ISSN: 2410-7727

Optimal Using of Machine Learning Algorithms Hyperparameters for Diabetes Prediction

الاستخدام الأمثل لمعاملات خوارزميات تعلم الآلة للتنبؤ بمرض السكري

Ismail I. Al-Ahmed[a],Yousif A. Al-Haj[b],

Marwan M. Al-Falah[c], Khadeja M. Al-Nashad[c], Naif M. Al-Falah[d]

a. Al-Saeeda University, Faculty of Engineering & Information Technology, Sana'a Yemen; ism116620@gmail.com.
b. Sana'a University, Faculty of Education, Humanities & Applied Science, Sanaa, Yemen; yalhag@gmail.com (Y.A.A).
c. Knowledge & Modern Science University (KMSU), Faculty of Information Technology & Engineering, Sanaa, Yemen.
d. Azal University for Human Development, Faculty of Information Technology & Engineering, Sanaa, Yemen.

2003
١٤٢٤
جامعة الأندلس
للعلوم والتقنية
Alandalus University For Science & Technology

Optimal Using of Machine Learning Algorithms Hyperparameters for Diabetes Prediction

Abstract:

Diabetes Mellitus (DM) is a growing health concern worldwide due to its widespread prevalence and chronic nature. Early diagnosis is crucial for effective management and improving patient outcomes. There are two forms of DM, type 1 which presents symptoms, and type 2, which is often asymptomatic in its early stages, making early detection challenging. To address this challenge, the Pima Indian Diabetes Dataset (PIDD) was utilized and employed machine learning algorithms including Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). A cross-validation model with K-fold stratified cross-validation equal to 10 was employed to evaluate the performance of the algorithms. Furthermore, hyperparameter tuning was performed to optimize the performance of the models. Our results showed that the logistic regression algorithm had the highest accuracy with a value of 79%. This study highlights the potential of using machine learning algorithms in the early detection and diagnosis of DM, especially in cases where traditional methods may be limited. Also the results of this study demonstrate the importance of hyperparameter tuning in improving the performance of machine learning algorithms for medical applications. Where the results of this study highlight the significant impact of hyperparameter tuning and feature engineering techniques on improving the accuracy of prediction models for diabetes. It is worth noting that the initial algorithms used in this study performed less effectively prior to the implementation of these techniques. These findings underscore the importance of careful algorithm tuning and advanced feature engineering in enhancing the efficacy of machine learning models for diabetes prediction. These results have important implications for the development of more accurate and reliable prediction models, which can aid medical professionals in providing timely diagnosis and effective treatment to patients with diabetes.

Keywords: Machine learning algorithms, Diabetes Mellitus (DM), Pima Indian Diabetes Dataset, Hyperparameter tuning.

الاستخدام الأمثل لمعاملات خوارزميات تعلم الآلة للتنبؤ بمرض السكري

د. إسماعيل إبراهيم الأحمد ، د. يوسف احمد الحاج سلطان،
مروان محمد مهدي الفلاح، خديجة محمد محمد النشاد ، نايف محمد الفلاح

الخلاصة:

يُعد مرض السكري (Diabetes Mellitus) قضية صحية متزايدة الانتشار عالمياً نظراً لانتشاره الواسع وطبيعته المزمنة. يُعتبر التشخيص المبكر له أمراً بالغ الأهمية لإدارة الحالة بفعالية وتحسين نتائج المرضى. ينقسم مرض السكري إلى نوعين: النوع الأول الذي يُظهر أعراضاً واضحة، والنوع الثاني الذي يكون عادةً بلا أعراض في مراحله المبكرة، مما يجعل الكشف المبكر عنه أمراً صعباً. من أجل التغلب على هذا التحدي، تم استخدام مجموعة بيانات بيما الهندية لمرض السكري (Pima Indian Diabetes Dataset) وتم استخدام خوارزميات تعلم الآلة، ومنها (Support Vector Machine)، و (Random Forest)، و (Logistic Regression). تم استخدام نموذج التقييم المتقاطع، مع تقسيم البيانات إلى عشرة مجموعات (K-fold stratified cross-validation) لتقييم أداء الخوارزميات. علاوة على ذلك، تم تنفيذ ضبط المعاملات (Hyperparameter tuning) لتحسين أداء هذه النماذج. لقد أظهرت النتائج أن خوارزمية الانحدار اللوجستي حققت أعلى معدل دقة بنسبة 79%. لذا تسلط هذه الدراسة الضوء على إمكانية استخدام خوارزميات تعلم الآلة في الكشف المبكر والتشخيص لمرض السكري، خاصة في الحالات التي تكون فيها الطرق التقليدية محدودة. كما توضح النتائج أهمية ضبط المعاملات لتحسين أداء خوارزميات تعلم الآلة في تطبيقات الطب. كما تؤكد هذه الدراسة أيضًا على تأثير ضبط المعاملات وتقنيات هندسة السمات المتقدمة في تحسين دقة نماذج التنبؤ بمرض السكري. إن الخوارزميات الأولية التي استخدمت في هذه الدراسة كانت أقل فعالية قبل تنفيذ هذه التقنيات. إن نتائج هذه الدراسة تبين أهمية ضبط معاملات الخوارزميات بعناية وتطوير تقنيات هندسة السمات المتقدمة لتعزيز فاعلية نماذج تعلم الآلة في التنبؤ بمرض السكري. كما تشير هذه النتائج إلى أهمية تطوير نماذج تنبؤ أكثر دقة وموثوقية، مما يُمَكِن الأطباء المهنيون من تشخيص مرض السكري في الوقت المناسب وتوفير العلاج الفعال للمرضى

I. INTRODUCTION

The world suffers from many medical problems and among the most serious diseases is DM because it causes other diseases in the patient's body, including(heart, nerve damage, blindness of the eyes, and kidney failure). The International Diabetes Federation reported 6.7 million deaths worldwide among adults with diabetes in 2021, accounting for 12.2% of all deaths globally [15]. DM is described as the greatest public health challenge as it is long-term and cannot be ended [7]. DM is divided into two types, the first type, is dependent on insulin and is produced due to a lack of insulin and requires taking insulin daily, and this type appears in children or people before the age of 30 years, the second type is not dependent on insulin and its symptoms do not appear in its early stages, so it is difficult to detect it Except in advanced stages, it arises due to obesity and lethargy, and it appears in adults and the elderly, and in high-income areas [6]. Due to the emergence of artificial intelligence and its high ability to predict diseases, ML algorithms used that depend on collecting information about DM (big data). Research has shown that the second pattern is what helped ML to predict the disease and make appropriate decisions [1]. In recent years, researchers have tended to develop ML-based systems for clinical decision support for diabetics. According to the International Diabetes Federation (IDF), the number of DM in 2045 AD will reach about 783 million. The reason for the spread of diabetes in recent years is the tendency of people to overeat sugars and the tendency of people towards independent work [10]. The causes of the disease are the inability of the pancreas to produce a sufficient amount of insulin. This is in the first type. As for the second type, its causes are because the body cannot benefit from the insulin produced. Among the symptoms of the disease are severe hunger and thirst and frequent urination. If the level of glucose in the blood is more than 126 mg / dL This refers to diabetes, as there is no permanent treatment for DM [7].

In this study, the aim was enhancing the accuracy of diabetes prediction using machine learning algorithms. To achieve this, the Pima Indian Diabetes Dataset (PIDD) was utilized and employed three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). The performance of these algorithms was optimized through hyperparameter tuning using methods such as GridSearchCV. K-fold stratified cross-validation to cross-validate the model was used and compared the performance of the algorithms under both default parameters and optimal hyperparameters. The results of this study will provide valuable insights into the effectiveness of using machine learning algorithms and hyperparameter tuning in the early detection and diagnosis of Diabetes Mellitus.

The scientific paper was divided as follows:

The second part contains related works, the methodology in the third part, the results in the fourth part, finaly the fifth part contains conclusion and future works.

II. RELATED WORKS

Recently, researchers have published many research papers on diabetes, relying on machine learning algorithms to predict this disease. In this section, some research papers related to our manuscript were discssed, which discussed the same Pima Indian Diabetes Dataset (PIDD).

In Paper [1], the authors applied some machine learning algorithms LR, KNN, SVM, and RF where they found that the LR algorithm showed better performance than other algorithms with an accuracy of 83%, it was also concluded that glucose and BMI are strongly associated with diabetes. using association rule mining.

In the same manner as the aforementioned paper, Paper No. [2] used a group of machine learning algorithms, and they concluded that the SVM algorithm was the best among the rest of the algorithms. In addition, they developed a smart web application to predict diabetes based on their findings.

In the paper [3], the researchers used a new algorithm called the Ontology Model, and through their application they concluded that it was the best in accuracy by 77.5%, beating the LR and SVM algorithms. The difference was slight, and the authors believed that this was due to using rules extracted from machine learning algorithms and integrating them using SWRL into the ontology. Along with some improvements they've made.

In paper No. [4], four machine learning algorithms were discussed, namely ANN, RF, NB, C5.0, and SVM in predicting Diabetes Mellitus and comparing them through Metric Measures say Accuracy, Precision, Sensitivity, Specificity, and F1 Score. As a result of this work, the C5.0 algorithm and Logistic Regression were equal in the result based on the accuracy measures. While the SVM algorithm finally came after the NB and ANN algorithms, respectively.

In the paper [5], the researchers implemented eight machine learning algorithms, LR, KNN, SVM, Gradient Boost (GB), DT, The Multilayer Perception MLP, RF, and Gaussian Naïve, to predict diabetes cases within the PIDD dataset. The performance measure and comparison of these algorithms were done by MAE, RMSE, ROC, Test Accuracy, Precision, and Recall obtained from the test set. In this study, it was concluded that Logistic Regression and Gradient Boost classifiers achieved the highest accuracy of 79% compared to other classifiers.

In this paper [6], two of the dataset were used to compare each other according to the accuracy of several machine learning algorithms, models such as XG Boost, Ada Boost, Gradient Boosting, Voting, Stacking classifier, and many more classifiers are being used. For PIDD, the Voting classifier gave an accuracy of 85%. As for Sylhet Diabetes Hospital Bangladesh Dataset (SDHBD), the XG Boost and some classifiers gave an accuracy of 98%. Through hyperparameter optimization, the XH Boost classifier gave about 99% accuracy.

During research paper No. [7], the researchers used the WEKA tool for pre-processing the data set, and through the use of feature reduction, three features were dropped and five input features (Glucose, BMI, Insulin, Pregnancy, and Age), and one output feature (outcome) were relied upon in the PIMA dataset. Here seven different machine learning algorithms were used including DT, KNN, RF, AB, NB, LR, and SVM all these algorithms showed good results greater than 70%, NN model was built with different hidden layers with many epochs and it was observed that two hidden layers provide us with a higher accuracy of 88.6%.

The researchers focused in paper [8] on health care and how to raise the level of response to any emergency for patients with diabetes in real-time, as the main purpose of the study was two basic things: the first is to use the Multilayer Perceptron (MLP) based algorithm in the classification of diabetes and use the deep learning based on LSTM to predict diabetes. The second thing is to propose a diabetes monitoring system based on the Internet of Things (IoT), as the main purpose of the system is to help users monitor their vital information using Bluetooth Low Energy (BLE) based sensor devices via their cell phones. The proposed approach achieved an accuracy of 87% when assessing the PIDD dataset.

The paper [9], relied on only two algorithms, the RF and SVM algorithm, where the accuracy was 83% and 81%, respectively. Only four important features were chosen, and the dimensionality reduction increases the test set accuracy of the RF and SVM, as it is assumed that this treatment is the reason for increasing the accuracy rates for these two classifiers.

In paper [10] predicting diabetes through some algorithms that were used in most scientific papers are LR, RF, DT, SVM, KNN, and NB. Then they used two different hyperparameter techniques called Randomized Search and TPOT (autoML) to analyze the comparison shown between default and optimized parameters for each algorithm and as the best result, the RF algorithm gave a high accuracy of 98% and was assumed to be reliable in predicting diabetes.

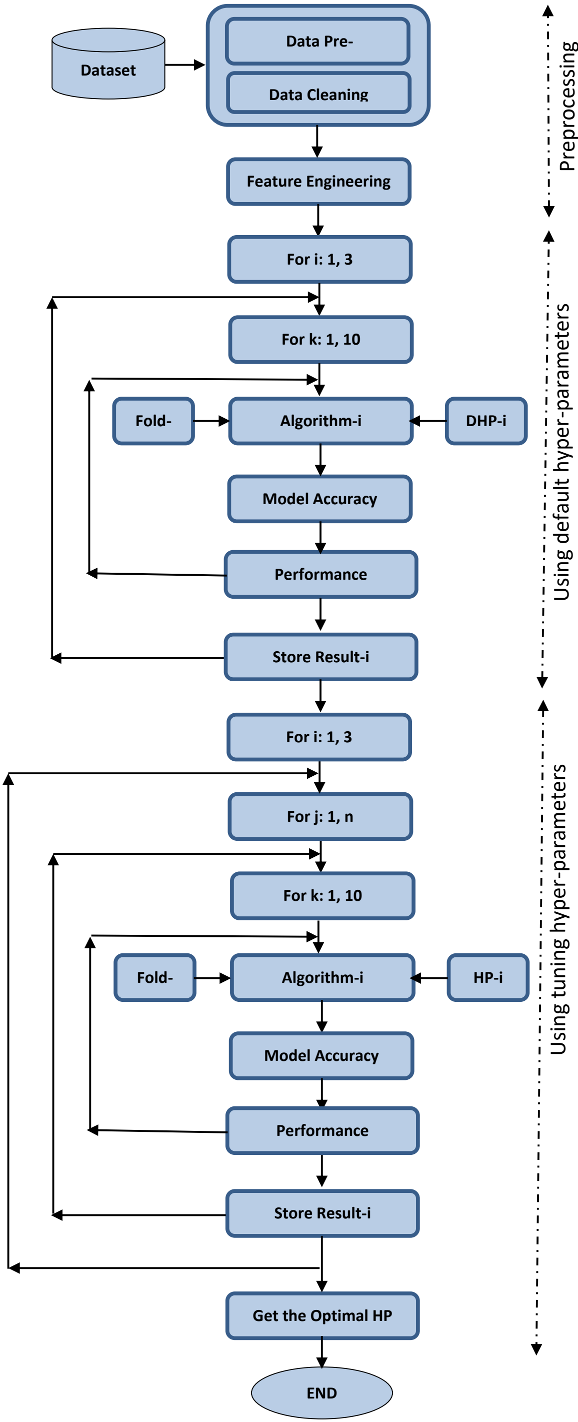
III. METHODOLOGY

The framework is divided into several different stages, as shown in the flow diagram in Figure 1. Python Jupiter Note was used for implementation. Various library packages were used such as NumPy, Pandas, Seaborn, Matplotlib, and Scikit used in data analysis, and helped us develop this research work.

The algorithm of the main model's steps:

1. Preprocessing the dataset.
2. Process the dataset by the three algorithms LR, RF, and SVMs using he default hyperparameters.
3. Store the results of step2.
4. Process the dataset by the same algorithms in step2 using the tuned hyperparameters.
5. Store the results of step4.
6. Get the optimal hyperparameters due selecting the best result for each algorithm.
7. End.

Figure 1 shows the details of the previous algorithm.



Algorithm-i: 1. Logistic Regression.

2. Random Forest.

3. Support Vector.

Fold-k: Fold Cross Validation.

DHP: Default Hyper-Parameter.

THP-i: Tuned Hyper-Parameter.

n: The number of Hyper-Parameter tunes.

Figure 1: Scheme of the model's steps.

A. Dataset Description

The Pima Indian Diabetes Dataset (PIDD) is a dataset from the UCI Machine Learning Repository from the Kaggle repository [14] that contains medical data for women of Pima Indian heritage. The dataset includes data on 768 women, with 8 features for each woman, including information on their pregnancies, blood pressure, skin thickness, insulin level, body mass index (BMI), and diabetes pedigree function. The goal of the dataset is to predict whether or not a woman has diabetes, based on the features provided in the dataset. The data was collected by the National Institute of Diabetes and Digestive and Kidney Diseases and is widely used as a benchmark dataset for classification tasks. The dataset is simple and clean and can be used as a simple benchmark to compare different algorithms and models. The "Outcome" variable in the dataset indicates whether or not a woman has diabetes. The Outcome variable has two possible values: 0 and 1.

- 0 indicates that the woman does not have diabetes.
- 1 indicates that the woman has diabetes.

500 of whom are negative and 268 are positive diabetics, respectively as proven in Fig. 2.

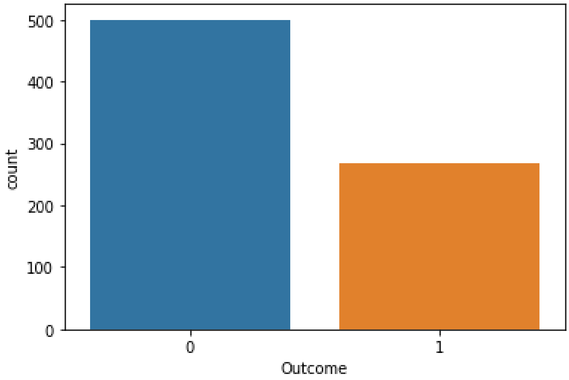
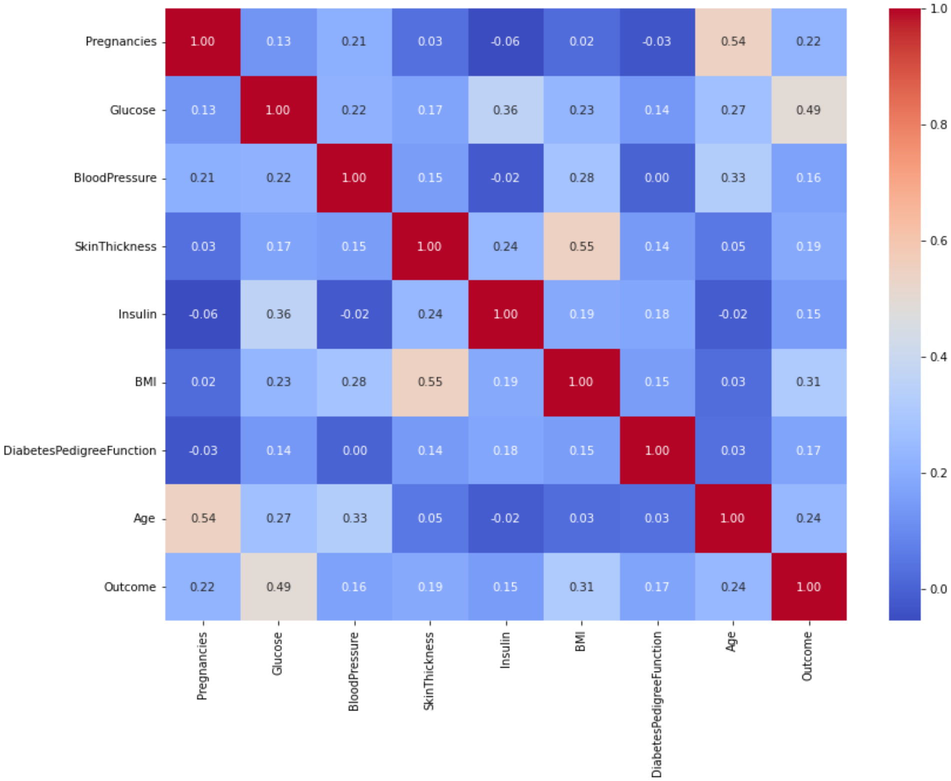


Figure 2: Number of Negative and Positive records.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.21	0.03	-0.06	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.22	0.17	0.36	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1.00	0.15	-0.02	0.28	0.00	0.33	0.16
SkinThickness	0.03	0.17	0.15	1.00	0.24	0.55	0.14	0.05	0.19
Insulin	-0.06	0.36	-0.02	0.24	1.00	0.19	0.18	-0.02	0.15
BMI	0.02	0.23	0.28	0.55	0.19	1.00	0.15	0.03	0.31
DiabetesPedigreeFunction	-0.03	0.14	0.00	0.14	0.18	0.15	1.00	0.03	0.17
Age	0.54	0.27	0.33	0.05	-0.02	0.03	0.03	1.00	0.24
Outcome	0.22	0.49	0.16	0.19	0.15	0.31	0.17	0.24	1.00

Figure 3: Heatmap.

B. Data Visualization:

Data visualization is the process of creating graphical representations of data to make it more easily understandable and interpretable. It can be a powerful tool for exploring and understanding the Pima Indian Diabetes Dataset (PIDD). The heat chart in Fig. 3 shows features such as pregnancy, Glucose, BMI, and Age are more correlated with Outcome.

To understand the relationship more, note Fig. 4 between Glucose, BMI, and Age, from that, there are some outliers in features.

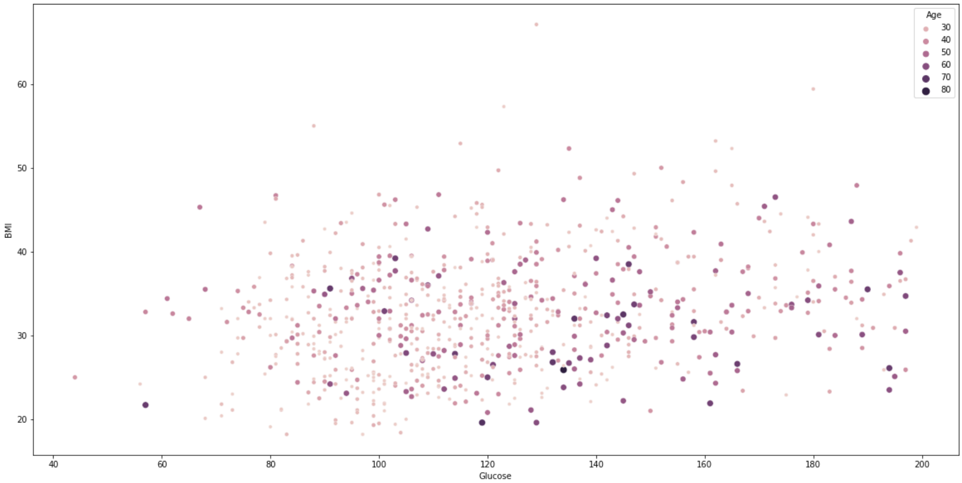


Figure 4: Explore Glucose vs BMI vs Age.

Also there is needing to remove this outlier in feature engineering to increase the effectiveness of the algorithms.

It's important to note that data visualization is not only about creating the plots but also about interpreting the data and the insights from the plots. Data visualization is a powerful tool for exploring and understanding the Pima Indian Diabetes Dataset (PIDD). This enabled it to be used to create graphical representations of data that are easier to understand and interpret.

C. Data Pre-processing

Data pre-processing is an important step in the analysis of any dataset, including the (PIDD). The goal of pre-processing is to clean and prepare the data for further analysis, such as feature engineering and model building. Several common pre-processing techniques can be applied to the PIDD. One of the first steps is to check for missing or null values in the dataset. In PIDD, it is common to have missing values in certain columns, such as 'Glucose', 'Blood Pressure', 'SkinThickness', 'Insulin', 'BMI', and 'DiabetesPedigreeFunction'. These missing values can be handled by imputing the mean or median value of the column.

Another step in pre-processing is to check for outliers in the dataset. Outliers can greatly impact the performance of machine learning models, so it's important to identify and handle them. One common method for outlier detection is to use box plots or scatter plots. These plots can help identify any data points that fall outside of a certain range and can then be removed or handled in another way. Another step is to check for duplicate rows in the dataset and remove them if any. Additionally, the PIDD dataset contains categorical data, such as 'Outcome' which is binary (0 or 1) and should be converted to numerical data using techniques like one-hot encoding. Finally, it is important to scale or normalize the data to ensure that all features are on the same scale. This can be done using techniques such as min-max scaling or standardization. Once these steps have been completed, the data is ready for further analysis, such as feature engineering and model building.

- Feature engineering is the process of transforming raw data into useful features for machine learning models. It can involve a variety of techniques such as normalization, scaling, and encoding categorical variables. In the Pima Indian Diabetes Dataset (PIDD), feature engineering can be used to improve the performance of machine learning models by creating new features or transforming existing features. For example, in PIDD, the BMI feature may be transformed by creating a new feature that indicates whether or not a person has a healthy BMI, based on the World Health Organization's [11] guidelines. Another example is to create a new feature that indicates the age of a woman in groups (e.g. young, middle-aged, and old) to help the model understand the relationship between age and diabetes. In reality, utilizing feature selection or reduction is the key to improving the classifiers’ performance [13].

- Outlier detection is the process of identifying and handling unusual observations in the dataset. Outliers can hurt the performance of machine learning models, so it's important to detect and handle them properly. In PIDD, outliers can be detected by using various statistical methods such as Z-score, IQR, and Mahalanobis Distance, and visualization techniques such as box plots and scatter plots. Once outliers are detected, several techniques can be used to handle them. One common method is to simply remove the outliers from the dataset. It's always good to check the domain knowledge, and also check the correlation between features before dropping the outlier values.

In summary, feature engineering and outlier detection are important techniques for improving the performance of machine learning models in the Pima Indian Diabetes Dataset (PIDD). Feature engineering can be used to create new features or transform existing features to make them more useful for machine learning models. Outlier detection can be used to identify and handle unusual observations in the dataset, which can negatively impact the performance of machine learning models. After dropping outliers from the PIDD dataset. Now this data may splited into test data and training data and then proceed with the modeling.

In addition to the data pre-processing steps I previously mentioned, another technique that can be applied to the Pima Indian Diabetes Dataset (PIDD) is data transformation. One common method of data transformation is to use quantile transformation. Quantile transformation maps the data to a standard normal distribution, which can help to stabilize the variance and reduce the effect of outliers. The process of data splitting is also an important step in the analysis of the PIDD. The goal of data splitting is to divide the data into training and testing sets so that models can be trained and evaluated on separate data sets.

One common method for data splitting is to use k-fold cross-validation. In this method, the data is divided into k subsets, and the model is trained and evaluated k times, with a different subset being used as the testing set each time. This helps to ensure that the model is not overfitting to the training data and will generalize well to new data. With the cross-validation method, the performance of the model can be measured more accurately, as it uses different sets of data for training and testing.

D. Hyperparameter Tuning

Hyperparameter tuning is an important step in the machine learning process, as it can greatly impact the performance of a model. The Pima Indian Diabetes Dataset (PIDD) is a binary classification problem, where the goal is to predict whether an individual has diabetes or not based on various features such as glucose levels, blood pressure, and body mass index.

When tuning hyperparameters for the SVM model on the PIDD, a common approach is to use gridsearchCV. GridsearchCV involves specifying a range of possible values for each hyperparameter and training models using all possible combinations of hyperparameters. The most important hyperparameters to tune for SVM are the regularization parameter (C) and the kernel. The regularization parameter controls the trade-off between maximizing the margin and minimizing the misclassification errors, and the kernel parameter specifies the type of kernel to be used (e.g. linear, polynomial, radial basis function, etc.).

The values presented in Table 1. highlight the default parameters for each of the machine learning algorithms applied in this study. To assess the efficacy of hyperparameter tuning, then these default parameters compared with the optimal parameters selected for improved model performance in subsequent analyses.

Table 1: The default parameters for each of the algorithms used.

Algorithm	Hyper-parameter	Value used
Logistic Regression	C	1.0
	penalty	L2
	solver	lbfgs
Support Vector Machine	C	1.0
	kernel	rbf
Random Forest	n_estimators	100
	criterion	gini
	max_depth	None
	max_features	auto

When tuning hyperparameters for the Logistic Regression model on the PIDD, a common approach is also to use grid search. The most important hyperparameter to tune for Logistic Regression is the regularization parameter (C) and solver. The regularization parameter controls the complexity of the model and the solver helps in choosing the appropriate optimization algorithm to be used.

Once the grid search is completed, the best set of hyperparameters can be chosen based on the performance of the model on a validation set. These optimal hyperparameters can then be used to train the final model on the entire dataset. It's worth mentioning that one of the most common methods for Hyperparameters tuning is also using RandomizedSearchCV, this method will sample a fixed number of random combinations of the hyperparameters and return the best combination based on the performance of the validation set. Using these methods, it is possible to find a set of optimal hyperparameters for the SVM, Random Forest, and Logistic Regression models on the PIDD dataset, which can greatly improve the performance of the models As shown in Table 2.

Table 2. Tuned hyperparameters (best parameters).

Algorithm	Hyper-parameter	Value used
Logistic Regression	C	10
	penalty	l2
	solver	liblinear
Support Vector Machine	C	10
	kernel	linear
Random Forest	n_estimators	200
	criterion	entropy
	max_depth	5
	max_features	log2

As shown in Table 2, where the best value for each parameter in the algorithms was mentioned, Table 3 shows the values that used for each parameter.

Table 3: The range of hyperparameter values for using each algorithm.

Algorithm	Hyper-parameter	range of Hyper-parameter
Logistic Regression	C	10,100,1.0 , 0.1 , 0.10
	Penalty	L2
	Solver	Newton-cg , Liblinear
Support Vector Machine	C	1, 10, 100, 1000
	Kernel	Linear, RBF
Random Forest	n_estimators	500,200
	Criterion	Gini, Entropy
	max_depth	4,5,6,7,8
	max_features	Auto, Sqrt , Log2

Using the Gridsearchcv function, the best values were chosen to increase the efficiency of the algorithms for each parameter, as mentioned in Table 2.

E. Machine Learning Algorithms

This paper focuses on improving the performance of three machine learning algorithms: logistic regression, random forest, and support vector machines (SVM). The method of improvement includes adjusting the ideal parameters through feature engineering, discovering outliers in the dataset, and projecting them to raise the performance of each algorithm. The gridsearchCV function is used to test the parameters to find the best ones through a random mixing process.

Logistic Regression (LR): A statistical method to predict the likelihood of an event based on input variables. The dependent variable in LR is the probability of the event and is bounded between 0 and 1.

Key Features:

- Bernoulli distribution for the dependent variable.
- Prediction made using max probability.
- Model fit measured by Concordance, KS-Statistics.

Uses a sigmoid function to map real numbers to a range between 0 and 1 • Classification made based on sigmoid output: greater than 0.5 is classified as YES, less than 0.5 as NO.

$$f(x)=\frac{1}{1+\mathrm{e}^{-(x)}}$$

In the mentioned paper, the best parameters found were C=10, Penalty=12, and Solver=Liblinear.

Random Forest (RF): Random Forest is a well-known machine learning algorithm that is part of the supervised learning category. It can be used to solve both classification and regression problems in ML. This algorithm is based on the principle of ensemble learning, which means that multiple classifiers are combined to solve a complex problem and improve the performance of the model. The RF algorithm is a classifier that takes multiple decision trees generated from different subsets of a given dataset, and the average of the predictions from these trees is used to improve the accuracy of the data set. This method of combining multiple decision trees helps to overcome the problem of overfitting and results in better accuracy.

There are several advantages to using the RF algorithm, such as:

- It requires less training time compared to other algorithms.
- It is efficient in predicting high-precision results even for large datasets.
- It can maintain accuracy even when a large percentage of data is missing.

RF is a versatile and efficient algorithm that can be used in various ML problems, especially for classification. Its ability to combine multiple classifiers and overcome the problem of overfitting makes it a popular choice among ML practitioners.

However, RF is not the best algorithm for regression tasks. The best parameters for this algorithm, as stated in a paper, are n_estimators = 200, Criterion = Entropy, max_depth = 5, and max_features = log2.

Entropy uses the probability of a given result of how the node is branching. But because of the logarithmic function used in calculating the Gini index, it is more mathematically intensive [12].

Support Vector Machines (SVMs): SVMs are a powerful and versatile tool for solving both classification and regression problems in machine learning. Despite their effectiveness, they come with some limitations that need to be considered before choosing to use this algorithm. To make the most out of SVMs, it is important to understand both their advantages and limitations. On the one hand, SVMs are known for their high precision and efficient memory usage, making them well-suited for working with high-dimensional data. However, this comes at the cost of longer training time, making SVMs less suitable for large datasets. Additionally, SVMs struggle with nested classes, making it challenging to handle these types of data.

While these limitations should be kept in mind when deciding to use SVMs, they can still be useful tools in the right context. By carefully considering the specific requirements of a problem, it is possible to determine whether SVMs is the best choice.

SVMs are a valuable addition to the machine learning toolkit, offering high precision and efficient memory usage. However, they come with limitations such as long training time and difficulty handling nested classes. The best parameters found in previous studies were (C=10, Kernel= Linear), and it is important to consider these limitations when deciding whether to use this algorithm.

IV. EXPERIMENT RESULTS

A. Developing Classification Model

Measuring the accuracy, precision, recall, and F1 score can provide insight into the model's performance. These metrics can be used to evaluate the model's ability to correctly classify instances of diabetes and non-diabetes in the dataset.

TP – Rate: True Positive (TP) rate is the proportion of actual positive cases that are correctly identified as positive by the model. It is calculated as shown in Equation 1 where TP is the number of true positive predictions and FN is the number of false negative predictions. A high TP rate indicates that the model has a high ability to correctly identify positive cases.

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

Equation 1: TP - Rate

FP – Rate: False Positive (FP) rate is the proportion of actual negative cases that are incorrectly identified as positive by the model. It is calculated as shown in Equation 2 where FP is the number of false positive predictions and TN is the number of true negative predictions. A low FP rate indicates that the model has a low rate of false alarms.

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

Equation 2: FP - Rate

Accuracy can be calculated as the proportion of correctly classified instances out of the total number of instances. However, it is important to note that the PIDD dataset is imbalanced, with a larger proportion of non-diabetic instances, which can affect the accuracy score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 3: Accuracy

Precision can be calculated as the proportion of true positives (correctly classified diabetic instances) out of all instances that were predicted as diabetic. High precision indicates that the model has a low rate of false positives.

$$Precision = \frac{TP}{TP + FP}$$

Equation 4: Precision

Recall can be calculated as the proportion of true positives out of all actual diabetic instances. High recall indicates that the model can correctly identify a large proportion of diabetic instances.

$$\text{Recal}\ \frac{TP}{TP+FN}$$

Equation 5: Recall

F1-score is a measure of the balance between precision and recall. It is particularly useful in imbalanced datasets, such as the PIDD, as it takes into account both precision and recall.

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Equation 6: F1-Score

When developing a classification model for the PIDD, it is important to consider the characteristics of the dataset and the specific problem. In this case, the recall may be more important than precision, as it is more important to identify all diabetic instances, even if it leads to more false positives. To improve the performance of the classification model on the PIDD, various techniques such as data pre-processing, feature engineering, using different algorithms, hyperparameter tuning, and cross-validation can be applied. It is important to evaluate different models and select the one that best suits the specific problem and dataset. The performance and accuracy of those algorithms are shown in Table 2.

Table 4: Performance of Algorithms.

Algorithm Name	Accuracy	Precision	Recall	F1-Score
Logistic Regression	79%	0.74	0.75	0.74
Support Vector Machine	78%	0.74	0.74	0.74
Random Forest	77%	0.73	0.74	0.74

B. Results and Discussion

When evaluating the performance of machine learning models on the (PIDD), it is important to consider various metrics such as accuracy, precision, recall, and F1-score.

SVM, Random Forest, and Logistic Regression are three popular classification algorithms that can be used for this dataset. SVM models have been known to perform well in high-dimensional spaces and when the number of features is greater than the number of samples. However, it is sensitive to the choice of kernel and the setting of the regularization parameter. Random Forest is an ensemble method that combines multiple decision trees to improve the overall performance. It is known to be robust to overfitting and handles missing data well. However, it can be computationally expensive. Logistic Regression is a simple and interpretable algorithm that is often used as a benchmark for binary classification problems.

The (PIDD) is a widely used dataset for evaluating machine learning models for predicting diabetes. In this study, classification models using the SVM was developed , Random Forest, and Logistic Regression algorithms and evaluated their performance using a variety of metrics including accuracy, precision, recall, and F1-score.

Experimental results showed that the Logistic Regression model had the highest overall performance, with an accuracy of 79.%, precision of 74%, recall of 75%, and F1-score of 74%. The SVM model had an accuracy of 78%, a precision of 74%, a recall of 74%, and an F1-score of 74%. The Random Forest model had an accuracy of 77%, a precision of 73%, a recall of 74%, and an F1-score of 74%. In terms of the True Positive (TP) rate and False Positive (FP) rate, all three models had similar performance.

Table 5: Compare previous studies

study	Algorithms	dataset
[1]	LR, KNN, SVM ,RF	PIDD
[2]	GB, LR, KNN, SVM, RF, DT, NB	PIDD, Local Dataset
[3]	SVM,KNN,ANN ,NB ,LR,DT	PIDD
[4]	SVM,DT,LR,NB	PIDD
[5]	GNB,MLP,LR,RF,DT ,SVM,GB,KNN	PIDD
[6]	ANN,SVM,DT,KNN,LR	Sylhet diabetes hospital Bangladesh, PIDD

[7]	SVM,LR, AB,DT,KNN,RF,NB	PIDD
[8]	MLP,LR,RF	PIDD
[9]	RF,SVM	PIDD
[10]	LR, RF,SVM,NB ,KNN,DT	PIDD
Proposed method	SVM, RF, LR	PIDD

V. CONCLUSION AND FUTURE WORK

C. In conclusion, the results from the implementation of hyperparameter tuning and feature engineering techniques revealed that optimizing the algorithms can lead to an increase in accuracy. Our findings indicate that the logistic regression model outperformed the Support Vector Machine and Random Forest algorithms in terms of overall performance, however, all three models demonstrated similar True Positive and False Positive rates. These results emphasize the significance of carefully tuning machine learning algorithms and *implementing* feature engineering techniques in the development of improved diabetes prediction models. Future research can investigate the use of other machine learning algorithms, such as neural networks, to compare their performance with the models presented in this study. Furthermore, further feature engineering and outlier detection methods could be applied to the dataset to further enhance the models' performance.

REFERENCES

[1] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of Healthcare Engineering, 2022*. https://doi.org/10.1155/2022/1684017.

[2] Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering, 2,* 229–241. https://doi.org/10.1016/j.ijcce.2021.12.001.

[3] el Massari, H., Sabouri, Z., Mhammedi, S., & Gherabi, N. (2022). Diabetes Prediction Using Machine Learning Algorithms and Ontology. *Journal of ICT Standardization, 10*(2), 319–338. https://doi.org/10.13052/jicts2245-800X.10212.

[4] Varma, K. M., & Panda, B. S. (2019). Comparative analysis of Predicting Diabetes Using Machine Learning Techniques. In *JETIR1907830 Journal of Emerging Technologies and Innovative Research* (Vol. 6). www.jetir.org.

[5] Patil, R., & Tamane, S. (2018). A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes. *International Journal of Electrical and Computer Engineering, 8*(5), 3966–3975. https://doi.org/10.11591/ijece.v8i5.pp3966-3975.

[6] Kumar, B. P. (2022). Diabetes Prediction and Comparative Analysis Using Machine Learning Algorithms. *International Research Journal of Modernization in Engineering Technology and Science, 04*(05). www.irjmets.com.

[7] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express, 7*(4), 432–439. Https://doi.org/10.1016/j.icte.2021.02.004.

[8] Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *Journal of Healthcare Engineering, 2021*. https://doi.org/10.1155/2021/9930985.

[9] Sivaranjani, S., Ananya, S., Aravinth, J., & Karthika, R. (2021). Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, 141–146. https://doi.org/10.1109/ICACCS51430.2021.9441935.

[10] Ali, C., Manal, K., & Atiyah, A. (2022). Predict Diabetes Using Voting Classifier and Hyper Tuning Technique. *Kurdistan Journal of Applied Research (KJAR) Kurdistan Journal of Applied Research*, *7*(2). https://doi.org/10.24017/Science.2022.2.10.

[11] *A healthy lifestyle - WHO recommendations*. (n.d.). Retrieved February 3, 2023, from https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations.

[12] Alhaj, Y. A., Al-Falah, M. M., Al-Arshy, A. M., al Nashad, K. M., Alabedeen, Z., al Nomi, A., al Badawi, B. A., al Khayat, M. S., al Haj, Y. A., al Falah, M. M., Al-Nashad, K. M., Al-Nomi, A., Al-Badawi, B. A., & Al-Khayat, M. S. (2022). *An Efficient Machine Learning Algorithm for Breast Cancer Prediction*. https://wwww.easychair.org/publications/preprint_download/48Pf

[13] Alhaj, Y. A., Dahou, A., Al-Qaness, M. A. A., Abualigah, L., Abbasi, A. A., Almaweri, N. A. O., Elaziz, M. A., & Damaševičius, R. (2022). A Novel Text Classification Technique Using Improved Particle Swarm Optimization: A Case Study of Arabic Language. *Future Internet*, *14*(7). https://doi.org/10.3390/fi14070194.

[14] *Pima Indians Diabetes Database | Kaggle*. (n.d.). Retrieved February 3, 2023, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[15] Kim, D., Alshuwaykh, O., Sandhu, K. K., Dennis, B. B., Cholankeril, G., & Ahmed, A. (2022). Trends in All-Cause and Cause-Specific Mortality Among Individuals With Diabetes Before and During the COVID-19 Pandemic in the U.S. *Diabetes Care*, *45*(6), e107–e109. https://doi.org/10.2337/DC22-0348.