Bachelor Thesis Proposal

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

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.

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 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, specifically Logistic Regression, Decision Tree, Random Forest and SVM.
- To assess the performance of each model using standard metrics.

- 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.

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.

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.
- 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, not multi-class or regression-based problems.

Theoretical Background

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 Cleaning: 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.

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 but the 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.

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.

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, Reducing Overfitting, Speeding Up Training Time and Enhancing Interpretability.

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

Key Terms and Concepts

- Machine Learning: According to Baloglu, Latifi, and Nazha (2021), machine learning is a subfield of artificial intelligence that focuses on the development and statistical analysis of algorithms that can learn from data. These algorithms are designed to make predictions and/or decisions on unseen data without hard coding rules for those tasks.
- Supervised Machine Learning: It refers to models that are trained using labeled datasets, where each input is paired with an output label. The models map inputs to outputs and generalize this knowledge to make predictions on new data. In contrast, unsupervised machine learning involves training on unlabeled data, where no output label is provided (Shruthi, 2022).
- Classification Machine Learning Models: A subset of supervised learning that involves predicting categorical outcomes. As outlined by DataScienceTribe (2023), classification tasks can be categorized into binary, multi class or multi label types.
- Early Detection: Referst to the identification of diseases at their initial stage, before significant symptoms are present. This allows for timely medical intervention and can lead to better outcomes for the patient (Setyati et al., 2024).

Diabetes and the Need for Early Detection

Diabetes mellitus is a chronic metabolic disorder that presents an elevated blood glucose levels, result of defects in insulin secretion or action. The most common form of diabetes is Type 1, Type 2 and gestational diabetes. According to the World Health Organization (2024), diabetes is one of the leading causes of death globally and associated with long-term complications like heart diseases, kidney failure among others.

Type 2 constitutes for 90% of all diabetes cases, developing gradually and remains undiagnosed for long periods of time, primarily due to its asymptomatic nature in its early stages. A delayed diagnosis is concerning because an early intervention can reduce the risk of complications and improve patient outcomes, since an early stage management involves lifestyle modifications or preventive medication (American Diabetes Association, 2023).

Traditional diagnostics rely on perdiodic blood glucose testing, HbA1c measurements and patient reported symptoms. While still used, these methods are more reactive and fail to identify patients in early stages of the disease (or those with a high risk), especially in populations with limited access to healthcare services and underdeveloped healthcare infrastructure. Moreover, conventional diagnostic can be time consuming and often underuse all the available patient data, which could include behavioral demographic and lifestyle information that may improve the diagnosis.

3.5 Role of Machine Learning in Healthcare

- Enhance diagnostic accuracy: Machine learning algorithms are highly effective in processing complex medical datasets, like imaging data and Electronic Health Records, to identify patterns typical of diseases. These models have shown promising results in detecting early signs of conditions like diabetic retinopathy, cardiovascular diseases and several types of cancer before obvious symptoms. Their diagnostic performance is comparable to that of traditional methods and sometimes better (Barth, S. and Flam, S, 2025).
- Personalize treatment plans: Analyzing a patient's medical history, behavioral data and lifestyle factors, these models can sypport the development of personalized treatment plans. These plans can enhance the outcomes and minimize side effects (Sarkar et al., 2020).
- Predictive analytics for disease prevention: Machine learning systems can process large scale datasets and estimate an individual's risk of developing certain diseases. These predictive capabilities can allow healthcare professionals to conduct early interventions and apply preventative measures (Kelley, 2024).

Methodology

Research Design

The proposed methodology will employ a quantitative, experimental research design to evaluate and compare the effectiveness of multiple supervised machine learning models 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 framework.

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.

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.

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.

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 and 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 between a 80/20 and a 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 and 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
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
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
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
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
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
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
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
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
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

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

12) **World Health Organization: WHO**, (2024). *Diabetes*. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes (Accessed: 26th May 2025).

13) **Barth, S. and Flam, S**, (2025). *Machine Learning in Healthcare: Guide to Applications & benefits*. Available at: https://www.foreseemed.com/blog/machine-learning-in-healthcare (Accessed: 28th May 2025).

14) **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/

15) **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/

16) **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

17) **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

18) **Centers for Disease Control and Prevention**, (2024). *Diabetes Basics*. Available at: https://www.cdc.gov/diabetes/about/index.html