Chronic Disease Detection: A Study on Diabetes Classification

Paulo Henrique da Silva Mota
Pre-Master Data Science (Student)
Jheronimus Academy of Data Science
S-Hertogenbosch, The Netherlands
p.h.dasilvamota@tilburguniversity.edu

Krishna Teja Atluri
Pre-Master Data Science (Student)
Jheronimus Academy of Data Science
S-Hertogenbosch, The Netherlands
k.t.atluri@tilburguniversity.edu

Joost Spruit
Pre-Master Data Science (Student)
Jheronimus Academy of Data Science
S-Hertogenbosch, The Netherlands
j.g.j.spruit@tilburguniversity.edu

***Abstract*—The effectiveness of the healthcare sector depends on general practitioners (GPs), who face high pressures due to demographic shifts and rising diabetes cases. To alleviate this scenario, we propose a machine learning solution to help the healthcare sector in accurately diagnosing diabetes whilst increasing efficiency and lowering workload for GP's. Guided by the CRISP-DM model, This approach achieves a 98 percent accuracy in classifying diabetes on patients. This Gradient Boosting model reveals the potential of machine learning to enhance healthcare, offering a promising solution for classifying the chronic disease diabetes.**

***Index Terms*—Classification, Machine Learning, Diabetes, Gradient Boosting, General practitioner**

I. Introduction

Currently, worldwide, the level of service in the healthcare sector is profoundly dependent on the work of general practitioners. General practitioners (GP), also known as family doctors, play a crucial role in the healthcare system. Their responsibilities are wide-ranging and encompass multiple aspects of general patient care. The main purposes of the GP's include serving as the first point of contact for patients seeking medical attention. Within the diagnosing phase of a patient, GP's will perform the assessment and monitoring for a wide range of health concerns including minor illnesses and injuries but also including chronic illnesses. Among these chronic illnesses, diabetes holds significant importance in a GP's evaluation of their patients. Diabetes is a chronic medical illness which can be characterized by elevated levels of glucose/sugar in the bloodstream. This occurs when the body fails to produce enough insulin, which is a hormone that regulates blood sugar levels. To diagnose diabetes, GP's typically will assess the medical history of the patient first. It will look for familial history of diabetes and gather information on other related conditions such as gender, age and smoking history. Next, a physical examination will take place to assess the overall health of the patient and to check for signs of complications associated with diabetes. Thirdly, a blood test will be performed by the GP to measure blood glucose levels after an overnight fast. Based on the results of these assessments, GPs can diagnose diabetes and initiate appropriate care for the patient.

Over the past few decades, the workload for GP's has experienced a steady increase, reflecting a broader trend in the healthcare sector worldwide. The main factors contributing to this scenario include demographic developments such as aging populations and rising life expectancies [7]. This leads to a greater prevalence of chronic diseases requiring the needed management and care by GP's. This is resulting in more complex patient cases and higher demands on their time and expertise. Moreover, other sociological changes as evolving patient expectations and increased awareness of preventive care have increased the workload for GP's [8]. As a result, GP's are challenged with high pressures to deliver their patient-centered care while also navigating increasingly complex healthcare systems and resource constraints. The described observations also lead to the fact that the healthcare sector is getting more expensive[9].

Also, currently, there is a concerning and consistent rise in both the number of cases and the prevalence of diabetes worldwide. This upward trend is explained by the complex interplay of the following factors. First, the shifts in lifestyle habits, such as unhealthy dietary patterns, as well as demographic changes, as aging populations and urban development contribute to the rising prevalence of the chronic disease diabetes[9]. Diabetes is emerging as a significant public health challenge, in the need of urgent action by the healthcare systems. The rise in the percental prevalence of diabetes is visible in figure 1[10].

Recent studies also show that a significant proportion of people with diabetes are unaware of this. According to the International Diabetes Federation (IDF), approximately one in two adults with diabetes are undiagnosed, thus meaning that they have the condition but have not been diagnosed by a GP. The main reason for this is the symptoms being overlooked. as well as inadequate access to healthcare services, particularly in underserved communities. Lack of awareness about diabetes risk factors and symptoms also contributes to delayed diagnosis whilst early detection and diagnosis of diabetes is crucial for the prevention of severe complications[11]. Untreated diabetes affects many major organs, including the heart, blood vessels, nerves, eyes, and kidneys. This highlights the importance of screening initiatives to identify undiagnosed individuals to be able to provide appropriate care[10].

The combination of the matters discussed above results in a problem for the healthcare sector. GP's being challenged with high pressures to deliver their patients the appropriate

care, rising levels of prevalence of diabetes and the fact that a significant proportion of people with diabetes are unaware of this explains the reasoning of creating a machine learning algorithm which indicates if the chronic disease diabetes is present at a patient whilst also alleviate the general workload for GP's [12].

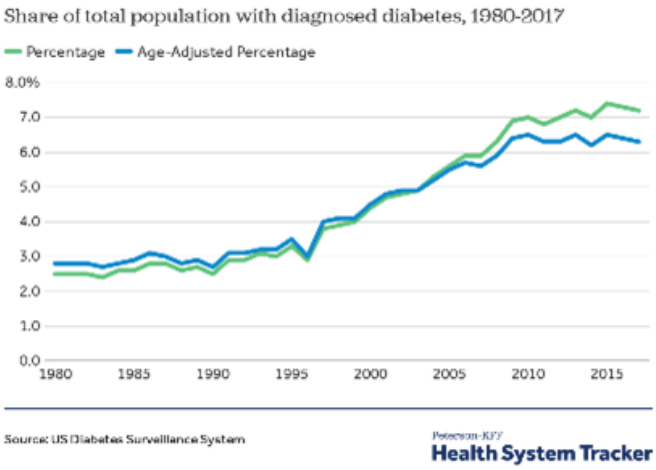

Fig. 1. Graph showing rising prevalence of diabetes

The problem stated above can be translated into a set of the research questions provided below. These research questions are based on the CRISP-DM model, visible in figure 2. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is a widely used framework for guiding data mining and machine learning projects. It consists of six phases: Business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Main research question:

"What methodologies and techniques should be used for developing a machine learning model to assist general practitioners in accurately diagnosing diabetes, while simultaneously alleviating their workload, considering key objectives, available data, data preprocessing, choice of algorithms, and hyperparameter tuning?"

Sub-research questions:

1) **Business understanding:** "What is the business case, which objectives and requirements are required for the machine learning model?"
2) **Data understanding:** "What dataset will be used for training the machine learning model, what are the characteristics of the features present in the dataset?"
3) **Data preparation:** "What data preprocessing techniques should be applied to prepare the input data for training the machine learning model?"
4) **Modelling:** "Which machine learning algorithm, is most appropriate for building a diabetes diagnostic model, and how can hyperparameter tuning be performed to optimise the performance?"
5) **Evaluation:** "What evaluation metrics should be used to assess the performance of the machine learning model, and how can cross-validation techniques be employed to validate the results?"

As is visible in the sub-research questions, for each phase of the CRISP-DM model a sub-research question is dedicated to it. First, within the business understanding phase, the model ensures alignment between project goals and objectives derived from the problem statement. Data understanding involves exploring and assessing the available dataset. Next, data preparation focuses on the preprocessing tasks to make the dataset suitable for the modelling phase. Then, the modelling phase is about selecting appropriate algorithms and techniques. Next, the evaluation phase assesses model performances using relevant metrics to be able to meet the criteria of the business. At last, the deployment involves integrating models into operational environments. So, CRISP-DM will be used for a structured approach to this machine learning project.

Fig. 2. figure showing the CRISP-DM structure

To briefly describe the results derived from this machine learning project, the sub-research questions are answered within this paragraph. By answering these questions, the main research question can be resolved as well. The first sub-research question regarding business understanding is illustrated in figure 3. The procedures for a GP is visualised and the target group for this machine-learning project is determined. Target group: "Patients visiting a General Practitioner". Also, the objective is set to create a machine-learning model with an accuracy of at least 90 percent. Reasoning for this is that in the healthcare sector is no room for error since people's health depends on it. Also, several other studies on this matter show that 90 percent accuracy is high for related matters.

For the data understanding and data preparation, sub-research questions 2 and 3, an appropriate dataset is found and

cleaned. This is covered in subchapter 4.1 Dataset description. To preprocess the data to prepare the data for the training of the machine learning model, the following actions are performed: Deleting NaN values and duplicates, balancing the dataset, and creating a split into a training set, test set, and validation set. This is also explained in chapter 4. Within the modeling phase of this project, four classification machine learning algorithms are trained and tested on the datasets. The results for these are visible in table 1. As is visible, the Gradient boosting has the highest F1-score.

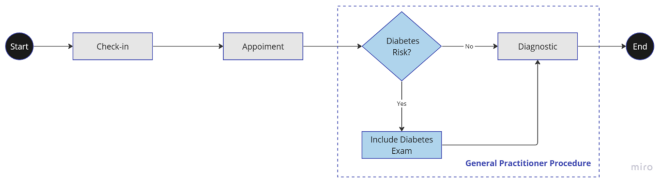

Fig. 3. Figure showing the GP's workflow for identifying diabetes

TABLE I
F1-SCORES FOR THE MACHINE LEARNING MODELS

Model	F1-score
Gradient Boost	0.98
Random Forest	0.92
KNN	0.95
Logistic Regression	0.91

In figure 4, the confusion matrix of the gradient boosting algorithm is provided. Within this figure, it is shown that of the 22484 classified patients handled by the algorithm, 588 are indicated as false negatives. This means that only for 2.6 percent of the patients classified, the machine learning model classified that these patients do not have diabetes while in reality, they do have diabetes. This result meets the objective of having an accuracy of at least 90 percent.

So, to conclude this section of the paper, the machine learning project meets the standards set by following the CRISP-DM model. The following should be noted regarding the greatness of the machine learning model. The main features that are used by the Gradient Boosting algorithm are the HB1AC and the Blood Glucose level. Together, they share an importance of 84 percent of the model. This is considered as the main weakness of this machine-learning project. A machine learning model predicting if a patient has diabetes based on simple demographic data is considered much more valuable. However, still, this machine learning model is of good value for the healthcare sector since it will predict the occurrence of having diabetes very accurately. In the sections Conclusion, Discussion, and Future work, this is elaborated.

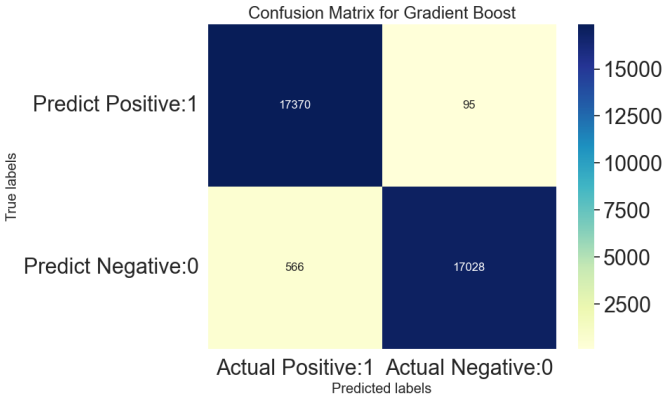

Fig. 4. Figure showing the confusion matrix of the gradient boosting algorithm

II. Related Work

A. State-of-The-Art (SoTA) Analysis

The State-of-The-Art (SoTA) in diabetes prediction using machine learning typically involves the application of various algorithms to predict the likelihood of an individual developing diabetes or to diagnose the disease based on patient data such as demographics, clinical measurements, and lifestyle factors. Here's an analysis of some key aspects:

1. **Data Sources and Features:**
- Researchers typically utilize datasets containing patient information such as age, gender, body mass index (BMI), blood pressure, glucose levels, family history, and lifestyle habits.
- Integration of electronic health records (EHR), wearable device data, genetic information, and other sources have become common to improve prediction accuracy.

2. **Feature Engineering and Selection:**
- Feature engineering techniques like normalization, scaling, and handling missing values are applied to preprocess the data.
- Feature selection methods such as wrapper methods, filter methods, and embedded methods are employed to identify the most relevant features for prediction.

3. **Algorithms and Models:**
- Various machine learning algorithms are employed, including logistic regression, decision trees, random forests, support vector machines (SVM), k-nearest neighbors (KNN), and ensemble methods like gradient boosting machines (GBM) and XGBoost.
- Deep learning techniques such as artificial neural networks (ANNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) are also explored for diabetes prediction.

4. **Evaluation Metrics:**
- Common evaluation metrics include accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), and area under the precision-recall curve (AUC-PR).

- Given the class imbalance in many diabetes datasets (i.e., more non-diabetic cases than diabetic cases), metrics like sensitivity and specificity are crucial.

5. **Challenges and Considerations:**

- Handling imbalanced datasets to prevent bias towards the majority class.
- Dealing with missing or noisy data, especially in real-world scenarios.
- Ensuring model interpretability and transparency, particularly in healthcare settings where decisions impact patient care.
- Generalization of models across diverse populations and settings.
- Ethical considerations regarding patient privacy and data security.

6. **Recent Advancements:**

- Incorporation of advanced techniques such as transfer learning and federated learning for better model performance and privacy preservation.
- Integration of multimodal data sources for more comprehensive patient profiling.
- Exploration of explainable AI methods to provide insights into model predictions and enhance trust among clinicians and patients.

7. **Open Challenges and Future Directions:**

- Robustness of models across different patient populations and healthcare settings.
- Integration of real-time monitoring and feedback systems to enable personalized interventions.
- Collaboration between researchers, healthcare providers, and policymakers to facilitate the adoption of machine learning-based diabetes prediction tools in clinical practice.

In conclusion, while machine learning has shown promising results in diabetes prediction, ongoing research is needed to address challenges related to data quality, model interpretability, and scalability for widespread adoption in healthcare settings.

B. Literature Review

Recent advancements in machine learning (ML) have led to various approaches being employed to predict diabetes risk with significant accuracy. This section outlines notable contributions in the domain of diabetes prediction using ML techniques.

Mahabub, Atik, in his study, "A Robust Voting Approach for Diabetes Prediction Using Traditional Machine Learning Techniques" (2019), explores a voting-based method that aggregates the decisions of several ML algorithms to improve prediction accuracy , such as AdaBoost, gradient boost, XGBoost, random forest, etc., to predict diabetes, considering several clinical parameters such as pregnancy, skin thickness, glucose, insulin, blood pressure, diabetes pedigree function, body mass index (BMI), age, and class variable. They achieved the highest accuracy rate of 84.42% with the multilayer perceptron algorithm.This method has shown effectiveness in enhancing the robustness of diabetes predictions by mitigating the weaknesses of individual models.

Ganie et al. (2023) presented an ensemble learning approach that harnesses the power of boosting techniques, specifically focusing on how ensemble methods can enhance the predictive performance by reducing bias and variance . Their work demonstrates the substantial gains in accuracy that can be achieved through sophisticated ensemble methods in genetic studies of diabetes. the effectiveness of five boosting algorithms were investigated, namely, XGBoost, CatBoost, LightGBM, AdaBoost, and Gradient Boosting, for predicting diabetes disease. According to their experimental findings, gradient boosting had the greatest accuracy rate of 96%.

Li et al. (2020) investigated the use of the XGBoost algorithm, a specific type of gradient boosting framework, for predicting diabetes . Their findings underscore the capability of XGBoost to handle large datasets with numerous features, making it a potent tool for medical predictions where dataset features can be extensive and complex. Their experiment results show that accuracy of diabetes prediction based the improved XGBoost algorithm with features combination is 80.2%.

Lai et al. (2019) explored various machine learning techniques to develop predictive models for diabetes mellitus. Their comparative analysis provides insights into the performance of different ML algorithms in accurately predicting the onset of diabetes, highlighting the effectiveness of machine learning in handling diverse and multidimensional medical data. They have obtained the AROC of 84.7% for GBM model, 88.0% for Logistic Regression Model, 87.1% for Random Forest Model, and 77.0% for Rpart model.

Mushtaq et al. (2022) discussed a voting classification-based approach that uses hypertuned ML techniques for diabetes prediction using the Pima diabetes dataset. Their research emphasizes the importance of hyperparameter tuning in achieving optimal model performance, demonstrating how fine-tuning ML models can significantly improve the prediction outcomes. The study implemented a dual-phase model selection strategy to construct the model. The voting classifier achieved the highest accuracy rate of 81.50%. Additionally, the Tomek links and Synthetic Minority Over-sampling Technique (SMOTE) were employed to balance the data and eliminate biases within the dataset. The authors recommended further research to assess the probability.

Tasin et al. (2022) used a combination of machine learning and explainable AI techniques to not only predict diabetes but also provide insights into the decision-making process of the models. The open-source Pima Indian and a private dataset of female Bangladeshi patients have been used in this work. SMOTE and ADASYN preprocessing techniques have been applied to handle the issue of imbalanced class problems. The XGBoost classifier achieved the best performance with 81% accuracy and an F1 score and AUC of 0.81 and 0.84, respectively, with the ADASYN approach.

Author	Year	Project Name	Algorithms used	Accuracy	Adoptions made
Li, Mingqi, Xiaoyang Fu, and Dongdong Li.	2020	Diabetes prediction based on XGBoost algorithm	XGBoost	80.20%	Gradient Boost
Mahabub, Atik.	2019	A robust voting approach for diabetes prediction using traditional machine learning techniques.	AdaBoost, gradient boost, XGBoost, random forest, etc.	84.42%	
Mushtaq, Zaigham, Muhammad Farhan Ramzan, Sikandar Ali, Samad Baseer, Ali Samad, and Mujtaba Husnain.	2022	Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques.	Voting Classifier (includes Random Forest, logistic regression, Support Vector Machine, KNN, Naive Bayes Theorem, and Gradient Boosting Classifier	81.50%	Voting Classifier
Shahid Mohammad Ganie	2023	An ensemble learning approach for diabetes prediction using boosting techniques	Gradient boosting algorithm	96.00%	Gradient boosting
Lai, Hang, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao.	2019	Predictive models for diabetes mellitus using machine learning techniques	Logistic Regression	88.00%	Logistic Regression

Fig. 5. An overview of various researches on which the project is based

This approach is particularly valuable in clinical settings where understanding the rationale behind a diagnosis or prediction is crucial for healthcare professionals.

These studies collectively highlight the diversity of machine learning techniques—from robust voting mechanisms and boosting strategies to sophisticated algorithms like XGBoost and transparent models using explainable AI—proving to be highly effective in the prediction of diabetes. They also indicate a broader trend towards integrating these technologies into practical healthcare applications, aiming to enhance predictive accuracy and improve patient outcomes in diabetes care.

III. Proposed Methodology and ML Techniques

A. Overview

Following the CRISP-DM methodology, the approach began with business understanding and data understanding, followed by the development of data preparation, modeling, and evaluation phases.

In the data understanding phasewe collected and assessed the available data, determining which features would serve as predictors and which would be the target variable for prediction. Through exploratory data analysiswe identified any anomalies or issues in the dataset, for building a preprocessing plan to be applied afterward. Moving into the data preparation stagewe crafted a preprocessing flow to handle duplicates, null values, outliers, and categorical variables, ensuring the dataset's quality and consistency.

Finally, in the modeling and evaluation phasewe partitioned the data into train, test, and validation sets. We trained four selected models—Random Forest, Gradient Boosting, KNN, and Logistic Regression—on the training set and evaluated their performance using various metrics. Through this processwe compared the models' performance and analyzed the feature importance for each, providing insights into their predictive capabilities and potential areas for refinement.

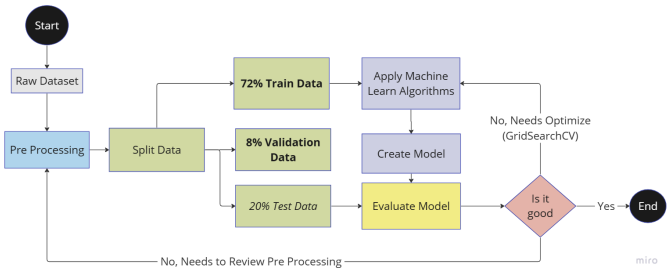

Fig. 6. The diagram shows the machine learning pipeline

B. Data Preparation

The dataset used, comprises medical and demographic information of patients, including their diabetes status (positive or negative), alongside various features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. Following an exploratory data analysiswe identified several issues: notably, data imbalance between diabetic and non-diabetic observations, as well as irrelevant labels, duplicates, and outliers in categorical values.

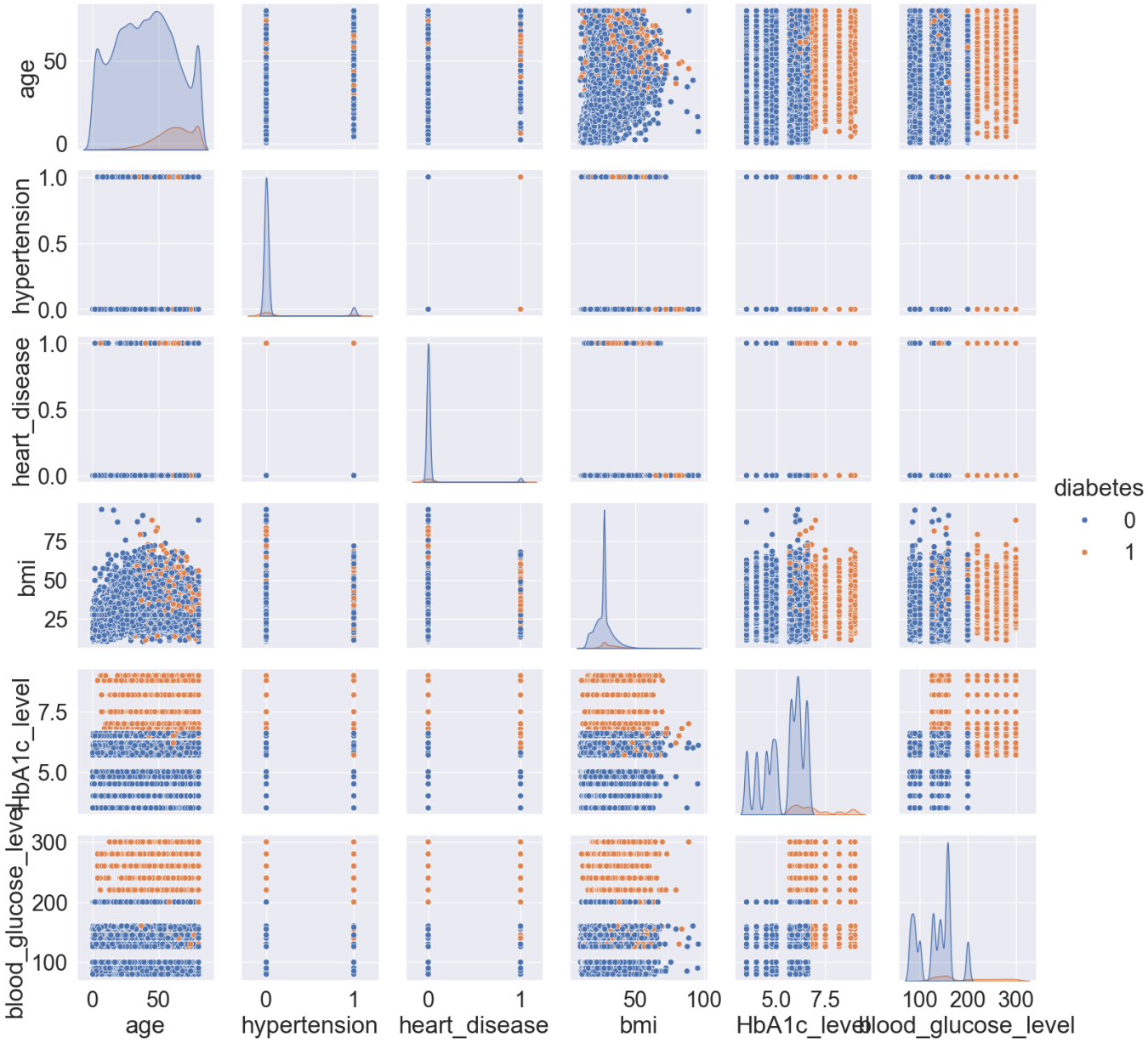

Fig. 7. Pair plot used in the EDA to analyze data

To address these issueswe undertook a thorough preprocessing phase to prepare the data for model training. The preprocessing steps involved removing duplicates and irrelevant categories such as ”other” for gender and ”No Info” for smoking history. The team then transformed categorical variables into numerical representations using dummy variables for smoking history and gender.

Outliers were identified and removed based on the Interquartile Range (IQR). Finally, to mitigate the imbalance in the dataset, we employed the Synthetic Minority Over-sampling Technique (SMOTE) for oversampling, ensuring a balanced representation of diabetic and non-diabetic instances.

Below you can see the new columns and correlation after the pre-processing steps carried out:

C. Model Selection

Based on the previous research carried out, we delved into several machine learning algorithms, all readily available within the scikit-learn package: Random Forest, Gradient Boost, KNN, and Logistic Regression. These algorithms, adept at classification tasks, were pivotal in addressing the problem statement.

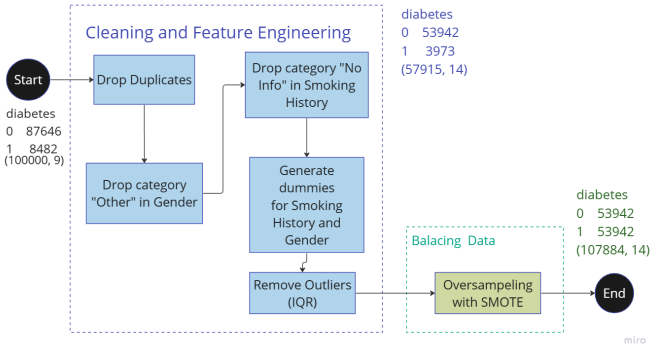

Fig. 8. The diagram depicts the pre-processing workflow, outlining how observations are handled for each binary classification.

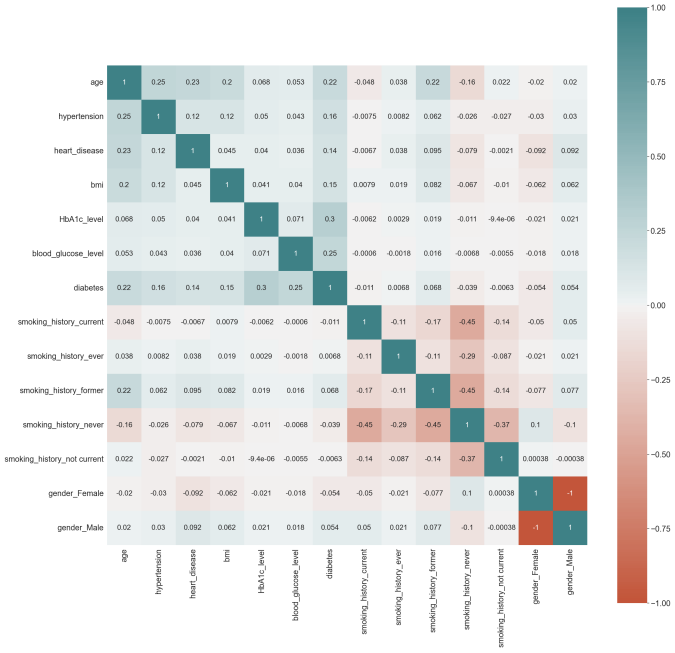

Fig. 9. Result of correlation metrics after the pre-processing

Random Forest is a group of learning techniques that combines many decision trees to make predictions. Each tree is trained on a different subset of the data, and the final prediction is based on the majority vote (classification) of the individual tree predictions. The main reason to select this algorithm is because it is robust and scalable. It handles high-dimensional data and is less prone to overfitting. It is a nice algorithm to use on features with mixed types.

Gradient Boosting builds models in sequence, where the new model fixes the mistakes of the previous ones. It optimizes the loss function by adding weak learners (most used decision trees). Furthermore, It performs in predictive accuracy and it can deal with complex interactions between features. However, it can be expensive, because it uses a lot of computing resources and requires careful tuning to prevent overfitting.

KNN classifies new data points based on their similarity to existing data points. The “neighbors” (closest data points)

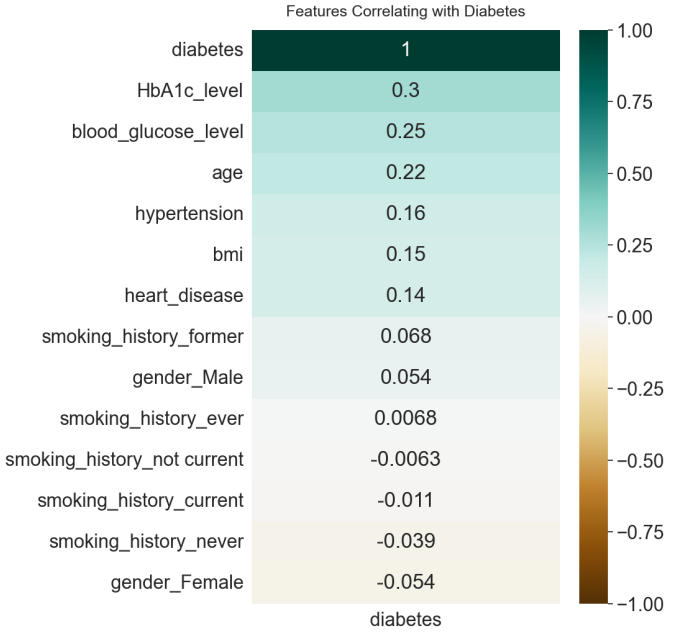

Fig. 10. The image depicts the relationship between the independent and dependent variables.

influence the classification. Its simplicity makes it effective with small datasets. It adapts well to complex or nonlinear decision boundaries. However, its performance may degrade as the feature space grows, and prediction can be slow for large datasets.

Logistic Regression is a statistical model that makes the probability of an instance belonging to a particular class. A linear model uses a logistic function to map input features to probabilities. It is efficient when the relationship between features and the target variable is approximately linear.

D. Model Training

The dataset underwent a partition into a training set comprising 80%, a test set comprising 20%, and a validation set derived from 10% of the training set. The chart below provides a visual breakdown of these sets in terms of both the percentage of the total dataset and the number of observations:

The training set served as the foundation for model training, while the test set and validation set were utilized for model evaluation.

During the training phasewe meticulously crafted hyperparameters tailored to each model. Additionally, we applied StandardScaler from Scikit-learn to the training dataset, enhancing its optimization. Subsequently, we employed GridSearchCV to meticulously traverse through hyperparameter space for each model, ultimately identifying the optimal parameters. This meticulous process ensured that each model was trained with its best possible configuration.

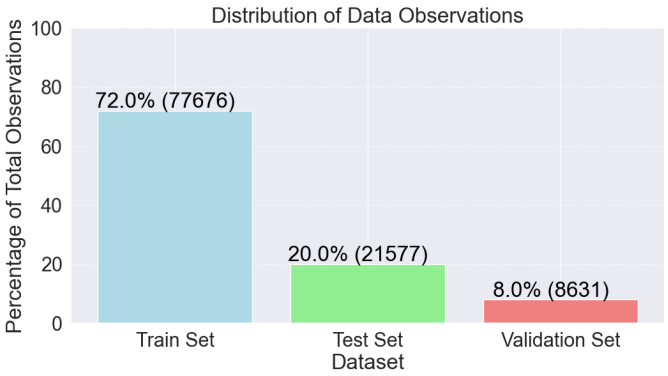

Fig. 11. The bar chart shows the distribution in percentage and observation counts.

TABLE II
BEST PARAMETERS FOR MODELS

Model	Best Parameters
Gradient Boost	{'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 150}
Random Forest	{'max_depth': None, 'n_estimators': 300}
KNN	{'n_neighbors': 7, 'p': 1, 'weights': 'distance'}
Logistic Regression	{'C': 100, 'penalty': 'l2'}

After training the model, we can evaluate its performance using a concrete example, as described below:

Input Features:

- Age: 45 years
- Gender: Female
- BMI: 30 kg/m
- Hypertension: Yes
- Heart Disease: No
- Smoking History: Never
- HbA1c Level: 6.5
- Blood Glucose Level: 150 mg/dL

To process the input features, We needed to preprocess them. Categorical values such as gender and smoking history are encoded into integers, generating dummy variables for each category. For gender, we assigns a value of 1 for the input gender (female) and 0 for the other gender (male). Similarly, for smoking history, we created dummy variables for each category, setting the input history (never smoked) to 1 and the others to 0. For hypertension and heart disease, which are binary columns indicating presence or absence, we converts them into 0 and 1 accordingly. The remaining features are kept in float format.

After preprocessing, the input features are transformed as follows:

- Age: 45
- Gender Female: 1
- Gender Male: 0
- BMI: 30
- Hypertension: 1
- Heart Disease: 0
- Smoking History Former: 0
- Smoking History Ever: 0
- Smoking History Not Current: 0
- Smoking History Current: 0
- Smoking History Never: 1
- HbA1c Level: 6.5
- Blood Glucose Level: 150.00

Finally, the model classifies whether a patient is having diabetes or not for the input patient, assigning a value of 1 or 0. In this scenario, the prediction was 0, indicating that it is not classified as having diabetes.

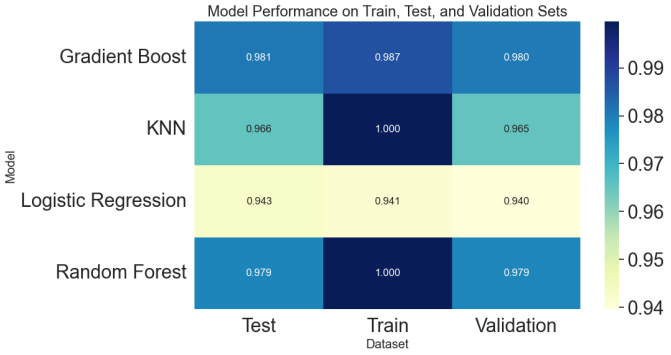

Fig. 12. The heatmap shows the accuracy for train, test and validation set

IV. EXPERIMENTAL EVALUATION

Within this chapter, the experimental evaluation will be covered. This is considered to be the main body of the machine learning project performed. Therefore, in the first subsection, the dataset which is used will be described. Next, the experimental settings will be elaborated on. In the third paragraph, the results of the machine learning models will be provided. The last sub-section covers a discussion paragraph regarding the results provided.

A. Dataset Description

Within this paragraph, the dataset used to train the machine learning model is described. This description includes information regarding the general information of the dataset. Moreover, the features provided in the dataset are elaborated and their corresponding properties and domain knowledge are discussed in this paragraph.

The dataset used to train, test, and validate the machine learning model is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose levels. To ensure confidentiality and privacy, detailed information regarding the dataset its specific location or continent, and the date of collection may not be publicly available or shared. The dataset contains approximately 100.000 rows of data.

Eight of the nine features covered within the dataset could potentially be useful for creating a classification machine learning algorithm. The ninth feature provides the diabetes

status. The domain knowledge for the other eight features and their properties are provided below:

- **Age**: One of the useful factors in predicting diabetes risk is expected to be age. As patients get older, their risk of developing diabetes increases simultaneously. Factors such as lower physical activity, changes in hormone levels, and a higher likelihood of developing other health conditions that contribute to diabetes are the main reasons for this. Within the dataset, age ranges from 0 to 80 years as a float value.
- **Gender**: Gender could also play a role in diabetes risks. For example, for women, gestational diabetes (diabetes during pregnancy) results in a higher risk of developing chronic diabetes. Additionally, other studies have suggested that men may have a slightly higher risk of diabetes compared to women. In the dataset used, gender exists of the following three categorical values: "Male", "Female" and "other".
- **Body Mass Index (BMI)**: This measures an indication of body fat based on a patient its height and weight. It is generally used as an indicator of overall weight status and is therefore helpful in predicting diabetes risks. Higher BMI is associated with a greater chance of developing diabetes. Too much body fat around the waist leads to insulin resistance and impair the ability to regulate blood sugar levels. In the dataset, this value ranges from 10.0 to 95.7 and is a float value
- **Hypertension**: Hypertension/high blood pressure, often coexists with diabetes. The two conditions share common risk factors and can contribute to the development of each other. Meaning, hypertension increases the risk of developing diabetes and vice versa. In the dataset, a Boolean value indicates whether a patient also has hypertension (1), or not (0).
- **Heart Disease**: Heart diseases in general are associated with an increased risk of diabetes as well. The relationship between heart diseases and diabetes is bidirectional such as with hypertension. So, having one of the condition increases the risk of developing the other because they share common risk factors. These are obesity, high blood pressure, and high cholesterol. In the dataset, a Boolean value indicates whether a patient also has heart diseases (1), or not (0).
- **Smoking History**: Smoking also is expected to influence the risk factor of diabetes. Smoking contributes to insulin resistance and impair glucose metabolism. However, it should be noted that quitting smoking significantly reduce the risks. In the dataset used, smoking history is indicated categorically in 6 distinct values ranging from "Current" to "Never".
- **HbA1c Level**: HbA1c (glycated hemoglobin) measures the average blood glucose level over the past 2 to 3 months. It provides information about long-term blood sugar control. Higher HbA1c levels indicate poorer glycemic control which is then again associated with an increased risk of diabetes. In the data used, this value ranges from 3.5 to 9.0 and is an float value.
- **Blood Glucose Level**: The amount of glucose/sugar present in the blood at a fasting state or after consuming carbohydrates, can indicate impaired glucose regulation and therefore also indicates higher risk of diabetes. In practice regular monitoring of blood glucose levels is important in the diagnosis and management of diabetes. In the dataset used, this value ranges from 80 to 300, and is an integer value.

All the features explained are expected to contribute to classifying whether a patient has diabetes or not. Especially when combining the features and analysing them with appropriate statistical and machine learning techniques.

B. Experimental Settings

Evaluation Criteria:

The evaluation metrics the team considered using in order to assess the performance of the predictive models are F1-score and the number of False Negatives predicted. The primary objective was to achieve a high F1-score while minimizing the occurrence of False Negatives. This emphasis on reducing False Negatives is paramount due to the critical nature of accurately identifying individuals at risk of diabetes. Misclassifying a person with a potential risk of diabetes as low risk could have significant consequences, underscoring the importance of the endeavor.

Specific Hypotheses Tested:

The research tests hypotheses related to the efficiency of machine learning algorithms in improving the diagnostic accuracy of diabetes in a clinical setting, specifically:

- Machine learning algorithms can accurately classify diabetes status, thus aiding general practitioners in diagnosing the condition more efficiently.
- Preprocessing techniques and hyperparameter tuning significantly enhance model performance.

Experimental Methodology:

The methodology follows the CRISP-DM framework and includes:

- Data Collection and Understanding: Identifying relevant features and anomalies within the diabetes dataset.
- Data Preparation: Implementing preprocessing steps such as handling missing values, removing duplicates, and encoding categorical variables.
- Model Building and Tuning: Training various machine learning models including Random Forest, Gradient Boosting, KNN, and Logistic Regression using the scikit-learn package. Hyperparameter tuning is conducted to optimize each model.
- Evaluation: Using a split dataset approach (training, testing, and validation sets), the models are evaluated based on the predefined metrics.

Dependent and Independent Variables:

- Dependent Variable: Diabetes status (binary classification: positive or negative).

- Independent Variables: Age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose levels.

Training and Test Data:

The Diabetes prediction dataset used was sourced from the Kaggle repository. The dataset has 100,000 rows of data, making it sufficiently large for robust machine-learning analysis. The dataset includes medical and demographic data from patients regarding their diabetes status, along with various health-related features. The data's richness and its comprehensive range of features relevant to diabetes make it a realistic and interesting choice for developing a predictive model.

In the preprocessing phase, the Synthetic Minority Oversampling Technique (SMOTE) was utilized to address class imbalance, enhancing the representativeness of minority classes. Subsequently, the data was divided into distinct sets for model training and evaluation: 80% was allocated as the training set, 20% served as the test set, and a validation set constituted 10% of the training set, corresponding to 8% of the total dataset. This partitioning strategy was designed to optimize the training process and ensure a comprehensive evaluation of the model's predictive performance.

Performance Data Collected:

The study collects performance data such as accuracy, precision, recall, F1-score, true positive (TP), false positive (FP), false negative (FN), and true negative (TN) across multiple models. The performance is visualized using heatmaps and other graphical representations to compare how each model performs on the test data.

Analysis and Presentation:

Performance data is analyzed using statistical methods and machine learning evaluation metrics. Results are presented through visualizations like confusion matrices and heatmaps, which highlight the efficiency of each model in various metrics, allowing for an easy comparison of their predictive capabilities.

Comparison to Competing Methods:

The research references several other studies and methodologies for diabetes prediction using machine learning and their performance metrics and methods. For example, previous works utilizing algorithms like Logistic Regression and Gradient Boosting are mentioned, with noted accuracies, allowing for a direct comparison of the newly developed models against existing benchmarks. This comparison helps to position the current study within the broader context of ongoing research and highlights the advancements or improvements made by the current approach.

This comprehensive setup not only aims to validate the effectiveness of machine learning models in clinical diabetes prediction but also enhances the workflow of general practitioners by potentially reducing diagnostic times and increasing accuracy.

C. Results

After training the different machine learning models, we conducted extensive evaluations using the best model identified through GridSearchCV, utilizing a separate test set for predictions. In this section, various visualizations and metrics to thoroughly assess and discuss the model results are presented.

All models were assessed using standard evaluation metrics such as precision, recall, f1 score, and accuracy. However, each model possesses unique characteristics that provide deeper insights beyond these scores. For instance, delving into feature importance and analyze the AUC curve specifically for the regression model.

Additionally, we'll highlight the models' ability to accurately predict both negative and positive cases of diabetes. It's crucial to emphasize the significance of minimizing false negatives, particularly in the healthcare sector, as they can pose significant patient risks.

The heatmap below illustrates the performance of all models across precision, recall, F1-score, and accuracy, with colors representing scores from lower to higher values. All models achieve satisfactory scores, meeting the requirement of achieving over 90% accuracy as stated in the problem statement.

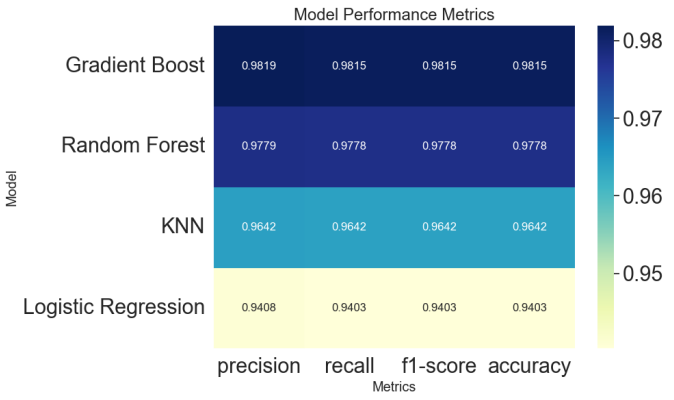

Fig. 13. The heatmap showcases the performance of all models across various metrics.

It is crucial to grasp not only the overall scores but also how performance translates into predicting yes or no for diabetes classification. The heat maps below illustrate the scores for the models predicting positive or negative outcomes for diabetes.

The team noted improved scores in predicting false outcomes compared to the overall heat map performance. Additionally, the 'recall' metric shows a significant increase compared to other metrics. On the other hand, for label 1, there is a decline in recall performance but an increase in precision. Finally, the F1-score remains consistent with the overall heat map in both cases, maintaining balance.

The team created confusion matrix charts for each model to see how well they deal with false and true positives. Within these visualizations, we can check in detail how the models perform with false negatives.

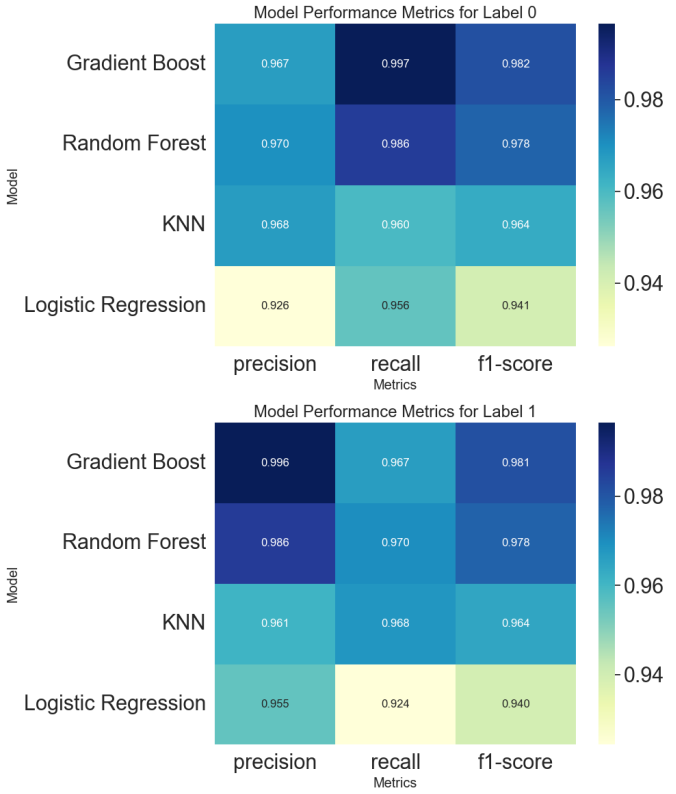

Fig. 14. The heatmap showcases the performance of all models for labels 0 and 1 across various metrics.

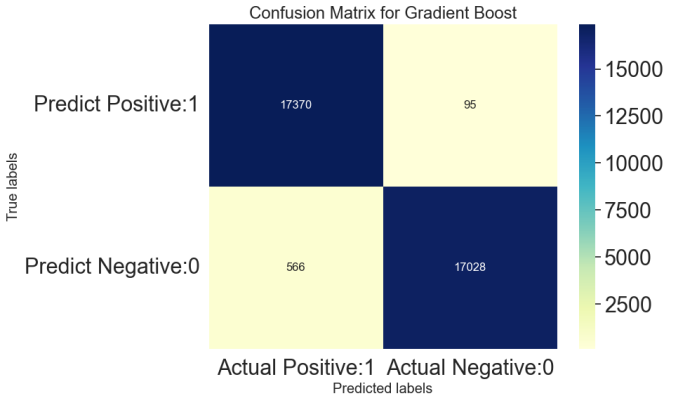

Fig. 15. Confusion Matrix of Gradient Boost Model Results

Additionally, we assessed the feature importance across all models except KNN. Both Gradient Boost and Random Forest models exhibited a consistent ranking of importance, albeit with varying magnitude.

Conversely, logistic regression yielded a distinct ranking, identifying the male feature as the second most important. Despite metrics suggesting relatively low importance compared to conventional high scores, a deeper dive into individual features unveils valuable insights. Furthermore, the important results align with the business expertise regarding the process and its features.

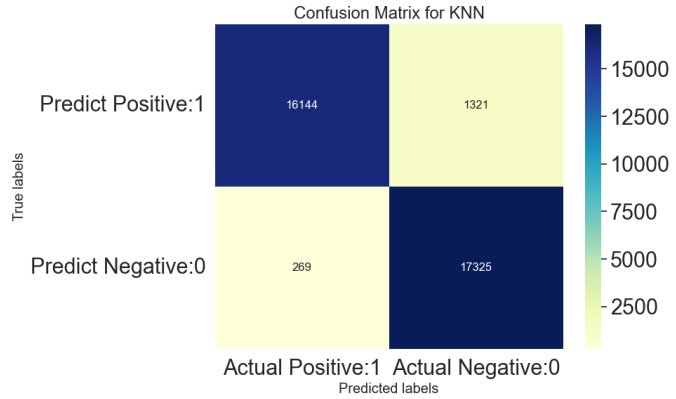

Fig. 16. Confusion Matrix of KNN Model Results

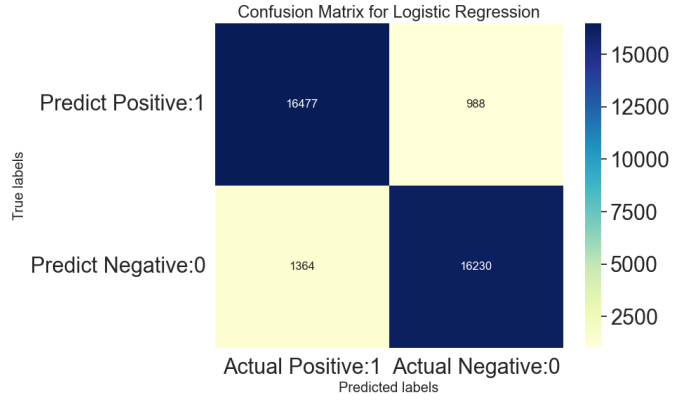

Fig. 17. Confusion Matrix of Logistic Regression Model Results

In addition to logistic regression, we examined two additional metrics: R-squared (R^2) and the Receiver Operating Characteristic (ROC) curve.

The R^2 score (coefficient of determination), quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. In the analysis, the R^2 score was calculated to be 0.77, indicating a strong level of explanatory power in the logistic regression model.

The ROC curve is a graphical plot that illustrates the performance of a binary classification model across various threshold settings. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for different threshold values. By analyzing the ROC curve, we can assess the trade-off between sensitivity and specificity for the logistic regression model.

With an ROC curve value of 0.99, the model demonstrates high discriminatory power, exhibiting a high true positive rate and low false positive rate across various threshold settings. This suggests that the model performs exceptionally well in distinguishing between the positive and negative classes, making it highly effective for binary classification tasks. Below, you can find the ROC curve plot, providing further insights into the performance of the logistic regression model.

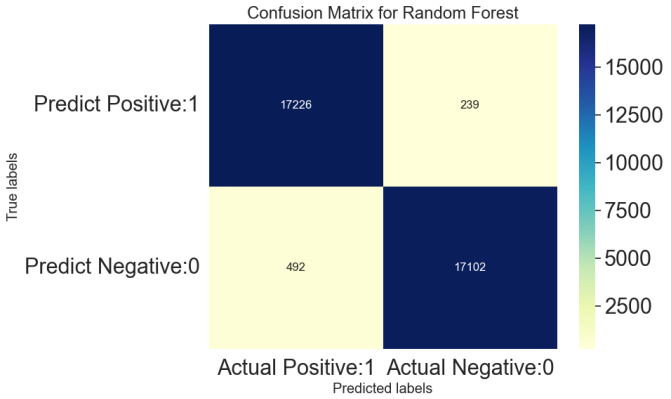

Fig. 18. Confusion Matrix of Random Forest Results

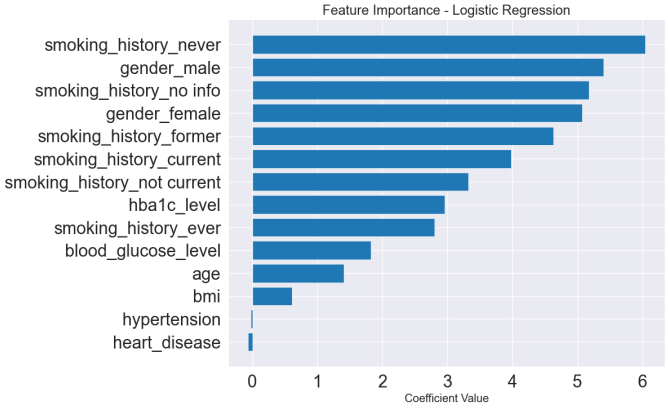

Fig. 20. Feature Importance for Logistic Regression

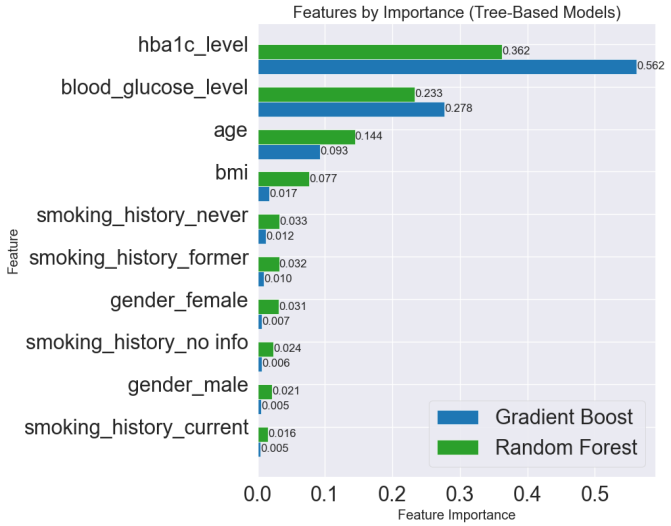

Fig. 19. Feature importance Gradient Boost and Random Forest

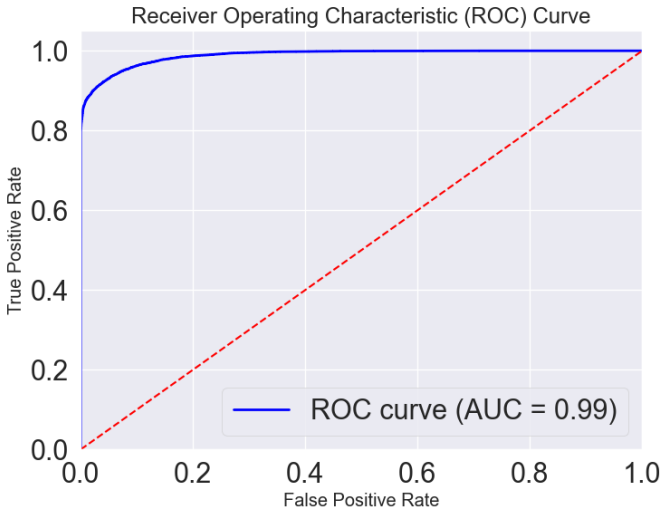

Fig. 21. ROC Curve for Logistic Regression

D. Discussion

After providing the results in the previous section, a discussion regarding these results will be provided within this section. The goal of this section is to provide a discussion that elaborates on the conclusions the results support about the strengths and weaknesses of the machine learning project. It will also tell how the results can be explained in terms of the underlying properties of the project. Within this project, the main goal of this machine learning project was to create a machine learning algorithm that classifies whether a patient visiting a GP has diabetes or not. Considering this, a model is created which meets the standards set. However, as with every project, also this project has weaknesses within the project. The main features that are used by the Gradient Boosting algorithm are the HB1AC and the Blood Glucose level. Together, they share an importance of 84 percent of the model. This is also visible in the figure showing the feature importance of the Gradient Boosting algorithm (figure?). This is considered as the main weakness of this machine-learning project. A machine learning model predicting if a patient has diabetes based on simple demographic data is considered much more valuable. Initially, as stated in the problem statement, creating a machine-learning model based on rather simple demographic data was part of the goal of the project. This would generate the biggest potential for decreasing the workload of GPs. The transformation of the machine learning model used features from only a demographic-based approach to also including HB1AC and blood glucose levels showing the nature of research and innovation in diabetes diagnosis. Still, a lot is to be researched and discovered regarding the chronic disease diabetes, which is getting more common generally. While a machine learning model based on also HB1AC and Blood Glucose levels represents a significant advancement in accuracy, it is important to acknowledge the potential value of a model based on demographic data for such projects. A machine learning model that identifies the risk of developing diabetes based on only demographic data could lower the workload of a GP much more, complementary to machine learning models which also need data on Blood Glucose and HB1AC. It is expected that, to build a well-performing

machine learning model with demographic data only, more features regarding demographic data are needed. Demographic factors such as age, gender, BMI, and smoking history but also other factors such as socioeconomic status and lifestyle behaviors play a crucial role in determining a patient its risk of having diabetes. By incorporating these factors into the diagnostic process, a demographic-based model could provide an assessment of diabetes risk as well.

V. Conclusion

The research conducts an in-depth investigation into the deployment of machine learning models to support general practitioners (GPs) in the diagnosis of diabetes. This study stands out by achieving a significant milestone with a Gradient Boosting model that reached an accuracy rate of **98%**, highlighting the robust potential of machine learning applications in healthcare. The following sections provide an elaborate discussion on the study's results, its critical insights, and its implications for future research and medical practice.

A. Detailed Results and Conclusions

This research achieved impressive outcomes, characterized by high predictive accuracy which underscores the efficacy of machine learning in medical diagnostics. The study's major achievements are summarized below:

• **Advanced Data Preprocessing**: By employing the Synthetic Minority Over-sampling Technique (SMOTE), the study effectively tackled the prevalent issue of class imbalance in medical datasets. This approach ensured a more equitable representation of diabetic and non-diabetic cases, thereby enhancing the model's learning accuracy.

• **Comprehensive Model Evaluation**: The Gradient Boosting model was assessed through various metrics, including precision, recall, and F1-score. These metrics were instrumental in verifying the model's high capability in accurately diagnosing diabetes with a notably low rate of false negatives. Reducing false negatives is critical in medical diagnostics, where missing a diagnosis can lead to severe complications for patients.

B. Significant Insights from the Study

The research elucidates several key insights into the integration of machine learning within healthcare diagnostics:

• **Enhancement of Diagnostic Accuracy and Efficiency**: The findings demonstrate the transformative potential of machine learning in refining diagnostic processes. Machine learning models, such as the one developed in this study, can significantly speed up the diagnostic process, reduce errors, and thus enhance the overall efficiency of medical services.

• **Operational Impact on General Practitioners**: The implementation of the model promises to substantially reduce the workload of GPs by automating the initial diagnostic procedures. This development allows GPs to dedicate more attention and resources to patient care and complex case management.

• **Robust Methodological Framework**: The structured approach adopted from the CRISP-DM framework exemplifies a rigorous methodological process in developing and evaluating predictive models. This comprehensive approach ensures that the models are not only accurate but also reliable and scalable across various healthcare settings.

C. Future Research and Applications

The implications of this study are extensive, setting a precedent for future technological advancements in healthcare:

• **Innovation in Predictive Modeling**: The success of the Gradient Boosting model provides a foundation for exploring other sophisticated machine learning algorithms. Future research could investigate hybrid models or advanced ensemble techniques that could potentially yield even higher accuracies and robustness in predictions.

• **Clinical Deployment and Expansion**: The study's insights pave the way for integrating these models into clinical practice, which could revolutionize the diagnosis and management of not only diabetes but also other chronic diseases.

• **Widening Scope of Machine Learning in Healthcare**: The methodology and successful application of machine learning demonstrated by this study encourage the broader adoption of these technologies across different healthcare domains. This could lead to innovations in diagnostic processes for a range of diseases, personalized treatment plans, and improved patient outcomes.

In conclusion, this research not only showcases the effectiveness of machine learning models in enhancing the diagnostics of diabetes but also sets a scalable framework for integrating advanced data-driven technologies in healthcare. The potential of these technologies to transform diagnostic accuracy and efficiency promises significant advancements in the management of chronic diseases, thereby improving patient care and healthcare delivery on a global scale.

VI. Future work

First of all, the current model performs well in a generic way from the patients' point of view. The main future work would be to analyze the data from patients in groups and develop a model for each group, with this solution we can have more precise results and give more context to the model.

We can utilize cluster algorithms such as KNN to partition the existing dataset into distinct groups. Subsequently, we can conduct exploratory data analysis within each cluster individually, following the same machine-learning methodology outlined in this report.

The challenge with this approach lies in implementing distinct pre-processing steps for each group. Several factors contribute to this complexity, including insufficient data for model training in certain groups and the inadequacy of current data features for specific groups. For instance, in the case of a teenager group, considerations such as sports routines and genetic history may be pivotal, given that teenagers typically have limited visits to healthcare facilities.

This approach will yield more specialized models tailored for real-world applications, enhancing their relevance and expanding their utility within business models. In addition,

create a percentage of risk instead just saying positive or negative.

VII. Abbreviations and Acronyms

References

[1] A. Mahabub, 'A robust voting approach for diabetes prediction using traditional machine learning techniques', SN Appl. Sci., vol. 1, no. 12, p. 1667, Nov. 2019, doi: 10.1007/s42452-019-1759-7.

[2] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, 'An ensemble learning approach for diabetes prediction using boosting techniques', Front. Genet., vol. 14, Oct. 2023, doi: 10.3389/fgene.2023.1252159.

[3] Li, Mingqi, Xiaoyang Fu, and Dongdong Li. "Diabetes Prediction Based on XGBoost Algorithm." IOP Conference Series: Materials Science and Engineering 768, no. 7 (March 2020): 072093. https://doi.org/10.1088/1757-899X/768/7/072093.

[4] Lai, Hang, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao. "Predictive Models for Diabetes Mellitus Using Machine Learning Techniques." BMC Endocrine Disorders 19, no. 1 (October 15, 2019): 101. https://doi.org/10.1186/s12902-019-0436-6.

[5] Mushtaq, Zaigham, Muhammad Farhan Ramzan, Sikandar Ali, Samad Baseer, Ali Samad, and Mujtaba Husnain. "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques." Mobile Information Systems 2022 (March 19, 2022): e6521532. https://doi.org/10.1155/2022/6521532.

[6] Tasin, Isfafuzzaman, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. "Diabetes Prediction Using Machine Learning and Explainable AI Techniques." Healthcare Technology Letters 10, no. 1–2 (December 14, 2022): 1–10. https://doi.org/10.1049/htl2.12039.

[7] 'How have diabetes costs and outcomes changed over time in the U.S.?', Peterson-KFF Health System Tracker. Accessed: May 12, 2024. [Online]. Available: https://www.healthsystemtracker.org/chart-collection/diabetes-care-u-s-changed-time/

[8] 'What Is Diabetes? - NIDDK'. Accessed: May 12, 2024. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[9] 'Understanding the work of general practitioners: a social science perspective on the context of medical decision making in primary care — BMC Primary Care — Full Text'. Accessed: May 12, 2024. [Online]. Available: https://bmcprimcare.biomedcentral.com/articles/10.1186/1471-2296-9-12

[10] D. Duong and L. Vogel, 'Overworked health workers are "past the point of exhaustion"', CMAJ, vol. 195, no. 8, pp. E309–E310, Feb. 2023, doi: 10.1503/cmaj.1096042.

[11] 'Global diabetes data report 2000 — 2045'. Accessed: May 12, 2024. [Online]. Available: https://diabetesatlas.org/data/

[12] 'Patient registration - RCM Glossary — MD Clarity'. Accessed: May 12, 2024. [Online]. Available: https://www.mdclarity.com/glossary/patient-registration

[13] 'sklearn.ensemble.RandomForestClassifier — scikit-learn 1.4.2 documentation'. Accessed: May 12, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[14] 'sklearn.neighbors.KNeighborsClassifier', scikit-learn. Accessed: May 12, 2024. [Online]. Available: https://scikitlearn/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[15] 'sklearn.ensemble.GradientBoostingClassifier', scikit-learn. Accessed: May 12, 2024. [Online]. Available: https://scikitlearn/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[16] 'sklearn.linear$_m$odel.$LogisticRegression'$, $scikit-learn.Accessed$: $May12, 2024.[Online].Available$: $https$: $//scikitlearn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html$