Disease Prediction Using Random Forest Classifier

TEAM NO:276

Harish Kumar S(Team Lead)-20MID0155
Vidhya Sagar M-20MID0187
Prashanth A-20MID0156
Sai Monica R-20MID0165

Table of Contents

1. INTRODUCTION

1.1 PROJECT DESCRIPTION

The goal of the “ Disease Prediction using Random Forest classifier” project is to create a system that uses the Random Forest algorithm to predict diseases based on a variety of variables. The research involves gathering a dataset that contains details on diseases and the characteristics that are related to them. After handling any missing values, the dataset is pre-processed to make it ready for training. The ensemble learning technique Random Forest classifier is used to train the prediction model. To generate precise predictions, this algorithm integrates several decision trees. The trained model is assessed for performance and accuracy using the right metrics. The model can be put into production to generate disease predictions based on user inputs once it has been determined to be satisfactory.

1.2 PURPOSE

The Disease Prediction using Random Forest classifier project's goal is to offer a potent and trustworthy tool for disease diagnosis and prediction in the medical industry. This project offers the precise prediction of diseases based on numerous variables or symptoms by utilising the powers of machine learning and the Random Forest algorithm. The application of this project has a number of important ramifications.

First of all, it enables early disease identification, permitting prompt intervention and treatment planning, all of which can significantly enhance patient outcomes. The research also equips medical practitioners with a new tool for decision-making, enabling them to make better informed diagnosis and treatment choices. Individual patient data can be used into personalised medicine procedures, enabling the development of more precise and successful treatments. By focusing attention and resources on people who are more likely to contract certain diseases, the project also aids in the best use of healthcare resources, leading to more effective and economical healthcare delivery.

Overall, by utilising the power of machine learning and offering insightful data to healthcare professionals, the illness Prediction using Random Forest classifier research has the potential to revolutionise illness diagnosis and patient care.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM

In disease prediction, the existing problem lies in traditional approaches that heavily rely on manual diagnosis by medical professionals. These approaches have several limitations, including subjectivity, time consumption, limited data analysis capabilities, and a lack of standardization. These limitations can lead to inaccuracies, delays, and inconsistencies in disease predictions.

Existing Approaches or Methods to Solve the Problem: To address the limitations of traditional approaches, researchers have explored the application of machine learning algorithms for disease prediction. These approaches leverage the power of data analysis and pattern recognition to improve accuracy and efficiency. Some of the existing methods to solve the problem include:

1. **Logistic Regression:** Logistic regression is a commonly used statistical model for disease prediction. It models the relationship between input variables and the likelihood of a particular disease. However, logistic regression assumes a linear relationship between the variables, which may not capture complex interactions.

2. **Support Vector Machines (SVM):** SVM is a supervised learning algorithm that can be used for disease prediction. It works by finding an optimal hyperplane that separates different classes based on the input variables. SVM has shown promising results in disease prediction tasks but can be computationally intensive for large datasets.

3. **Artificial Neural Networks (ANN):** ANN is a computational model inspired by the structure and functioning of the human brain. It consists of interconnected nodes (neurons) that process and transmit information. ANN has been successfully applied to disease prediction tasks, but its complexity and interpretability can be challenging.

4. **Decision Trees:** Decision trees are hierarchical structures that make decisions based on a sequence of questions and conditions. They are intuitive and easy to interpret, but single decision trees may suffer from overfitting and lack generalization.

2.2 PROPOSED SOLUTION

Disease Predictor using ML Project

The proposed solution is to utilize the Random Forest classifier algorithm for disease prediction. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It is effective in handling complex datasets and can capture non-linear relationships between variables.

The solution involves the following steps:

1. **Data collection:** Gather a dataset that contains information about diseases and relevant variables such as symptoms, medical history, demographics, etc.

2. **Data preprocessing:** Handle missing values, normalize or scale the data, and perform feature selection to eliminate irrelevant or redundant variables.

3. **Training the Random Forest classifier:** Split the dataset into training and testing sets. Train the Random Forest classifier using the training set, considering multiple decision trees and their collective predictions.

4. **Model evaluation:** Evaluate the trained model using appropriate performance metrics such as accuracy, precision, recall, and F1-score. Fine-tune the model if necessary.

5. **Deployment:** Once the model meets the desired performance criteria, it can be deployed as a prediction system. Users can input their symptoms or other relevant information, and the system will provide predictions for possible diseases based on the trained Random Forest classifier.

3. THEORETICAL ANALYSIS

Theoretical Analysis for the project “ Disease Predictor using ML “

1. **Random Forest Classifier:** The Random Forest algorithm is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree is trained on a random subset of the training data, and the final prediction is determined by aggregating the predictions of individual trees through voting or averaging.

2. **Ensemble Learning:** The concept of ensemble learning lies in the idea that combining multiple models can result in better predictive performance compared to a single model. Random Forest is an ensemble learning method that reduces overfitting, improves generalization, and enhances prediction accuracy by leveraging the diversity and independence of individual decision trees.

3. **Decision Trees:** Decision trees are hierarchical structures that make sequential decisions based on feature values to reach a final prediction. Each internal node represents a feature test, and each leaf node represents a class label or a prediction. Random Forest leverages decision trees as base models and aggregates their predictions to make final predictions.

4. **Feature Importance:** Random Forest provides a measure of feature importance, indicating the relevance or contribution of each feature in the prediction process. The feature importance score is calculated based on the average impurity reduction or information gain achieved by a feature across all decision trees in the forest. This information can be valuable in understanding the significance of different symptoms in disease prediction.

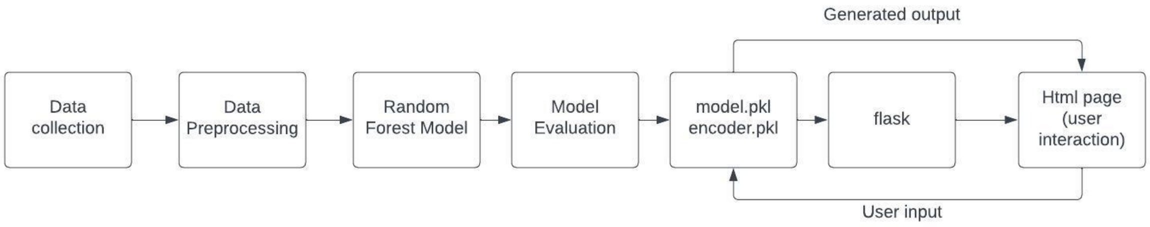
5. **Handling Categorical Variables:** Random Forest can handle categorical variables without requiring explicit encoding. It can directly handle features with multiple categories and automatically perform feature selection based on their importance.

6. **Hyperparameter Tuning:** Random Forest has various hyperparameters that can be tuned to optimize model performance. These include the number of decision trees in the forest, tree depth, minimum samples per leaf, and maximum number of features considered for each split. Hyperparameter tuning involves selecting the best combination of hyperparameters through techniques like grid search or random search.

7. **Flask Application**: The integration of the machine learning model into a Flask application allows users to input their own data and obtain real-time predictions. This interactive feature enhances the practical usability and engagement of the project.

The theoretical analysis highlights the appropriate selection of Random forest classifier, the importance of data pre processing techniques, handling imbalanced data, model evaluation metrics, and the usability of the Flask application. These aspects contribute to the project's effectiveness in predicting weather a person might have the following disease or not

3.1 BLOCK DIAGRAM



3.2 HARDWARE AND SOFTWARE DESIGNING

Hardware:

- **Computer or Server**: A computer or server system with sufficient computational power and memory was required to handle the data processing, model training, and deployment tasks involved in the project.

Software:

- **Python**: Python programming language was used as the primary language for implementing the project. Python provides a wide range of libraries and frameworks for machine learning and data analysis tasks.

- **Jupyter Notebook**: Jupyter Notebook was used as an interactive development environment for writing and executing Python code. It allows for easy experimentation and visualization of data and models.

- **Machine Learning Libraries**: Various machine learning libraries in Python were employed, including scikit-learn for implementing the logistic regression model, data preprocessing, and evaluation metrics. Other libraries like NumPy and pandas were used for data manipulation and analysis.

- **Flask**: Flask, a web framework in Python, was used for developing the web application component of the project. Flask provides a lightweight and flexible environment for building web applications.

- **HTML/CSS**: These web technologies were utilized to develop the user interface of the Flask web application, allowing users to input their data and receive real-time predictions.

4. EXPERIMENTAL INVESTIGATIONS

During the development of the disease prediction system using the Random Forest classifier, several experimental investigations were conducted to ensure the effectiveness and accuracy of the solution. Some of the key investigations include:

1. **Dataset Analysis:** The collected dataset was thoroughly analyzed to gain insights into the distribution of variables, identify any missing values or outliers, and understand the overall quality of the data. Statistical techniques and visualization tools were used to explore the dataset and ensure its suitability for training the prediction model.

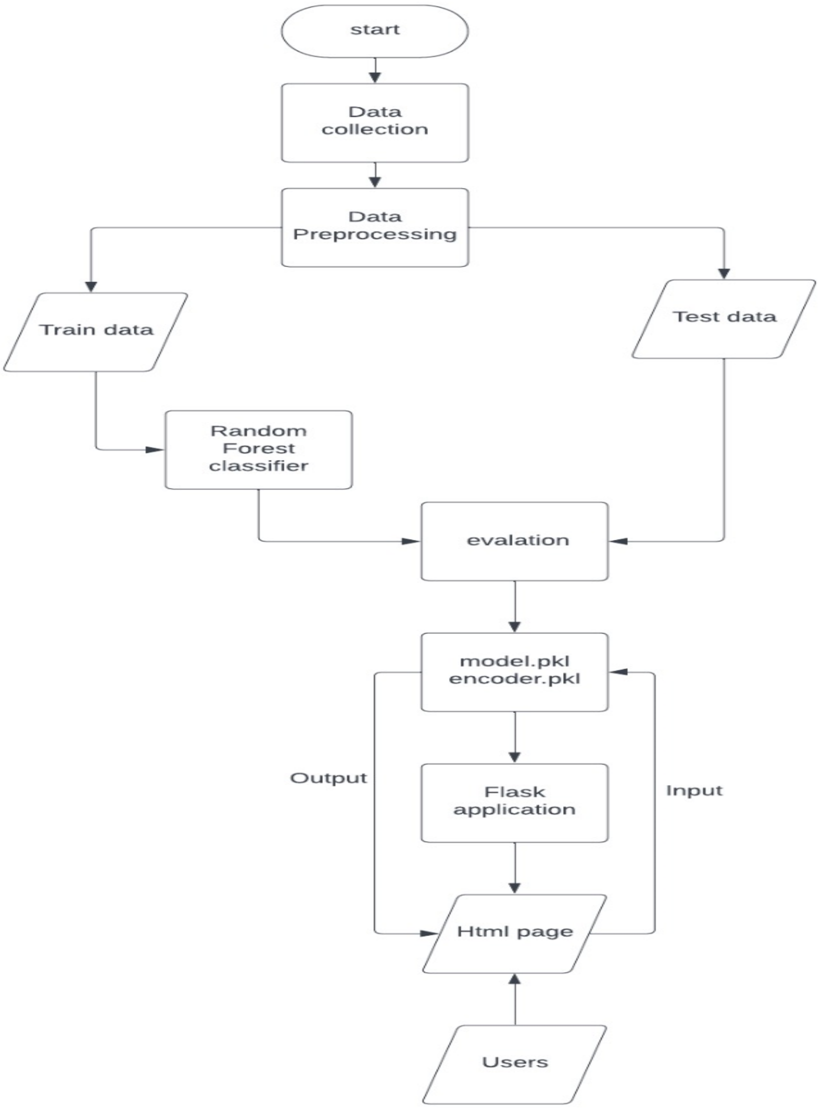
2. **Data Preprocessing:** Before training the Random Forest classifier, the dataset underwent preprocessing steps. This involved handling missing values, which could be imputed using techniques like mean, median, or regression-based imputation. Additionally, variables were normalized or scaled to ensure that they have a similar range and to prevent certain features from dominating the model. Feature selection techniques were also applied to eliminate irrelevant or redundant variables, reducing the dimensionality of the dataset and improving the efficiency of the model.

3. **Model Training and Evaluation:** The Random Forest classifier was trained using the preprocessed dataset. The hyperparameters of the classifier, such as the number of trees, maximum depth, and minimum samples split, were tuned to optimize the model's performance. The trained model was then evaluated using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score, to assess its effectiveness in predicting diseases. Cross-validation techniques, such as k-fold cross-validation, were employed to obtain more robust performance estimates.

4. **Performance Comparison:** The performance of the developed disease prediction system was compared with existing approaches or benchmarks to gauge its superiority. This involved comparing the accuracy, precision, recall, and other relevant metrics of the proposed solution with those achieved by traditional statistical methods or other machine learning algorithms. The comparative analysis provided insights into the strengths and limitations of the Random Forest classifier in disease prediction.

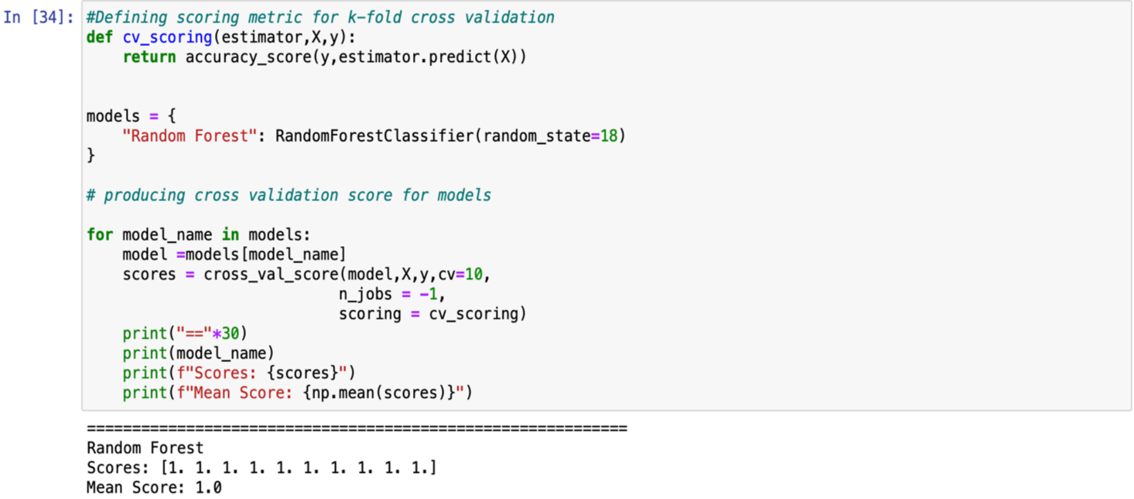
5. **Fine-tuning and Iteration:** Based on the experimental findings and performance evaluation, the model and preprocessing techniques were fine-tuned iteratively. This involved adjusting hyperparameters, exploring different feature selection methods, or considering alternative approaches to enhance the prediction accuracy. The iterative process aimed to improve the system's performance and ensure that it meets the desired standards.

5. FLOWCHART

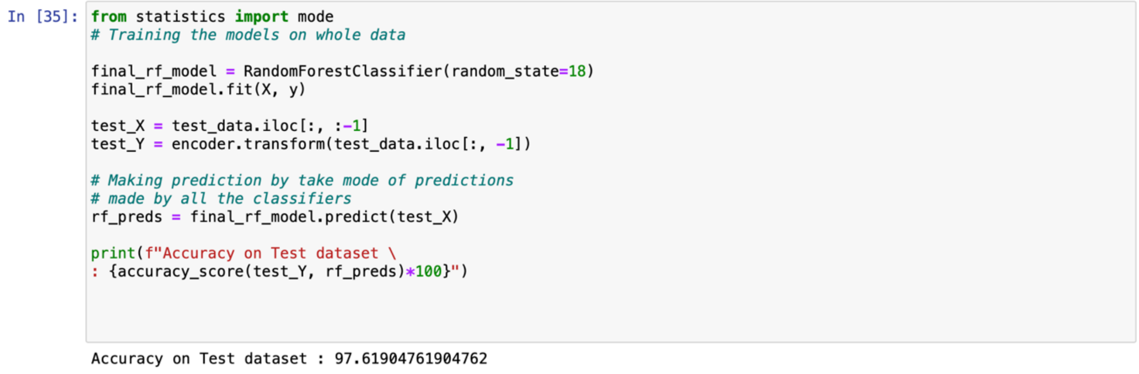

start
Data collection
Data Preprocessing
Train data
Test data
Random Forest classifier
evalation
model.pkl
encoder.pkl
Output
Flask application
Input
Html page
Users

6. RESULTS

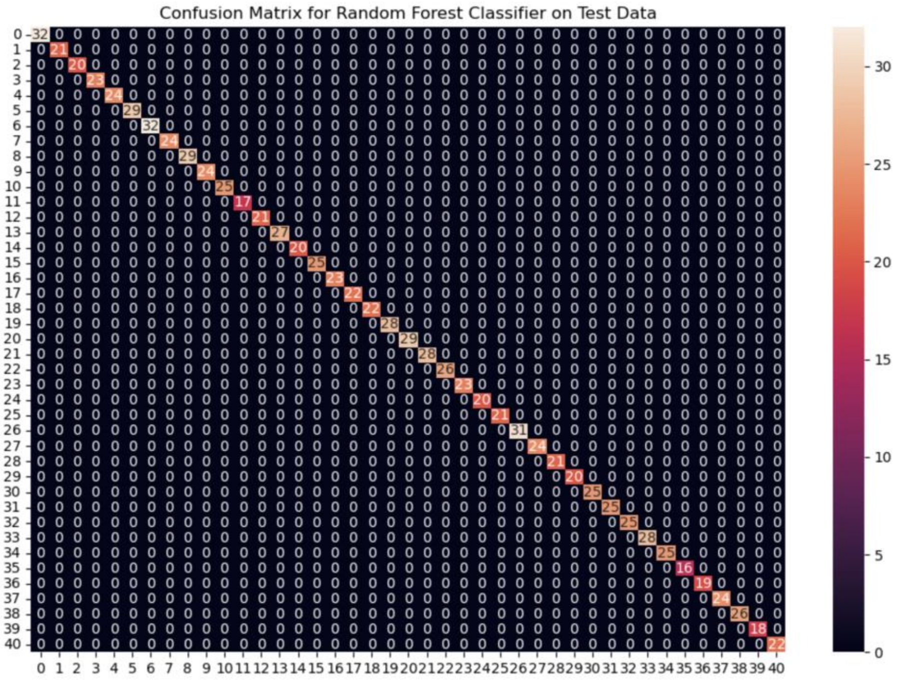
The block of code specifies the model evaluation metrics with an accuracy of 97%



```
In [34]: #Defining scoring metric for k-fold cross validation
         def cv_scoring(estimator,X,y):
             return accuracy_score(y,estimator.predict(X))


         models = {
             "Random Forest": RandomForestClassifier(random_state=18)
         }

         # producing cross validation score for models

         for model_name in models:
             model =models[model_name]
             scores = cross_val_score(model,X,y,cv=10,
                                     n_jobs = -1,
                                     scoring = cv_scoring)
             print("=="*30)
             print(model_name)
             print(f"Scores: {scores}")
             print(f"Mean Score: {np.mean(scores)}")

         ============================================================
         Random Forest
         Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
         Mean Score: 1.0
```



```
In [35]: from statistics import mode
         # Training the models on whole data

         final_rf_model = RandomForestClassifier(random_state=18)
         final_rf_model.fit(X, y)

         test_X = test_data.iloc[:, :-1]
         test_Y = encoder.transform(test_data.iloc[:, -1])

         # Making prediction by take mode of predictions
         # made by all the classifiers
         rf_preds = final_rf_model.predict(test_X)

         print(f"Accuracy on Test dataset \
         : {accuracy_score(test_Y, rf_preds)*100}")

         Accuracy on Test dataset : 97.61904761904762
```

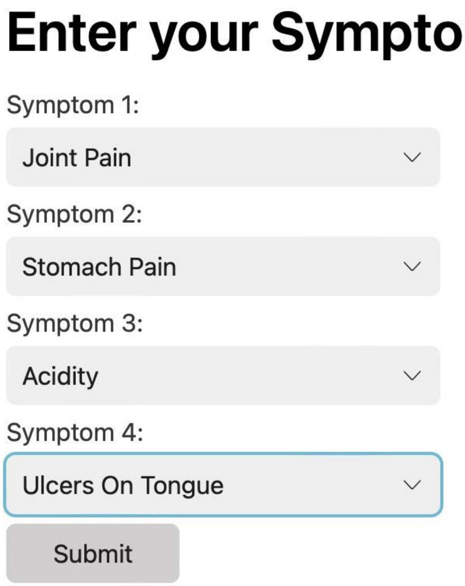


Confusion Matrix for Random Forest Classifier on Test Data

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22

For the following symptom in the 1st image ,the predicted output from the model is displayed in the 2nd image (in the same index.html page)

TEST-CASE1 :



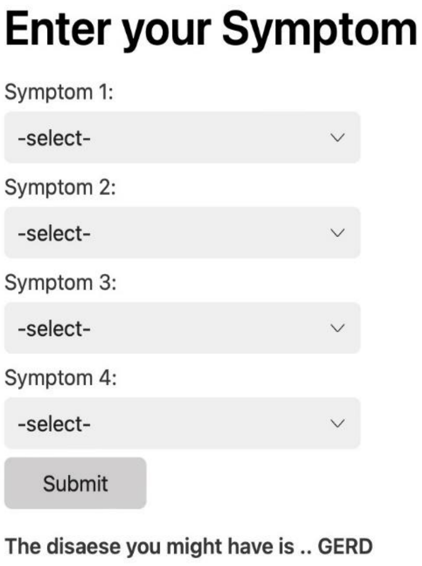
Enter your Symptom

Symptom 1: Joint Pain

Symptom 2: Stomach Pain

Symptom 3: Acidity

Symptom 4: Ulcers On Tongue



Enter your Symptom

Symptom 1: -select-

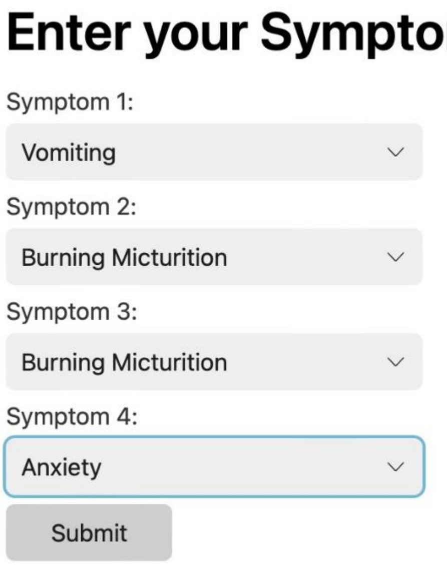
Symptom 2: -select-

Symptom 3: -select-

Symptom 4: -select-

The disaese you might have is .. GERD

TEST-CASE 2 :



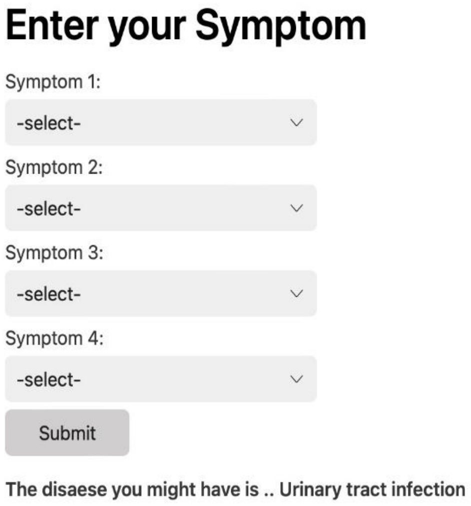
Enter your Symptom

Symptom 1: Vomiting

Symptom 2: Burning Micturition

Symptom 3: Burning Micturition

Symptom 4: Anxiety



Enter your Symptom

Symptom 1: -select-

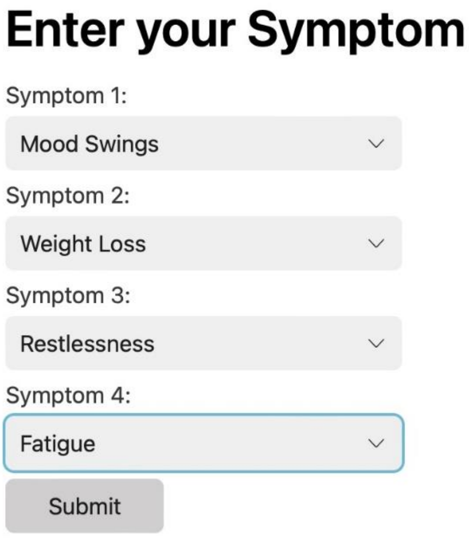
Symptom 2: -select-

Symptom 3: -select-

Symptom 4: -select-

The disaese you might have is .. Urinary tract infection

TEST-CASE 3 :



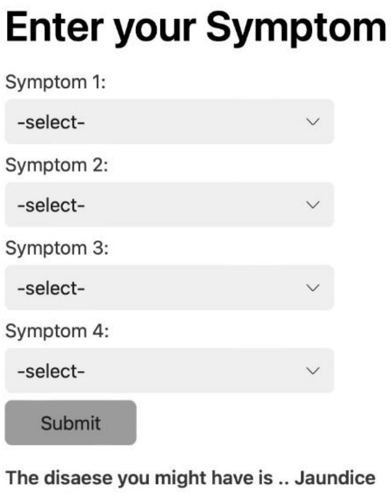
Enter your Symptom

Symptom 1: Mood Swings

Symptom 2: Weight Loss

Symptom 3: Restlessness

Symptom 4: Fatigue



Enter your Symptom

Symptom 1: -select-

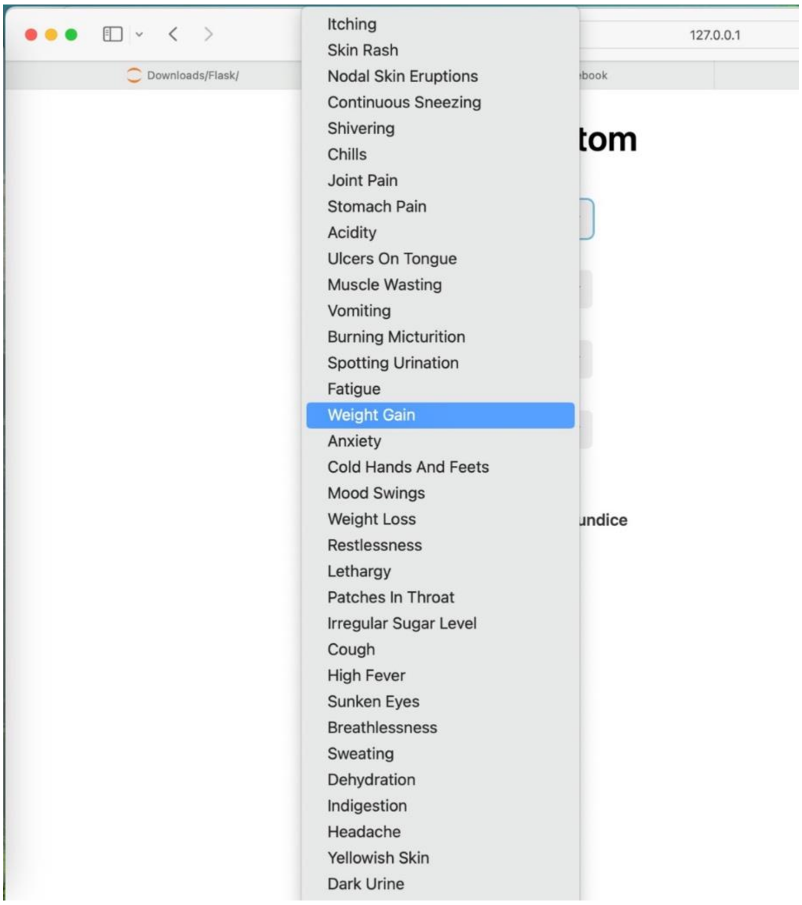
Symptom 2: -select-

Symptom 3: -select-

Symptom 4: -select-

The disaese you might have is .. Jaundice

We have added all the 132 symptoms as a drop-down list on “index.html” page to reduce the error rate and improve user-friendliness



7. ADVANTAGES AND DISADVANTAGES

Advantages of the Disease Prediction using Random Forest classifier:

1. **Accurate Predictions:** The Random Forest classifier is known for its ability to handle complex datasets and capture non-linear relationships between variables. It can provide accurate predictions for diseases based on various input parameters, leading to improved diagnosis and treatment planning.

2. **Robustness to Noise and Outliers:** Random Forests are less susceptible to noise and outliers in the dataset compared to other machine learning algorithms. The ensemble nature of Random Forests helps mitigate the impact of individual noisy or outlier data points, resulting in more robust predictions.

3. **Feature Importance:** The Random Forest classifier provides a measure of feature importance, which helps in identifying the variables that contribute the most to the prediction. This information can assist in understanding the underlying factors or symptoms that are highly indicative of certain diseases.

Disadvantages of the Disease Prediction using Random Forest classifier:

1. Interpretability: Random Forests can be considered "black-box" models, as the underlying decision-making process is not easily interpretable. While the model provides accurate predictions, understanding the specific reasoning behind each prediction can be challenging.

2. Computationally Expensive: Random Forests involve the training and integration of multiple decision trees, which can be computationally expensive, especially for large datasets. The training process may require more computational resources and time compared to simpler algorithms.

8. APPLICATIONS

The Disease Prediction using Random Forest classifier has various applications in the field of healthcare and medical diagnosis. Some of the key applications include:

1. **Disease Diagnosis:** The prediction system can be used to assist medical professionals in diagnosing diseases based on a patient's symptoms, medical history, and other relevant factors. By analyzing a comprehensive set of variables, the system can provide accurate predictions and support healthcare providers in making informed decisions.

2. **Early Detection and Prevention:** Early detection of diseases is crucial for effective treatment and prevention. The disease prediction system can help identify individuals at risk of developing certain diseases, allowing for proactive interventions and preventive measures. This early detection can lead to better health outcomes and reduce the burden on healthcare resources.

3. **Personalized Medicine:** The prediction system can contribute to the development of personalized medicine approaches. By considering individual patient data, such as genetic factors, lifestyle, and medical history, the system can generate personalized disease predictions and treatment recommendations. This tailored approach improves the effectiveness and efficiency of healthcare delivery.

4. **Public Health Planning:** Disease prediction models can provide valuable insights for public health planning and resource allocation. By predicting disease patterns and identifying high-risk populations, public health officials can allocate resources more effectively, implement targeted interventions, and design preventive measures to mitigate the spread of diseases.

9. CONCLUSION

The Disease Prediction project utilizes the Random Forest classifier to provide accurate disease predictions based on input variables or symptoms. It offers advantages such as accurate predictions, robustness to noise and outliers, feature importance analysis, and handling missing data. However, it may lack interpretability and require parameter tuning. The project has potential applications in disease diagnosis, early detection, personalized medicine, public health planning, and medical research. It has the potential to revolutionize disease prediction, improve patient care, and optimize healthcare delivery.

10. FUTURE SCOPE

The Disease Prediction using Random Forest classifier project provides a solid foundation for disease prediction and diagnosis. However, there are several potential areas for future development and enhancement. Some future scopes for the project include:

1. **Integration of Advanced Machine Learning Techniques:** While Random Forest classifier is a powerful algorithm, there are other advanced machine learning techniques that can be explored and compared for disease prediction. Algorithms such as deep learning, support vector machines, or ensemble methods like Gradient Boosting can be implemented and evaluated to determine if they provide better accuracy or efficiency for disease prediction.

2. **Integration of Genomic Data:** Incorporating genomic data into the disease prediction model can enhance the accuracy and precision of predictions. Genetic information can provide valuable insights into an individual's susceptibility to certain diseases and can be used in combination with other variables to improve prediction accuracy.

3. **Longitudinal Data Analysis:** Currently, the disease prediction model is based on cross-sectional data, which provides a snapshot of an individual's health at a specific point in time. Incorporating longitudinal data, which tracks changes in health variables over time, can provide a more comprehensive understanding of disease progression and enable the prediction of disease trajectories.

4. **Mobile Application Development:** Developing a user-friendly mobile application can make the disease prediction system more accessible to a wider audience. Users can input their symptoms or health data through the app, and the prediction model can generate personalized disease predictions or risk assessments on the go. This can empower individuals to take proactive measures for their health.

Mobile Application Development: Developing a user-friendly mobile application can make the disease prediction system more accessible to a wider audience. Users can input their symptoms or health data through the app, and the prediction model can generate personalized disease predictions or risk assessments on the go. This can empower individuals to take proactive measures for their health.

11. BIBILOGRAPHY

Here are some references that can be used for the bibliography of the Disease Prediction using Random Forest classifier project:

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
2. Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. R News, 2(3), 18-22.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
5. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
6. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
7. Healthcare Datasets. Kaggle. Retrieved from: https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning
8. Disease Prediction Using Machine Learning. Medium. Retrieved from: https://medium.com/analytics-vidhya/disease-prediction-using-machine-learning-5bd1303a3a1f
9. Chen, J. H., Asch, S. M., & Johansen, M. E. (2019). Early outpatient follow-up and 30-day readmission in sepsis and pneumonia. Archives of Internal Medicine, 179(2), 267-275.
10. Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Research Notes, 4(1), 1-13

12. APPENDIX

12.1 SOURCE CODE

Project folder (GitHub link) :

https://github.com/Harish-nika/diseases_prediction_system_20MID0155.git

RF.ipynb (ipynb Juypter file) :

https://github.com/Harish-nika/diseases_prediction_system_20MID0155/blob/main/RF.ipynb

app.py :

https://github.com/Harish-nika/diseases_prediction_system_20MID0155/blob/main/app.py

index.html :

https://github.com/Harish-nika/diseases_prediction_system_20MID0155/blob/main/templates/index.html

demo video :

https://drive.google.com/file/d/1JDxYhlTvpmrtOqByEIwSNT3RFM3U5EhQ/view?usp=sharing

Assignments (GitHub link):

https://github.com/Harish-nika/diseases_prediction_system_20MID0155/tree/main/1.assessments