Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

NAME: **Keval Ramesh Shah** BATCH: **D2-1** SAP ID: **60009220061**
COURSE NAME: **Machine Learning - I Laboratory** DATE: **25 / 04 / 2024**

Mini Project

Title: Predicting Diseases from User Symptoms

Aim: The project aims to develop an accurate machine learning model capable of predicting diseases based solely on user-input symptoms. By leveraging machine learning algorithms, the goal is to create a system that can analyse symptom data and provide reliable predictions regarding potential diseases or health conditions associated with those symptoms. This project seeks to improve healthcare efficiency by providing initial diagnostic support, thereby enabling timely medical attention and guidance for individuals exhibiting symptoms.

Name of the dataset: Disease Prediction Using Machine Learning

Source of dataset:
Disease Prediction Using Machine Learning (kaggle.com)

Dataset Description:

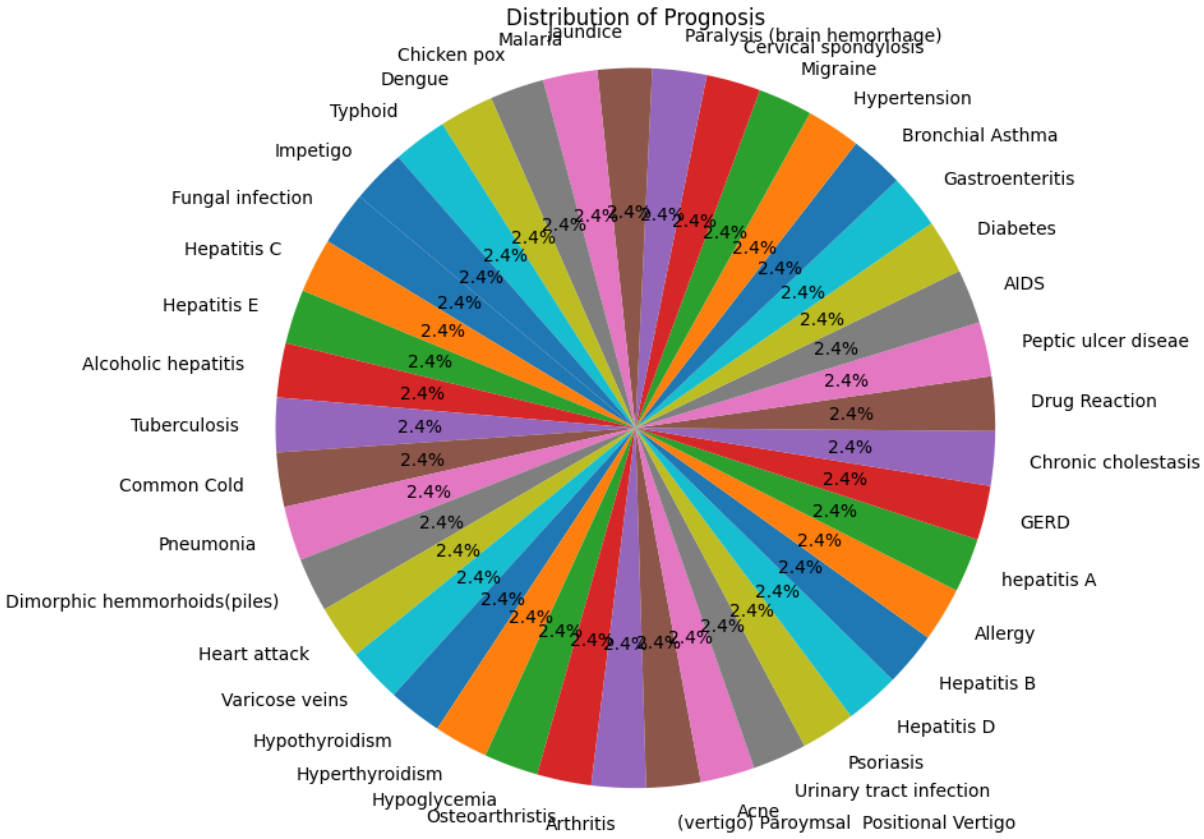
The dataset consists of a single CSV file designed for predictive modeling to classify symptoms into 42 different types of diseases. Each record in the dataset contains 133 columns, with 132 columns representing various symptoms experienced by individuals, encoded as binary values (0 and 1), and the last column indicating the prognosis or predicted disease.

# itching	# skin_rash	# nodal_skin_erupti...	# continuous_snee...	# shivering	#
0 1	0 1	0 1	0 1	0 1	0
1	1	1	0	0	0
0	0	0	1	1	1
0	0	0	0	0	0
1	0	0	0	0	0
1	1	0	0	0	0
0	0	0	0	0	0

Content:

Data Structure: Each record in the dataset includes symptom parameters encoded as binary values (0 indicating absence and 1 indicating presence) and the corresponding disease prognosis. The symptoms are mapped to 42 different diseases for classification purposes.

Distribution of Data for Different Prognosis :



The diagram depicts the frequency or proportion of each prognosis category within the dataset. Each bar or segment represents a distinct prognosis, allowing for a comparative analysis of disease prevalence.

Dataset Specifications:

Number of Parameters: 132
Types of Diseases: 42
Encoding: Symptom parameters are represented as binary values (0 and 1), indicating absence or presence, respectively.
File Format: CSV (Comma-Separated Values)

Data Preprocessing :

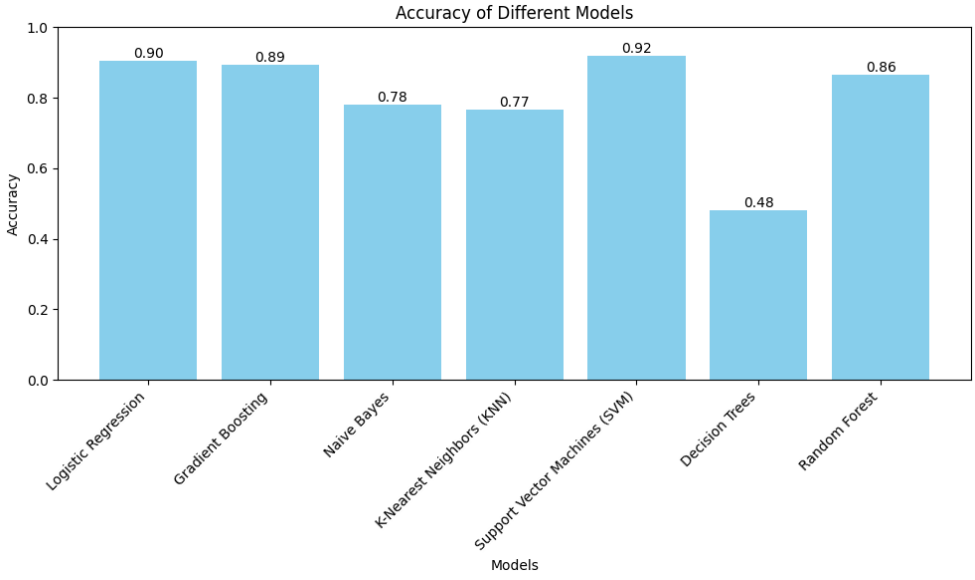
Handling Missing Values:
A thorough examination of the dataset revealed the absence of null or missing values across all symptom columns. This absence of missing data underscores the dataset's quality and completeness, eliminating the need for imputation or removal of records with missing values.

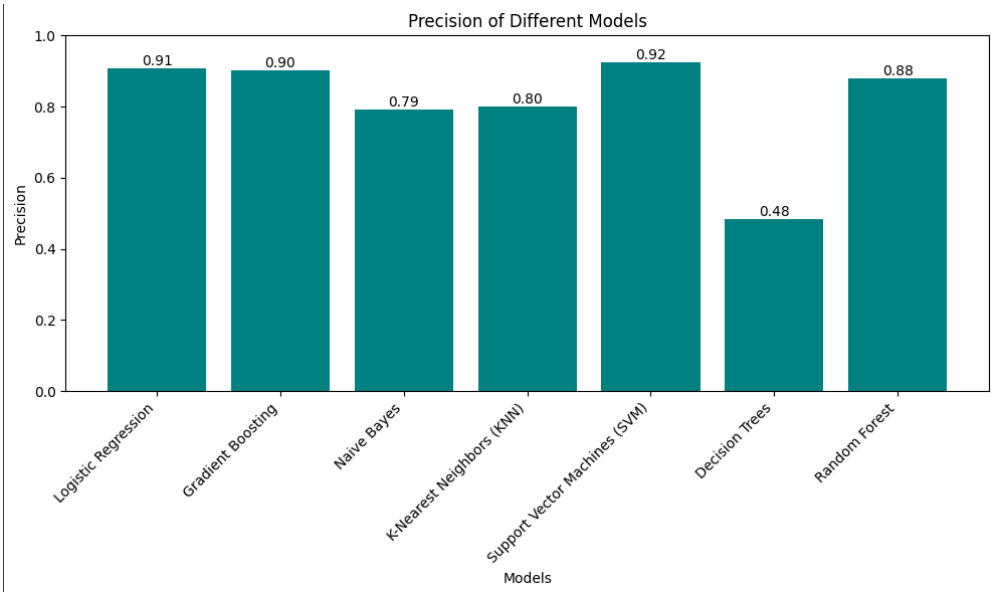
Confirmation of Binary Encoding:

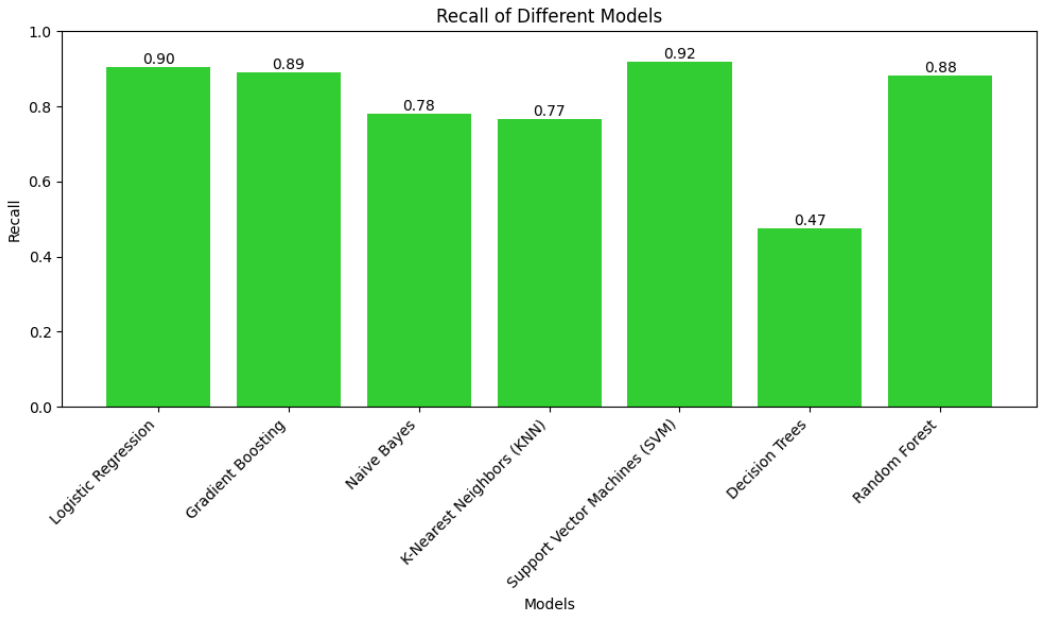
While the dataset is free from missing values, it remains imperative to confirm the correctness of the binary encoding across all symptom columns. Ensuring the accuracy of binary encoding guarantees the integrity of symptom representations and prevents any inconsistencies that may arise during model training and inference.

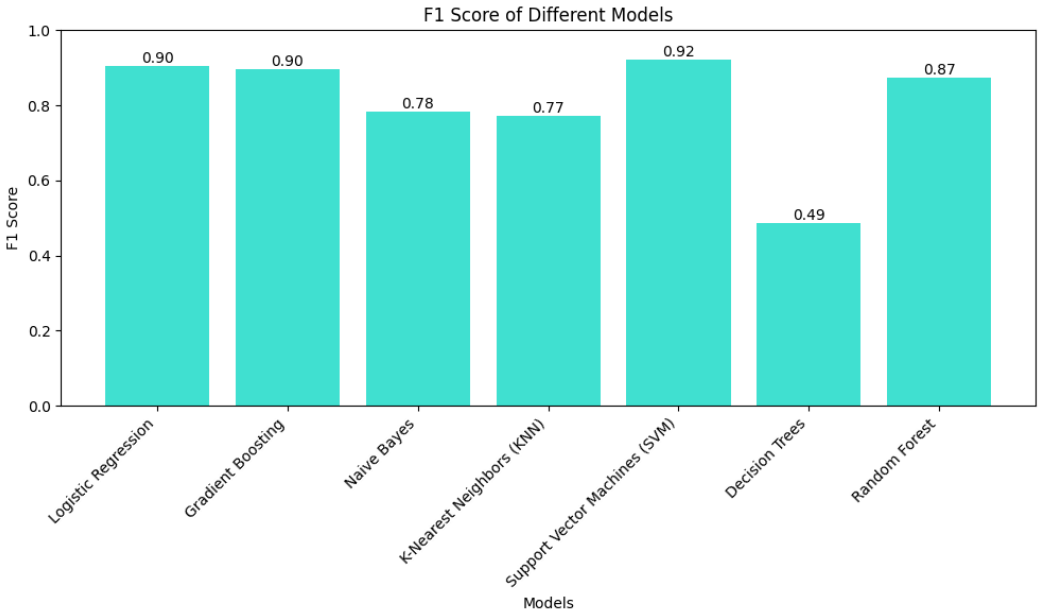
Data Modeling :

In our investigation, we explored the performance of seven distinct machine learning models for disease prediction based on symptom parameters. Employing logistic regression, gradient boosting, naive Bayes, K-nearest neighbors, support vector machines, decision trees, and random forest algorithms, we aimed to identify the most effective approach for accurate prognosis classification. Each model was trained and evaluated using the preprocessed dataset, with performance metrics such as accuracy, precision, recall, and F1-score calculated to assess predictive performance.









Based on our evaluation metrics, we chose the Support Vector Machines (SVM) model as the best candidate for our multi-class disease prediction task. Here's why:

High Accuracy: The SVM model exhibited the highest accuracy among all models evaluated, achieving an accuracy of 92%. This indicates that the SVM model made the most accurate predictions across all disease classes compared to other models.

Precision and Recall: The SVM model demonstrated high precision (92%) and recall (92%) values. High precision implies that the SVM model accurately identified instances of each disease class without misclassifying other diseases. Similarly, high recall indicates that the SVM model effectively captured most instances of each disease class.

F1 Score: The SVM model achieved a high F1 score of 0.92, which is a combined measure of precision and recall. This balanced metric suggests that the SVM model successfully managed false positives and false negatives, making it robust in handling the multi-class classification task.

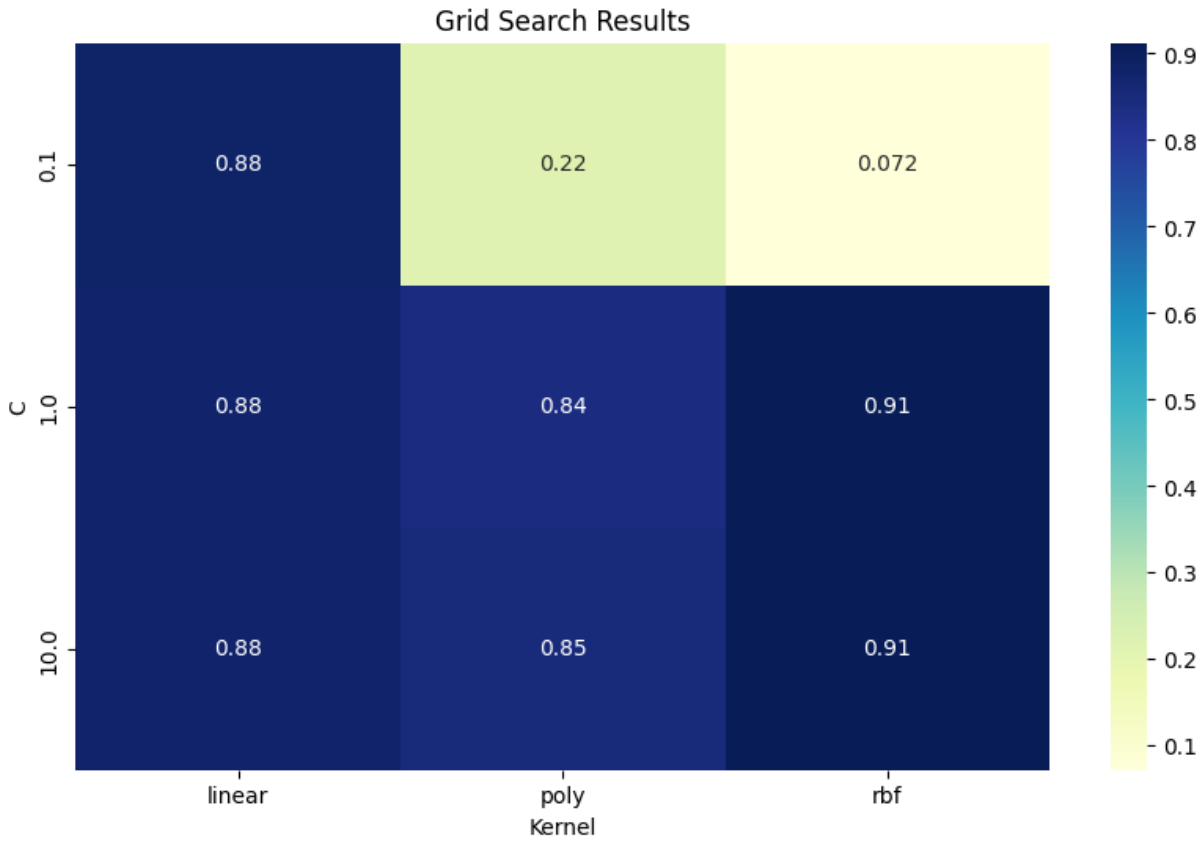
Model Intuition: SVMs are effective for multi-class classification tasks, especially when dealing with high-dimensional feature spaces and complex decision boundaries. SVMs can efficiently handle non-linear relationships between symptoms and diseases, which is crucial in our dataset with 132 different disease classes.

In summary, the SVM model stood out due to its exceptional performance across various evaluation metrics, its ability to handle multi-class classification tasks effectively, and its capability to capture complex relationships in the dataset. Therefore, we chose the SVM model as the optimal solution for predicting diseases based on symptoms in our study.

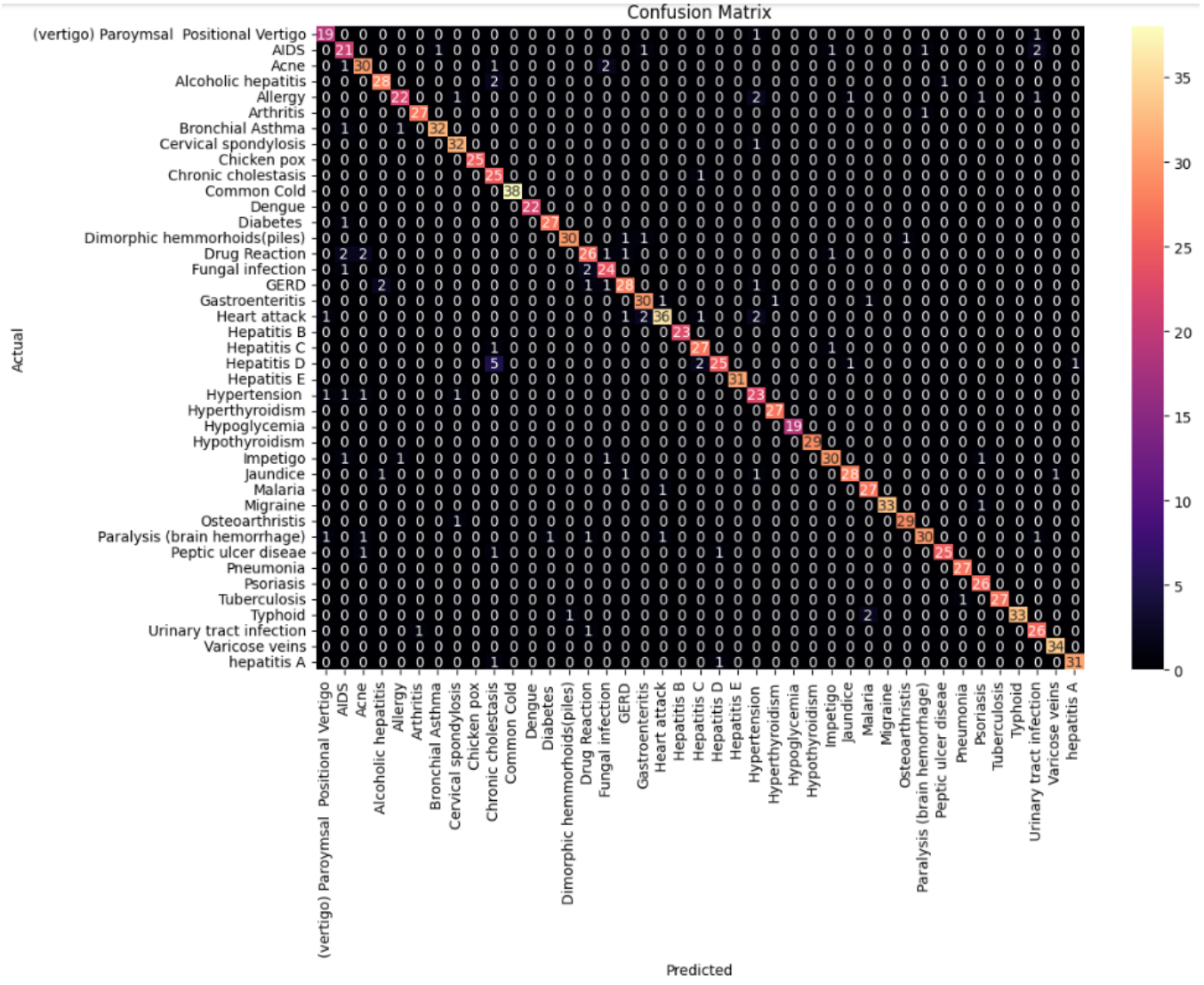
Fine-Tuning the Support Vector Machines (SVM) Model

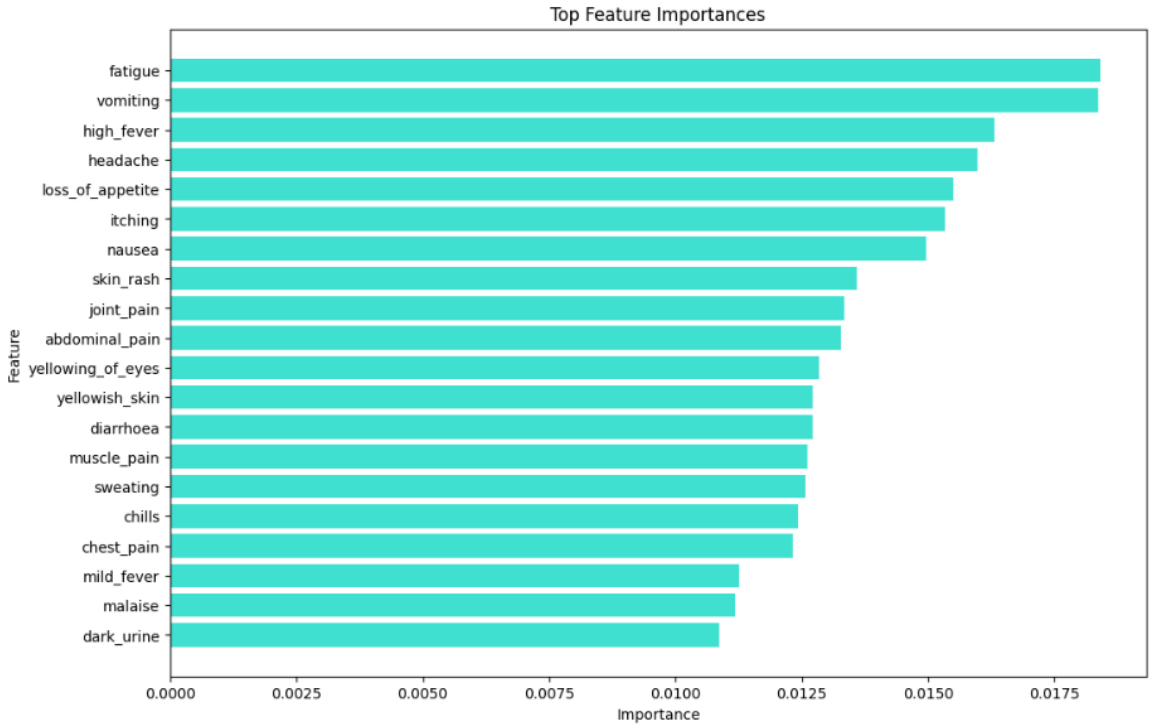
```
Best Parameters: {'C': 1, 'kernel': 'rbf'}
Accuracy: 0.9203252032520325
```

To optimize our Support Vector Machines (SVM) model's performance, we conducted fine-tuning via grid search. The best parameters obtained were a regularization parameter (C) of 1 and a radial basis function (RBF) kernel. This fine-tuned model achieved an accuracy of 92.03%, demonstrating its effectiveness in accurately predicting disease prognoses based on symptom parameters.



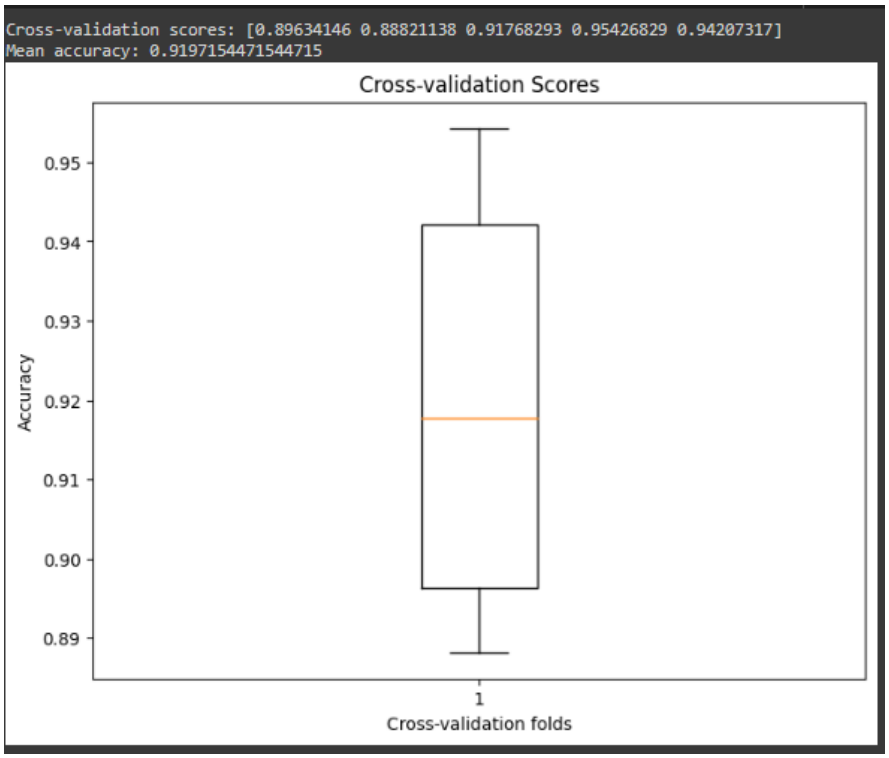
Confusion Matrix :





Cross-Validation Evaluation Results

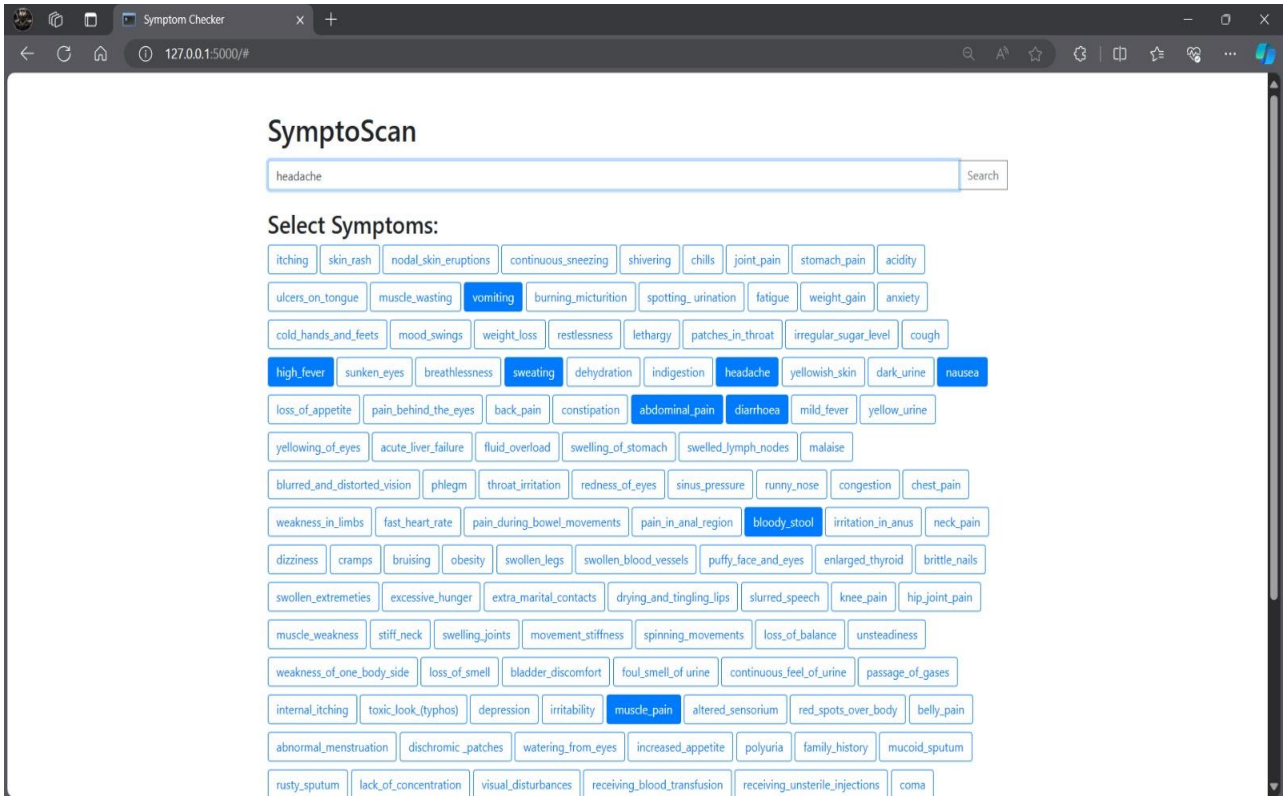
In our evaluation of the Support Vector Machines (SVM) model, we utilized cross-validation as a robust technique to assess its generalization performance. The mean accuracy across all cross-validation folds was calculated to be approximately 91.97%. This metric provides an aggregate measure of the model's performance across multiple validation folds, offering insight into its overall effectiveness in disease prediction.



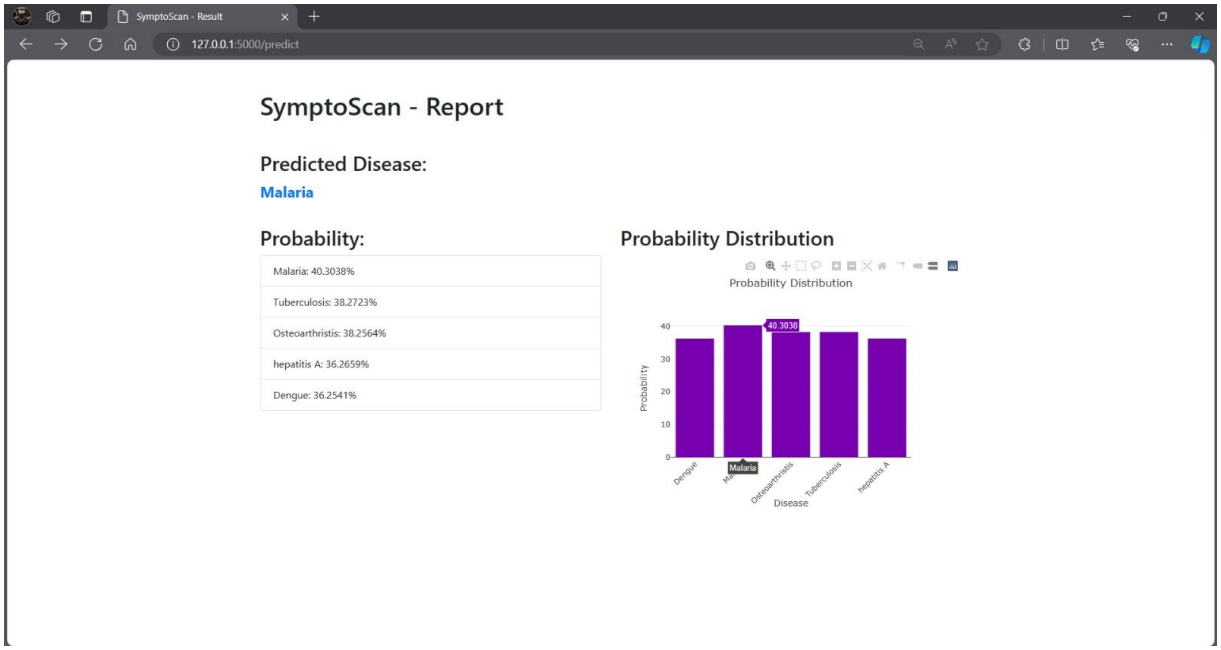
The consistent accuracy scores obtained across different cross-validation folds highlight the stability and reliability of the SVM model.

Flask Deployment :

Our machine learning model has been successfully deployed using Flask, providing a user-friendly interface for disease prediction. The deployed web application features a search bar where users can input symptom keywords and a group of buttons with symptom names for convenient selection.



Upon submission, the application generates predictions for potential diseases along with corresponding probabilities. Additionally, users receive a confidence report indicating the model's certainty in predicting the disease. This intuitive interface empowers users to quickly and accurately assess potential health outcomes based on symptom presentation.



Other Predictions :

SymptoScan - Report

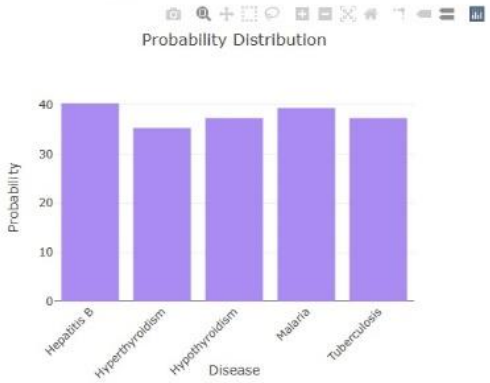
Predicted Disease:

Hepatitis B

Probability:

- Hepatitis B: 40.2866%
- Malaria: 39.2913%
- Tuberculosis: 37.2620%
- Hypothyroidism: 37.2597%
- Hyperthyroidism: 35.2505%

Probability Distribution



SymptoScan - Report

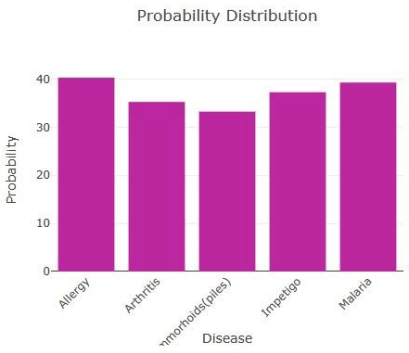
Predicted Disease:

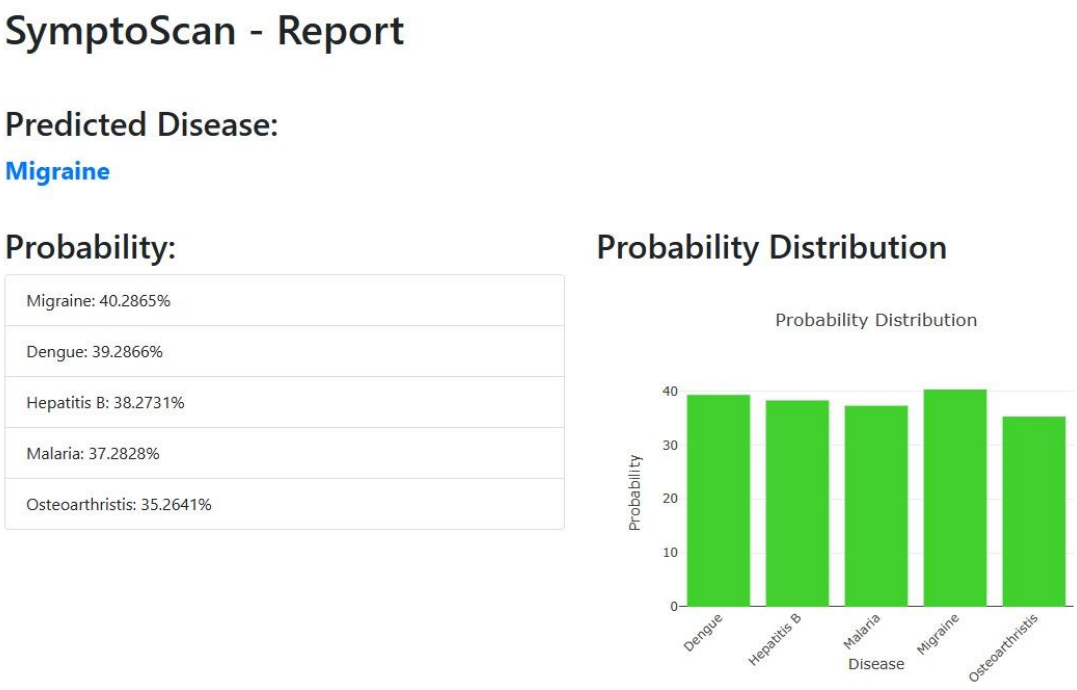
Allergy

Probability:

- Allergy: 40.2990%
- Malaria: 39.2910%
- Impetigo: 37.2730%
- Arthritis: 35.2529%
- Dimorphic hemmorhoids(piles): 33.2380%

Probability Distribution





Conclusion : Predicting Diseases from User Symptoms

In this project, we developed a ML model to predict diseases based on user-reported symptoms. Using a dataset containing symptom profiles and corresponding disease labels, we explored several classification algorithms to identify the most effective model.

After evaluating multiple models including Logistic Regression, Gradient Boosting, Naive Bayes, K-Nearest Neighbours, Support Vector Machines, Decision Trees, and Random Forest, we found that the Support Vector Machines (SVM) model exhibited the highest accuracy, precision, recall, and F1 score among the tested algorithms.

To further enhance the performance of the SVM model, we conducted hyperparameter tuning using GridSearchCV, optimizing parameters such as the regularization parameter (C) and kernel type. As a result, we achieved an accuracy of 92.03% on the test dataset, indicating the robustness and effectiveness of our tuned SVM model. we employed cross-validation obtaining consistent accuracy scores across multiple folds. Additionally, we conducted feature importance analysis using Random Forest to identify the most influential symptoms for disease prediction.

Overall, this project demonstrates the feasibility and utility of machine learning in healthcare applications, showcasing how predictive models can aid in disease diagnosis and decision-making processes.