**< Disease Prediction Using Machine Learning>**

*Submitted for   
Statistical Machine Learning CSET211*

Submitted by:

*(<E23CSEU0677>) Pulkit Malik   
(<E23CSEU0667>) Sanskar Sengar*

*Batch -23 (G-06)*

Submitted to  
DR. Sanchali Das

July-Dec 2024  
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A close-up of a logo

Description automatically generated

# INDEX

|  |  |
| --- | --- |
| **Sr.No** | **Content** |
| 1 | Abstract |
| 2 | Introduction |
| 3 | Related Work (If Any) |
| 4 | Methodology |
| 5 | Hardware/Software Required |
| 6 | Experimental Results |
| 7 | Conclusions |
| 8 | Future Scope |
| 9 | GitHub Link of Your Complete Project |

# Abstract

In this project, we aimed to predict diseases based on the symptoms observed in patients using machine learning techniques. The dataset used consists of 133 columns, where 132 represent different symptoms and one represents the disease prognosis. Various machine learning classifiers such as Support Vector Classifier (SVC), Gaussian Naive Bayes (NB), and Random Forest were trained on the data. By using K-Fold cross-validation and combining the predictions of all three classifiers, we developed an ensemble model to enhance accuracy. The system provides predictions on disease outcomes based on input symptoms and was evaluated using accuracy scores and confusion matrices.

# Introduction

The goal of this project is to build an accurate machine learning model for predicting diseases based on symptoms. Early detection of diseases based on symptoms can significantly improve treatment outcomes. Machine learning has been used to automate the diagnosis process, especially with large datasets of symptoms and diseases available. The dataset used for this project includes 132 symptom features, and the target column contains the disease prognosis. This project combines the predictions of multiple classifiers to create a robust disease prediction model.

# Methodology

The project was carried out in the following steps:  
  
1. **Data Gathering**: The dataset used was obtained from Kaggle and consists of 133 columns, where 132 represent symptoms and one represents disease prognosis.  
2**. Data Preprocessing**: The target column (prognosis) was encoded using `LabelEncoder`, and any null values were removed from the dataset.  
3. **Model Building**: Three classifiers were chosen for the project: Support Vector Classifier (SVC), Gaussian Naive Bayes (NB), and Random Forest (RF). These models were trained on the data.  
4. **K-Fold Cross-Validation**: To evaluate model performance, K-Fold Cross-Validation was used, where the dataset was split into K subsets.  
5. **Ensemble Model**: The predictions from all three models were combined using the mode of their predictions to improve robustness.  
6. **Prediction Function**: A function was created to take input symptoms and predict the disease based on the ensemble model.

# Hardware/Software Required

Hardware:  
- A computer with at least 8GB RAM and a multi-core processor is recommended.  
  
Software:  
- Python 3.x  
- Libraries: `numpy`, `pandas`, `matplotlib`, `seaborn`, `scikit-learn`, `scipy`  
- Development Environment: Jupyter Notebook or any Python IDE (e.g., PyCharm, VSCode)  
- Operating System: Windows, macOS, or Linux

# Experimental Results

The performance of each model was evaluated based on accuracy and confusion matrices:  
  
1. \*\*Model Performance\*\*: The SVC and Naive Bayes models achieved \*\*100% accuracy\*\* on both training and testing datasets. The Random Forest model achieved an accuracy of \*\*98.6%\*\* on the testing dataset.  
  
2. \*\*Accuracy Comparison\*\*:  
 - \*\*SVC\*\*: 100%  
 - \*\*Naive Bayes\*\*: 100%  
 - \*\*Random Forest\*\*: 98.6%

# Confusion Matrices

The confusion matrices for each model were visualized using heatmaps. These matrices show the number of correct and incorrect predictions for each disease class.  
  
- \*\*SVC Confusion Matrix\*\*: The SVC model had very few misclassifications, showing high performance.  
- \*\*Naive Bayes Confusion Matrix\*\*: The Naive Bayes model also had minimal misclassifications.  
- \*\*Random Forest Confusion Matrix\*\*: The Random Forest model showed slightly more misclassifications but still performed well.

# Conclusions

This project demonstrated the power of machine learning for disease prediction based on symptoms. By using multiple classifiers and combining their predictions, we created a robust disease prediction system. The results show that ensemble learning significantly enhances prediction accuracy. However, future work can explore larger datasets and more advanced algorithms like deep learning for even better performance.

# Future Scope

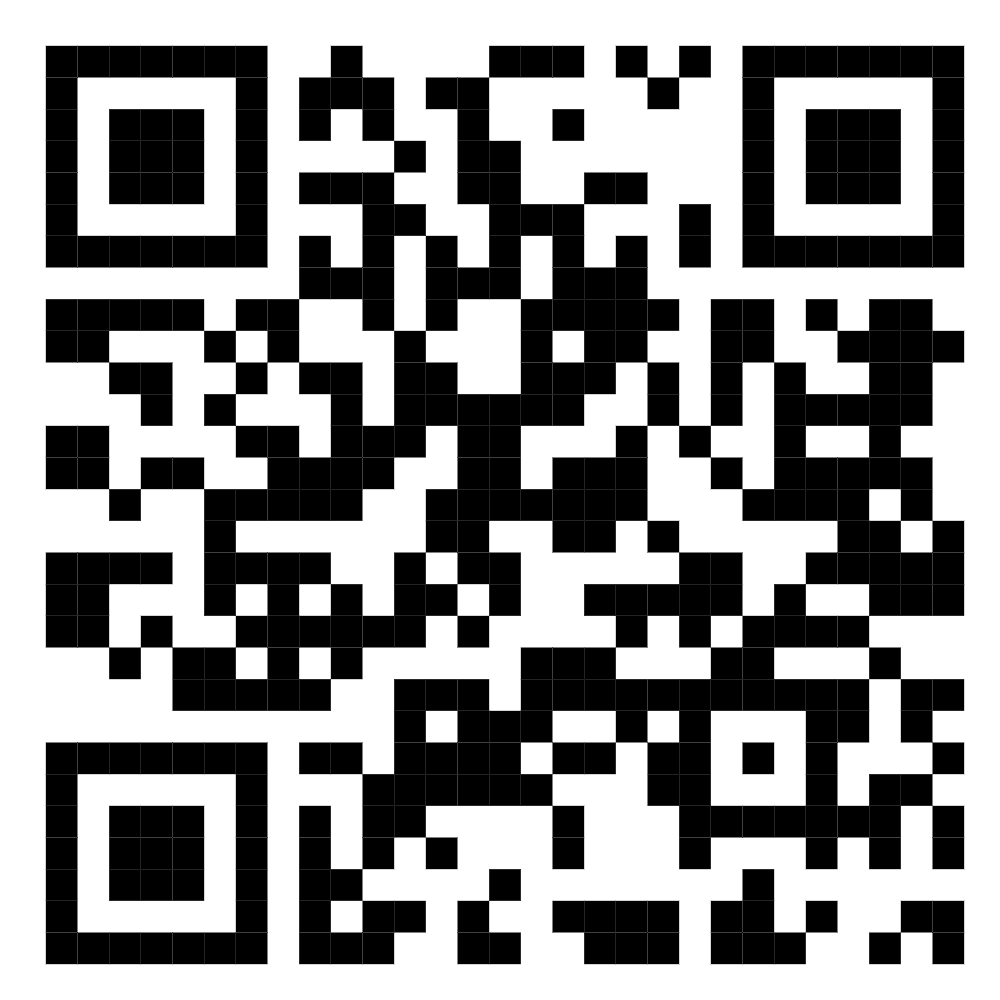
1. \*\*Larger Datasets\*\*: Future work can involve testing the model on more diverse and larger datasets to improve generalization.  
2. \*\*Advanced Algorithms\*\*: Deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) can be explored for more effective feature extraction.  
3. \*\*Real-time Prediction\*\*: Integrating the model into a real-time healthcare system to predict diseases instantly based on input symptoms.

# GitHub Link of Your Complete Project

****

**Pulkit Malik :**

GitHub Link: [https://github.com/Pulkitmalik-677/Sml-Project-Sem-3-.git](https://github.com/Pulkitmalik-677/Sml-Project-Sem-3-.git)  
  
**Sanskar Sengar :**

https://github.com/SanskarSinghSengar/Disease-Prediction-Using-Machine- Learning/blob/main/SML\_project\_Final%20(2).ipynb