A logo for a university

Description automatically generated

**Bachelor Thesis**

**Bachelor of Science (BSc.)**

**Department of Tech and Software**

**Major: Software Engineering**

**Application of Machine Learning Techniques for the Early Detection of Diabetes: A Comparative Study of Classification Models**

Sebastian Russo

Matriculation Number: 79117092

First supervisor: XXXXXXXX

Second supervisor: XXXXXXXX

Submitted on: XX.XX.2025

**Statutory Declaration**

I hereby declare that I have developed and written the enclosed Bachelor Thesis completely by myself and have not used sources or means without declaration in the text. I clearly marked and separately listed all the literature and all the other sources which I employed when producing this academic work, either literally or in content. I am aware that the violation of this regulation will lead to the failure of the thesis.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

XX.XX.2025, Potsdam, Germany Author’s signature

**Abstract DO AT THE END**

Write at maximum one page abstract, describe the problem, its importance and then your contributions.

**Table of Contents DO AT THE END, THE VERY LAST THING**

**Staturtory declaration**

**Abstract**

**List of Abbreviations**

**Glossary**

**List of Tables**

**List of Figures**

**Chapter 1.** **Introduction 1**

1. 1 Background 1

1. 2 Problem Statement 1

1. 3 Research Questions 1

1. 4 Thesis Objectives 1

1. 5 Contributions 1

1. 6 Thesis Organization 1

**Chapter 2.** **Theoretical Background 2**

2. 1 Introduction 2

2. 2 Data Science and ML Theory 2

2. 3 Supervised Classification Algorithms 2

2.3.1 Logistic Regression 2

2.3.2 Decision Tree 2

2.3.3 Random Forest 2

2.3.4 Support Vector Machines (SVM) 2

2. 4 Model Evaluation Metrics 2

2.5 Feature Selection and Importance in Machine Learning 2

2.6 Challenges of Real-World Deployment of ML in Healthcare 2

**Chapter 3.** **Literature Review 3**

3. 1 Purpose of the Literature Review 3

3. 2 Structure of the Literature Review 3

3. 3 Key Terms and Concepts  3

3.4 Diabetes and the Need for Early Detection 3

3.5 Role of Machine Learning in Healthcare 3

3.6 Comparative Studies of ML Models for Diabetes Prediction 3

3.6.1 Commonly Used Machine Learning Models 3

3.6.2 Performance Comparison Across Studies 3

3.6.3 Key Insights from Comparative Studies 3

3.7 Interpretability and Model Transparency in Healthcare 3

3.8 Identified Gaps in the Literature 3

**Chapter 4.** **Methodology 4**

4. 1 Overview 4

4. 2 Proposed Framework/Model/Technique 4

4. 3 Methodology 4

*4.3.1* *Dataset 4*

*4.3.2* *Data Pre-Processing 4*

*4.3.3* *Feature Engineering 4*

*4.3.4* *Feature Selection 4*

*4.3.5* *Model Training 4*

4. 4 Evaluation Criteria 4

4. 5 Benchmark Algorithms 5

**Chapter 5.** **Results and Discussion 6**

5. 1 Results 6

5. 2 Discussion 6

**Chapter 6.** **Conclusion and Future Work 7**

6. 1 Conclusion 7

6. 2 Future Work 7

**References 8**

**List of Abbreviations**

|  |  |
| --- | --- |
| * AI | Artificial Inteligence |
| * ANN | Artificial Neural Networks |
| * AUC | Area Under the (ROC) Curve |
| * AVG | Average |
| * BMI | Body Mass Index |
| * BRFSS | Behavioral Risk Factor Surveillance System |
| * CDC | Center for Disease Control and Prevention |
| * CPU | Central Processing Unit |
| * DL | Deep Learning |
| * EDA | Exploratory Data Analysis |
| * EHR | Electronic Health Records |
| * FN | False Negative (Incorrect negative cases) |
| * FP | False Positive (Incorrect positive cases) |
| * GDPR | General Data Protection Regulation of the European Union |
| * HIPAA | Health Insurance Portability and Accountability Act of 1996 of the US |
| * ML | Machine Learning |
| * OS | Operating System |
| * RAM | Random Access Memory |
| * RBF | Radial Basis Function |
| * ROC | Receiver Operating Characteristic |
| * ROC-AUC | Receiver Operating Characteristic Area Under the Curve |
| * SML | Supervised Machine Learning |
| * SMOTE | Synthetic Minority Over-Sampling |
| * SVM | Support Vector Machine |
| * T1D | Type 1 Diabetes |
| * T2D | Type 2 Diabetes |
| * TN | True Negative (Correct negative cases) |
| * TP | True Positive (Correct positive cases) |
| * UML | Unsupervised Machine Learning |
| * WHO | World Health Organization |

Table 1. List of abbreviations

**Glossary**

* Accuracy: Metric that measures overall correctness of a ML model (right predictions).
* ANN: Computing models that are inspired by the human brain, typically used for tasks like classification and pattern recognition
* AUC: Metric that indicates a ML model's ability to differentiate between classes. It is derived from the ROC curve.
* BMI: Health metric derived from height and weight to classify the body weight status.
* DL: Subset of ML model’s that utilize multi-layered neural networks to perform complex data tasks.
* F1-Score: Harmonic mean of precision and recall, balancing both metrics in one score.
* ML: Algorithms capable of learning patterns from a given data to make predictions and/or decisions.
* Precision: Metric that indicates how many of the positive predictions were actually correct.
* Recall: Metric that shows how well the model identifies actual positive cases.
* ROC: Plot that shows the performance of a binary classifier derived from the comparison of true and false positive rates.
* ROC-AUC: Sincle score that summarizes the ROC curve where a higher value indicates a better performance.
* SML: ML models that are trained exclusively on labeled data to predict outcomes.
* SMOTE: Method to balance class distribution by generating synthetic minority class samples.
* SVM: Supervised ML model that finds the best boundary to separate classes.
* UML: ML models that are trained exclusively on unlabeled data and have to find patterns without predefined outputs.

**List of Tables DO AT THE END, VERY LAST**

**List of Figures DO AT THE END, VERY LAST**

**Introduction DO AT THE END**

This chapter describes the background of the problem, and the motivation behind selecting the research area carried out in this thesis. It should also provide an overview of the thesis structure.

**Background**

**Problem Statement**

**Research Questions**

**Thesis Objectives**

**Contributions**

**Thesis Organization**

**Theoretical Background**

**2.1 Introduction**

The theoretical background outlines, as its name implies, the theoretical foundation that supporst the present study on the application og machine learning (ML) techniques for early diabetes detection. This research has been based on the principles of data science, particularly for supervised machine learning and classification theory behind them. Through the establishement of a clear understanding of the average data science pipeline, the utilized classification algorithms and the different model evaluation methods, it is possible to provide a conceptual foundantions that are necessary to implement and analyze machine learning models in the context of healthcare and diagnostics.

**2.2 Data Science and ML Theory**

Data science consist in an interdisciplinary field that utilizes and combines scientific methods/approaches, statistics, math, analytics, specialized algorithms and systems to extract information and insights from structured and/or unstructured data sets (Zarbin, Lee, Keane, & Chiang, 2021). Within this context, machine learning (a subfield of artificial intelligence) is focused on enabling systems to learn certain patterns from data (usually datasets) without hardcoding explicitly the program to do these tasks (Badillo et al., 2020).

The data science lifecycle, adapted and expanded from Wing (2019), usually includes the following stages:

* Data collection: Gathering and obtaining structured and/or unstructured data from various sources. Ethical principles are critical, especially when it comes to common people’s private data.
* Pre-processing and data cleaning: Consist on removing inconsistencies, handling missing values and transform data into usable formats that the algorithms expect.
* Feature engineering: From all features that may be obtained from the data, selecting and transforming variables greatly improve the model performance.
* Machine learning model training: Diferent types of algorithms have been used so the implemented ML models are capable to learn from the training set of data.
* Machine learning model evaluation: Subsequently to the training of the ML model, it is necessary to evaluate and ensure that the model’s performance is satisfactory by utilizing specific metrics to validate its effectiveness.
* Interpretation and deployment: Making the output of the model understandable and applicable to real-world decision-making and real case scenarios.



Figure 1. Data science lifecycle

Regarding healthcare applications, this life cycle is crucial to develop accurate and interpretable predictive model that are appropriate to support clinical decisions, especially in the early diagnosis of diseases like diabetes.

**2.3 Supervised Classification Algorithms**

The supervised learning is a category of machine learning algorithms in which the model is trained on a dataset that is organized with input and output pairs. The ML algorithm model then starts learning how to map the input features (or X) to a target label (or Y), so, after being trained, the ML model is capable to predict Y for new, unseen, uncategorized data (unlabeled data). For the present study, the focus is completely on binary classification, where the goal is to predict if the patient is diabetic or not based on the given eatures, the following overview of Supervised classification algorithms has been adapted from Richards (2022).

**2.3.1 Logistic Regression**

Logistic regression models specifically the probability that a given input belogs to a particular class using the sigmoid function:

Equation 1. Logistic Regression Formula

The results are interpretable and the model performs well with linearly separable data.

**2.3.2 Decision Tree**

Decision Trees split the given data based on feature values to form a structure similar to a tree, where each of the nodes represent a decision and each “leaf” is a prediction. They are fairly easy to interpret, handling both classification and regresstion tasks and support numerical and categorical data alike. An imoportant note is that this ML model is very prone to overfitting the data, especially if the given dataset is small or too noisy (presence of outliers).

**2.3.3 Random Forest**

Random Forest is an ensemble learning method that constructs multiple decision trees during training and then it outputs the majority of class (classification) as the final result, capable of handling non-linear relationships well and provides feature importance.

**2.3.4 Support Vector Machines (SVM)**

SVM realizes the optimal hyperplane that separates data into classes bz maximizing the margin between support vectors. Related to non-linear data, the kernel works as a Radial Basis Function (aka RBF) that is used to project the given data into higher dimensions if necessary.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Type | Interpretability | Training time | Handles Non-Linear | Robust to Outliers |
| Logsitic Regression | Linear | High | Fast | No | No |
| Decision Tree | Non-Linear | |  | | --- | |  |  |  | | --- | | Medium-High | | Fast | Yes | No |
| Random forest | Ensemble (Bagging) | Medium | Moderate | Yes | Yes |
| SVM | Non-linear (kernel) | Low-Medium | Slow (for large data) | Yes (kernel) | Yes |

Table 2. ML algorithms comparative table

**2.4 Model Evaluation Metrics**

The evaluation of ML models in healthcare must go beyond just simple accuracy and satisfactory results. Misclassification, especially false negatives, can have severe consequences for the patients in medicals context, therefore several metrics are used. The following overview of model evaluation metrics is adapted from Rainio, Teuho, and Klén (2024):

* Accuracy: Proportion of total correct predictions (True positives and True negatives).

Equation 2. Accuracy formula

* Precision: Proportion of positive identifications (Positive predictions) that were actually correct (True).

Equation 3. Precision formula

* Recall (Sensitivity): Proportion of actual positives correctly identified (True positives).

Equation 4. Recall formula

* F1-Score: Consist in the harmonic mean of precision and recall times 2.

Equation 5. F1-score formula

* ROC-AUC: also known as Receiver Operating Characteristic Area Under the Curve, this metric measures the trade-off observed between true positive rate and the false positive rate. The area under the curve provides an extra/aggregated performance measure for analysis.

The selected metrics offer a set of comprehensive insights into any model’s performance, particularly working with imbalanced datasets that are very common in medical researches.

**2.5 Feature Selection and Importance in Machine Learning**

Whithin the supervised machine learning category (moreover in medical diagnostics), feature selection constitutes a critical step that directly impacts the model performance, interpretability and generalization of unseen data. According to the review by Mahadeo and Dhanalakshmi (2022), Feature selection consist to the process of identifying and separating the most significant input variables (aka features or X) from the dataset because they are the input variables that contribute the most meaningfully to the prediction capabilities of the model when “deciding” the target outcome (aka predictions or Y), for the case if the present study, diabetes.

The reasons why Feature Selection matters (Adapted from Naheed et al., 2020) is:

* Improve model accuracy: Consist in eliminating the irrelevant or redundant features that may just introduce noise and tamper with the accuracy.
* Reduce overfitting: By simplifying the model and ensure the focus is towards the most meaningful features only.
* Training Speed up: Simply lowering the dimensions of the data can increase the speed it takes to train any ML model.
* Enhance interpretatability: It allows to better understand the influence of features and how they correlate to the output, especially in healthcare (indicators like BMI, age, glucoes etc,) it is essential for clinical acceptance.

**2.6 Real-World Challenges for Deployment of ML models in Healthcare**

Even though Machine Learning models often present an elevated accuracy in controlled environments and experiments, their deployment into a real world setting and specifically in a clinical environment presents a great deal of challenges that must be addresssed not only to ensure the model’s effectiveness and ethical use, but also when it comes to legal implications. Some of them include:

* Data privacy and Security: Healthcare data is highly sensitive information and subject to a great deal of regulations such as HIPAA in the United States or GDPR in the European Union, ensuring (non-negotiably) data is anonymous, securely stored and used responsibly (Md Shahin Ali et al. 2024).
* Integration with Clinical workflows: ML models must be integrated in a seamless matter into already existing hospital systems like Electronic Health Records (EHRs) because tool’s used by the medical staff can not disrupt their workflow or add complexity because it risks being ignored or underused (Khattak et al., 2023).
* Possible generalization across populations: ML models trained on specific datasets (for example the BRFSS used in this study) does not mean that the same model will perform as correctly in other populations (since BRFSS is based on the American population) because of biases or differences in dermographics related to the lifestyle of the patients or the available healthcare infrastructure, making external validation essential (Petersen et al., 2022).
* Interpretability and trust: The professionals working at a hospital or research instituttion must be able to easily understand and completely trust the predictions of the ML models, especially because it involves life-altering decisions like a diagnosis. The so called “Black-box” models are commonly avoided unless paired with interpretability tools for this very reason (Petersen et al., 2022).
* Legal and ethical concerns: Given that AI is still a very recent topic and a larguely unregulated one to this date, a lot of questions that concern liability and fairness must be considered. Unfair models can reinforce health disparities and misdiagnoses (Petersen et al., 2022).
* ML model’s maintenance: Healthcare data evolves over time and at an accelerated pace due to changes in clinical practices, diagnostic criteria or population health trends. Any ML model must be preiodically retrained and monitored to remain at an acceptable level of accuracy and relevance (Khattak et al., 2023).

**Literature Review**

**3.1 Purpose of the Literature Review**

The purpose of the literature review is to critically review pre-existing research related to the application of Machine Learning (ML) techniques for the early detection of diabetes. This chapter identifies the strengths, weaknesses and discoveries of previous related studies, with a focus on the usage of supervised ML algorithms for binary classification tasks in healthcare, specifically in the pre-emptive detection of diabetes disease. The review aims to highlight gaps in the current research, particularly in comparative analyses of classification models, model interpretability and real-world integration of ML tools in clinical settings, providing insights that has acted as the foundation for the current study and justify its methodology and research questions.

**3.2 Structure of the Literature Review**

This chapter is organized in a thematical way to provide a coherent analysis of the field. Beginning with an overview of diabetes and the importance of early detection. Then exploring the role of Machine Learning to provide a coherent analysis of the field of medical diagnostics, followed by a detailed examination of various ML models used for diabetes prediction. Then discuss the model performance comparison, interpretability in healthcare ML applications and challenges related to real-world deplolyment of ML tools in live clinical environments. Lastly, a summarize the gaps identified and positioning the current research within the broader context of the existing literature.

**3.3 Key Terms and Concepts**

* Machine Learning (ML): It is a field in the study of artificial intelligence (AI). It is concerned on developing and statistical studying of algorithms capable of learning from data in order to make predictions and/or decisions based on unseen data without requiring giving explicit instructions nor hard-coding for those specific tasks.
* Supervised Machine Learning (SML): It is a specific type of machine learning model that is trained exclusively on a labeled datasets, meaning that, each data point within the dataset is always associated with an output label (pairs); in contrast, Unsupervised Mahcine Learning models (UML) are trained with unlabeled datasets (there are no associated output labels). Therefore, supervised ML models learn how to map the inputs to the correct outputs and make accurate generalizations with unseen data.
* Classification Machine Learning Models: A category within the Supervised Machine Learning models is the Classification ML models, which are used for predicting categorical outcomes. There are 3 possible classification ML models (Binary, Multi-class and Multi-label), for this particular research study the proposed supervised ML models fall into the Binary classifcation, in order to distinguish between diabetic/prediabetic and non-diabetic individuals.
* Interpretability: In Machine Learning, interpretability consists of the extent to which a human can understand the decisions and/or predicitons performed by a trained model. This aspect is critical, especially when it comes to healthcare to ensure transparency, clinical adoption and trust.
* Early detections: This concep consist in the identification of any disease or chronic conditions at an initial stage before significant symptoms appear, which allows for timely intervention and improved outcomes for the patient.

**3.4 Diabetes and the Need for Early Detection**

The disease known as “Diabetes mellitus” is a chronic metabolic disorder characterized by the elevated levels of glucose in blood, resulting from defects in insulin secretion, insulin action or both. The most usual forms include Type 1 diabetes (T1D), Type 2 diabetes (T2D) and gestational diabetes. According to the World Health Organization (WHO, 2021), diabetes is found among the leading causes of death worldwide and often associated with long-term complications such as cardiovascular diseas, kidney failure, blindness and lower-limb amputation.

Type 2 diabetes accounts for about over 90% of all cases, often developing gradually and remain undiagnosed for years due to the fact is larguely asymptomatic in nature (especially in early stages). Unnoticed diagnoses constitues a major concer, given that an early intervention to any disease before clinical symptoms start showing can greatly reduce the risk of future complications and improve the quality of life of the patient, often times involving simple lifestyle changes or medications (American Diabetes Association, 2023).

Traditional screening methods, that are still used to this day, rely on peridodic blood testing on glucose levels, HbA1c levels and patien-reported symptoms. However, these methods are very much reactive and may miss a patient in its early-stage or high-risk individuals, especially in populations with limited acces to healthcare and deficient healthcare infrastructure. Moreover, diagnostic processes are time-consuming and may not even leverage the full potential of patiend data, that should include behavioral, demographic and lifestyle information.

Thus, the early detection is essential for initial timely treatment and also enabling preventative measures and strategies. Studies have shown that individuals diagnosed at early stages are more likely to respond positively to medical interventions, producing a better management of the disease and reducing healthcare costs (Zimmet et al., 2016). However, the ever increasing volume and complexity of health data has outpaced the ability of conventional diagnostic methods to process the aforementioned data efficiently.

Within the context of this research study, the integration of machine learning models in clinical settings offer a primising alternative for the early detection of diabetes (or other diseases). According to a study by Varma & Soni (2020), machine learning algorithms are capable of analyzing large datasets with, often times, a higher accuracy and in a faster matter than traditional statistical methods. Focusing on variables (features) like BMI, age, gender and physical activity among others, ML models were capable of identifying individuals with a high risk of diabetes before clinical symptoms arise and become apparent.

While diabetes as a disease continues to increase, especially in low and middle income countries, there is a clear necessity for a scalable, data-driven diagnostic tools and software solutions of the same nature that becomes increasingly urgent as time goes on. The early detection not only mitigates the adverse effects of the dieseas and improve the patient’s overall health, but it also alleviates the broader burden put on public health systems. Therefore, researches that study and advance on ML-based solutions for diabetes disease prediction contributes significantly to both clinical and global health priorities.

**3.5 Role of Machine Learning in Healthcare**

There are several roles Machine Learning models can perform withih the healthcare industry, some of them consist on:

* Enhance diagnostic accuracy: ML algorithms excel in analyzing complex data, in this case medical data such as imaging and electronic health records to identify patterns indicative of diseases. The capabilites of ML models for detecting early signs of diseases like diabetic retinopathy, cardiovascular diseases and various types of cancer before critical symptoms start showing and with an accuracy comparable to traditional methods and even higher than them (Foresee Medical, 2025).
* Personalize treatment plans: With the analyzis of a patient's medical history and lifestyle factors, a ML model can assist in developing personalized treatment strategies tailored to each patient individually, improving the outcomes and minimizing adverse effects of the patient (Sarkar et al., 2020).
* Predictive analytics for disease prevention: Being capable of processing vast amounts of data to predict the likelihood of disease development on a individual, meaning that this predictive capabilities enable healthcare providers to implement the previous point and perform early interventions (Kelley, 2024).
* Increase operational efficiency and reduce costs: Through the automation and acceleration of processes that traditional methods for diagnostics would occupy more resources and take a longer time, it is possible to reduce the operational costs, minimize human error and allocate resources more efficiently for healthcare institutions and patients alike (Kelley, 2024).
* Advacements in drugs: With the theoretically superior analysis capabilities, machine learning models can potentially also accelerate the discovery of drugs or the improvement of already existing ones by analyzing biological data to identify potential therapeutic targets and predict the efficacy of drug compounds (Kelley, 2024).

**3.6 Comparative Studies of ML Models for Diabetes Prediction----\*\*\*\***

As previously stated, the application of ML models in the prediction of diseases has gained a significant attention in recent years. Numerous studies have compared the performance of different ML algorithms on datasets, including those related to diabetes aiming to identify the most accurate and reliable model. These studies usually focus on binary classification tasks, where the objective is to distinguish between diabetic and non-diabetic based on various features related to the patient’s health features.

**3.6.1 Commonly Used Machine Learning Models**

A variety of machine learning algorithms have been explored for diabetes prediction, including but not limited to:

* Logistic regression (LR): Simple yet widely used algorithm for binary classification tasks. It works exceptionally well with linearly separable data and provides interpretable results, making it potentially suitable for clinical settings where transparency is critical (Zhang, 2025).
* Decision tree (DT): A tree-based algorithm model that splits the data into numerous branches based on the feature values, leading to a decisioon at each “leaf” or node, which is the prediciton. Its intuitive, visual structure makes it one of the easiest to interpret, which is beneficial in clinical applications, but its prone to overfitting with noisy data (Abedini, 2020).
* Random forest (RF): An ensemble learning method that is composed of multiple decision trees and then aggregates their results. This model often perform well thanks to its ability to handle large datasets gracefully and also because of its robustness to overfitting, making it an appealing choice for diabetes prediction (Zhang, 2025).
* Support vector machines (SVM): Known for its capability to handle complex, highl-dimenional data, this algorithm has been shown to perform well in diabetes prediction tasks, particularly when using the kernels to transform non-linearnly separable data into higher-dimensional spaces (Zhang, 2025).
* Neural networks (ANN): A more complex model than traditional classifiers since this one is a Deep learning model. Despite its complexity, it has been applied to diabetes prediction tasks and yielding high performance, but its “black-box” nature poses a great challenge in terms of interpretability and understandibiltiy (Abedini, 2020).
* Ensemble methods (various): Techniques like Gradient Boosting and Voting Classifiers combine the outputs of several models to enhance the prediction accuracy, often outperforming single classifiers by reduing biases and variances (Mushtaq et al., 2025).

**3.6.2 Performance Comparison Across Studies**

Many academical studies have compared and implemented the models located in subsection 3.6.1 for the diabetes prediction and even other diseases, finding interesting variations in performance based on different aspects like dataset characteristics, feature selection and evaluation metrics. Therefore, the following consist in different key findings from comparative studies:

Zhang (2025) evaluated the models of Logistic regression, SVM, Random fores and XGBoost on the Prima Indians diabetes dataset (dataset initially considered for this thesis). Among the aforementioned models, XGBoost achieved the highest accuracy (85%) and ROC-AUC (91%), having BMI, glucose and age as key features. It is worth noting that, the Random forest model also performed well but due to its lower interpretability compared to simpler models (albeit less accurate) like the Logisitic regression, the Random Forest approach might limit a seamless clinical implementation.

Li et al. (2021) performed a similar study using the BRFSS 2015 dataset (dataset selected for this study). Their results showed that Neuronal networks provided a superior accuracy performance but they were more prone to overfitting. Meanwhile, SVM models achieved a somewhat healthy balance between accuracy and and interpretability.

On the other hand, Husain and Khan (2018) applied several ML models to the NHANES 2013-2014 datset and developed an ensemble model using a majority voting technique. Utilizing a dataset with 10,172 samples and 54 features, the ensemble model proved to be the overall best predictor in terms of accuracy performance by achieving an ROC-AUC of 0.75. These findings highlight the potential of ensemble learning to enhance diabetes prediction at early stages.

**3.6.3 Key Insights from Comparative Studies**

* Accuracy vs Interpretability: While more advanced AI models, like DL neural networks or Random forest, are able to achieve high predictive accuracy in a healthcare context, their lack of transparency usually limits a practical use in a clinical setting. According to a study by Tonekaboni et al. (2019), a review and a survey of clinical staff, the authors highlighted that the interpretability of a model played a highly important role in building trust and implementation, where medical staff showed a preference for less complex models like Logistic regression and support for the SVM model due to their more explainable and clear nature for the outputs and results, even if the trade off meant a slight reduction in accuracy. This proves the need for explainable AI techniques that individuals with no technical education on the matter can still understand the results.
* Class imbalance handling: Many studies like Salmi et al. (2024) focus on addressing class imbalance since in medicine and real life scenarios, it is impossible to get balanced data and datasets. Therefore, techniques like SMOTE (Synthetic Minority Over-Sampling) or class weighting can mitigate the effects of any imbalances and avoid bias predictions towards majority class
* Feature selection: The importance of selecting the most significant relevant features from the dataset is very important, given that unnecesary features will only clutter the model’s training and performance.

**3.7 Interpretability and Model Transparency in Healthcare**

The transparecy and interpretability of ML models are paramount for fostering trust, ensure accountability and facilitate adoption by healthcare professionals in a medical setting like hospital or institutions. As previously stated, more complex models like DL networks are more accurate but their complexity hinders their preference by clinical professionals.

A study published by Luo et al. (2024) discussed the diferent trade-offs between different of machine learning models, specifically in terms of interpretability and accuracy. The aforementioned study supports the claim that, while “white-box” models potentially output a lower accuracy, their transparency and understandability garner more trust among users despite “black-box” models yielding a better performance in accuracy.

It becomes clear that, while accuracy it is an unexpendable aspect of ML models to predict the likeness of a disease like diabetes, it is also important to balance it with a clear interpretability to ensure that AI systems are both effective and trustworthy in a healthcaer environment.

**3.8 Identified Gaps in the Literature**

Despite the considerable amount of research dedicated to diabetes and other diseases prediction using machine learning (ML) several key gaps remain in the literature review. These gaps highlight areas where current studies may be insufficient and pinpoint where future studies are needed to further enhance the effectiveness of machine learning models in healthcare applications.

First and foremost, there is a clear lack of consensus on the “best” model for Diabetes prediction (or other diseases for that matter). A very recurring challenge is the lack of a definitive answer in this regard. While promising models reange from simpler models like Logistic regression all the way to more complex like a Deep learning models, the variation in results across studies suggest that the most optimal approach may depend on various different factors (dataset, feature selection and evaluation criteria) but the absence of a universal standard consensus makes it almost impossible to adopt a specific machine learning system in the clinical field, not to mention that certain models excel in specific scenarios, but a comprehensive, context-aware framework approach when selecting a model is still lacking a strong foundation to support the claim it is “the best”.

Following the previos point, a lot of research between machine learning and the medical field seem to prioritize predictive accuracy over interpretability or usability in real-life scenarios, which are not optional for clinical adoption since doctors need the most reliable but also understandable results. A model can not just perform well, but also be clear and easy to understand in their context and even to patient that most of the times have no real education in the medical field nor software engineering field. Unfortunately, it seems like a clear trade-off is that high-performing models (ANN or Ensemble methods) suffer from a “black-box” nature that makes them difficult for practitioners to intrepret and trust while “white-box” models (Logistic regression or Decision tree) are easy to interpret but lack the accuracy of their “black-box” counterpoarts. Studies that combine predictive performance and model transparency are limited but essential to close this gap.

Another aspect, particularly inevitable due to the nature of the medical field, is the class imbalance of data and datasets. In real-world healthcare datasets, the distribution of data is most of the time skewed, where, for example, there are more individuals diagnosed with diabetes compared to those who are not and as previously revised, class imbalance can lead to biased predictions where the model overestimates the likelihood of the majority class. There are techniques to mitigate this effect (SMOTE or Class weighting), they are not capable of consistetly apply them across all studies, leading to less reliable results in some cases.

Another significant gap is related to the limited number of reporductible and real-world-valid studies. Numerous studies for different ML models have proposed, tested and have come to conclusions by using well known datasets of the likes of Pima Indians diabetes or NHANES, but this comes at the cost of the lack of validation of using more diverse, real-world data. Furthermore, many sutides do not provided enough details on the model training, testing or hyperparemeter tuning their models undergo, difficulting the capacity to replicate or validate their reulst. Real-world in the healthcare field is crucial not only because it is necessary to ensure that models generalize well outside controlled environments and can be deployed effectively in clinical practice, but also because countless patient’s can be negatively affected by incorrect predictions (False Positives and False Negatives) that can even lead to their deaths.

**Methodology**

**4.1 Research Design**

This study employs a quantitative, experimental research design in nature to evaluate and compate the effectiveness of several supervised machine learning classification models in predicting diabetes. The methodology of the study is structured around the typocal data science lifecycle, encompassing data acquisition, preprocessing, model development, model testing, model evaluation and results interpretation. A comparative approach is adopted in the analysis to determine the relative performance of selected algorithms using a consistent dataset with all of them and a consistent evaluation framework (with specific evaluation metrics).

**4.2 Dataset and Data Collection**

The chosen data set for the present study, comes from the Diabetes Binary Health Indicators BRFSS2015 dataset, which is publicly available on Kaggle and originally prepared by the user Alex Teboul (2021). The aforementioned dataset is actually derived from the Behavrioal Risk Factor Surveillance System (BFRSS) made in 2015, an annual health-related telephone survey conducted by the Center for Disease Control and Prevention (CDC). In total, the dataset contains the responses of 253,680 individuals, with 21 predictor variables and a binary target variable (Diabetes\_binary). In the description of the dataset, it is stated that the dataset is already pre-cleaned, however, this dataset still has undergone additional preprocessing steps to ensure suitability for ML tasks. The 21 predictor variables and the binary target variable are as follows:

|  |  |  |
| --- | --- | --- |
| Variables | Description | Values / Encoding |
| Diabetes\_binary | Diabetes status | 0 = No diabetes, 1 = Prediabetes or Diabetes |
| HighBP | High blood pressure | 0 = No, 1 = Yes |
| HighChol | High cholesterol | 0 = No, 1 = Yes |
| CholCheck | Cholesterol checked in past 5 years | 0 = No, 1 = Yes |
| BMI | Body Mass Index | Continuous numerical value |
| Smoker | Smoked at leas 100 cigarettes in their lifetime (or 5 packs) | 0 = No, 1 = Yes |
| Stroke | Ever had a stroke? | 0 = No, 1 = Yes |
| HeartDiseaseorAttack | History of heart disease or attack (CHD or MI) | 0 = No, 1 = Yes |
| PhysActivity | Physical activity in past 30 days | 0 = No, 1 = Yes |
| Fruits | Fruit consumption | 0 = No, 1 = Yes |
| Veggies | Vegetable consumption | 0 = No, 1 = Yes |
| HvyAlcoholConsump | Adult men >=14 drinks a week. Adult women>=7 drinks a week | 0 = No, 1 = Yes |
| AnyHealthcare | Has any kind of healthcare coverage | 0 = No, 1 = Yes |
| NoDocbcCost | Could not see doctor due to cost in the last 12 years | 0 = No, 1 = Yes |
| GenHlth | General health status | 1 = Excellent to 5 = Poor |
| MentHlth | Days mental health not good | 0–30 days (0 for days with bad MentHlth and 30 for all days with bad MentHlth) |
| PhysHlth | Days physical health not good | 0–30 days (0 for days with bad PhysHlth and 30 for all days with bad PhysHlth) |
| DiffWalk | Serious difficulty walking or climbing stairs | 0 = No, 1 = Yes |
| Sex | Gender | 0 = Female, 1 = Male |
| Age | 13-level age category (\_AGEG5YR see codebook) | Ordinal scale 1 to 13 |
| Education | Education level (EDUCA see codebook) | Ordinal scale 1 to 6 |
| Income | Income scale (INCOME2 see codebook) | Ordinal scale 1 to 8. |

Table 3. Data set 22 columns

The dataset is very appropriate for predictive modelling due to its large size, dirversified features and real-world relevance and origin. The survey responses capture key behavioral and dermographic factors that contribute to diabetes risk. It is also worth noting that the dataset is imbalanced, where the majority of responses are labeled as non-diabetic, which might seem as a drawback at first glance, however in real-life scenarios the data will most likely be imbalanced, therefore an ML model must be able to tackle this issue.

According to the CDC (2024), 1 in 5 diabetics and almost 8 in 10 prediabetics are unaware of their condition until clinical symptoms become apparent, therefore, this dataset provides a vital foundation for developing appropriate predictive models, enabling early diagnosis and reducing overall healthcare burden.

**4.3 Data Preprocessing**

Despite the dataset being “cleaned”, it is imperative to still undergo common preprocessing steps to optimize each model performance, given that the dataset consist of an imbalanced type with 21 features and a single binary target variable where all data entries seem to be floating point numbers:

1. Handle possible missing values: Inspect dataset to look for any remaining null or invalid entires and apply **imputation (median, mean, mode or regression imputation)** or **dropping** entries as required for better results.
2. Remove duplicate values: **Drop** all duplicate values given that they skew training and lead to overfitting.
3. Encoding Ordinal features: Identify and encode ordinal categorcal variables, therefore, it is possible to ensure that the encoded values still preserve a correct logical order. It is worth noting that this step was never implemented since dataset already has encoded data in all of its columns
4. Data splitting: **Split** data into training and testing sets, also **stratify** to tackle imbalance between the sets.
5. Check and handle class imbalance in training set: Utilize **SMOTE** to oversmaple the minority class with synthetic samples or, alternatively, utilize **Class weighting** for models that support it.
6. Feature scaling: On the training set, **standardize** continuous features and **Robust scaling** (if necessary) for outliers but ensure not to scale binary or already-encoded categorical features.
7. Data shuffling: **Shuffle** training set to ensure it is as random as possible and leave test set as it is to simulate real-world scenarios as close as possible.

**4.4 Machine Learning Algorithms**

Four Supervised classification algorithms have been selected for implementation and evaluation for this study (each subject to further refinement based on a exploratory analysis approach):

* Logistic Regression or LR
* Decision Tree Classifier
* Random Forest Classifier
* Support Vector Machines (SVM)

Each of the aforementioned models have been chosen for their interpretability, performance records and diversified algorithm approaches, whichh constitutes linear, tree-based, ensemble and kernel-based.

**4.5 Model Training**

In order to train and validate each machine learning model and ensure a robust model performanc, the following startegies have been followed:

* Train-test split: Utilize between 80/20 and 70/30 split ratio to divide the dataset into training and testing sets.
* Cross-validation: K-fold cross-validation (k=5 or 10) have been utilized to reduce the possibility of overfitting the models and ensure generalization across samples.

**4.6 Evaluation Metrics for validation**

To assess and evaluate each classification model comprehensively and fairly, the following metrics have been used:

* Accuracy = (TP + TN) / (TP + TN + FP + FN)
* Precision = TP / (TP + FP)
* Recall = TP / (TP + FN)
* F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)
* ROC-AUC (Area Under the Receiver Operating Characteristic Curve)
* Confusion Matrix (to analyze true positives, true negatives, false positives and false negatives).

The listed metrics have provided an accurate and rounded view on each of the model’s effectiveness, especially in the context of medical diagnosis and the aforementioned class imbalance the dataset is stated to have. Also, these metrics also served as a clear comparison point between each model.

**4.7 Feature Importance and Model Interpretability**

To evaluate how the different features in the dataset have contributed to the prediction and facilitaed clinical decision making:

* Feature importance scores have been analyzed (for the likes of tree-based models).

**4.8 Tools and Technologies**

The implementation of has been carried out using programming language on VSCode code-editor (often referred as an IDE) using the following libraries:

* Scikit-learn: For Machine Learning implementation and evaluation.
* Pandas and NumPy: For data manipulation and preprocessing from the dataset and analysis.
* Matplotlib and Seaborn: For data visualization (plotting and graphing) and results presentation in a comprehensible matter.

**4.9 Ethical Considerations**

The present research study uses a publicly available data that contains no personal information that can lead to identifying any of the subjects that participated in the BFRSS dataset by the CDC. Therefore, the usage of this dataset poses minimal ethical risk, however, ethival diligence is still maintained in the following ways:

* ML model fairness: The perofrmance of each Supervised classification model across different demopraphic groups (for example sex ir age) will be analyzed to detect and minimize biases.
* Resposible use of AI technologies: Each insight found within this study was contextualized within the clinical realities and contexts, supporting the role of Machine learning models as a tool to help in the diagnosis of diabetes disease and not a replacement of the expert judgement of the human professionals in the field.

**4.10 Summary**

The present chapter outlines the methodology that was adopted during this research study, from data acquisition to the ML model evaluation. By applying and comparing multiple Supervised machine learning classification models on a large scale health dataset, aiming to identify optimal strategies to detect diabetes and/or prediabetes in its early stages. There is a special emphasis on the interpretability, performance and clinical effectiveness of the models to ensure practical contributions to the healthcare domains.

**Results and Discussion**

**5.1 Hardware and Software specifications**

The performance and efficiency of the training upon on a ML model can be greatly increased by the hardware and software envirmont in which the script is executed and the research conducted. Topics like processing speed, available memory and system architecture can heavily influence how quickly and effectively a model can be trained (Scikit-learn Developers, 2024), especially on models that are very intensive like SVMs. During this research, all experiments where conducted on a machine with a system running Windows 11 OS and the following hardware specifications:

* CPU: Intel(R) Core(TM) i7-1255U 12th Gen processor (3.52 GHz)
* RAM: 16GB from 2 SK Hynix DDR4-3200 (8GB each)
* GPU: Intel(R) Iris(R) Xe integrated graphics
* Storage: 1TB SSD

Even though the system provides a somewhat adequate performance for datasets of moderate size, it is important to take into account that certain, more complex models and/or larger datasets might require an upgrade from these specifications because if not, it would result in longer training times or convergence issues due to lack of GPU acceleration or not even be able to meet the computational demands and overrunning the available RAM. Consequently, for this and future studies, these specifications must be considered when analyzing the results or attempts to replicate this work on different machines or environments.

**5.2 Full scripts workflows**

The following constitutes the general workflow each script of the implemented ML models implement:

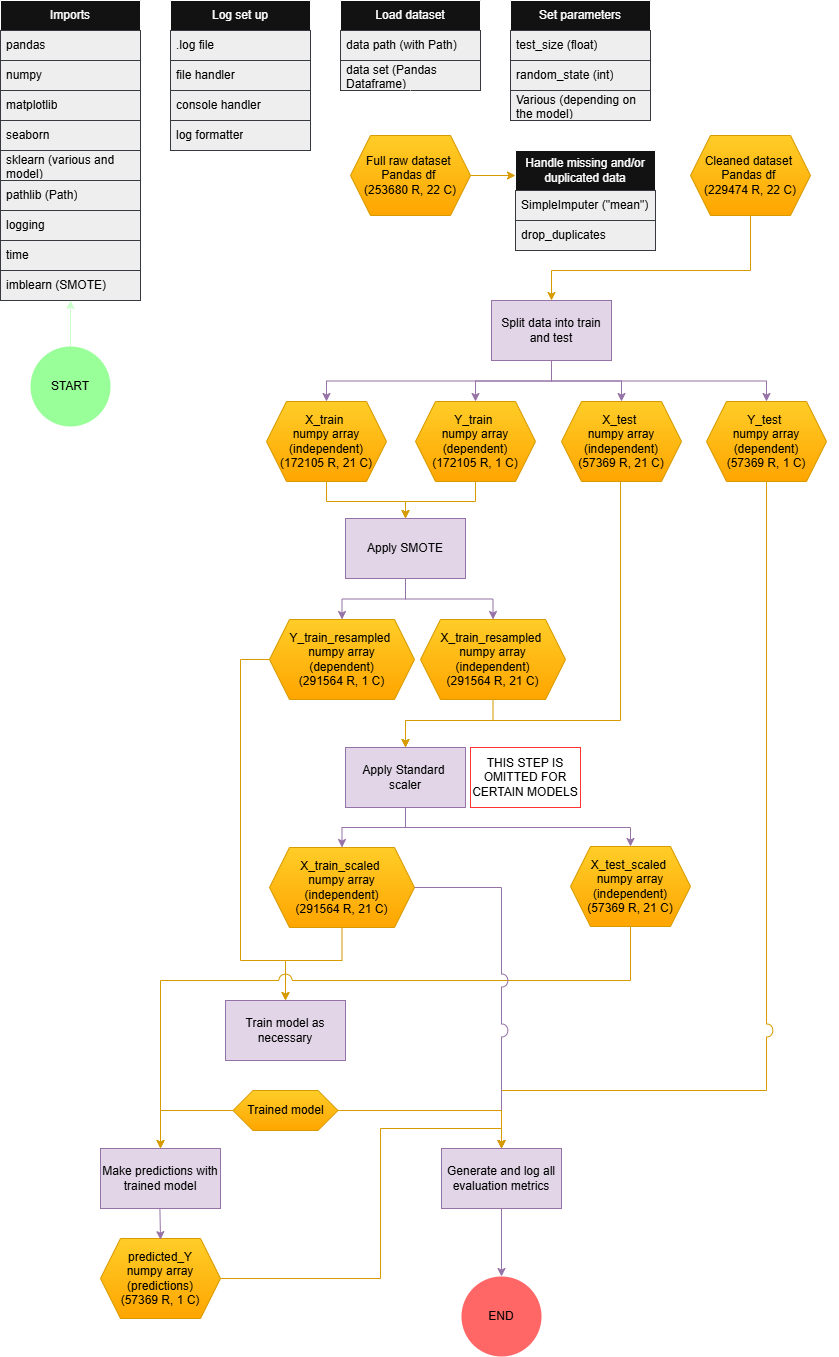


Figure 2. General workflow of ML model scripts

**5.3 Data set analysis EDA (Exploratory Data Analysis)**

The following plots and idk represent the analysis of the raw data provided by the Control and Prevention (CDC) and their Behavrioal Risk Factor Surveillance System (BFRSS) made in 2015 and prepared and published by the user Alex Teboul (2021) in Kaggle:

* **Data set overview**

|  |  |
| --- | --- |
| Number of columns | Number of rows |
| 22 (21 features and 1 objective variable) | 253680 (total number of entries) |

Table 4. Number of columns and rows of raw data set

* **Target variable distribution**

A graph of a number of blue squares

AI-generated content may be incorrect.

Figure 3. Target variable distribution

As seen in this plot, it is possible to visualize a clear class imbalance in which the majority classification is 0 or no diabetes (218334 out of 253680), whereas 1 or diabetes/prediabetes represent a clear minority in the dataset (35346 out of 253680). This ration constitues ≈86.07 of no diabetes while ≈ 13.93% of diabetes and/or prediabetes.

* **Histograms of all independent variables**

A screenshot of a graph

AI-generated content may be incorrect.

Figure 4. Histogram of all independent variables

There are several key takeaways from these series of histograms. Firstly there is a clear imbalance a significant imbalnace for binary health conditions like stroke or heart diseases where the majority of reponders claimed to not have these conditions (0 or no being the dominant values). Then, there is a high proportion of individuals have undergone cholesterols checks and enganged in physical activities, which suggests a baseline level of health awareness and preventive behavior among population. When it comes to self-reported health metrics, such ass mental and physical health, shows that while most reponders report low frequencies, there is a long tail of individuals that experience extended periods of health challenges. Lastly, on a social level, the sociodemographic distribution like education, income and age are skewed toward higher levels, indicating that the selected dataset may over represent older individuals with a higher socio-economical status.

* **Summary statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Diabetes\_binary | HighBP | HighChol | CholCheck | BMI | Smoker |
| Count | 253680 | | | | |  |
| Mean | 0.139333 | 0.429001 | 0.424121 | 0.962670 | 28.382364 | 0.443169 |
| STD | 0.346294 | 0.494934 | 0.494210 | 0.189571 | 6.608694 | 0.496761 |
| Min | 0 | 0 | 0 | 0 | 12 | 0 |
| 25% | 0 | 0 | 0 | 1 | 24 | 0 |
| 50% | 0 | 0 | 0 | 1 | 27 | 0 |
| 75% | 0 | 1 | 1 | 1 | 31 | 1 |
| Max | 1 | 1 | 1 | 1 | 98 | 1 |

Table 5. Summary statistics part 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Stroke | HeartDiseaseorAttack | PhysActivity | Fruits | Veggie | DiffWalk |
| Count | 253680 | | | | | |
| Mean | 0.040571 | 0.094186 | 0.756544 | 0.634256 | 0.811420 | 0.168224 |
| STD | 0.197294 | 0.292087 | 0.429169 | 0.481639 | 0.391175 | 0.374066 |
| Min | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 0 | 0 | 1 | 0 | 1 | 0 |
| 50% | 0 | 0 | 1 | 1 | 1 | 0 |
| 75% | 0 | 0 | 1 | 1 | 1 | 0 |
| Max | 1 | 1 | 1 | 1 | 1 | 1 |

Table 6. Summary of statistics part 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | HvyAlcoholConsump | AnyHealthcare | NoDocbcCost | GenHlth | MentHlth |
| Count | 253680 | | | | |
| Mean | 0.056197 | 0.951053 | 0.084177 | 2.511392 | 3.184772 |
| STD | 0.230302 | 0.215759 | 0.277654 | 1.068477 | 7.412847 |
| Min | 0 | 0 | 0 | 1 | 0 |
| 25% | 0 | 1 | 0 | 2 | 0 |
| 50% | 0 | 1 | 0 | 2 | 0 |
| 75% | 0 | 1 | 0 | 3 | 2 |
| Max | 1 | 1 | 1 | 5 | 30 |

Table 7. Summary of statistics part 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sex | Age | Education | Income |
| Count | 253680 | | | |
| Mean | 0.440342 | 8.032119 | 5.050434 | 6.053875 |
| STD | 0.496429 | 3.054220 | 0.985774 | 2.071148 |
| Min | 0 | 1 | 1 | 1 |
| 25% | 0 | 6 | 4 | 5 |
| 50% | 0 | 8 | 5 | 7 |
| 75% | 1 | 10 | 6 | 8 |
| Max | 1 | 13 | 6 | 8 |

Table 8. Summary of statistics part 4

These summaries provide a clearer insight into the distribution of each column, supporting the previous histograms plots and the key takeaways.

* **Missing values and duplicates**

No missing values were found, however, there were 24206 duplicated rows found that must be dealt with to avoid artificially inflating the importance of certain patterns or features.

* **Independet variables correlation with dependent variable**

|  |  |
| --- | --- |
| Feature | Correlation with Diabetes\_binary |
| GenHlth | 0.293569 |
| HighBP | 0.263129 |
| DiffWalk | 0.218344 |
| BMI | 0.216843 |
| HighChol | 0.200276 |
| Age | 0.177442 |
| HeartDiseaseorAttack | 0.177282 |
| PhysHlth | 0.171337 |
| Stroke | 0.105816 |
| MentHlth | 0.069315 |
| CholCheck | 0.064761 |
| Smoker | 0.060789 |
| NoDocbcCost | 0.031433 |
| Sex | 0.031430 |
| AnyHealthcare | 0.016255 |
| Fruits | -0.040779 |
| Veggies | -0.056584 |
| HvyAlcoholConsump | -0.057056 |
| PhysActivity | -0.118133 |
| Education | -0.124456 |
| Income | -0.163919 |

Table 9. Independent variables correlation to dependent variable

**A graph of a bar graph

AI-generated content may be incorrect.**

Figure 5. Independent variables correlation to Diabetes\_binary

To identify the factors the most associated to diabetes disease in the selected dataset, Pearson correlation coefficients have been computed between all numerical and binary features of the dataset and the target variable Diabetes\_binary.

The correlation between independent variables and the objective variable oscillates between -1 and 1, where -1 constitutes to the least correlated feature and 1 the most correlated feature. In this context we can see that the General health is the most correlated feature to Diabetes\_binary while Income is the least correlated feature. However, it is important to to note that these are correlations and not causal relationships, while they might align to existing medical knowledge, further modelling and/or controlled studies are required to establish a true causation.

* **Correlation matrix heatmap**

**A graph with numbers and letters

AI-generated content may be incorrect.**

Figure 6. Correlation matrix between all columns of dataset

The present correlation matrix illustrates how each individual independent feature relate to diabetes and/or prediabetes presence. It is notable that, poor general health (GenHlth), high blood pressure (HighBP) and mobility issues (DiffWalk) show the strongest positive correlations to diabetes or prediabetes that can confirm widely accepted medical risk factors. On a different note, features like higher income, better education and increased physical activity are weakly negatively correlated, suggesting that the aforementioned may be protective factors. Once again these correlations are not causal and should be interpreted in the context of a broader multi-variables analyses.

* **Countplot for categorical and binary features (Correlation, not causal)**

A graph of diabetes distribution

AI-generated content may be incorrect.

Figure 7. High blood pressure distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often suffer from high blood pressure, whereas people that do not suffer from diabetes often times do not suffer from high blood pressure either.

A graph of diabetes

AI-generated content may be incorrect.

Figure 8. High cholesterol distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often suffer from high cholesterol, whereas people that do not suffer from diabetes often times do not suffer from high cholesterol either.

A graph of smoker distribution by diabetes

AI-generated content may be incorrect.

Figure 9. Smoker distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes (sligthly) more often are smokers, whereas people that do not suffer from diabetes often times are not smokers, although in this case, the distribution is more even than with cholesterol and blood pressure.

A graph of stroke distribution

AI-generated content may be incorrect.

Figure 10. Stroke distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often have not suffered a stroke at least once in their life, also, people that do not suffer from diabetes have not suffered from a stroke at least once in their life either.

A graph of a number of people with diabetes

AI-generated content may be incorrect.

Figure 11. Heart disease or attack distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often have not suffered a heart attack nor do they possess a heart disease, also, people that do not suffer from diabetes havenot suffered from a heart attack nor do they possess a heart disease either.

A graph of diabetes status

AI-generated content may be incorrect.

Figure 12. Physical activity distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often practice some sort of physical activity, also, people that do not suffer from diabetes practice some sort of physical activity as well.

A graph of fruit distribution

AI-generated content may be incorrect.

Figure 13. Fruit consumption distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also consume fruits on a regular basis, also, people that do not suffer from diabetes consume fruits on a regular basis as well.

A graph of a number of veggies distribution

AI-generated content may be incorrect.

Figure 14. Vegetables consumption by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also consume vegetables on a regular basis, also, people that do not suffer from diabetes consume vegetables on a regular basis. In this case, the implications have a higher disparity compared to fruit consumption as well.

A graph of diabetes

AI-generated content may be incorrect.

Figure 15. Heavy alcohol consumption distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also do not heavily consume alcohol, also, people that do not suffer from diabetes do not heavily consume alcohol either.

A graph of diabetes status

AI-generated content may be incorrect.

Figure 16. Any healthcare distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also have some sort of healthcare in their lifes, also, people that do not suffer from diabetes still have some sort of healthcare in their lifes as well.

A graph of a number of bars

AI-generated content may be incorrect.

Figure 17. No visit to the Doctor due to cost distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often were unable to go for a medical visit due to its cost, also, people that do not suffer from diabetes were unable to go for a medical visit due to its cost as well.

A graph of blue and orange bars

AI-generated content may be incorrect.

Figure 18. General health distribution by Diabetes presence

This plot ranges from 1=Excellent general health to 5=Poor health. It implies that people that suffer the most from diabetes or prediabetes have a poorer general health, also, people that do not suffer the most from diabetes have a general excellent general health.

A graph of a diabetes distribution

AI-generated content may be incorrect.

Figure 19. Difficulty walking distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often do not have difficulties walking or going up stairs, also, people that do not suffer from diabetes do not have difficulties walking or going up stairs as well.

A graph of a number of people

AI-generated content may be incorrect.

Figure 20. Sex distribution by Diabetes presence

This plot maps 0 = Female and 1 = Male. It implies that people that suffer from diabete or prediabetes more often are Females, also, people that do not suffer from diabetes are Female as well. Although when it comes to suffering from diabetes or prediabetes, the distribution is very even in comparison to the non-diabetes distribution.

A graph of age distribution

AI-generated content may be incorrect.

Figure 21. Age distribution by Diabetes presence

This plot has 13 levels of age ranges in which 1 = 18–24y, 2 = 25–29y, 3 = 30–34y, 4 = 35–39y, 5 = 40–44y, 6 = 45–49y, 7 = 50–54y, 8 = 55–59y, 9 = 60–64y, 10 = 65–69y, 11 = 70–74y, 12 = 75–79y and 13 = 80 or older. It implies that people that suffer from diabetes or prediabetes more often are between 55 to 70 years old, also, people that do not suffer from diabetes are between 45 and 70 years old.

**A graph of diabetes status

AI-generated content may be incorrect.**

Figure 22. Education distribution by Diabetes presence

This plot has 6 levels of educations and they represent 1 = Never attended, 2 = Grades 1 to 8, 3 = Grades 9 to 11, 4 = Grade 12 or GED, 5 = College or Technical school (no degree) and 6 = College graduate (4-year degree or higher). It implies that people suffer from diabetes or prediabetes also have at least graduated High School, also, people that do not suffer from diabetes have at least graduated High School as well.

A graph of a number of people with diabetes

AI-generated content may be incorrect.

Figure 23. Income distribution by Diabetes presence

This plot has 8 levels of educations and they represent 1 = less than 10k $, 2 = 10k to 15k $, 3 = 15k to 20k $, 4 = 20k to 25k $, 5 = 25k to 35k $, 6 = 35k to 50k $, 7 = 50k to 75k $ and 8 = 75k $ or more. It implies that people that suffer from diabetes or prediabetes more often have a higher income, also, people that do not suffer from diabetes have a higher income as well, in a more notorious way.

* **KDE for continuous features (Correlation, not causal)**

A graph of a number of bmi

AI-generated content may be incorrect.

Figure 24. BMI distribution by Diabetes presence

In this plot, the values usually oscillate between 10 and 60 derived from the patient’s weight in kg divided by the patient’s height in meters to the power of 2, where the lower the value the more underweight and the bigger the value the more overweight. For reference, anything below 16 is consideted severely underweight (malnourished) and anything equal or above 40 is considered morbidly obese. It implies that the people that suffer the most from diabetes or prediabetes range between a BMI of 20 and 40, same when it comes to people that do not suffer from diabetes.

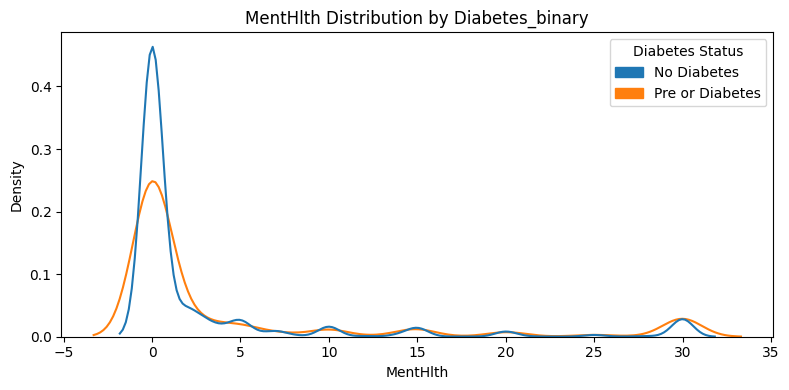


Figure 25. Days with bad Mental health by Diabetes presence

**A graph of a number of people

AI-generated content may be incorrect.**

Figure 26. Days with bad Physical health by Diabetes presence

In the 2 present plots, they range from 0 to 30, representing the number of days in which the responder considers their mental and/or physical health has not been good, where 0 means that there was no days with negative health and 30 to represent all days with negative healt. Both plots imply the the people that suffer the most diabetes or prediabetes did not have a single day of bad mental or physicial health, same with the people that do not suffer from diabtes. However, it is worth noting that does that claimed to have 30 days of bad mental or physical health present a surge in which the people suffer from diabetes or prediabetes more often than not.

**5.4 Logistic Regression Model Results**

The Logistic Regression Model constitutes the simplest among the selected ML models for this research (Not to say it is simple as a concept). The tuning of the aforementioned model is as follows:

|  |  |  |
| --- | --- | --- |
| Parameter | Value | Description |
| class\_weight | “balanced” | Adjust the weights inverse to the class frequencies to handle imbalance |
| test\_size | 0.25 | When splitting the independent and dependent variables, 25% is testing and 75% is for training |
| max\_iter | 1000 | Maximum number of optimization iterations before forcing stop |
| random\_state | 42 | Selected seed for random number generation to ensure reproductibility |
| solver | lbfgs | Optimization algorithm (default) |
| penalty | l2 | Regularization technique to reduce overfitting (default) |
| c | 1 | Inverse of regularization strength (default) |

Table 10. Logistic Regression model parameters for tuning

Regarding the data preparation before training the model is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Step (In order) | Description | Method/Tool used | Result/Notes |
| Missing Value Handling | Checked any null values | SimpleImputer (strategy='mean') | No missing values detected |
| Duplicate Removal | Checked any duplicated rows | drop\_duplicates() | 24206 duplciates removed |
| Train/Test Splitting | Split dataset into training and testing subsets | train\_test\_split (test\_size=0.25) | 75% training, 25% testing; stratified |
| Class Imbalance Handling | Addressed imbalance in target variable | SMOTE (random\_state=42) | Class ratio balanced. 0 and 1: 145782 |
| Feature Scaling | Standardized features for uniform scale | StandardScaler() | Applied to both training and test sets using .fit\_transform() and .transform() |
| Data shuffling | Randomly shuffled training data | shuffle() (random\_state=42) | Shuffled training set after resampling and scaling |

Table 11. Data preparation for Logistic Regression model training

These parameters and preparation steps for the data before training the Logistic Regression wielded the best results and performance for the model. The results of the model are as follows (tables and figures):

|  |  |
| --- | --- |
| Metric | Value |
| Training time (s) | 0.2 to 0.3s on average |
| Acurracy (0-1) | 0.7154 |
| Precision (0-1) | 0.3186 |
| Recall (0-1) | 0.7560 |
| F1-Score (0-1) | 0.4483 |
| ROC-AUC (0-1) | 0.8076 |
| AVG Cross-Validation F1 (0-1) | 0.7504 ± 0.0017 (std deviation) |
| 5-Fold Cross-Validation F1 (0-1) | [0.7478, 0.7497, 0.7504, 0.7521, 0.7519] |

Table 12. Logistic Regression model evaluation Metric results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification report | | | | |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No diabetes) | 0.94 | 0.71 | 0.81 | 48595 |
| 1 (Diabetes or Prediabetes) | 0.32 | 0.76 | 0.45 | 8774 |
| Accuracy |  |  | 0.72 | 57369 |
| Macro avg | 0.63 | 0.73 | 0.63 | 57369 |
| Weighted avg | 0.85 | 0.72 | 0.75 | 57369 |

Table 13. Logistic Regression model Classfication Report

**A blue squares with black numbers

AI-generated content may be incorrect.**

Figure 27. Logistic Regression model Confusion Matrix

The evaluation metrics prove that the Logistic Regression ML model perfroms well when it comes to separating the classes, especially considering the class imbalance confimerd in the EDA. The ROC-AUC of 0.81 implies that the model is good at ranking the positive and negative cases (predictions). However, the precision is 0.32, which is low but common when recall its prioritized via techniques like SMOTE and class\_weight balancing. The precision is high for class 0 (no diabetes) but low for class 1 (diabetes/prediabetes), recall is strong for class 1 (0.76) so the model detects most diabetes cases, the macro avg gives equal weight to both classes so performance is balanced and the weighted avg favors the majority class (0), inflating overall scores.

The confusion matrix reinforces the idea of a high recall (few missed positives) and low precision (many false alarms). Therefore, this model is recall-oriented (ideal when the idea is not to miss potential diabetes cases) but with the expense of precision that can raise false alarms that might be acceptable in a medical screening contest and lastly the ROC-ACU and CV values prove the model’s stability and generalized learning.

**5.5 Decision Tree Model Results**

The Decision Tree Model constitutes a strong model that generally delivers a good performance while not increasing its complexity. The tuning of the aforementioned model is as follows:

|  |  |  |
| --- | --- | --- |
| Parameter | Value | Description |
| class\_weight | “balanced” | Adjust the weights inverse to the class frequencies to handle imbalance |
| test\_size | 0.25 | When splitting the independent and dependent variables, 25% is testing and 75% is for training |
| max\_depth | 16 | Maximum permitted depth of the decision tree to prevent overfitting |
| random\_state | 42 | Selected seed for random number generation to ensure reproductibility |

Regarding the data preparation before training the model is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Step (In order) | Description | Method/Tool used | Result/Notes |
| Missing Value Handling | Checked any null values | SimpleImputer (strategy='mean') | No missing values detected |
| Duplicate Removal | Checked any duplicated rows | drop\_duplicates() | 24206 duplciates removed |
| Train/Test Splitting | Split dataset into training and testing subsets | train\_test\_split (test\_size=0.25) | 75% training, 25% testing; stratified |
| Class Imbalance Handling | Addressed imbalance in target variable | SMOTE (random\_state=42) | Class ratio balanced. 0 and 1: 145782 |
| Feature Scaling | SKIPPED | SKIPPED | SKIPPED |
| Data shuffling | Randomly shuffled training data | shuffle() (random\_state=42) | Shuffled training set after resampling and scaling |

Table 14. Data preparation for Decision Tree model training

These parameters and preparation steps for the data before training the Logistic Regression wielded the best results and performance for the model. The results of the model are as follows (tables and figures):

|  |  |
| --- | --- |
| Metric | Value |
| Training time (s) | 2.3 to 2.6 on average |
| Acurracy (0-1) | 0.8164 |
| Precision (0-1) | 0.3966 |
| Recall (0-1) | 0.3848 |
| F1-Score (0-1) | 0.3906 |
| ROC-AUC (0-1) | 0.7663 |
| AVG Cross-Validation F1 (0-1) | 0.8580 ± 0.0020 (std deviation) |
| 5-Fold Cross-Validation F1 (0-1) | [0.8549, 0.8575, 0.8607, 0.8599, 0.8568] |

Table 15. Decision Tree model evaluation Metric results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification report | | | | |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No diabetes) | 0.89 | 0.89 | 0.89 | 48595 |
| 1 (Diabetes or Prediabetes) | 0.40 | 0.38 | 0.39 | 8774 |
| Accuracy |  |  | 0.82 | 57369 |
| Macro avg | 0.64 | 0.64 | 0.64 | 57369 |
| Weighted avg | 0.81 | 0.82 | 0.82 | 57369 |

Table 16. Decision Tree model Classfication Report

A blue squares with white text

AI-generated content may be incorrect.

Figure 28. Decision Tree model Confusion Matrix

The evaluation metrics of the Decision Tree model indicate a higher overall accuracy (0.81) compared to that achieved by the Logistice Regression model. However, the model’s other metrics like precision and recall for the postivie class (diabetes/prediabetes) are relatively low at 0.4 and 0.38 respectively. This implies that the model is good at correctly identifying the negative class (non-diabetes) with both precision and recall at 0.89 but struggles more with the positive class (diabetes/prediabetes).

The ROC-AUC score of 0.77 implies that there is a moderate discrimination capacity between both classes but is lower than the Logistic regression’s 0.81, a slightly less robust ranking predictions. The confusion matrix further confirms this, the model procedures a moderate number of false negatives (FN, missed diabetes cases) and false positives (FP), a balanced but modes performance on detecting positive cases.

Despite this, the model proves to be very stable and with a generalizable performance during training with a 5-fold Cross-validation F1 scores (0.86). Overall, the Decision Tree model shows a strong overall accuracy and stable learning but may require further complementary techniques to improve its recall to diabetes cases, which is a key factor in diagnostics because missing positives are catastrophic.

**5.6 Random Forest Model Results**

The Random Forest Model constitutes an ensemble model which basically implements mani Decision Tree models. The tuning of the aforementioned model is as follows:

|  |  |  |
| --- | --- | --- |
| Parameter | Value | Description |
| class\_weight | “balanced” | Adjust the weights inverse to the class frequencies to handle imbalance |
| test\_size | 0.25 | When splitting the independent and dependent variables, 25% is testing and 75% is for training |
| max\_depth | 16 | Maximum permitted depth of the decision tree to prevent overfitting |
| random\_state | 42 | Selected seed for random number generation to ensure reproductibility |
| n\_estimators | 100 | Number of trees in the forest to build for the ensemble. |
| min\_samples\_split | 15 | Minimum samples required to split a node to control tree growth. |
| min\_sample\_leaf | 5 | Minimum samples required at a leaf node to prevent overfitting. |
| max\_features | “sqrt” | Limit number of features considered for each split; "sqrt" (default) |
| n\_jobs | -1 | Uses all available CPU cores to parallelize training and improve performance. |

Table 17. Random Forest model parameters for tuning

Regarding the data preparation before training the model is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Step (In order) | Description | Method/Tool used | Result/Notes |
| Missing Value Handling | Checked any null values | SimpleImputer (strategy='mean') | No missing values detected |
| Duplicate Removal | Checked any duplicated rows | drop\_duplicates() | 24206 duplciates removed |
| Train/Test Splitting | Split dataset into training and testing subsets | train\_test\_split (test\_size=0.25) | 75% training, 25% testing; stratified |
| Class Imbalance Handling | Addressed imbalance in target variable | SMOTE (random\_state=42) | Class ratio balanced. 0 and 1: 145782 |
| Feature Scaling | SKIPPED | SKIPPED | SKIPPED |
| Data shuffling | Randomly shuffled training data | shuffle() (random\_state=42) | Shuffled training set after resampling and scaling |

Table 18. Data preparation for Random Forest model training

These parameters and preparation steps for the data before training the Logistic Regression wielded the best results and performance for the model. The results of the model are as follows (tables and figures):

|  |  |
| --- | --- |
| Metric | Value |
| Training time (s) | 9 to 10 on average |
| Acurracy (0-1) | 0.8347 |
| Precision (0-1) | 0.4519 |
| Recall (0-1) | 0.3798 |
| F1-Score (0-1) | 0.4127 |
| ROC-AUC (0-1) | 0.8077 |
| AVG Cross-Validation F1 (0-1) | 0.8889 ± 0.0014 (std deviation) |
| 5-Fold Cross-Validation F1 (0-1) | [0.8874, 0.8873, 0.8896, 0.8916, 0.8887] |

Table 19. Random Forest model evaluation Metric results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification report | | | | |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No diabetes) | 0.89 | 0.92 | 0.90 | 48595 |
| 1 (Diabetes or Prediabetes) | 0.45 | 0.38 | 0.41 | 8774 |
| Accuracy |  |  | 0.83 | 57369 |
| Macro avg | 0.67 | 0.65 | 0.66 | 57369 |
| Weighted avg | 0.82 | 0.83 | 0.83 | 57369 |

Table 20. Random Forest model Classfication Report

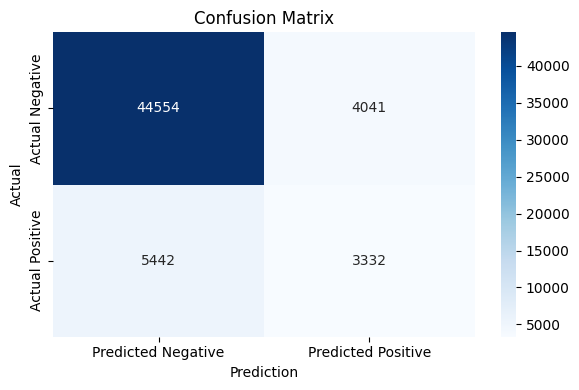


Figure 29. Random Forest model Confusion Matrix

The evaluation metrics of the Random Forest model demontratest a strong overall performance in general thanks to its ensemble nature, showing an accuracy of 0.83 (slightly higher than Decision Tree and Logistic Regression models). Precision and recall for the positive class (diabetes/prediabetes) are 0.45 and 0.38 respectively, which improves over Decision Tree precision but recall reamins modest at best. Indicating the ensemble model is better at identifying positive cases (TP) without significantly increasing false positives (FP).

It also performs very well on the negative class (non-diabetes), achieving a precision of 0.89 and recall of 0.92, a highly reliable model in ruling out diabetes when the prediction comes as negative and the ROC-AUC socre of 0.81 reflects a solid ability to distinguish the 2 classes, matching that of Logistic Regression model and surpassing Decision Tree. Furthermore, the Confusion Matrix supports this, suggesting the model is more conservative in flagging diabetes (likely due to imbalanced class distributions) but it’s still excellent generalization and stability during training, evidenced by the 5-folf Cross-Validation F1 socres averaging 0.8889, making it robust and less prone to overfitting.

While recall on positive class could still be improved, this model strikes a strong balance between accuracy, generalization and interpretability.

**5.6 SVM Model Results**

**5.7 Direct comparison between all models**

**Discussion**

Please discuss the results. Give rationale for the results, why some are better than others, what can be the possible reasons. Why are some results not as expected? Justify the results, if possible. Provide context from the literature.

**Conclusion and Future Work DO AT THE END**

**NOTES**

\*Individual (white box) models are less accurate but easier to develop, easier to train, more interpretable and easier to seamlessly implement in healtcare.

\*Ensemble methods are more accurate but more complex to develop, more complex and longer to train, less interpretable and harder to implement in healthcare, also a lot of them are prone to overfitting

**Conclusion**

**Future Work**

**References**

1. DATA SET<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>
2. WHAT IS DATA SCIENCE <https://tvst.arvojournals.org/article.aspx?articleid=2776501>
3. WHAT IS MACHINE LEARNING <https://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/cpt.1796>
4. DATA SCIENCE LIFECYCLE, THEORETICAL BACKGROUND .2 <https://assets.pubpub.org/hzllsv80/e3005d85-fea9-4b60-8ba0-f5409ca0905e.pdf> OR OR OR <https://hdsr.mitpress.mit.edu/pub/577rq08d/release/4> OR OR OR A DIFFERENT AUTHOR <https://arxiv.org/abs/2301.13761>
5. SUPERVISED LEARNING ALGORITHMS <https://link.springer.com/chapter/10.1007/978-3-030-82327-6_8>
6. MODEL EVALUATION METRICS <https://www.nature.com/articles/s41598-024-56706-x#Sec2>
7. CHALLENGES REAL WORLD IMPLEMENTATION <https://arxiv.org/abs/2107.09546> AND <https://arxiv.org/abs/2405.13832> AND <https://arxiv.org/abs/2305.02474>
8. WHAT IS FEATURE SELECTION <https://www.sciencedirect.com/science/article/pii/S1319157819304379>
9. WHY IS FEATURE SELECTION IMPORTANT [https://www.ingentaconnect.com/contentone/tsp/cmes/2020/00000125/00000001/art00017#](https://www.ingentaconnect.com/contentone/tsp/cmes/2020/00000125/00000001/art00017)
10. WHAT IS DIABETES AND FACTS OF IT <https://www.who.int/news-room/fact-sheets/detail/diabetes>
11. DIABETES 2 AND PROVE IT IS CURABLE <https://diabetes.org/newsroom/american-diabetes-association-2023-standards-care-diabetes-guide-for-prevention-diagnosis-treatment-people-living-with-diabetes>
12. MORE DIABETES <https://pubmed.ncbi.nlm.nih.gov/27388988/>
13. ML AND DIABETES <https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques>
14. ML IN PREDICTING DISEASES <https://www.foreseemed.com/blog/machine-learning-in-healthcare>
15. TREATMENT PERSONALIZED <https://proceedings.mlr.press/v136/sarkar20a/sarkar20a.pdf>
16. ROLE ML IN HEALTHCARE <https://pg-p.ctme.caltech.edu/blog/ai-ml/machine-learning-in-healthcare-applications-use-cases-careers>
17. COMMONLY USED ML MODELS <https://www.paradigmpress.org/jimr/article/view/1532> AND <https://www.academia.edu/43751015/Classification_of_Pima_Indian_Diabetes_Dataset_using_Ensemble_of_Decision_Tree_Logistic_Regression_and_Neural_Network> AND <https://www.scirp.org/reference/referencespapers?referenceid=3961876> AND <https://www.mdpi.com/1999-4893/16/11/503> AND <https://link.springer.com/chapter/10.1007/978-981-13-1810-8_10#citeas>
18. KEY INSIGHTS <https://proceedings.mlr.press/v139/wu21c.html> AND <https://arxiv.org/abs/1905.05134> AND <https://link.springer.com/article/10.1007/s10462-024-10884-2> AND <https://www.sciencedirect.com/science/article/abs/pii/S1566253524001908>
19. 1 IN 5 DIABETES <https://www.cdc.gov/diabetes/about/index.html>
20. **Scikit-learn Developers. (2024).** *Scikit-learn Documentation*. Version 1.4.2. <https://scikit-learn.org/stable/>