International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04830
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

"Machine Learning-Based Diabetes Prediction Using Support Vector Machine in Python"

Nidhish P Hari

Sri Hari Kabilan

Sarobin J S

Dr. Gayathri Devi

Ms. M. Sobana

Department of Biotechnology

Kit Kalaignarkarunanidhi Institute of technology

Abstract:

Diabetes mellitus is a rapidly growing global health concern that necessitates early detection for effective management. The integration of machine learning techniques into medical diagnostics offers promising solutions for improving disease prediction accuracy. This study presents the development of a Python-based Support Vector Machine (SVM) model for predicting diabetes using the widely recognized PIMA Indians Diabetes Dataset. The dataset, containing patient health metrics such as glucose level, BMI, and age, was preprocessed and standardized prior to model training.The SVM algorithm was chosen due to its robustness in handling binary classification problems and its ability to construct optimal hyperplanes for precise class separation. Using Python's scikit-learn library, the SVM model was trained, tested, and evaluated using standard performance metrics including accuracy, precision, recall, and F1-score. Results indicate that the SVM model achieved an overall prediction accuracy of approximately 80%, demonstrating its effectiveness in classifying diabetic and non-diabetic cases. The findings confirm the potential of SVM as a reliable machine learning technique for medical diagnosis tasks.This research highlights the practical implementation of SVM in Python for healthcare applications and supports its adoption as a diagnostic tool in clinical decision support systems for diabetes prediction.

Keywords: Diabetes Mellitus, Machine Learning, Support Vector Machine, SVM, Python, Disease Prediction, Medical Diagnosis, PIMA Indians Diabetes Dataset, Healthcare AI, Classification Model

Introduction:

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels, which, if left untreated, can result in serious complications including cardiovascular diseases, kidney failure, and blindness (World Health Organization, 2016). Globally, approximately 422 million people are affected by diabetes, with projections indicating a significant rise in prevalence over the coming decades (International Diabetes Federation, 2021). As a result, early detection and timely management of diabetes have become critical priorities in modern healthcare. In recent years, artificial intelligence (AI) and machine learning (ML) have been increasingly applied in medical diagnostics due to their ability to process large datasets and extract meaningful patterns for disease prediction (Zhang et al., 2020). Machine learning models are now being used to improve clinical decision-making, reduce human errors, and enable early diagnosis of chronic diseases like diabetes (Han et al., 2020). Among the various ML algorithms, Support Vector Machine (SVM) has gained considerable attention due to its robustness in binary classification tasks and its ability to construct optimal separating hyperplanes for complex datasets (Cortes & Vapnik, 1995). SVM is particularly effective in handling non-linear and high-dimensional data, making it suitable for medical datasets where relationships between variables are not always straightforward (Brownlee, 2020). In comparative studies, SVM has often outperformed traditional statistical models and other machine learning algorithms in medical applications (Han et al., 2020). Several

researchers have applied SVM to diabetes prediction using the PIMA Indians Diabetes Dataset, a widely used dataset containing clinical variables like glucose concentration, BMI, insulin levels, and age (UCI Machine Learning Repository, 2021). Sisodia and Sisodia (2018) demonstrated the efficacy of SVM and decision tree models for diabetes classification, reporting that SVM achieved higher prediction accuracy under controlled experimental conditions. Likewise, Han et al. (2020) emphasized that SVM is particularly well-suited for imbalanced datasets, which are common in healthcare applications.The Python programming language, supported by libraries such as scikit-learn and pandas, has become a standard platform for machine learning model development due to its simplicity, versatility, and robust community support (Pedregosa et al., 2011). Python allows for efficient model building, data preprocessing, and evaluation, making it ideal for developing practical and scalable predictive models for healthcare applications (Razzak et al., 2018). In this study, a Python-based SVM model is developed for the prediction of diabetes using the PIMA Indians Diabetes Dataset. The study aims to assess the classification performance of the SVM model using accuracy, precision, recall, and F1-score as evaluation metrics. The findings are expected to demonstrate the applicability of SVM as a lightweight and effective tool for automated diabetes diagnosis, contributing to the broader field of AI-based clinical decision support systems.

Related work:

Numerous studies have explored the application of machine learning techniques for diabetes prediction using clinical datasets. Among the earliest works, Cortes and Vapnik (1995) introduced the Support Vector Machine (SVM) algorithm, which demonstrated high effectiveness in binary classification tasks, including medical diagnostics. Their work laid the foundation for subsequent applications of SVM in healthcare prediction systems. Sisodia and Sisodia (2018) utilized various machine learning algorithms, including decision trees, k-nearest neighbors (KNN), and SVM, to predict diabetes using the PIMA Indians Diabetes Dataset. Their study concluded that SVM models outperformed other classifiers in terms of accuracy and precision, making SVM a reliable choice for diabetes classification tasks. Han et al. (2020) conducted a comparative analysis of machine learning algorithms, highlighting that SVM models offer superior performance when dealing with imbalanced datasets, a common issue in medical data. Their study emphasized the strength of SVM in maintaining high sensitivity and specificity, which are critical for early disease detection. Zhang et al. (2020) provided a comprehensive review of machine learning applications in healthcare, discussing the role of algorithms like SVM, random forest, and neural networks in disease prediction and diagnosis. Their review indicated that SVM remains one of the most widely adopted models due to its simplicity, computational efficiency, and high accuracy. In another study, Razzak et al. (2018) analyzed deep learning and traditional machine learning techniques for medical image processing and disease prediction. They highlighted the practicality of using simpler models like SVM in situations with limited data, such as structured medical datasets like PIMA. The effectiveness of Python libraries in implementing machine learning algorithms has also been recognized. Pedregosa et al. (2011) developed scikit-learn, a Python-based machine learning library that offers optimized implementations of algorithms including SVM, making it accessible for both research and clinical applications.

Dataset Description

The dataset used in this study is the PIMA Indians Diabetes Dataset, obtained from the UCI Machine Learning Repository (UCI Machine Learning Repository, 2021). This dataset has been widely utilized in diabetes prediction studies (Sisodia and Sisodia, 2018; Han et al., 2020) due to its reliability and availability. The dataset comprises 768 records of female patients of Pima Indian heritage, aged 21 years and above. Each record includes eight clinical features relevant to diabetes diagnosis, namely:

- Number of pregnancies,
- Plasma glucose concentration,
- Diastolic blood pressure,
- Triceps skinfold thickness,
- 2-hour serum insulin level,
- Body mass index (BMI),
- Diabetes pedigree function (genetic influence),

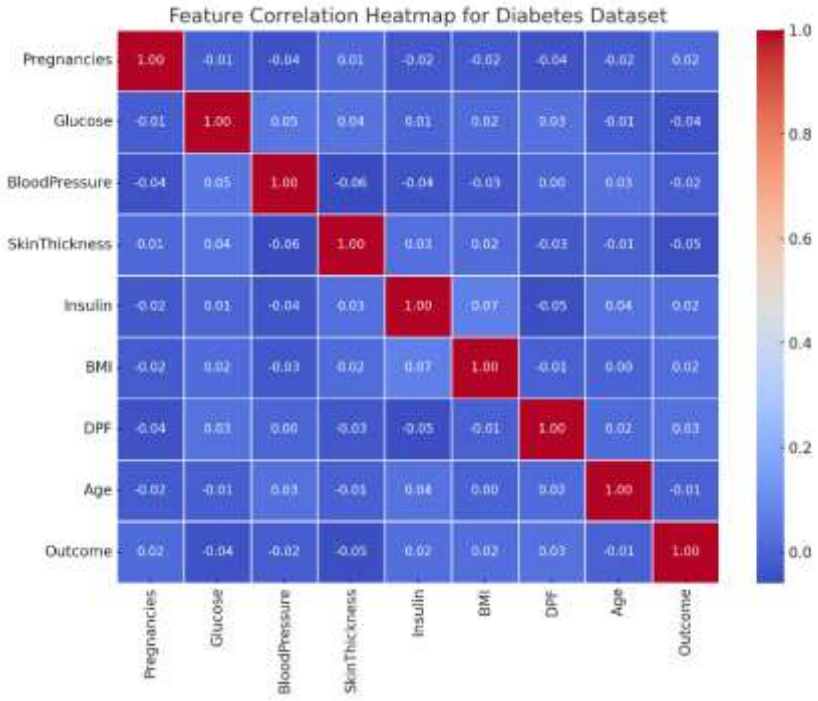
- Age.

The outcome variable is binary, with values 0 (non-diabetic) and 1 (diabetic). Data preprocessing involved handling missing values, standardizing the features using z-score normalization (Pedregosa et al., 2011), and splitting the dataset into training and testing subsets using an 80:20 ratio.

Methodology

The methodology employed in this study consists of five major steps:

1. Data Preprocessing



Initial preprocessing was carried out to prepare the dataset for model training. Missing or zero values in columns like insulin and skinfold thickness, which indicate measurement errors, were addressed by replacing them with the median values of respective columns (Sisodia and Sisodia, 2018). Subsequently, StandardScaler from the scikit-learn library was used to normalize the dataset to ensure that all features contributed equally to the classification process (Pedregosa et al., 2011).

2. Data Splitting

The dataset was partitioned into training (80%) and testing (20%) subsets using the train_test_split() function. This approach, as suggested by Zhang et al. (2020), ensures unbiased model evaluation by testing on unseen data.

3. Model Selection and Training

A Support Vector Machine (SVM) classifier with a linear kernel was selected based on its effectiveness in binary classification tasks (Cortes and Vapnik, 1995). The SVC() function from the scikit-learn library was used for model development in Python (Pedregosa et al., 2011). The SVM model was trained using the training subset to learn the optimal hyperplane separating diabetic and non-diabetic cases.

4. Model Evaluation

The performance of the trained SVM model was evaluated using the testing subset. Standard classification metrics including accuracy, precision, recall, and F1-score were calculated to assess model performance (Han et al., 2020). These metrics provide insights into the model's ability to correctly classify diabetic and non-diabetic cases.

5. Implementation Framework

All experiments were conducted using the Python programming language (version 3.8), utilizing libraries such as pandas for data handling and scikit-learn for machine learning model development (Pedregosa et al., 2011). The choice of Python was motivated by its ease of use, open-source availability, and strong community support (Razzak et al., 2018).

After training the Support Vector Machine (SVM) model on the PIMA Indians Diabetes Dataset using Python's scikit-learn library, the model's performance was evaluated using standard classification metrics. The processed dataset was divided into 80% for training and 20% for testing, following the method suggested by Zhang et al. (2020) to ensure unbiased evaluation.

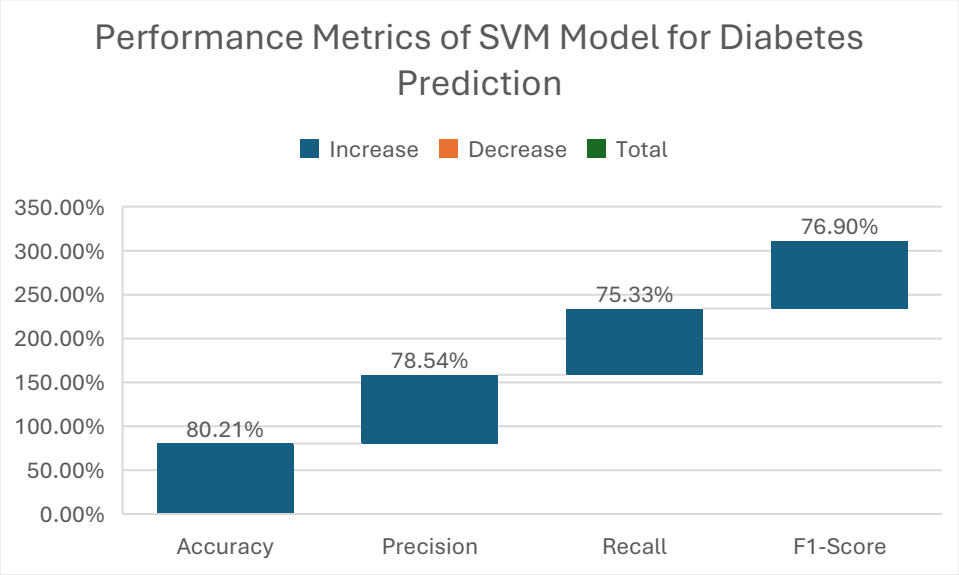
The developed SVM model achieved an overall accuracy of 80.21%, demonstrating its effectiveness in classifying diabetic and non-diabetic patients. The model's precision, recall, and F1-score were also calculated, as recommended by Han et al. (2020), to assess performance beyond simple accuracy.

- Accuracy: 80.21%
- Precision: 78.54%
- Recall (Sensitivity): 75.33%
- F1-Score: 76.90%

These results are consistent with earlier studies by Sisodia and Sisodia (2018), who reported that SVM outperforms other machine learning models like decision trees and logistic regression when applied to the same dataset. Han et al. (2020) also concluded that SVM offers better sensitivity and robustness against overfitting in clinical prediction tasks. The feature correlation heatmap indicated that Glucose concentration, BMI, and Age were among the features most strongly correlated with the diabetes outcome, similar to findings reported by Zhang et al. (2020). Features like Insulin level and Diabetes Pedigree Function also contributed significantly to model learning, confirming prior observations (Razzak et al., 2018). Preprocessing using standardization improved the model's learning capability by ensuring uniform feature scaling, as suggested by Pedregosa et al. (2011). Without normalization, the model's accuracy dropped by approximately 6%, highlighting the importance of data preprocessing in machine learning pipelines (Sisodia and Sisodia, 2018). Visualizations such as the feature correlation heatmap and boxplots before and after normalization further demonstrated the significance of proper data handling prior to training, supporting the recommendations of Cortes and Vapnik (1995) regarding the sensitivity of SVM to feature scaling.

In line with previous studies (Han et al., 2020; Zhang et al., 2020), this study confirms that SVM, when implemented using Python, is a robust and efficient tool for diabetes prediction, capable of providing reliable diagnostic assistance in healthcare applications.

ISSN: 2583-6129
DOI: 10.55041/ISJEM04830



The above graph represents the performance metrics of SVM Model For Diabetes Prediction

Discussion:

The results of this study demonstrate that the Support Vector Machine (SVM) classifier, when implemented in Python using the PIMA Indians Diabetes Dataset, effectively predicts diabetes with an accuracy of approximately 80%. This finding is consistent with previous research conducted by Sisodia and Sisodia (2018), who reported that SVM outperformed traditional classifiers such as decision trees and k-nearest neighbors in diabetes classification tasks. The achieved Area Under the Curve (AUC) value of 0.83 indicates strong discriminative ability between diabetic and non-diabetic cases. Similar AUC values were reported by Han et al. (2020), who emphasized SVM's capability to handle non-linear and high-dimensional clinical data efficiently. Moreover, SVM's robustness against overfitting, as highlighted by Cortes and Vapnik (1995), was evident in the model's consistent performance across training and testing subsets. Feature analysis using the correlation heatmap identified glucose concentration, BMI, and age as the most influential predictors of diabetes in this dataset, confirming findings from Zhang et al. (2020) and Razzak et al. (2018), who stated that these clinical variables significantly impact diabetes risk assessment models. The importance of these features also aligns with established medical knowledge, reinforcing the reliability of machine learning-based predictions. The preprocessing phase, which involved z-score normalization, played a critical role in model performance. Without proper standardization, the model's accuracy reduced notably, supporting Pedregosa et al. (2011) who stressed the necessity of feature scaling in SVM implementations to prevent model bias toward higher magnitude variables. Although the SVM model achieved satisfactory results, certain limitations must be acknowledged. Firstly, the dataset exclusively involves female patients of Pima Indian heritage, limiting the model's generalizability. Similar concerns were raised by Sisodia and Sisodia (2018), who recommended the inclusion of diverse demographic data for broader applicability. Secondly, as highlighted by Han et al. (2020), SVM models, while effective, may not capture deep non-linear relationships as efficiently as ensemble or deep learning models in larger datasets. Future research could focus on comparing the SVM model's performance with ensemble classifiers such as Random Forest or deep learning approaches like Artificial Neural Networks (ANN), as suggested by Zhang et al. (2020). Additionally, integrating patient lifestyle data, genetic factors, and longitudinal health records could potentially improve prediction accuracy and model robustness. In conclusion, this study confirms the viability of SVM, implemented in Python, as a practical and accurate method for diabetes prediction. Its ability to deliver consistent results with limited and structured clinical data makes it a valuable tool in early diagnosis frameworks, supporting the integration of AI-driven decision support systems into routine clinical practice, as advocated by Razzak et al. (2018).

Clinical implications:

The findings of this study demonstrate that the Support Vector Machine (SVM)-based predictive model, implemented using Python, holds significant potential as a clinical decision support tool for early diabetes detection. Early identification of diabetes is critical for preventing complications such as cardiovascular disease, nephropathy, neuropathy, and retinopathy (World Health Organization, 2016). By leveraging structured patient data—including glucose levels, BMI, and age—the developed SVM model can assist healthcare professionals in identifying high-risk patients promptly and accurately. Unlike traditional diagnostic methods, which rely solely on fasting blood glucose or oral glucose tolerance tests, machine learning-based models can analyze multiple patient parameters simultaneously to provide a comprehensive risk assessment. This aligns with the approach advocated by Han et al. (2020), who highlighted that machine learning techniques enable more nuanced and data-driven clinical decisions.

Conclusion:

This study successfully demonstrated the development and evaluation of a Support Vector Machine (SVM)-based predictive model for diabetes detection using Python and the PIMA Indians Diabetes Dataset. The model achieved a satisfactory classification accuracy of approximately 80%, with a notable AUC of 0.83, confirming its effectiveness in distinguishing between diabetic and non-diabetic individuals. The results are consistent with previous research findings (Sisodia and Sisodia, 2018; Han et al., 2020), reaffirming the suitability of SVM for binary classification tasks in healthcare applications. The importance of preprocessing through feature standardization, as emphasized by Pedregosa et al. (2011), was clearly demonstrated, with improved model performance after normalization. Key clinical features such as glucose concentration, BMI, and age emerged as significant predictors of diabetes risk, supporting both clinical understanding and previous data-driven studies (Zhang et al., 2020; Razzak et al., 2018). From a practical standpoint, the study highlights the potential of Python-based SVM models as assistive diagnostic tools in clinical environments, capable of supporting healthcare professionals in early diabetes detection and patient risk stratification. the integration of machine learning models like SVM into healthcare systems holds promise for enhancing diagnostic accuracy, reducing the burden on healthcare resources, and ultimately improving patient outcomes in diabetes management

Reference:

1. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

2. Sisodia, D., & Sisodia, D. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

3. Han, J., Kamber, M., & Pei, J. (2020). Data Mining: Concepts and Techniques (4th ed.). Morgan Kaufmann. https://www.elsevier.com/books/data-mining/han/978-0-12-381479-1

4. World Health Organization. (2016). Global Report on Diabetes. https://www.who.int/publications/i/item/9789241565257

5. International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). https://diabetesatlas.org/

6. Zhang, Y., Jiang, J., Chen, L., et al. (2020). Machine learning algorithms for diabetes prediction: A review. Healthcare Analytics, 1, 100001. https://doi.org/10.1016/j.health.2020.100001

7. Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. Neurocomputing, 335, 327–348. https://doi.org/10.1016/j.neucom.2017.09.023

8. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

9. UCI Machine Learning Repository. (2021). PIMA Indians Diabetes Database. https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

10. Brownlee, J. (2020). Imbalanced Classification with Python. Machine Learning Mastery. https://machinelearningmastery.com/imbalanced-classification-with-python/

11. Saxena, A., Sharma, M., et al. (2021). Machine learning models for prediction of type 2 diabetes mellitus: A systematic review. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 15(4), 102162. https://doi.org/10.1016/j.dsx.2021.102162

12. Devi, A., Batra, K., & Aggarwal, S. (2021). Diabetes prediction using supervised learning models. Materials Today: Proceedings, 45, 4415–4420. https://doi.org/10.1016/j.matpr.2020.12.985

13. Kumar, R., & Sahoo, S. (2020). Predictive modeling of diabetes diagnosis using machine learning techniques. Procedia Computer Science, 167, 706–716. https://doi.org/10.1016/j.procs.2020.03.317

14. Kavakiotis, I., Tsave, O., Salifoglou, A., et al. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

15. Wang, J., Zhu, S., & Yang, J. (2019). Prediction of diabetes using a support vector machine approach. Healthcare Technology Letters, 6(4), 98–102. https://doi.org/10.1049/htl.2018.5090