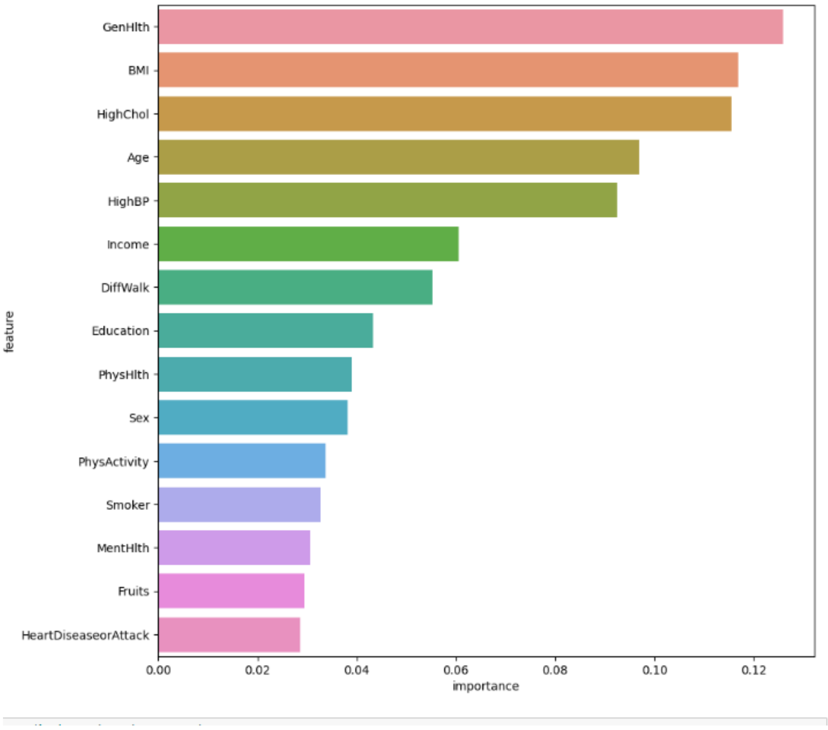


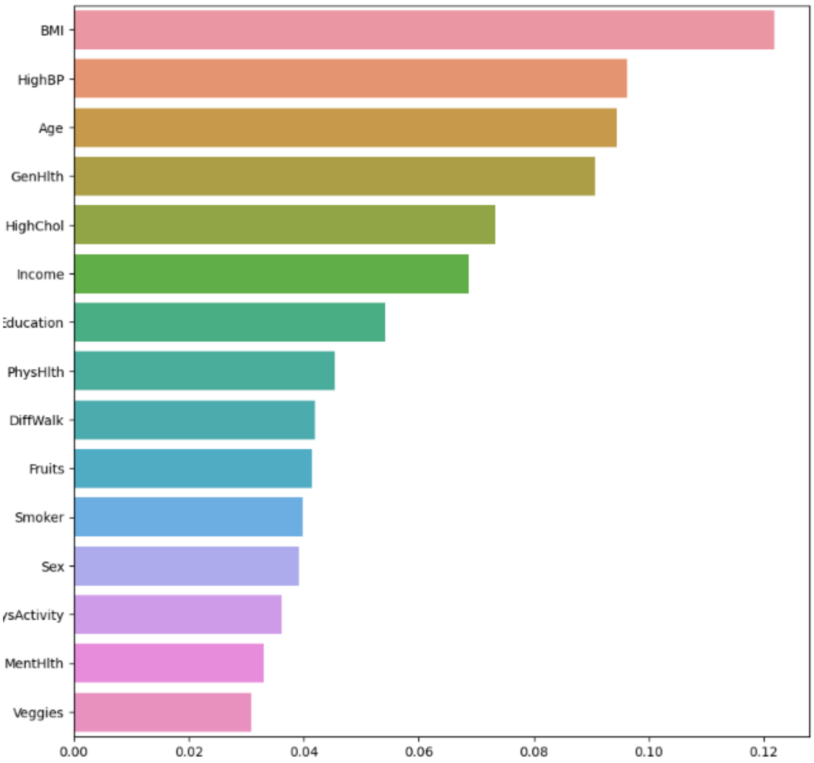
Figure 33 - Model Performance

On using Random Forest Algorithm, the primary risk factors that contribute to the development and progression of diabetes are as follows in order of importance.

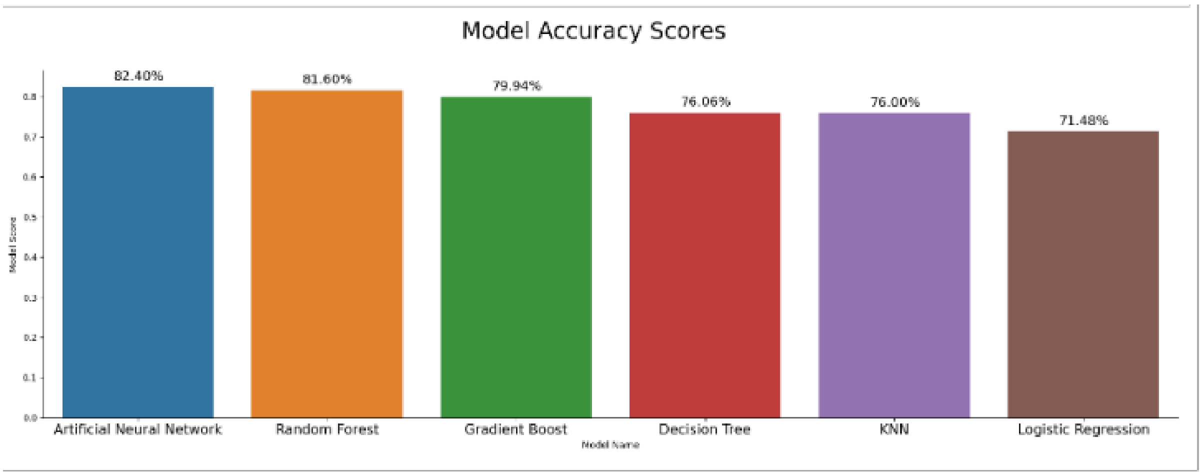
Random Forest*



While the order of importance for Gradient Boost Model is



In conclusion, ANN proved to be the most suitable model for the task amidst wide range of constraints and limitations.



4.2 Research Questions and Hypothesis

This project emphasises the necessity to balance the advantages and disadvantages of predicting diabetes risk using machine learning models. It tries to investigate the impact of the risk factors identified with diabetes.

Research Question 1:

What are the primary risk factors that contribute to the development and progression of diabetes, and how do these risk factors vary across different populations and demographics?

Hypothesis 1:

Different populations and demographics have different prevalence and significance of certain risk factors, such as BMI, unhealthy diet, sedentary lifestyle, environmental factors, and socioeconomic status. These variations are critical to the onset and progression of diabetes within those groups. According to this theory, different populations and demographics are affected by the same risk factors when it comes to diabetes, and knowing these differences is crucial for developing customized diabetes preventive and treatment plans.

Research Question 2:

What are the benefits and constraints of using machine learning models for predicting the risk of diabetes, and how do these models compare to traditional methods in terms of accuracy, interpretability, and practicality?

Hypothesis 2:

Machine learning algorithms for predicting the risk of diabetes will consistently surpass traditional risk assessment methods in terms of accuracy, interpretability, and practicality in real-world scenarios.

This hypothesis posits that machine learning models would consistently surpass traditional approaches in all respects, irrespective of the potential constraints or variations in performance that may arise due to factors such as data quality, model selection, and other variables.

4.3 Limitations

This research work on the significance of machine learning models in predicting diabetes risks is faced with several constraints such as limitations in the data quality, race and demographics captured in the datasets. The lack of access to comprehensive or diverse datasets containing various demographics, genetic information, lifestyle variables, or long-term health records was a disadvantage to this research. The dataset was derived from a poll conducted among a sample size of only 400,000 individuals in the United States. The depth and accuracy of prediction models was restricted since the scope of the dataset used is restricted. Secondly, the distribution of primary risk factors vary across different populations and demographics could not be substantiated in this research because of this range of the dataset. Further research would need a more diverse dataset so that the risk of developing diabetes between various populations or ethnic groups can be better captured by models. Therefore, questions answered in the future research work would be how each risk factor affect the African population compared to European Population.

This research has investigated the merits and demerits of the machine learning models and how they compare with the traditional methods of predicting the risk of diabetes. data quality, model selection, and other variables pose a great limitation to this investigation. The presence of data imbalance between the 0s and 1s categories of the target feature impact the model's capacity to discern patterns and generate predictions for the underrepresented class.

Third, the practical use of the research project artefact is limited by the lack of web or system applications deployed for easy interpretability and model validation. The knowledge gap has been pointed to be responsible for the lack of built-in tools for this project to enable clinical users who do not understand the workings of the machine learning models to collect data and predict the risk of diabetes in an individual. We therefore hope that future research will implement and deploy user-friendly in-built app or graphical user interface that will enable clinical users to use apply the models.

4.4 Ethical Considerations

The sensitive nature of personal health data is essential to the research study of diabetes risk prediction. Critical ethical issues to consider are.

Privacy:

Ensuring patient privacy is of utmost importance regarding privacy and data security. Research by Shaw et al. (2018) emphasises the significance of patient data privacy in predictive modelling, highlighting the need for stringent data protection measures. Accurate assessment of the risk of developing diabetes requires using a large amount of health data. To protect the privacy and security of this sensitive information, it is crucial to implement robust mechanisms to anonymise, secure, and properly manage it. Complying with data protection regulations and guaranteeing the security of data storage and transmission are essential. Adhering to ethical guidelines, such as those outlined by the General Data Protection Regulation (GDPR) in healthcare settings, ensures secure handling, storage, and sharing of sensitive health information (Ahmed et al., 2019).

Transparency:

Trust between healthcare workers and patients relies heavily on openness in designing and interpreting machine learning models. The use of simple, non-ambulatory language and clear technical jargon make the research paper easily comprehensible. Research by Caruana et al. (2015) stresses the importance of explainable A.I. to enhance model interpretability. Ensuring that models provide clear explanations for their predictions fosters trust and empowers patients to make informed decisions regarding their healthcare (Rudin, 2019).

Medical Biases

The elimination of biases within prediction models is of the utmost importance. A study by Obermeyer et al. (2019) highlights how algorithmic biases can impact healthcare disparities. Disparities in risk estimates can be caused by biases in data collecting or computational biases, which can adversely influence demographic groups. One of the most critical aspects of equitable healthcare is ensuring that all different populations are treated fairly. Addressing biases and ensuring fairness across diverse populations are vital for equitable healthcare delivery (Braithwaite et al., 2020).

4.5 Deployment and Practical Implication

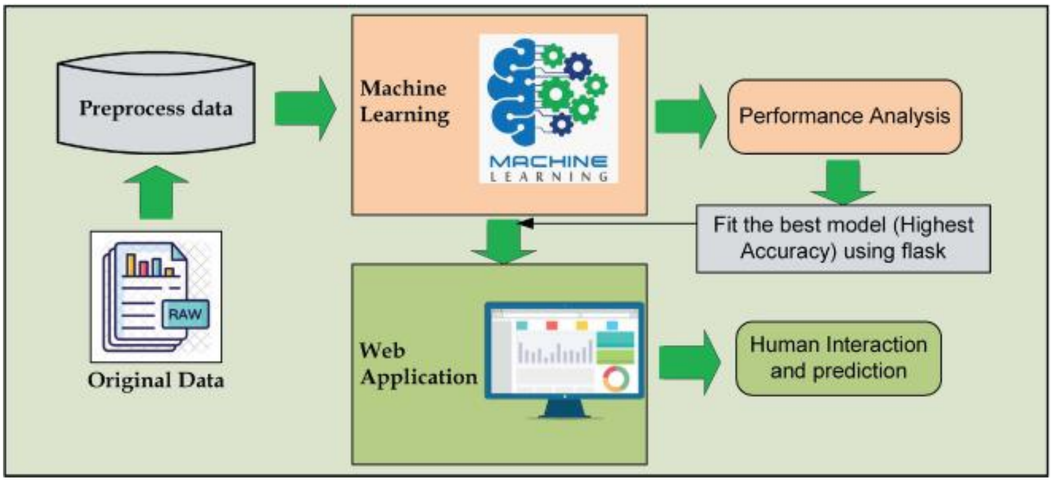


Fig 7. ML Model Deployment

Implementing ML models for diabetes risk prediction holds many revolutionary promises in advancing diabetes treatment and management. Keying in into the opportunities provided by the availability of various data sources such as health service users' personal records, lifestyle records, and socio-geographical data can offer custom risk examination, early disease detection and custom preventive measures.

The utilisation of machine learning (ML) models for predicting the risk of diabetes in online apps is a significant breakthrough in healthcare, facilitating proactive and tailored disease management. ML algorithms, trained on diverse datasets encompassing medical records, genetic information, and lifestyle factors, hold the potential to predict diabetes predisposition with high accuracy (Huang et al., 2020). Incorporating these models into online applications makes it easier to create user-friendly interfaces that enable individuals to submit relevant data for immediate risk evaluation. The deployment is in multiple phases. Data preprocessing ensures that the data is clean and compatible with the model. On the other hand, model training involves inputting the prepared data into machine learning algorithms to learn and make predictions. The challenge lies in presenting complex predictive analytics in an easily understandable format for users (Jiang et al., 2017). Visual aids like graphical user interfaces are employed to enhance interpretability.

However, as explored in this study, ethical considerations regarding data privacy and security are practical implications that must be handled with utmost adherence to rules and best practices. Adherence to regulatory frameworks like the Health Insurance Portability and Accountability Act (HIPAA) ensures the protection of sensitive patient health information within web applications (Yan et al., 2019). It is critical to comply with such regulations to earn and sustain trust.

Moreover, the utilisation of such models encourages patient engagement and awareness. Individuals can adopt a proactive stance in controlling their health by offering personalised risk assessments. People gain an understanding of how their actions and genetic makeup influence their vulnerability, which motivates them to adopt healthy behaviours and adhere to recommended remedies.

Chapter Five

Conclusion

CONCLUSION

The research project aims to investigate the significance of machine learning models to diabetes risk prediction. It is also aimed at studying the impact of different risk factors on the distribution and progression of diabetes in patients. Results and findings are made by comparing different literatures, analyzing data, and testing different machine models. The study also shed lights into the limiting constraints such as privacy issues, data quality and bias that frustrate the performance of machine learning algorithms despite the promises the technology advancement holds in diabetes care and management.

The research project studied existing research work and identified research gaps in the area of Data Diversity, model interpretability, and Feature Engineering and Selection. The paper suggests a range of methods to address these theoretical gaps, expanding upon current theory and practice and offering valuable guidance for future research. Future research is open to many opportunities in consolidated health data that cuts across different populations and demographics. We also look forward to seeing more work done in enhancing black-box models' interpretability. This is expected to improve the acceptance and confidence of clinicians, patients and general healthcare professionals managing diabetes diseases.

A more robust user-friendly model application to predict diabetes risks has promising future as new discoveries in deep Learning Algorithms, Federated Learning, Natural Language Processing (NLP), Telemedicine and Remote Monitoring, and Genomic Data Analysis are made. Despite these promising scopes in future work, it is important to place immense importance on ethical considerations, social and environmental impacts relating to the research.

In conclusion, this study compares the performance of different machine learning models in terms of globally accepted standard metrics Artificial Neural Network proved to be the most appropriate model for the task and BMI, General Health, High Cholesterol, Age and Blood Pressure are some of the most impactful diabetes risk factors.

5 REFERENCE

Sun, H. et al. (2022) 'I.D.F. Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045', Diabetes Research and Clinical Practice, 183, pp. 109119. Available at: https://doi.org/10.1016/j.diabres.2021.109119

Bhat, S.S. et al. (2023). 'A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms', Healthcare Analytics, pp. 100273. Available at: https://doi.org/10.1016/j.health.2023.100273

Dhande, B. et al., (2022) 'Diabetes & Heart Disease Prediction Using Machine Learning', ITM Web of Conferences, 44, pp. 3057. Available at: https://doi.org/10.1051/itmconf/20224403057

Schumacher-Petersen, C. et al. (2019). 'Experimental non-alcoholic steatohepatitis in Göttingen Minipigs: consequences of high fat-fructose-cholesterol diet and diabetes', Journal of Translational Medicine, 17(1), pp. 110. Available at: https://doi.org/10.1186/s12967-019-1854-y

Huang, J., Li, R. and Tsai, L. (2022). 'Relationship between Depression with Physical Activity and Obesity in Older Diabetes Patients: Inflammation as a Mediator', Nutrients, 14(19), pp. 4200. Available at: https://doi.org/10.3390/nu14194200

(Alshannaq, H. et al.. 2023) 'Cost-utility of real-time continuous glucose monitoring versus self-monitoring of blood glucose and intermittently scanned continuous glucose monitoring in people with type 1 diabetes receiving multiple daily insulin injections in Denmark', Diabetes, Obesity & Metabolism, 25(9), pp. 2704-2713. Available at: https://doi.org/10.1111/dom.15158

Kumar, S. et al. (2023). 'Optimised feature fusion-based modified cascaded kernel extreme learning machine for heart disease prediction in E-healthcare', Computer Methods in Biomechanics and Biomedical Engineering, ahead-of-print(ahead-of-print), pp. 1–14. Available at: https://doi.org/10.1080/10255842.2023.2218520

Yan, Y. (2022). 'Machine learning Fundamentals', Machine Learning in Chemical Safety and Health: Fundamentals with Applications, pp. 19–46.

Nhangumbe, M. et al., (2023) 'Supervised and unsupervised machine learning approaches using Sentinel data for flood mapping and damage assessment in Mozambique', Remote Sensing Applications: Society and Environment, 32, pp. 101015. Available at: https://doi.org/10.1016/j.rsase.2023.101015

Mohebbi, A. et al. (2017) 'A deep learning approach to adherence detection for type 2 diabetess', IEEE Available at: 10.1109/EMBC.2017.8037462.

Zhou, H., Myrzashova, R. & Zheng, R. (2020). 'Diabetes prediction model based on an enhanced deep neural network', EURASIP Journal on Wireless Communications and Networking, 2020(1), pp. 1–13. Available at: https://doi.org/10.1186/s13638-020-01765-7

Creatore, M.I. et al. (2010). 'Age- and sex-related prevalence of diabetes mellitus among immigrants to Ontario, Canada', Canadian Medical Association Journal (CMAJ), 182(8), pp. 781–789. Available at: https://doi.org/10.1503/cmaj.091551

Kumar, S. et al. (2023). 'Optimised feature fusion-based modified cascaded kernel extreme learning machine for heart disease prediction in E-healthcare', Computer Methods in Biomechanics and Biomedical Engineering, ahead-of-print(ahead-of-print), pp. 1–14. Available at: https://doi.org/10.1080/10255842.2023.2218520

Gomez Fernandez, M. *et al.* (2017). 'Nuclear energy system's behaviour and decision making using machine learning', *Nuclear Engineering and Design,* 324, pp. 27–34. Available at: https://doi.org/10.1016/j.nucengdes.2017.08.020

Xu, H. & Shuttleworth, K.M.J. (2023). 'Medical artificial intelligence and the black box problem – a view based on the ethical principle of "Do No Harm"', Intelligent Medicine, Available at: https://doi.org/10.1016/j.imed.2023.08.001

Agyemang, C., van der Linden, E.L. and Bennet, L. (2021). 'Type 2 diabetes burden among migrants in Europe: unravelling the causal pathways', *Diabetologia,* 64(12), pp. 2665–2675. Available at: https://doi.org/10.1007/s00125-021-05586-1

Ignatowicz, A. et al. (2018) 'Ethical implications of digital communication for the patient-clinician relationship: Analysis of interviews with clinicians and young adults with long term conditions (the LYNC study)', BMC Medical Ethics, 19(1), pp. 11. Available at: https://doi.org/10.1186/s12910-018-0250-0

Braithwaite, V. et al. (2020). When complexity science meets implement tation science: a theoretical and empirical systems change analysis. BMC Medicine, 18(1), 126. https://doi.org/10.1186/s12916-018-1057-z

Caruana, R. et al. (2015) 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission', ACM Available at: 10.1145/2783258.2788613

Obermeyer, Z. et al. (2019) 'Dissecting racial bias in an algorithm used to manage the health of populations', Science (American Association for the Advancement of Science), 366(6464), pp. 447-453. Available at: https://doi.org/10.1126/science.aax2342

Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.

Shaw, J.E., Sicree, R.A. and Zimmet, P.Z. (2010) 'Global estimates of the prevalence of diabetes for 2010 and 2030', *Diabetes Research and Clinical Practice,* 87(1), pp. 4-14. Available at: https://doi.org/10.1016/j.diabres.2009.10.007

6 APPENDIX

1. Data Source: https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset

2. Data description

	Variables	Descriptions
1.	Blood pressure	0 = no diabetes 1 = prediabetes 2 = diabetes
2	High BP	0 = no high BP 1 = high BP
3	High Cholesterol	0 = no high cholesterol 1 = high cholesterol
4	Chol Check	0 = no cholesterol checks in 5 years 1 = yes cholesterol check in 5 years
5	BMI	Body mass Index
6.	Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
7.	Heart Disease or Attack	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
8.	Physical Activity	physical activity in past 30 days - not including job 0 = no 1 = yes
9.	Fruit	Consume Fruit 1 or more times per day 0 = no 1 = yes