"Diabetes & Heart Disease Prediction System"

Neha Mahajan
B.Tech.A.I.&D.S
Dr D. Y. Patil School of Science & Technology, Dr D. Y. Patil Vidyapeeth, Pune, India.
23150027.dypsst@dpu.edu.in

***Abstract*— The "Multiple Disease Prediction System" is an advanced application designed to predict the likelihood of multiple diseases such as diabetes, and heart disease within a single, unified platform. By utilizing machine learning techniques, particularly Support Vector Classifier(SVC) and Logistic Regression algorithms, the system efficiently analyzes health data to provide accurate predictions. Built with a front-end interface using the Pickel framework, the system is user-friendly and accessible. It allows individuals to input their health data and receive immediate insights into their risk for various diseases. The project involved importing necessary libraries, preprocessing data, training and evaluating models on datasets from platforms like Kaggle, and deploying the model through Pickel. This comprehensive approach not only simplifies the process of disease prediction but also enhances early diagnosis and health management, potentially leading to better patient outcomes by enabling timely medical intervention. The key feature of our code is after risk prediction the advantage of Prevention tips the customer can get.**

Keywords:
Multiple Disease Prediction System, Machine Learning, Support Vector Machine (SVM), Logistic Regression, Pickel Framework, Data Preprocessing, Early Diagnosis, Health Management, Timely Medical Intervention.

I.INTRODUCTION

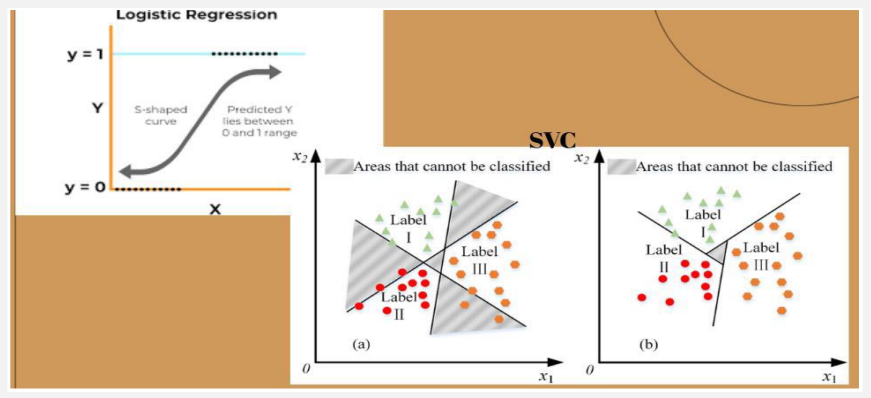
The fact that diseases are faced currently allows for a serious number of tools needed to detect them in their early stages without causing delays in diagnosis. This challenge is commonly faced by the "Multiple Disease Prediction Web Application" providing an efficient platform for the prediction of diseases according to symptoms described by the users. Early detection is crucial for diseases such as heart disease and diabetes, where prompt treatment can save many lives [1]. This approach also provides an opportunity to tailor healthcare services for every patient's needs, which is invaluable in modern pluralistic healthcare [2].

This application realizing effective prediction models uses Python with the frameworks machine learning, including TensorFlow, Scikit-Learn, and NLTK. These tools help the software predict a wide variety of diseases accurately and rapidly as it trains from vast datasets [3] Moreover, integration with natural language processing will enable the application to interpret symptoms in everyday language. Hence, even experts may not apply this, thus having a nonexpert user-friendly experience while maintaining the precision required for a medical application [4].

Applying machine learning technologies such as random forests, decision trees, and deep learning methods in the application, can predict results, and it can adapt to various kinds of diseases [5]. Hybrid models, for instance, have already begun showing promising results in cardiac disease prediction, and they may revolutionize the healthcare sector [6]. In the same way, diabetes predictive machine learning systems prove effective at analyzing patient data enabling early stages detection of people in potential risk [7] [8].

The web-based nature of the application guarantees accessibility benefits. In this respect, users can input symptoms from any internet-connected device for immediate predictions [9]. Its user-friendly interface ensures that even those with limited healthcare experience can easily input symptoms and access relevant information, providing a personalized experience [10]. This project is highly valued in its application both to patients and healthcare professionals, based on simplicity, accessibility, and accuracy in diagnostics.

The "Multiple Disease Prediction Web Application" bridges the gap between machine learning research and practical healthcare solutions with usability in consideration. Rapid processing and analysis of user input by the system accelerate diagnoses, increase the workload on medical staff, and empower users to take proactive health management actions. This initiative demonstrates the capability of sophisticated technology to use smarter and more efficient healthcare services



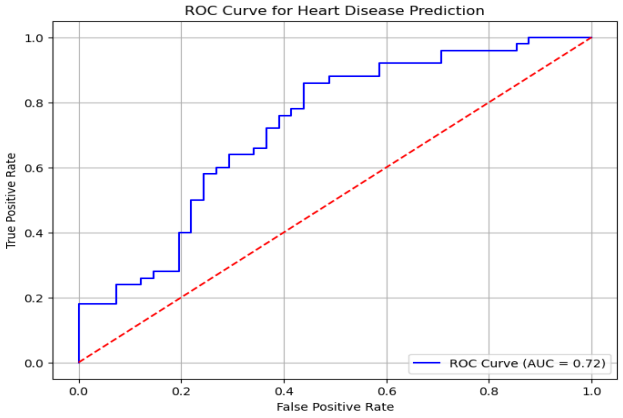
LOGISTIC REGRESSION MODEL

Overview of the Logistic Regression Model
Logistic regression is a supervised learning algorithm used for binary classification problems. It is a statistical model that estimates probabilities using a logistic function, also known as the sigmoid function.
Logistic Regression Model (for Heart)
Overview of the Logistic Regression Model:-
Logistic regression is a supervised learning algorithm used for binary classification problems. It is a statistical model that estimates probabilities using a logistic function, also known as the sigmoid function. The logistic function maps any real-valued number to a value between 0 and 1, which can be interpreted as a probability.
- Logistic Regression Operations and Workings: 1. Data Preparation: The dataset is prepared by splitting it into features (X) and the target variable (y).
2. Model Training: The logistic regression model is trained using the training data. The model learns the relationship between the features and the target variable.
3. Logistic Function: The logistic function is used to estimate the probability of the target variable being 1 (positive class) given the features.
4. Optimization: The model's parameters are optimized using an optimization algorithm, such as gradient descent.
5. Prediction: The trained model is used to make predictions on new, unseen data.

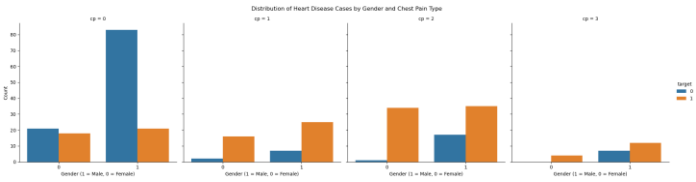
ROC Curve (Receiver Operating Characteristics)

>-Performance Assessment: ROC curves visualize logistic regression model performance for heart disease prediction.
-AUC Metric: AUC(Area Under the Curve.) measures model discrimination, with values closer to 1 indicating better accuracy.

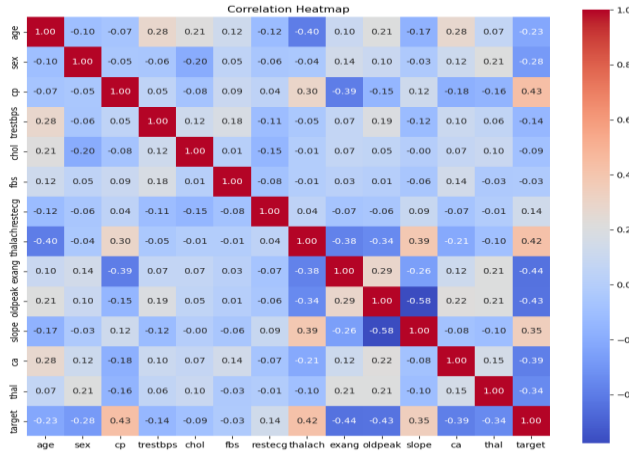
-Threshold Optimization: ROC analysis helps choose the best classification threshold to balance sensitivity and specificity.



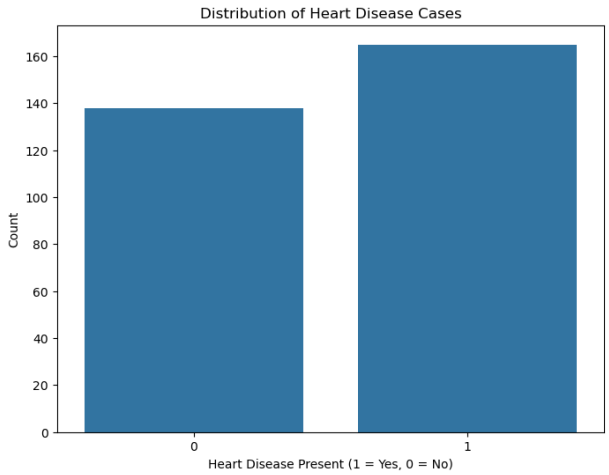
Catplot to visualize multiple categorical variables :-



Correlation Heatmap:-



Count Plot
(Heart Disease cases):-



Accuracy :- Logistic regression model :- 80 %

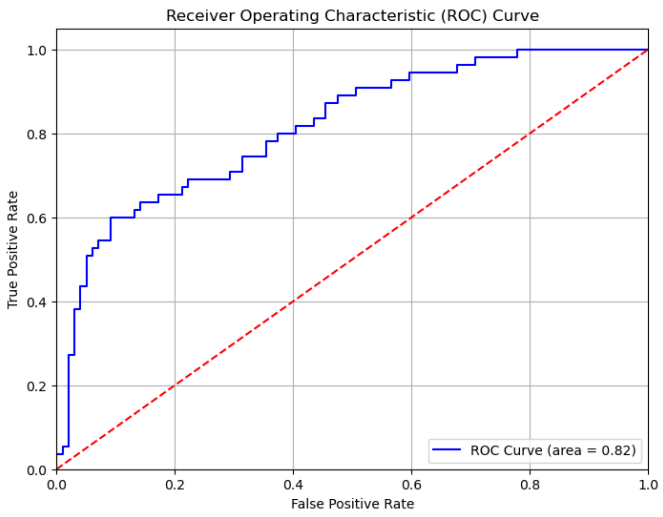
SUPPORT VECTOR CLASSIFIER

SVC (Support Vector Classifier):-
Diabetes Overview of the Support Vector Classifier (SVC)
Support Vector Classifier (SVC) is a supervised learning algorithm used for binary classification problems. It works by finding the optimal hyperplane that separates the data into different classes. It's especially effective in high-dimensional spaces and can handle complex, non-linear data.

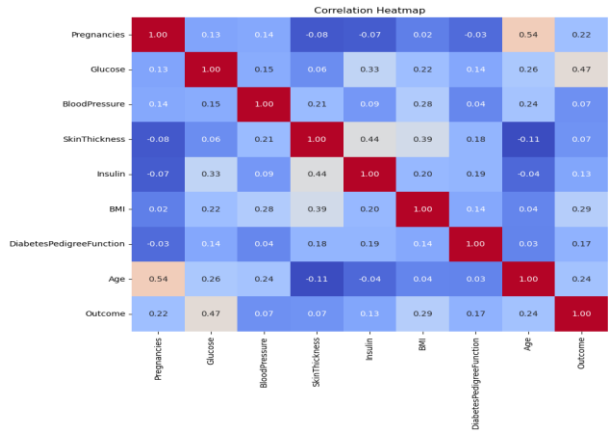
SVC Operations:-
-Data Preparation: The dataset is split into features (X) and the target variable (y), representing whether a person has diabetes (1) or not (0).
-Model Training: SVC learns to separate the classes by finding the best hyperplane that divides the data into diabetic and non-diabetic groups.
Kernel Function: Different kernel functions (linear, RBF, etc.) are applied to transform the data, enabling better separation of non-linear relationships.
-Optimization: The model maximizes the margin between the closest data points (support vectors) and the hyperplane, ensuring optimal classification.
-Prediction: After training, SVC predicts whether new data corresponds to diabetes or not.

Uses of SVC:
-Binary Classification: Ideal for predicting diabetes (0 or 1).
-Handles High Dimensions: Works well with health data features like glucose, BMI, etc.
-Non-Linear Boundaries: Can model complex relationships using kernels like RBF.
-Robust Generalization: Effective with unseen data, important for real-world diabetes prediction.

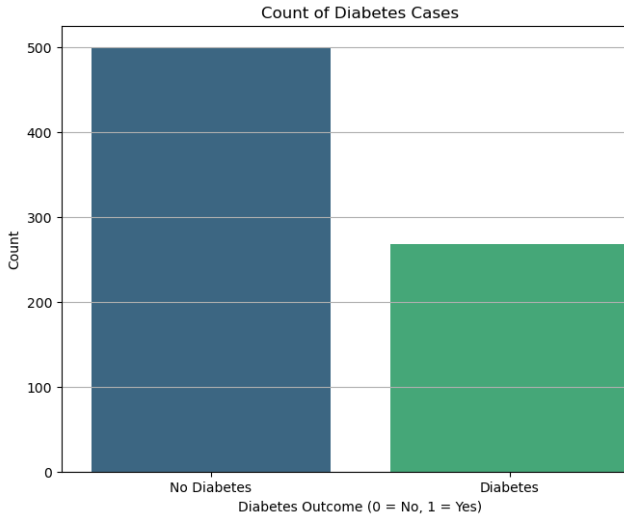
ROC Curve
(Reciever Operating Characteristics):-
-Visualization: ROC curves provide a visual representation of a model's performance, helping to understand how well the SVC model distinguishes between positive and negative classes at different classification thresholds.
-Comparison: ROC curves can be used to compare the performance of multiple SVC models or other classification algorithms, identifying the model with the best overall performance.
-Bias Detection: ROC curves can help detect potential biases in the model. If the curve is skewed towards one class, it might indicate that the model is performing better for one class than the other.



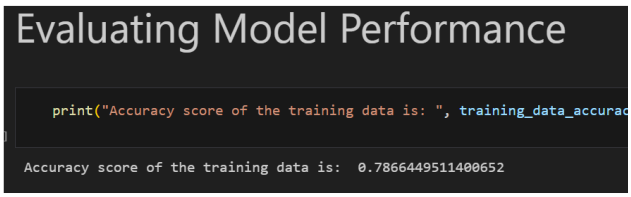
Heatmap for Correlation matrix:-



Count Plot (Diabetic or non-diabetic):-



Accuracy for diabetes(SVC) :-78%



LIMITATIONS.

1. Limitations of Logistic Regression Model

• Logistic regression assumes a linear relationship between the features and the target variable, which may not always be the case.

• Logistic regression is sensitive to outliers in the data, which can affect the model's performance.

• Logistic regression is not suitable for multi-class classification problems, as it is designed for binary classification.

• Logistic regression assumes that the observations are independent, which may not always be the case.
• Logistic regression is not robust to missing values in the data, which can affect the model's performance.

2. Limitations of Support Vector Classifier (SVC) Model

• SVM can be computationally expensive, especially when dealing with large datasets.

• SVM is sensitive to the choice of kernel, which can affect the model's performance.

• SVM is not suitable for noisy data, as it can be affected by outliers and noise.

• SVM can be affected by the curse of dimensionality, which can lead to overfitting.

• SVM can be affected by the curse of dimensionality, which can lead to overfitting.

CONCLUSION AND FUTURE SCOPE

-Model Performance: Both logistic regression and SVC performed well, with logistic regression achieving a slightly higher accuracy of 80% compared to SVC's 78%.

--Model Selection: While both models are effective, logistic regression might be a slightly better choice based on the accuracy results in this specific case.

--Further Analysis: It's important to consider other factors beyond accuracy, such as interpretability, computational cost, and domain knowledge when selecting a model for a particular application.

Future Scope: -

• Explore deep learning models for disease prediction

• Leverage transfer learning for disease prediction

• Develop multi-class classification models for predicting multiple diseases

• Handle imbalanced data to improve model performance

• Improve model interpretability through explainability techniques

• Deploy models in real-world settings for practical evaluation

• Use data augmentation to increase dataset size and improve model performance

• Optimize model performance through hyperparameter tuning

REFERENCES

[1] M. C. T. &. G. S. S, ""Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, 2019. [Online]. Available:

https://doi.org/10.1109/ACCESS.2019.2923707. [Accessed 2024].

[2] C. &. V. V. Cortes, “Support-vector networks,” Machine Learning, 1995. [Online]. Available: https://link.springer.com/article/10.1007/BF00994018. [Accessed 2024].

[3] V. Vapnik, “Statistical learning theory,” John Wiley & Sons, 1998. [Online]. Available: https://www.wiley.com/en-us/Statistical+Learning+Theory-p9780471030034. [Accessed 2024].

[4] I. &. E. A. Guyon, “ An introduction to variable and feature selection,” Journal of Machine Learning Research, 2003. [Online]. Available: https://dl.acm.org/citation.cfm?id=944919. [Accessed 2024].

[5] T. e. a. Hastie, “The elements of statistical learning: Data mining, inference, and prediction,” Springer, 2017. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-84858-7. [Accessed 2024].

[6] M. I. &. M. T. M. Jordan, “Machine learning: Trends, perspectives, and prospects,” Science, 349(6245), 2015. [Online]. Available: https://science.sciencemag.org/content/349/6245/255. [Accessed 2024].

[7] P. Domingos, “A few useful things to know about machine learning,” Communications of the ACM, 2012. [Online]. Available: https://dl.acm.org/citation.cfm?id=2347736. [Accessed 2024].

[8] M. &. J. K. Kuhn, “Applied predictive modeling,” Springer, 2013. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4614-6849-3. [Accessed 2024].

[9] G. e. a. James, “An introduction to statistical learning: With applications in R,” Springer, 2013. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4614-7138-7. [Accessed 2024].

[10] J. &. B. Y. Bergstra, “Random search for hyper-parameter optimization,” Journal of Machine Learning Research, 2012. [Online]. Available: https://dl.acm.org/citation.cfm?id=2188385. [Accessed 2024].