TEESSIDE UNIVERSITY

Master Of Science in Data Science

Machine Learning CIS4035-N

"Early Prediction of Gestational Diabetes Mellitus: The Machine Learning Approach"

Author: Tran Thuy Vy - S3369268

Email: S3369268@live.tees.ac.uk

Word Count: 2477 (Excluding references)

Submission Date: 13/5/2024

Table of Contents

1. Abstract **3**

2. Introduction **3**

3. Literature Review **4**

4. Data Exploration and Features Selection **4**

4.1. Data Selection *4*

4.2. Data Exploration and Pre-Processing *5*

4.3. Feature Selection *8*

5. Experiments **9**

5.1. Train-Test Split *9*

5.2. Models Training *9*

5.2.1 Logistic Regression 10

5.2.2 Random Forest Classifier 10

5.2.3 Support Vector Machine (SVM) 10

5.3. Models Evaluation *10*

5.3.1 Evaluation Metrics 10

5.3.2 Confusion Matrix 11

5.4. Hyperparameter Optimization and Tuning *11*

6. Results **11**

7. Discussion **13**

8. Consideration of Professional, Ethical, and Legal issues **14**

9. Conclusion and Future Work **14**

10. References **14**

1. Abstract

This study utilizes machine learning (ML) techniques to predict GDM risk during early pregnancy using the Pima Indian Diabetes dataset. Three ML algorithms: Logistic Regression (LR), Random Forest Classifier (RF), and Support Vector Machines (SVM); were employed for GDM prediction due to their effectiveness in binary classification tasks. Results show that SVM exhibited superior performance with high precision and recall, highlighting its suitability for real-world applications. Further improvements through fine-tuning and feature selection could optimize SVM's predictive capabilities, enabling accurate GDM diagnosis and prediction.

2. Introduction

Gestational diabetes mellitus (GDM) is a pregnancy-related form of diabetes affecting approximately one in six pregnancies globally (Goyal et al., 2020). It poses significant health risks for both mothers and infants, including excessive birth weight, preeclampsia, and increased risk of type 2 diabetes later in life (Kc et al., 2015). Early prediction of GDM is essential for timely intervention and personalized care to improve health outcomes for both mothers and infants. The current standard method involves screening with an Oral Glucose Tolerance Test (OGTT) between 24 and 28 weeks, but this may miss the opportunity to prevent fetal development impact and improve birth outcomes (Sletner et al., 2017).

This study utilizes machine learning (ML) techniques to predict GDM risk early in pregnancy, using the Pima Indian Diabetes dataset from Kaggle, containing 2000 health records of women aged 21 and older from the Pima Indian community with eight key clinical attributes. Logistic Regression (LR), Random Forest Classifier (RF), and Support Vector Machine (SVM) were selected and employed due to their effectiveness in binary classification tasks, especially in medical data. LR is well-suited for binary classification tasks like GDM prediction and offers interpretable results, RF handles complex interactions and nonlinear relationships, and SVM is effective for binary classification tasks, particularly when the data is not linearly separable. These algorithms have been widely used and have shown to be effective in various medical diagnosis tasks, including diabetes prediction and can manage large datasets while providing insights into key predictive factors.

This study aims to develop accurate and practical ML classifiers that can help healthcare professionals identify individuals at risk of developing GDM early in pregnancy. The findings from this study could lead to more personalized and effective prenatal care strategies, benefiting maternal and fetal health outcomes.

3. Literature Review

GDM detection during pregnancy is important for ensuring the health of both mother and baby. ML offers promising future for developing non-invasive and efficient prediction models. Many prediction models have been developed and implemented by various researchers using variants of ML algorithms. This systematic review explores three widely used classification algorithms for GDM prediction: LR, RF, and SVM. The analysis of their strengths and limitations in the context of GDM classification aims to aid in selecting the most suitable algorithm for this specific application with medical datasets.

In one study, Aishwarya and Vaidehi tested various ML algorithms including SVM and LR. Their findings showed that LR achieved as the top performer, achieving an impressive accuracy of 96% on various dataset (Mujumdar et al.,2019). In a different study, a LR model was also investigated and tested on 128 patients and training it on 459 individuals, achieving an impressive classification accuracy of 92% (Qawqzeh et al.,2020).

In a separate investigation, a new approach: a 'Stacking Classifier' which combines multiple classifiers (Naive Bayes, KNN, Linear Discriminant Analysis, RF) was proposed to improve overall performance and reduce misdiagnosis. This ensemble method, with RF as the final classifier (Meta classifier), achieves a significantly higher accuracy (97.35%) compared to individual classifiers (ranging from 74.60% to 78.57%) (Maria Ali et al.,2022). On another hand, Tejas and Pramila opted for LR and SVM algorithms to develop a diabetes prediction model. They conducted data pre-processing to enhance the quality of results, revealing that SVM outperformed LR with an accuracy of 79% (Joshi et al.,2018).

A diabetes prediction model using three different ML algorithms: RF, Decision Tree, and Naïve Bayes (NB), was developed by Yuvaraj and Sripreethaa and implemented in Hadoop-based clusters. Extensive pre-processing techniques were applied to the dataset, resulting in the RF algorithm achieving the highest accuracy rate of 94% in the predictive model (Yuvaraj et al.,2019).

A recent study highlighted an impressive performance of various ML models and an Ensemble Model on 2 datasets. For Dataset 1, LR achieved 75.32% accuracy, SVM achieved 74.89% and NB achieved 74.03%. In contrast, for Dataset 2, higher accuracies were observed across all algorithms, with LR at 88.89%, SVM at 88.03%, and NB at 88.89%. (Priyanka et al.,2021)

4. Data Exploration and Features Selection

4.1. Data Selection

The Pima Indian Diabetes dataset has been widely used in diabetes classification research, despite the availability of larger and more complex datasets. With its binary

outcome variable, the dataset naturally supports supervised learning methods like logistic regression. However, researchers have explored various ML algorithms beyond logistic regression to develop classification models. In this study, the focus is on analyzing the Pima Indian Dataset including 2000 data points, sourced from Kaggle, using three classification algorithms: Logistic Regression (LG), Random Forest Classifier (RF), and Support Vector Machine (SVM).

Table 1 shows overview of the dataset, it records eight causal characteristics and the corresponding classification. The binary classification Outcome variable takes (0 or 1) values, where 0 indicates a negative test for GDM, and 1 implies a positive test. Based on medical domain knowledge, key features such as plasma glucose level, number of pregnancies, and age are essential for GDM prediction.

No	Feature	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration at 2 Hours in an oral glucose tolerance test (OGTT)
3	BloodPressure	Diastolic Blood Pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour Serum insulin (μh/ml)
6	BMI	Body mass index [weight in kg/ (Height in m)]
7	DiabetesPedigreeFunction	Diabetes pedigree function
8	Age	Age (years)
9	Outcome	Binary value indicating non-diabetic /diabetic

TABLE 1.OVERVIEW OF PIMA INDIAN DIABETES DATASET

4.2. Data Exploration and Pre-Processing

Data Exploration and pre-processing are important steps in developing a reliable ML model since the quality and quantity of processed data directly impact the reliability and accuracy of ML model outcomes (Alberto Meola et al.,2023).

As seen in Figure 1, a bar plot displaying count of patients with and without GDM. 1316 individuals are without GDM, while 648 patients have been diagnosed with GDM. This provides a clear overview of the distribution of GDM status within the dataset, indicating a larger proportion of non-GDM individuals compared to those diagnosed with GDM.

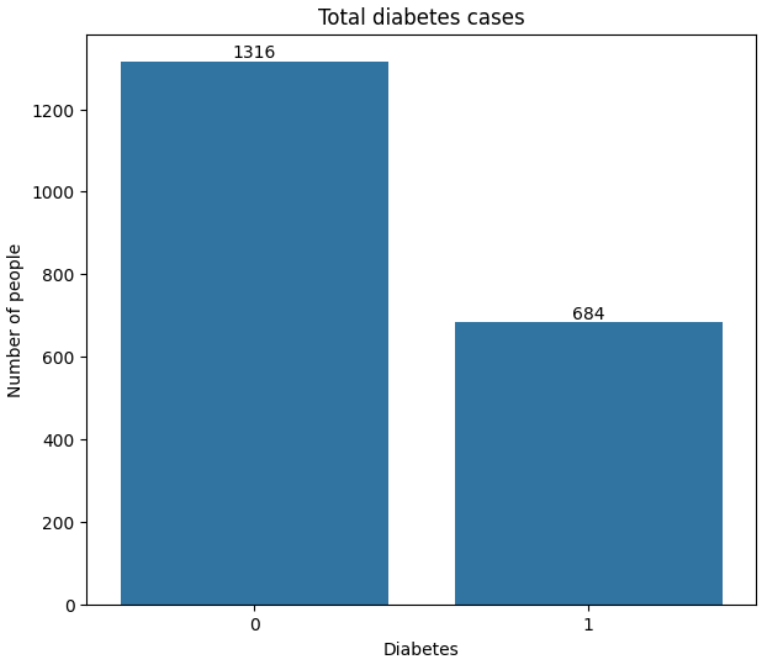

FIGURE 1. COUNT OF NON-GDM AND GDM CASES

Table 2 summarizes the skewness of each attribute in a dataset. Skewness is a measure of asymmetry in a distribution. Blood Pressure stands out with a strong negative skewed. Insulin, Diabetes Pedigree Function, and Age exhibit a high positive skew. The remaining attributes, including Pregnancies and Outcome, show a moderate positive skew. Fortunately, SkinThickness, Glucose, and BMI appear close to a normal distribution.

Attribute	Skewness value
Pregnancies	0.982366
Glucose	0.158806
BloodPressure	-1.854476
SkinThickness	0.207228
Insulin	1.996084
BMI	-0.090455
DiabetesPedigreeFunction	1.811979
Age	1.181267
Outcome	0.666633

TABLE 2. ATTRIBUTE DISTRIBUTIONS

Figure 2, the heatmap illustrates feature correlations, Age and Pregnancies exhibit a strong positive correlation (0.54), while Glucose correlates moderately with Insulin (0.32) and BMI (0.23). Skin Thickness is strongly positively correlated with Insulin (0.45) but slightly negatively correlated with Age (-0.11). BMI shows moderate to strong positive

correlations with Blood Pressure (0.28), Insulin (0.45), and Skin Thickness (0.39), indicating potential dependencies between these variables.

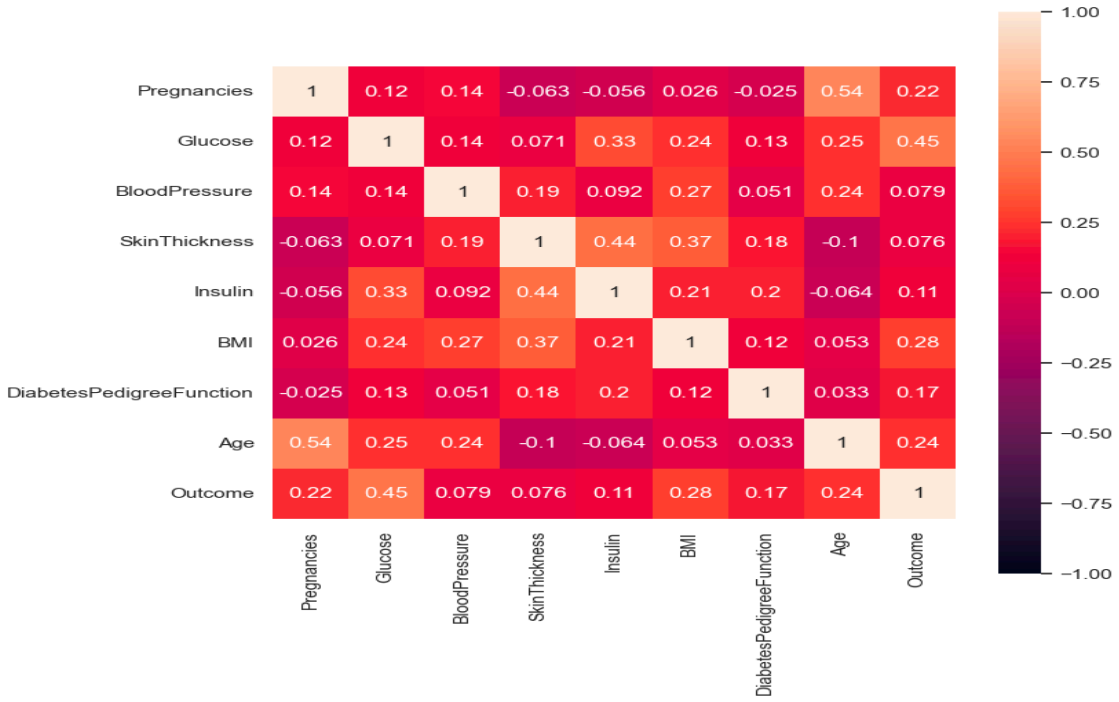

FIGURE 2. CORRELATION VALUES AND HEATMAP, SHOWING THE CORRELATION BETWEEN THE FEATURES

Figure 3 shows histograms displaying distributions of all attributes in the dataset.

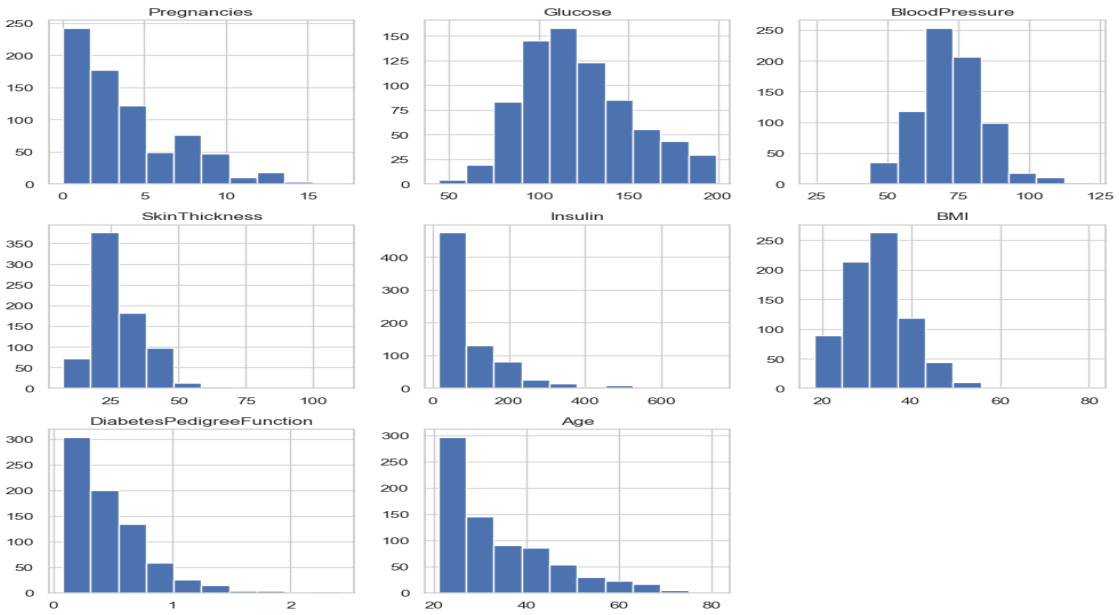

FIGURE 3. HISTOGRAMS, SHOWING DISTRIBUTIONS OF ALL ATTRIBUTES

The original dataset has 2000 rows and 9 columns without missing values; however, it had many duplicated records. The dataset was refined to 744 rows and 9 columns (491 non-GDM and 253 GDM) after removing duplicated rows.

```
# Check duplicated rows
duplicated_count = df.duplicated().sum()
if duplicated_count:
    #Remove duplicated rows
    print(f"Removed {duplicated_count} duplicated row(s)")
    df.drop_duplicates(keep='first', inplace=True)
    #removes duplicated rows from the df while keeping the first occurrence
    # Recheck number of rows and columns of the dataset
    data = df.shape
    print(f"--------------------")
    print(f"Number of Rows : {data[0]} \nNumber of Columns : {data[1]}")
else:
    print("No duplicated rows")
```

✓ 0.0s

```
Removed 1256 duplicated row(s)
--------------------
Number of Rows : 744
Number of Columns : 9
```

FIGURE 4. DUPLICATED ROWS REMOVAL

Within the dataset, invalid zero values can directly impact the accuracy of ML models. The next step involves addressing these values by imputing them with either mean value or median value depending on each attribute's distribution.

```
# Replacing the 0 value from ['Glucose','BloodPressure','SkinThickness','Insulin','BMI']
# by either mean value or median value depending upon distribution:
# BloodPressure, Insulin have skewed distributons => replace by Median value
# SkinThickness, Glucose, BMI have normal distributions => replace by Mean value
df_new = df.copy(deep= True)
df_new['Glucose'].replace(0,df_new['Glucose'].mean(axis=0),inplace=True)
df_new['BloodPressure'].replace(0,df_new['BloodPressure'].median(axis=0),inplace=True)
df_new['SkinThickness'].replace(0,df_new['SkinThickness'].mean(axis=0),inplace=True)
df_new['Insulin'].replace(0,df_new['Insulin'].median(axis=0),inplace=True)
df_new['BMI'].replace(0,df_new['BMI'].mean(axis=0),inplace=True)
```

FIGURE 5. MEAN AND MEDIAN IMPUTATION

4.3. Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model and enhancing classification methods by lowering computational costs and potentially improving model performance (Labory et al., 2024).

Attribute	Pearson Correlation Coefficient
Pregnancies	0.220942

Glucose	0.453939
BloodPressure	0.078808
SkinThickness	0.075562
Insulin	0.114655
BMI	0.278123
DiabetesPedigreeFunction	0.174688
Age	0.242077

TABLE 3. CORRELATION COEFFICIENTS BETWEEN EACH FEATURE AND OUTCOME

The Pearson correlation method is commonly employed to identify key traits or features by calculating correlation coefficients between input and output attributes. As seen in Table 3, Glucose levels show the strongest positive correlation with Outcome. Additionally, moderate positive correlations exist with BMI and age. Other factors like pregnancies, insulin levels, and blood pressure exhibit weaker correlations with Outcome, suggesting their lesser impact compared to glucose, BMI, and age.

5. Experiments

All three the proposed ML algorithms chosen to be used in this research paper are classification models. These algorithms were chosen based on their strengths in binary classification for medical data. LR is suitable due to its interpretability in binary classification tasks (diabetic/non-diabetic), RF can capture complex relationships in the data without explicitly assuming linearity. It utilizes decision trees, making it robust to outliers and capable of handling a mix of feature types (categorical and numerical), and SVM tackles the high dimensionality of medical data, often containing numerous features.

5.1. Train-Test Split

The dataset was split into training and testing sets using a standard split ratio method. For this project, an 80:20 split ratio was adopted, allocating 80% (595 samples) of the dataset for training and reserving the remaining 20% (149 samples) for testing purposes. The same training and testing sets were used for all three models.

5.2. Models Training

Training supervised machine learning models is an essential step in implementing them effectively, where algorithms learn from labeled data to make accurate predictions. This process involves presenting the algorithm with input data and corresponding output labels, teaching it to map features to correct outputs based on learned patterns. Figure 6, 7, 8 shows the implementation of 3 ML models.

5.2.1 Logistic Regression

```
from sklearn.linear_model import LogisticRegression

logistic_reg = LogisticRegression(solver='liblinear
logistic_reg.fit(X_train, y_train)
```

✓ 0.0s

```
LogisticRegression
LogisticRegression(solver='liblinear')
```

FIGURE 6. IMPLEMENTATION OF LOGISTIC REGRESSION CLASSIFIER

5.2.2 Random Forest

```
from sklearn.ensemble import RandomForestClassifier

random_forest = RandomForestClassifier(random_state=42)
random_forest.fit(X_train, y_train)
```

✓ 0.1s

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

FIGURE 7. IMPLEMENTATION OF RANDOM FOREST CLASSIFIER

5.2.3 Support Vector Machine (SVM)

```
from sklearn.svm import SVC
svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)
```

✓ 2.1s

```
SVC
SVC(kernel='linear')
```

FIGURE 8. IMPLEMENTATION OF SVM CLASSIFIER

5.3. Models Evaluation

5.3.1 Evaluation Metrics

Four evaluation metrics were employed to assess the results, which include accuracy, precision, recall, F-score. These metrics are calculated using a confusion matrix, which displays the actual and predicted outcome classes based on the testing set.

Accuracy refers to the percentage of all samples that have been predicted correctly. It is the ratio of the correctly classified prediction to the total number of predictions made. Precision metric shows what percent of predictions are correct. Recall describes what

percent of positives are correctly identified and F1 score is the percent of positive predictions that are correct.

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

$$Recall = \frac{TP}{TP + FN} \qquad F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative. TP and TN represent the cases when the actual outcome and the result are the same, whereas FP and FN are the cases when the opposite results are obtained.

5.3.2 Confusion Matrix

Confusion matrix shows the values of the actual outcome classes and the predicted outcome classes on the testing set.

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

FIGURE 9. CONFUSION MATRIX TEMPLATE

5.4. Hyperparameter Optimization and Tuning

Hyperparameters directly impact model structure, function, and performance, necessitating tuning for optimal results in machine learning. In this study, GridSearch Cross-Validation (GridSearchCV) was employed to automatically identify the best hyperparameter values, improving model performance by streamlining parameter selection and optimization.

6. Results

Before tuning, the models showed moderate accuracy ranging from 0.74 to 0.78, which improved slightly after tuning, with LR and RF achieving an accuracy of 0.77. Precision and Recall are critical metrics in medical diagnostics like GDM prediction. Before tuning, SVC had the highest precision (0.64) and recall (0.63). After tuning, SVC maintained its strong performance, while LR and RF showed improvements in both precision and recall.

The F1 Score, which is a balanced measure of precision and recall, improved across all models after tuning, with SVC consistently performing well in this regard.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.74	0.59	0.50	0.54
Random Forest Classifier	0.74	0.58	0.57	0.57
SVC	0.78	0.64	0.66	0.64

TABLE 4. EVALUATION METRICS BEFORE HYPERPARAMETER TUNING

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.77	0.63	0.59	0.61
Random Forest Classifier	0.77	0.62	0.63	0.62
SVC	0.78	0.64	0.63	0.64

TABLE 5. EVALUATION METRICS AFTER HYPERPARAMETER TUNING

Based on the evaluation, SVC appears to be the most suitable model for GDM prediction both before and after hyperparameter tuning. It consistently achieves the highest scores in accuracy, precision, recall, and F1 score. The tuned SVC model retains its strong performance, making it a robust choice for this classification task.

Figure 10, 11, 12 show confusion matrixes of 3 models after Hyperparameter Tuning

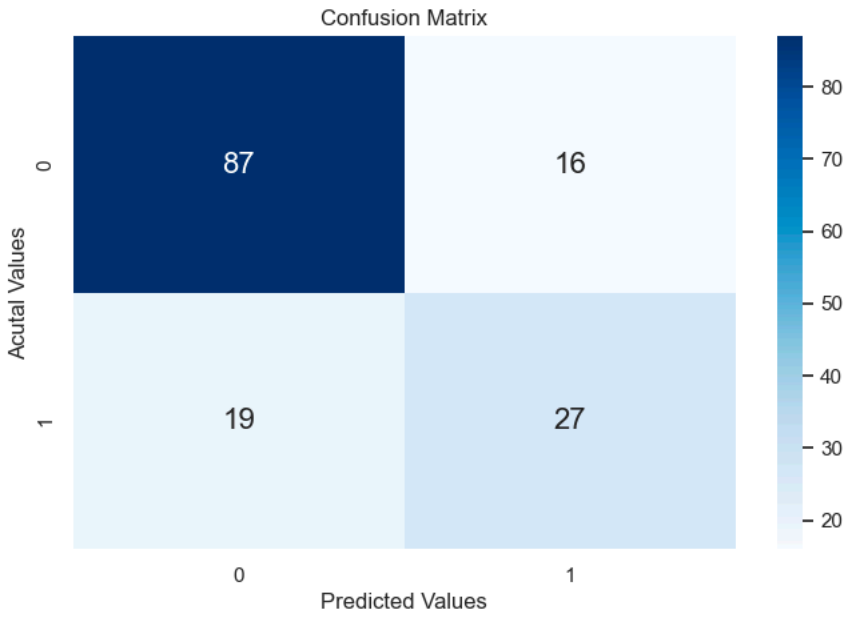

FIGURE 10. LOGISTIC REGRESSION MODEL CONFUSION MATRIX AFTER HYPERPARAMETER TUNING

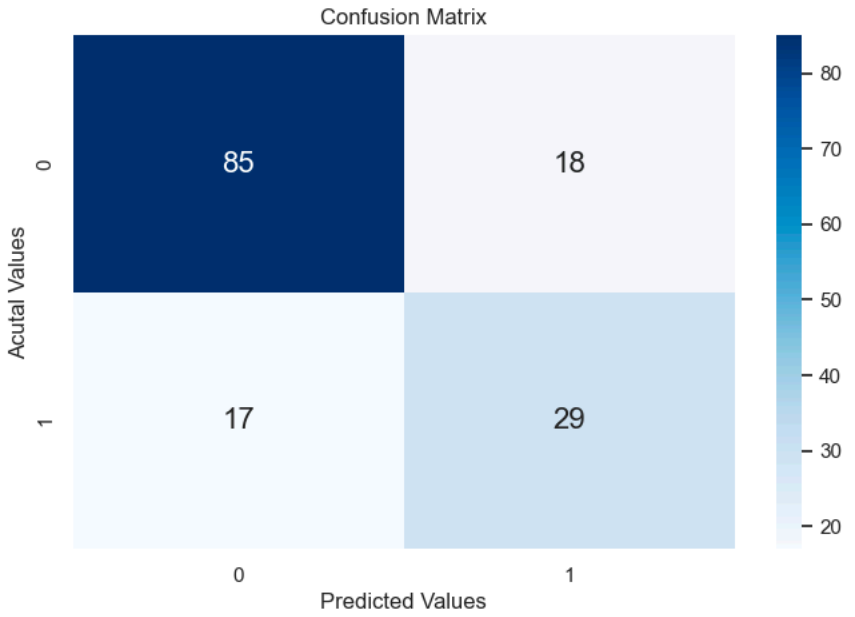

FIGURE 11. RANDOM FOREST MODEL CONFUSION MATRIX AFTER HYPERPARAMETER TUNING

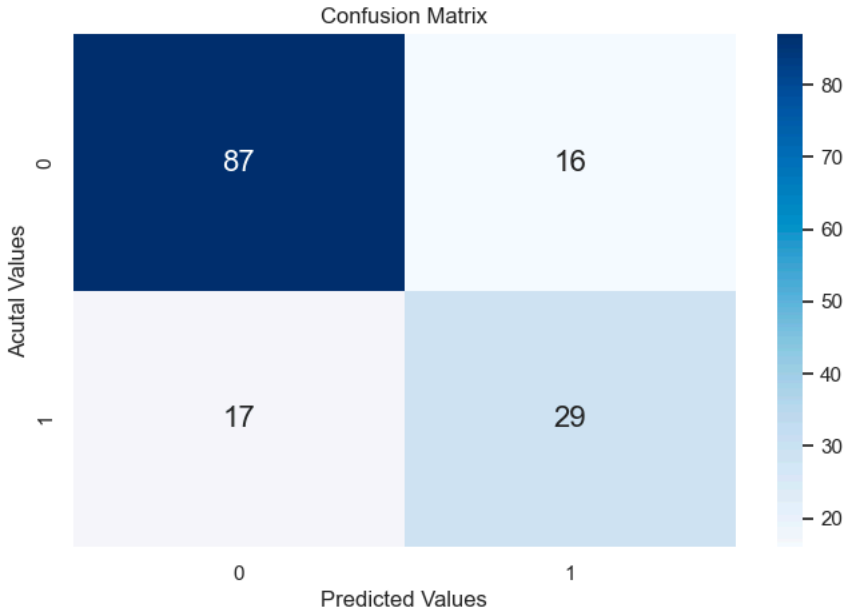

FIGURE 12. SVC MODEL CONFUSION MATRIX AFTER HYPERPARAMETER TUNING

7. Discussion

The discussion centers on the findings and implications of the machine learning project for diabetes prediction. The performance of different algorithms: LR, RF, and SVC, was analyzed before and after Hyperparameter Tuning. The SVC model emerged as the most promising due to its high accuracy, precision, recall, and F1 score. The importance of Model Optimization Techniques, especially Hyperparameter Tuning in improving model

performance was highlighted, along with the significance of precision and recall in medical diagnostics. The results underscore the need for robust ML models in healthcare applications like GDM prediction.

8. Consideration of Professional, Ethical, and Legal issues

In professional, ethical, and legal contexts, responsible AI implementation in healthcare is essential. This includes ensuring data privacy, confidentiality, and adherence to ethical guidelines throughout the project. Transparency in model development and result interpretation is crucial for healthcare professionals and patients alike. Understanding legal regulations and compliance related to medical data usage and model deployment is vital for maintaining ethical conduct and patient safety.

9. Conclusion and Future Work

In conclusion, the study demonstrates the effectiveness of ML models for GDM prediction. The SVC model showed superior performance, emphasizing its suitability for real-world applications. Further fine-tuning and feature selection could significantly enhance SVC's performance, making it an even more robust tool for accurate GDM diagnosis and prediction.

Moving forward, future work will focus on exploring advanced feature engineering techniques and leveraging ensemble methods to optimize SVC's predictive capabilities. Ethical considerations will continue to guide research efforts, ensuring patient privacy, fairness, and transparency in AI-driven healthcare solutions. This research contributes to advancing predictive analytics in healthcare and sets the stage for continued refinement of ML models in medical diagnostics.

10. References

Goyal, A., Gupta, Y., Singla, R., Kalra, S., & Tandon, N. (2020). American Diabetes Association "Standards of Medical Care-2020 for Gestational Diabetes Mellitus": A Critical Appraisal. Diabetes Therapy: Research, Treatment and Education of Diabetes and Related Disorders, 11(8), 1639–1644. https://doi.org/10.1007/s13300-020-00865-3

Kc, K., Shakya, S. and Zhang, H., 2015. Gestational diabetes mellitus and macrosomia: a literature review. Annals of Nutrition and Metabolism, 66(Suppl. 2), pp.14-20.

Sletner, L., Jenum, A.K., Yajnik, C.S., Mørkrid, K., Nakstad, B., Rognerud-Jensen, O.H., Birkeland, K.I. and Vangen, S., 2017. Fetal growth trajectories in pregnancies of European and South Asian mothers with and without gestational diabetes, a population-based cohort study. PloS one, 12(3), p.e0172946.

Yuvaraj, N., SriPreethaa, K.R. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Comput 22 (Suppl 1), 1–9 (2019). https://doi.org/10.1007/s10586-017-1532-x

Mujumdar, A. and Vaidehi, V., 2019. Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, pp.292-299.

Joshi, T.N. and Chawan, P.M., 2018. Logistic regression and svm based diabetes prediction system. International Journal For Technological Research In Engineering, 5, pp.4347-4350.

Qawqzeh, Y.K., Bajahzar, A.S., Jemmali, M., Otoom, M.M. and Thaljaoui, A., 2020. Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling. BioMed Research International, 2020.

Priyanka Rajendra, & Shahram Latifi (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.

Alberto Meola, Manuel Winkler, & Sören Weinrich (2023). Metaheuristic optimization of data preparation and machine learning hyperparameters for prediction of dynamic methane production. Bioresource Technology, 372, 128604.

Justine Labory, Evariste Njomgue-Fotso, & Silvia Bottini (2024). Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. Computational and Structural Biotechnology Journal, 23, 1274-1287.

Maria Ali, Muhammad Nasim Haider, Saima Anwar Lashari, Wareesa Sharif, Abdullah Khan, & Dzati Athiar Ramli (2022). Stacking Classifier with Random Forest functioning as a Meta Classifier for Diabetes Diseases Classification. Procedia Computer Science, 207, 3459-3468.