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: Dr. Rand Kouatly

Second supervisor: Dr. Souad El Hassanie

Submitted on: 07.2025

**Abstract**

A very common chronic disease that afflicts a large portion of the global population is Diabetes. The early detection in the early stages of this disease constitutes a determining factor for preventing severe health complications derived from idiosincrasies of the disease and reduce the strain it consitutes to the healthcare system and the patients alike. The proposed study aims explore and compare the performance of different supervised machine learning models, particularly Logistic Regression, Decision Tree, Random Forest and Support Vector Machines. The models will be trained baed the dataset the “Diabetes Binary Health Indicatros BRFSS2015”, that contains over 250,000 entries with 21 health-related predictor variables and 1 objective variable.

Standard classification metrics like accuracy, precision, recall, F1-score, ROC-AUC, K-fold cross validation and Confusion matrix are considered to be used to compare the performance of each of the proposed models. These metrics will be analyzed by themselves and within the context of a clinical setting.

The proposed study and it ajdioajdia aim to contribute to the field of intelligent healthcare diagnostic systems to potentially enhance early intervention strategies in healthcare.

**Introduction**

**Background**

Chronic diseases like diabetes are expected to become more prevalent across societies worldwide, in developed and developing countries. This growing trend creates an urgent need for intelligent system capable of supporting an early detection and timely intervention. Traditional diagnostic approcahes, while clinically accepted, usually fail to fully make full use of the vast amount of structured health data available and that can be obtained.

Machine Learning, a subfield of Artificial Intelligence, offers a promising solution by enabling automated, data driven predictions that could potentially assist health professionals in identifying individuals at risk more efficiently and effectively. In the are of Software Engineering, the integration of Machine Learning models in diagnostics tools represents a significant step towards intelligent healthcare applications.

Although, the selection of the most appropiate Machine Learning model will involve the careful consideration of its general performance (accuracy or interpertability) and how feasible would it be to implement them in a clinical setting. The proposed study aims to evaluate and compare multiple supervised Machine Learning models on a healthcare dataset that is publicly available to compare and determine which of the models is the most effective and practical use in clinical decision support system, if such conclusion can be reached at all.

**Problem Statement**

Diabetes, as a chronice disease, continues to rise in its global prevalence, posing serious health risks for patients while burdening healthcare infrastructures. The early detection of such diseases is critcal for effective intervention and management. Traditional diagnostic methods, can be time consuming and may not fully address the potential of the available healthcare data. With the growing availability of health-related datasets and advancements in machine learning, it is now feasible to explore Machine Learning approaches for early and accurate diabetes prediction. Nevertheless, many challenges remain, especially in determining which Machine Learning model performs best. This will involve the evaluation the model’s acurracy, interpretability and applicability in real world environments.

**Research Questions**

1. Which Machine Learning algorithms provide the most accurate results in predicting diabetes based on health datasets?
2. How do different features (like cholesterol, BMI, age, sex etc.) influence the predictive power of ML models?
3. Can ensemble models outperform traditional single classifiers in diabetes prediction?
4. What challenges arise in implementing ML-based diagnostic tools in real-world healthcare systems?
5. How can the interpretability of ML models influence their adoption in clinical decision-making in diagnostics?

**General Objective**

* To explore and compare the performance and effectiveness of various machine learning classification models in the early detection of diabetes using a real-world health dataset.

**Specific Objectives**

* To identify and preprocess the relevant diabetes-related dataset.
* To implement and evaluate multiple ML algorithms, for example Logistic Regression, Decision Tree, Random Forest and SVM.
* To assess the performance of each model using standard classification metrcis.
* To analyze the feature importance and model interpretability in the context of medical decision-making.
* To discuss the potential integration of ML models into healthcare systems for diagnostic support (not replacement).

**Rationale of the Study**

The proposed study will contribute tot the fields of healthcare informatics and software engineering by studying how existing mahcine learning models could enhance diabetes prediction and clinical decision making. Since AI driven solutions become more embedded in medical softwares, it is essential to evaluate which models offer the best support for diagnostic tasks.

With the exploring how data drive methods can improve the early detection of diabetes, the proposed research intends to support positive patient outcomes and encourage technological innovation in the medical field. The findings may also serve as a base for future healthcare policies surrounding intelligent diagnostic systems.

**Scope**

* The study will focus on supervised classification machine learning models applied to the Diabetes Binary Health Indicators BRFSS2015 dataset.
* The dataset contains pre-cleaned, structured data with 21 total predictor variables and one binary target label.
* Evaluation will be based on the previously mentioned, pubicly available dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS) by the CDC and curated by the kaggle user Alex Teboul.
* Model training will follow a generalized pipeline; however, model-specific parameters will be individually tuned as needed to optimize the performance of each algorithm.
* Evaluation metrics will include standar classification metrics like Accuracy, Precision, Recal, F1-score, ROC-AUC, K-fold cross validation, Classification report, Training time and Confusion Matrix.

**Delimitation**

* The study will not involve real-time or live clinidal data.
* The study will only the aforementioned dataset; no additional datasets will be combined or explored.
* It will not address diabetes treatment or progression prediction.
* It will focus on binary classification (diabetic/prediabetic vs non-diabetic), not multi-class (diabetic vs prediabetic vs non-diabetic for instance) or regression-based problems.

**Theoretical Background**

**Introduction**

This chapter outlines the theoretical foundation that will support the proposed study on the application of Machine Learning techniques for early diabetes detection. The research will be grounded in the principles of data science, with a particular focus on supervised machine learning and the underlying theories of classification models. With a clear understanding of the typical data science pipeline, exploratory data analysis, the classification models to be implemented and the model evaluation techniques or metrics, this chapter will provide the conceptual framework necessary to guide the implementation and analysis of Machine Learning models in the healthcare system and in a diagnostic context

**Data Science and ML Theory**

Data Science will be the interdisciplinary foundation of the proposed research. It combines scientific methods, statistical analysis, mathematics, analytics, specialized algorithms and computational systems to obtain meaningful insights from structured and/or unstructured datasets (Zarbin, Lee, Keane, & Chiang, 2021). Within this framework, Machine Learning will be applied to enable systems to learn patterns from data without explictly programming instructions to perform specific tasks to recognize such patterns (Badillo et al., 2020).

The study will follow a data science lifecycle adapted and expanded from Wing (2019), which typically includes:

* Data Collection: The relevant structured and/or unstructured data will be gathered from reliable, credible sources and following ethical considerations, particularly those involving personal health data, will be taken into account throughout this process.
* Preprocessing and Data Cleanin**g**: Raw data will be appropiately prepared for analysis by addressing common factors like missing values, correcting inconsistencies, outliners and transforming data into suitable formats required by the machine learning models.
* Feature Engineering: Only meaningful features will be selected from the dataset to improve overal performance and relevance.
* Machine Learning Model Training: Several supervised machine learning models will be implemented and trained using a portion of the dataset and the tested with a smaller portion of the same dataset. Each model might require different tuning.
* Model Evaluation: After training and testing, the models will be evaluated with standard performance metrics to ensure their accuracy, reliability, and generalizability.
* Interpretation and Deployment Considerations: The interpretability of model outputs will be examined to ensure their applicability in clinical settings and decision making processes.

In the context of healthcare, this lifecycle will, most likely, prove essential for the development of accurate predictive model suited for clinical decision support.

**Supervised Classification Algorithms**

Supervised learning constitutes a category of machine learning, in which the model is trained on a dataset that is organized with input and output pairs. The supervised model will learn how to map the input features to a target label, so after training, the model is capable to predict a target for a new, uncategorized data (Igual and Seguí, 2024). For this thesis proposal, the focus will be only on binary classification, in which the goal is to predict if the patient is diabetic or not based on an set of features of the datasets. The supervised classification algorithms that will be implemented are in the following sections.

**Logistic Regression**

According to Richards (2022), Logistic regression models specifically calculates the probability that a given input belogs to a particular class using the sigmoid function.

**Decision Tree**

Zhou (2022), states that Decision Trees split the given data based on feature values in order to form a structure similar to a tree, where each of the nodes represent a decision and each “leaf” constitutes a prediction, based on a feature. They are fairly easy to interpret, handling both classification and regression tasks and support numerical and categorical data alike, however, an imoportant note is that this model is prone to overfitting, especially if the dataset is too small or noisy. It uses the Gini impurity, Entropy and Information gain formulas as its mathematical foundation.

**Random Forest**

Random Forest is an ensemble method that constructs multiple decision trees during training and outputs the majority of class as the final result. This model is can handle non-linear relationships and provides feature importance. Since this model is an ensemble of Decision Tree, the formulas and math foundation that support this model are the same as Decision Tree (Rahman, Md.A. et al*,* 2023).

**Support Vector Machines (SVM)**

As described in the Scikit-learn oficial documentation, SVM realizes an optimal hyperplane that separates the given data into classes by maximizing the margin between support vectors. When it comes to non-linear data, certain models can use the hardware’s kernel as a Radial Basis Function (also known as RBF), which is used to project the given data into higher dimensions. Other SVM models, like LinearSVC, use a linear kernel that is optimized for linearly separable data, directly learning the linear decision boundary equation and using the Hinge Loss formula.

**Machine Learning Models comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 | Linear (LinearSVC) or Non-linear (SVC with kernel) | Low-Medium | Fast (LinearSVC), Slow on large data (SVC) | Yes (SVC only) | Yes (SVC), No (LinearSVC without calibration) |

ML algorithms comparative table

**Model Evaluation Metrics**

The evaluation of Machine Learning models in the healthcare industry cannot be simple satisfactory results. Misclassification, especially false negatives can have severe consequences for the patients. Several metrics will be used to evaluate each of the implemented models. The following list of metrics is adapted from Rainio, Teuho and Klén (2024):

* Accuracy: Proportion of total correct predictions (True positives and True negatives).
* Precision: Proportion of positive identifications (Positive predictions) that were actually correct (True).
* Recall (Sensitivity): Proportion of actual positives correctly identified (True positives).
* F1-Score: Consist in the harmonic mean of precision and recall times 2.
* ROC-AUC: This metric measures the trade-off observed between true positive rate and the false positive rate. The area under the curve provides an extra performance measure for analysis.

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

**Feature Selection**

Within the domain of supervised machine learning, feature selection will represent a crucial step that directly impacts the model’s performance, interpretability and ability to generalize unseen data. As highlighted by Mahadeo, Dhanalakshmi and Dhanalakshmi (2022), feature selection refers to the process of identifying and isolating the most relevant features from the dataset. These features are the ones that contribute the most to the model’s ability to accurately predict a target outcome, which in the case of this proposed study will the presence or not of diabetes.

Feature selection will form an integral part of the data science pipeline and dataset preparation before model training. According to Naheed et al. (2020), the advantages of feature selection are:

* Improving Model Accuracy: Remove irrelevant or redundant features that may introduce noise into the training process.
* Reducing Overfitting: Simplifying the model to focus only on the most meaningful features the risk of overfitting can be minimized.
* Speeding Up Training Time: Reduce dimensionality of the dataset is anticipated to decrease the time required to tran the machine learning algorithms.
* Enhancing Interpretability: Understanding the features that contribute the most to predictions can be essential in healthcare settings. For example, features like BMI, age and glucose levels are highly interpretable and clinically relevant, that may support acceptance and trust among healthcare professionals.

In the proposed study, feature selection will be applied to each model, that allows it, to improve the model performance and practical applicability of its results.

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

Despite Machine Learning models often demonstrating high accuracy in controlled environments, their deployment in real world environments like clinical will face numerous challenges. Addressing these challenges is critical for the ethical, legal and regulatory landscape. Some key idetified challenges for this proposal include:

* Data Privacy: Healthcare data is highly sensitive and regulated by laws such as HIPAA and GDPR in the US and European Union respectively. Ensuring that data is securely stored, used responsibly, and anonymized will be a non-negotiable requirement for this and future research (Ali, S et al., 2024).
* Interpretability and Trust: Healthcare professionals must be able to understand and trust the predictions of the models, especially when the predictions influence diagnoses. “Black-box” models will be approached cautiously unless paired with interpretability tools to enhance transparency (Petersen et al., 2022).

**Literature Review**

**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. This review aims to highlight gaps in the current research, particularly in the comparative analyses of classification models, model interpretability and real-world integration of ML tools in clinical settings, providing insights that acted as the foundation for the current study and justify its methodology and research questions.

**Key Terms and Concepts**

* Machine Learning (ML): Based on Baloglu, O., Latifi, S.Q. and Nazha, A. (2021), 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, or in another words there are no associated output labels (Shruthi H, 2022). 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 which consist in Binary, Multi-class and Multi-label (article in DataScienceTribe 2023), 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: According to Erasmus, A., Brunet, T.D.P. and Fisher, E. (2021), in the topic of 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 concept 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 (Setyati, R. et al, 2024).

**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 diseases, kidney failure, blindness and even 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, traditional diagnostic processes are very time-consuming and may not even leverage the full potential of patient medical data, that should include behavioral, demographic and lifestyle information.

**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).

**Interpretability and Model Transparency in Healthcare**

The transparecy and interpretability of machine learning 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 deep learning neural networks are more accurate but their inherent 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.

**Identified Gaps in the Literature**

Despite the considerable amount of research dedicated to diabetes and other diseases prediction using machine learning 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.

**Methodology**

**Research Design**

The proposed methodology will employ a quantitative, experimental research design to evaluate and compare the effectiveness of multiple supervised machine learning mocels in predicting diabetes (classification). The methodology will be structured around the standard data science lifecycle, with steps like data acquisition, preprocessing, model development, model testing, evaluation and interpretation of results. A comparative approach will be adopted to determine the relative performance of the models, using the same dataset and a unified evaluation framework based on predefined metrics.

**Dataset and Data Collection**

The proposed dataset will be the Diabetes Binary Health Indicators BRFSS2015, which is publicly available on Kaggle and originally prepared by Alex Teboul (2021). This dataset originates from the Behavioral Risk Factor Surveillance System (BRFSS) from 2015, an annual health telephone survey conducted by the Center for Disease Control and Prevention. It holds responses from 253,680 individuals, with 21 predictor variables and a binary targe variable indicating diabetes status.

|  |  |  |
| --- | --- | --- |
| Variables | Description | Values / Encoding |
| Diabetes\_binary | Diabetes status | 0 = No diabetes, 1 = Diabetes or Prediabetes |
| 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. |

Data set 22 columns

Despite the dataset being described as pre-cleaned, additional preprocessing will be implemented to ensure suitability for training. The variables include behavioral, demographic and specific health indicators relevant to diabetes prediction. The dataset’s large size, feature diversity and real world origin makes it appropriate for predictive modeling, but it is important to note that the dataset is imbalanced with a majority of non-diabetic responses. This imbalance reflects real clinical scenarios. According to the

The dataset’s large size, feature diversity, and real-world origin make it highly appropriate for predictive modeling. It is important to note that the dataset is imbalanced, with a majority of non-diabetic responses. This imbalance reflects real clinical scenarios and will be addressed accordingly in model development. According to the Center for Disease Control and Prevention (2024), many individuals are unaware of their diabetic or prediabetic status until apparent clinical symptoms appear.

**Data Preprocessing**

Although the dataset’s initial cleaning, it will undergo additional preprocessing steps to optimize model performance, including:

1. Handle possible missing values: Identify and handle missing values by imputation or removal, as appropriate.
2. Remove duplicate values: Remove duplicate entries to prevent the model training from being skewed.
3. Encoding of ordinal features: Taking non-numerical features (like ordinal) and encode them to preserve logic to preserve logical order where necessary.
4. Data splitting: Split data into training and testing sets, also apply stratification to tackle the stated class imbalance between both sets.
5. Check and handle class imbalance in the training set: Address class imbalance on the training set with SMOTE oversampling or clas weight.
6. Feature scaling: Apply feature scaling, like standardization or robust scaling, to continuous variables while preserving binary/categorical features.
7. Data shuffling: Shuffle training data to ensure maximum randomness to simulate real-world conditions.

**Machine Learning Algorithms**

Four supervised classification algorithms are to be implemented and evaluated during the experimentation of the thesis. The following models are selected for their interpretability, performance and diverse methodological approaches:

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

The aforementioned models represent linear, tree-based, ensemble and kernel-based (in some types) approaches, respectively.

**Model Training**

Model training will incorporate the following strategies:

* Between 80/20 and 70/30 split ratio to divide the dataset into training and testing sets.
* K-fold cross-validation (k=5 or 10) will be utilized to reduce the possibility of overfitting the models and ensure generalization across samples.

**Evaluation Metrics for validation**

Model performance will be evaluated under the following metrics:

* Accuracy
* Precision
* Recall
* F1-Score
* ROC-AUC
* Confusion Matrix
* Cross-validation results

These metrics will provide a balanced assessment of the effectiveness of each model, considering the dataset’s class imbalance and clinical context.

**Feature Importance and Model Interpretability**

Feature importance scores will be analyzed, particularly for tree-based models, in order to understand how input variables contribute to predictions and support clinical decision-making.

**Tools and Technologies**

The implementation of the models will be carried out using Python as the programming language on the VSCode code-editor using the following libraries:

* Scikit-learn
* Pandas
* NumPy
* Matplotlib
* Seaborn
* Imblearn
* Logging
* Time
* Pathlib

**Ethical Considerations**

The proposed study will use publicly available, anonymized data to minimize ethical risks. Regardless, ethical diligence will be maintained by:

* Assessing model fairness across demographic groups to identify and mitigate bias.
* Contextualizing Machine Learning insights within clinical realities with an emphasis that these tools will support human expert judgment, not replace it.

**References**

1. **Teboul, A**. (2021). *Diabetes Health Indicators Dataset*. Kaggle. Available at: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset> (Accessed and downloaded: 12th May 2025).
2. **Centers for Disease Control and Prevention (CDC)** (2022). Behavioral Risk Factor Surveillance System. Kaggle. Available at: <https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system> (Accessed: 12th May 2025).
3. **Wing, J.M**, (2019). *The Data Life Cycle*, in: Harvard Data Science Review, 1(1). Available at: <https://hdsr.mitpress.mit.edu/pub/577rq08d/release/4> (Accessed: XX May 2025).
4. **Badillo, S, et al.** (2020). *An introduction to machine learning*, in: Clinical Pharmacology & Therapeutics. Wiley-Blackwell, 107(4), pp. 871–885. Available at: <https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.1796> (Accessed: xx May 2025).
5. **Naheed, N. et al**, (2020). *Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review*, in: Computer Modeling in Engineering & Sciences. Tech Science Press, 125(1), pp. 314–344(31). Available at: <https://www.ingentaconnect.com/contentone/tsp/cmes/2020/00000125/00000001/art00017#Refs> (Accessed: XX May 2025).
6. **Zarbin, M.A, et al.** (2021). *Data science*, in: Translational Vision Science & Technology. ARVO Journal, 10 (8), p. 20. Available at: <https://tvst.arvojournals.org/article.aspx?articleid=2776501> (Accessed: 16th May 2025).
7. **Richards, J.A**, (2022). *Supervised classification techniques*, in: Remote Sensing Digital Image Analysis. Springer eBooks, 6, pp. 263–367. Available at: <https://link.springer.com/chapter/10.1007/978-3-030-82327-6_8> (Accessed: XX May 2025).
8. **Zhou, V**, (2022). *Random forests for complete beginners*, in: victorzhou.com. Available at: <https://victorzhou.com/blog/intro-to-random-forests/#23-training-a-decision-tree-when-to-stop> (Accessed: XX May 2025).
9. **Balakrishnan, K., Dhanalakshmi, R. and Mahadeo, U**, (2021). *Analysing stable feature selection through an augmented marine predator algorithm based on opposition‐based learning*, in: Expert Systems, 39(1). Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12816> (Accessed: XX May 2025).
10. **Petersen, E. *et al*,** (2022). *Responsible and Regulatory conform Machine Learning for Medicine: A Survey of Challenges and solutions*. IEEE Access, 10, pp. 58375–58418. Available at: <https://ieeexplore.ieee.org/document/9783196> (Accessed: XX May 2025).
11. **Rahman, Md.A. et al**, (2023). *Machine Learning-Based Approach for Predicting Diabetes employing Socio-Demographic Characteristics*, in Algorithms, 16(11), pp. 503. Available at: <https://www.mdpi.com/1999-4893/16/11/503> (Accessed: XX May 2025).
12. **Sandhu, S. et al**, (2020). *Integrating a machine learning system into clinical workflows: Qualitative study*, in: Journal of Medical Internet Research, 22(11), p. e22421. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7714645/> (Accessed: XX May 2025).
13. **Schinkel, M. et al**, (2023). *Detecting changes in the performance of a clinical machine learning tool over time*, in: EBioMedicine, 97. Available at: <https://pubmed.ncbi.nlm.nih.gov/37793210/> (Accessed: XX May 2025).
14. **Igual, L. and Seguí, S.** (2024). *Supervised learning*, in: Introduction to Data Science. Springer eBooks, pp. 67–97. Available at: <https://link.springer.com/chapter/10.1007/978-3-031-48956-3_5> (Accessed: XX May 2025).
15. **Rainio, O., Teuho, J. & Klén, R**, (2024). *Evaluation metrics and statistical tests for machine learning*. Scientific Reports 14(1). Available at: <https://www.nature.com/articles/s41598-024-56706-x> (Accessed: XX May 2025).
17. **Centers for Disease Control and Prevention**, (2024). *Diabetes Basics*. Available at: <https://www.cdc.gov/diabetes/about/index.html> (Accessed: XX May 2025).