PROJECT TITLE : ”Accurate Prediction of Diseases based on Symptoms of Patients using Machine Learning Algorithms”

Batch Number: CSE-G87

Roll Number	Student Name
20191CSE0710	Yashwanth S
20191CSE0760	Abrar Hussain Dar
20191CSE0746	Mallarapu Vaishnavi
20191CSE0732	Akshay N

Under the Supervision of,

Mr. Mohan Kumar A V
Assistant Professor
School of Computer Science & Engineering
Presidency University

Abstract

- Accurate disease prediction plays a pivotal role in healthcare, enabling timely diagnosis and effective treatment.
- Traditional diagnostic methods based on symptom analysis are limited by human capacity and often result in misdiagnosis or delayed intervention.
- Leveraging the power of machine learning algorithms, this project aims to develop a system for accurate prediction of diseases based on patients' symptoms, contributing to improved healthcare outcomes.
- Through a comprehensive literature review, previous studies on disease prediction using machine learning algorithms were examined, identifying gaps and limitations in existing research. A dataset comprising medical records, symptom descriptions, and patient information was collected, adhering to privacy and ethical guidelines

Abstract

- The results showcased the effectiveness of machine learning algorithms in accurately predicting diseases based on symptoms.
- Comparative analysis of the algorithms revealed their respective strengths and weaknesses, aiding in the selection of the most suitable algorithm for disease prediction
- In conclusion, accurate disease prediction based on symptoms using machine learning algorithms presents a promising approach to enhance healthcare delivery.
- By leveraging the power of data and machine learning, this project aims to improve disease diagnosis, ultimately leading to better patient care and improved healthcare outcomes

Introduction

- Accurate disease prediction plays a crucial role in healthcare by enabling timely diagnosis, effective treatment, and improved patient outcomes.
- Traditionally, medical professionals rely on their expertise and knowledge to identify diseases based on patients' symptoms.
- However, the human capacity to accurately diagnose complex diseases solely through symptom analysis is limited, often leading to misdiagnosis or delayed treatment. This is where the power of machine learning algorithms comes into play.
- In this project, we aim to develop a system for accurate prediction of diseases based on symptoms using machine learning algorithms

Literature Review

Existing Method Disadvantages

- Low Accuracy
- High Complexity
- Highly inefficient
- Limited Data Availability
- Interpretability Challenges

:

Literature Review

- Predicting High-Risk Prostate Cancer Using Machine Learning Methods

The evaluations of classifiers are presented. In Holdout methods, 25% of the data were used for testing and the rest were used to train the classifiers. The performance classifiers vary for accuracy and AUC. ADABoost was the best algorithm for this dataset, given that it had the equal best AUC in holdout and is only 0.002 off the best in cross-validation. Its accuracy in both was also no more than 0.076 from the top accuracy, and was higher than that of decision tree, which was the only algorithm withbetter AUC. Therefore, ADABoost is the machine learning algorithm used for this model in the remaining predictions on PoPC-labelled data.

- Prediction of Prostate Cancer using Machine Learning Algorithms

In this analysis, we used machine learning algorithms on a dataset of prostate cancer patients to predict which patients will have terminal prostate cancer and which people would not be incapacitated, based on the data for each patient's particular characteristics. Our goal was to consider several layout models and choose the most effective one. Five calculations—the K-Nearest Neighbour, Support Vector Machines, Logistic Regression, Naive Bayes, and Random Forest—were used to create our analysis

- Early Detection of Breast Cancer Using Machine Learning Techniques

Most researchers, according to Figure 2, have focused on mammography pictures since they are faster and safer than other methods of detecting breast cancer. Figure 3 compares the algorithms and ML techniques used in the evaluated literature listed in Table 1 for the identification of breast cancer. SVM is shown to be the approach that is employed the most. Figure 4 illustrates the outcomes of ML-based breast cancer detection.

Proposed Methods

- **K-Nearest Neighbors (KNN)**
- **Decision Tree (DT)**
- **Random Forest (RF) Algorithm**
- **AdaBoost Classifier**
- **Naïve Bayes (NB) Classifier**

Objectives

- Develop a Disease Prediction System
- Improve Accuracy and Precision
- Handle Large and Diverse Datasets
- Enhance Interpretability
- Validate and Evaluate Performance
- Create a User-Friendly Interface
- Document and Report Findings
- Contribute to Healthcare Practice

Methodology

A: K-Nearest Neighbors (KNN):

KNN is a non-parametric algorithm that classifies data points based on their proximity to other labelled data points. KNN determines the class of an unclassified point based on the majority vote of its nearest neighbours. KNN can be used for disease prediction based on symptoms by calculating the similarity between patients' symptom profiles and assigning the most common disease label among the nearest neighbours.

B: Decision Tree (DT):

Decision trees are hierarchical tree-like structures that make decisions based on a series of rules and conditions. Each internal node represents a decision based on a specific symptom, while the leaf nodes represent the predicted disease. Decision trees are interpretable and easy to understand, making them suitable for disease prediction based on symptoms.

C:Random Forest (RF) Algorithm:

Random forest is an ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest is trained on a random subset of the data and features, reducing the risk of overfitting. Random forest algorithms can handle large and complex datasets and provide robust predictions for disease classification based on symptoms.

Methodology

D: AdaBoost Classifier:

AdaBoost (Adaptive Boosting) is an ensemble learning method that combines multiple weak classifiers to create a strong classifier. In the context of disease prediction based on symptoms, AdaBoost can be a valuable algorithm to consider.

E: Naïve Bayes (NB) Classifier:

Naive Bayes is a probabilistic classifier based on Bayes' theorem and assumes independence among features. Despite its simplifying assumptions, Naive Bayes algorithms are computationally efficient and can handle high-dimensional data. Naive Bayes models are particularly useful when the dimensionality of symptom data is large, and there is a need for real-time predictions.

System Design

- The development of a system is a process in which a device is implemented utilising various approaches and design ideas.
- Software Development Life Cycle – SDLC:
  The Software Development Life Cycle (SDLC) is a structured approach to software development that consists of a series of phases or stages. Each phase has specific objectives, deliverables, and activities that contribute to the overall development process. Here are the typical phases of the SDLC:

System Design

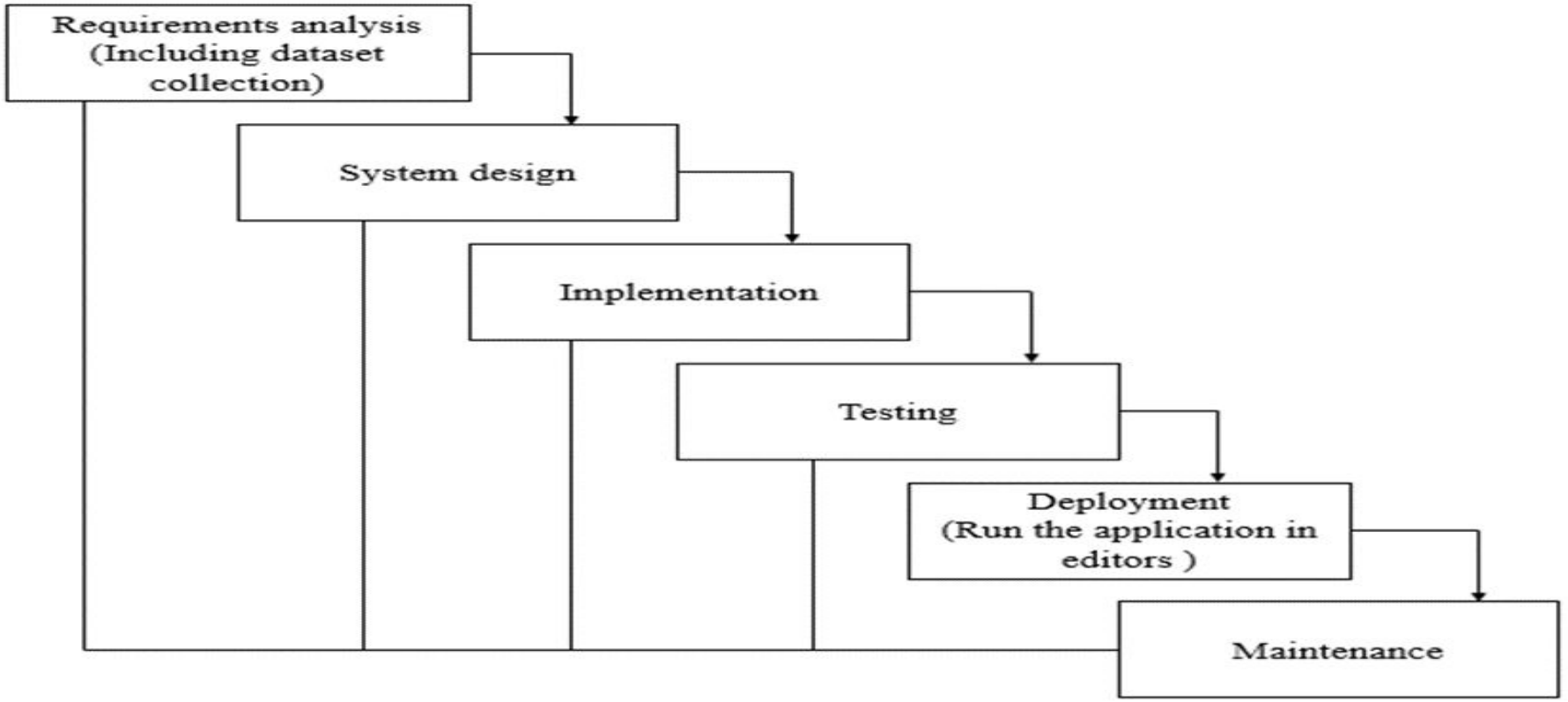


Fig 1.1: Waterfall Model

Detailed Design

Input Design:

In an information system, input is the raw data that is processed to produce output. During the input design, the developers must consider the input devices such as PC, MICR, OMR, etc.

Output Design:

The design of output is the most important task of any system. During output design, developers identify the type of outputs needed, and consider the necessary output controls and prototype report layouts.

Detailed Design

- UML Diagram:

  UML (Unified Modelling Language) diagrams can be used to visualise different aspects of your project's design

- Use Case Diagram:

Fig 1.2: Use Case Diagram

Sequence Diagram

- Depicts the steps involved in receiving symptom inputs, pre-processing the data, and invoking machine learning algorithms for prediction

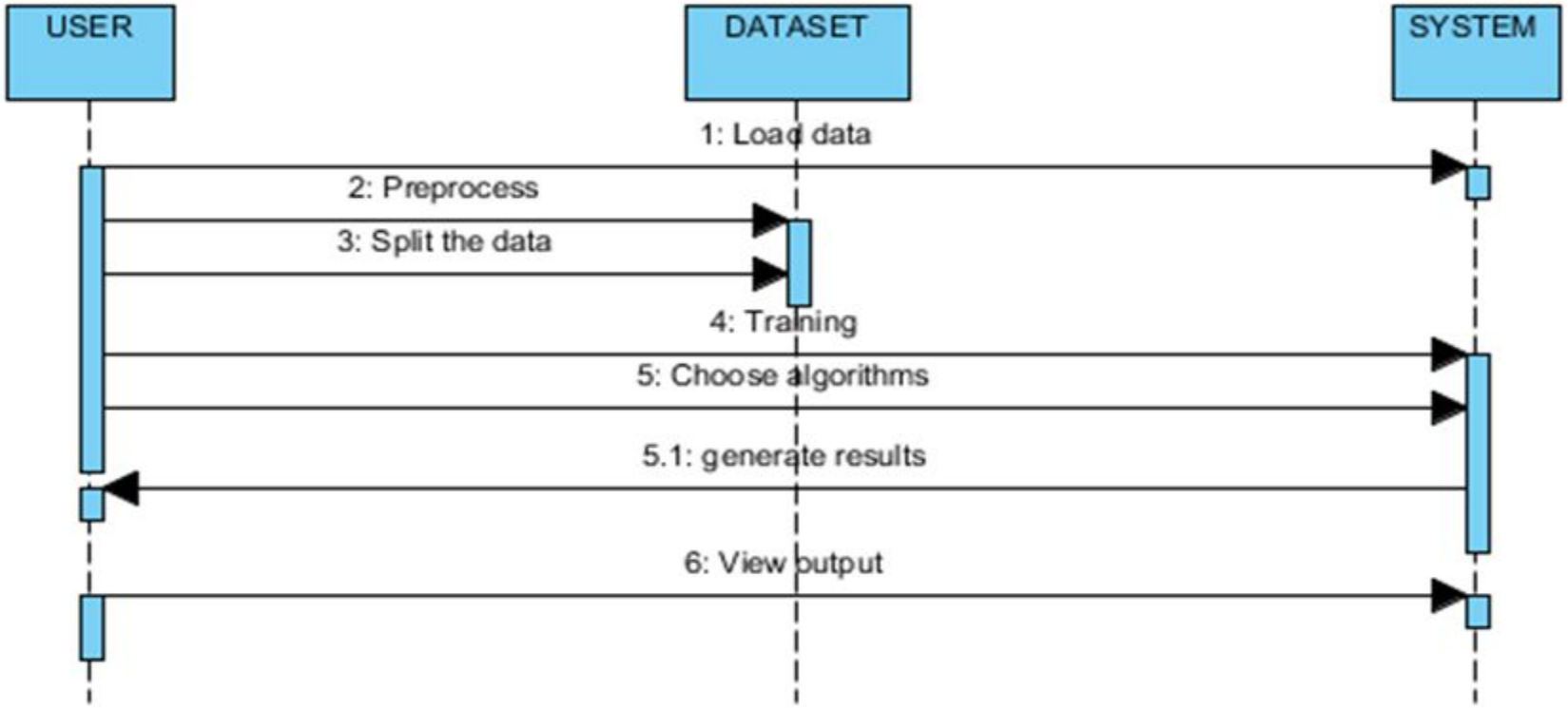


Fig 1.3:Sequence Diagram

Activity Diagram

- Activity diagrams represent the flow of activities or processes within your system. They are useful for visualising the workflow and decision points

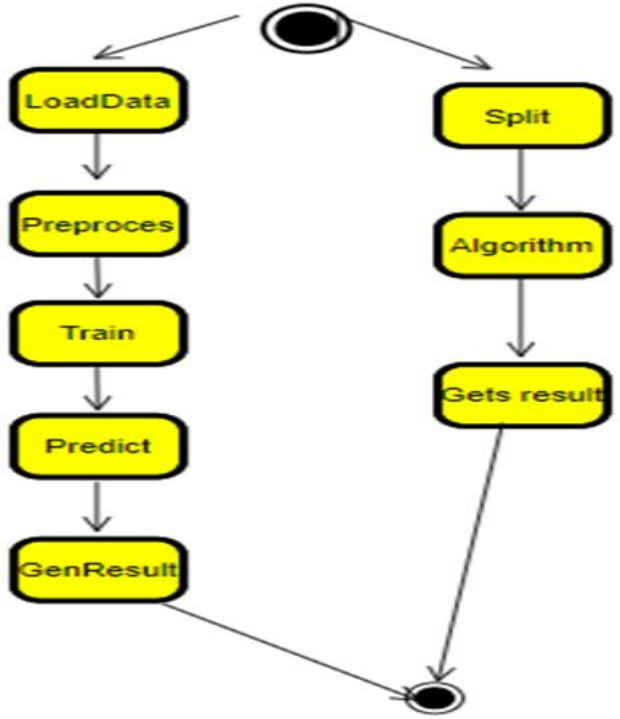


Fig 1.4:Activity Diagram

Database Design

- ER DIAGRAM

An Entity–relationship model (ER model) describes the structure of a database with the help of a diagram, which is known as Entity Relationship Diagram (ER Diagram).

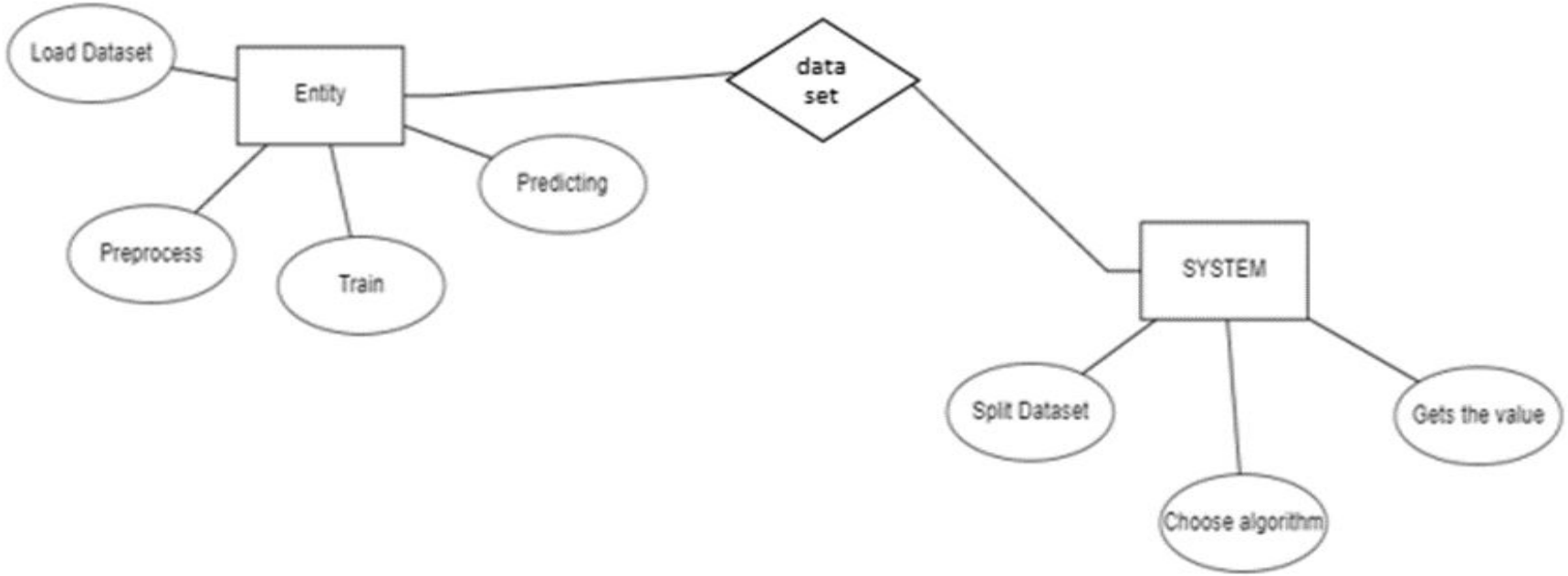


Fig 1.5:ER Diagram

Implementation

Algorithm Description:

➢ **Decision Trees**:

- Decision trees are tree-like models that make predictions by learning simple decision rules from the input features.
- They split the data based on the most informative features at each node, creating a tree structure where each leaf node represents a class label or a prediction.

➢ **Random Forest**:

- Random forests are an ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

➢ **AdaBoost Classifier**:

- AdaBoost is an ensemble learning method that combines weak classifiers to create a strong classifier.

Implementation

- **Naïve Bayes (NB) Classifier**:

  Naive Bayes is a probabilistic classifier based on Bayes' theorem and assumes independence among features. Despite its simplifying assumptions, Naive Bayes algorithms are computationally efficient and can handle high-dimensional data

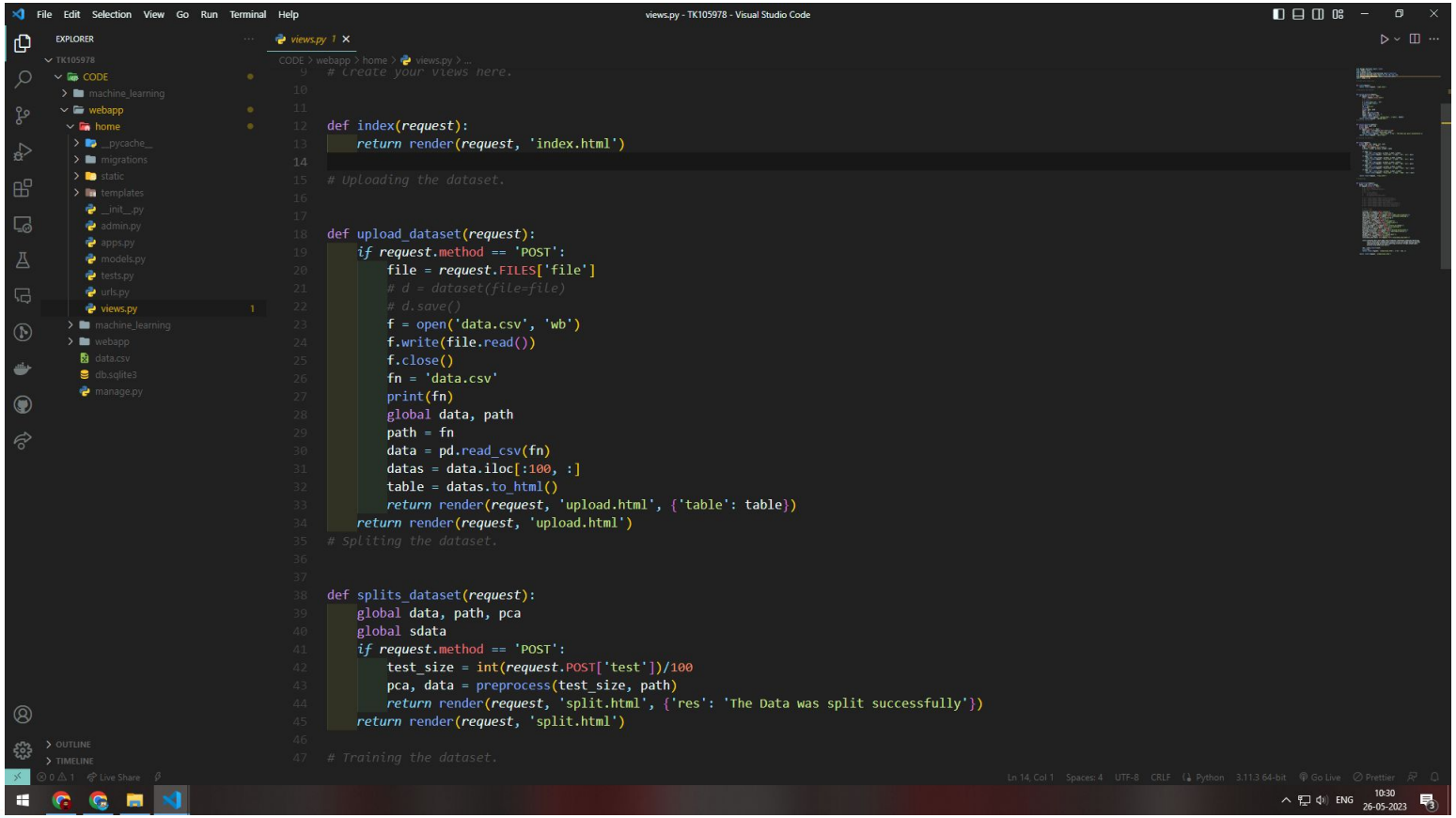
- **K-Nearest Neighbors (KNN)**:

  KNN is a non-parametric algorithm that classifies data points based on their proximity to other labelled data points. .

Implementation

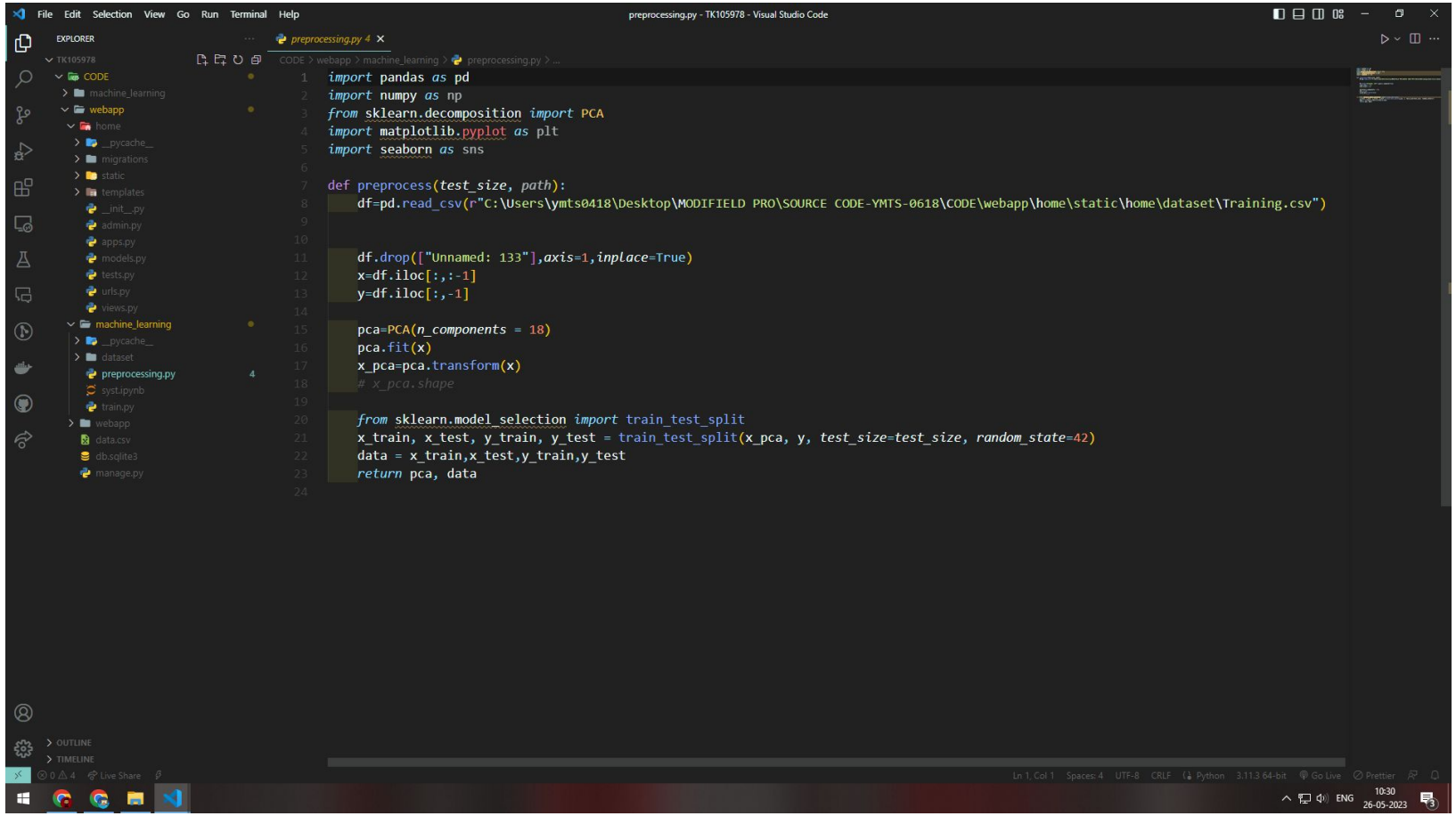
Source Code Description:

- **views.py**



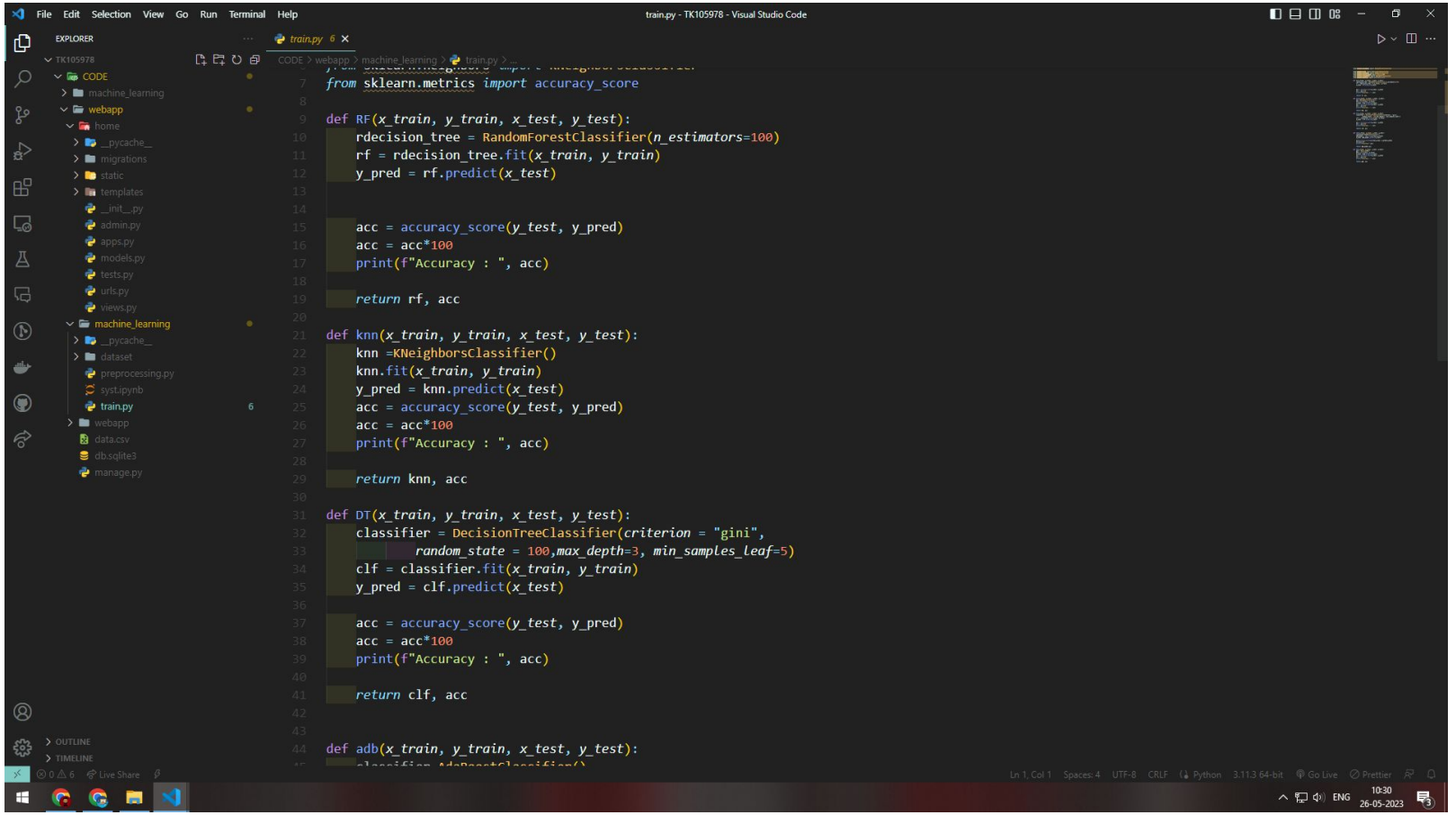
Implementation

- **processing.py**



Implementation

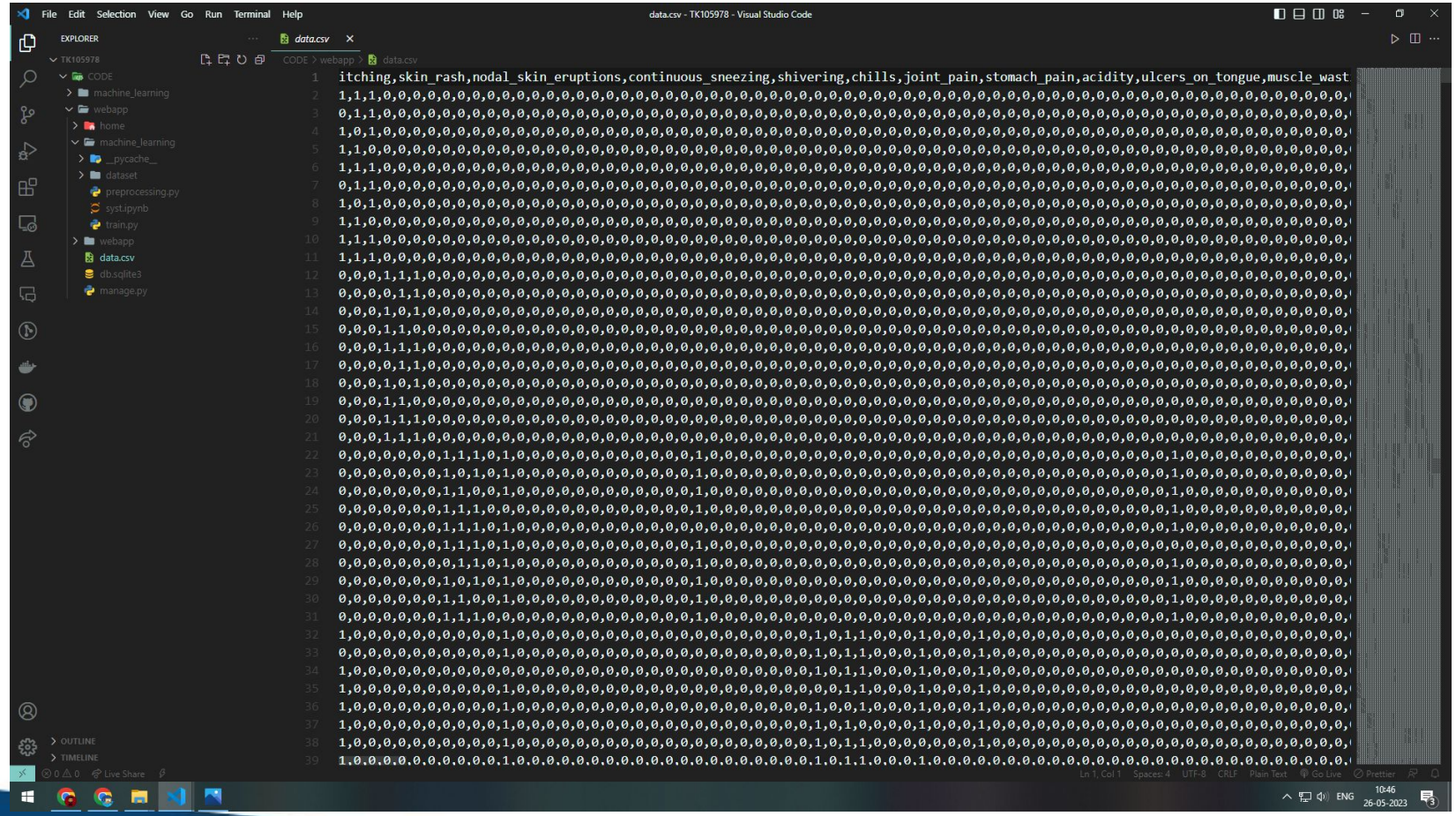
- **train.py**



Implementation

Data Dictionary:

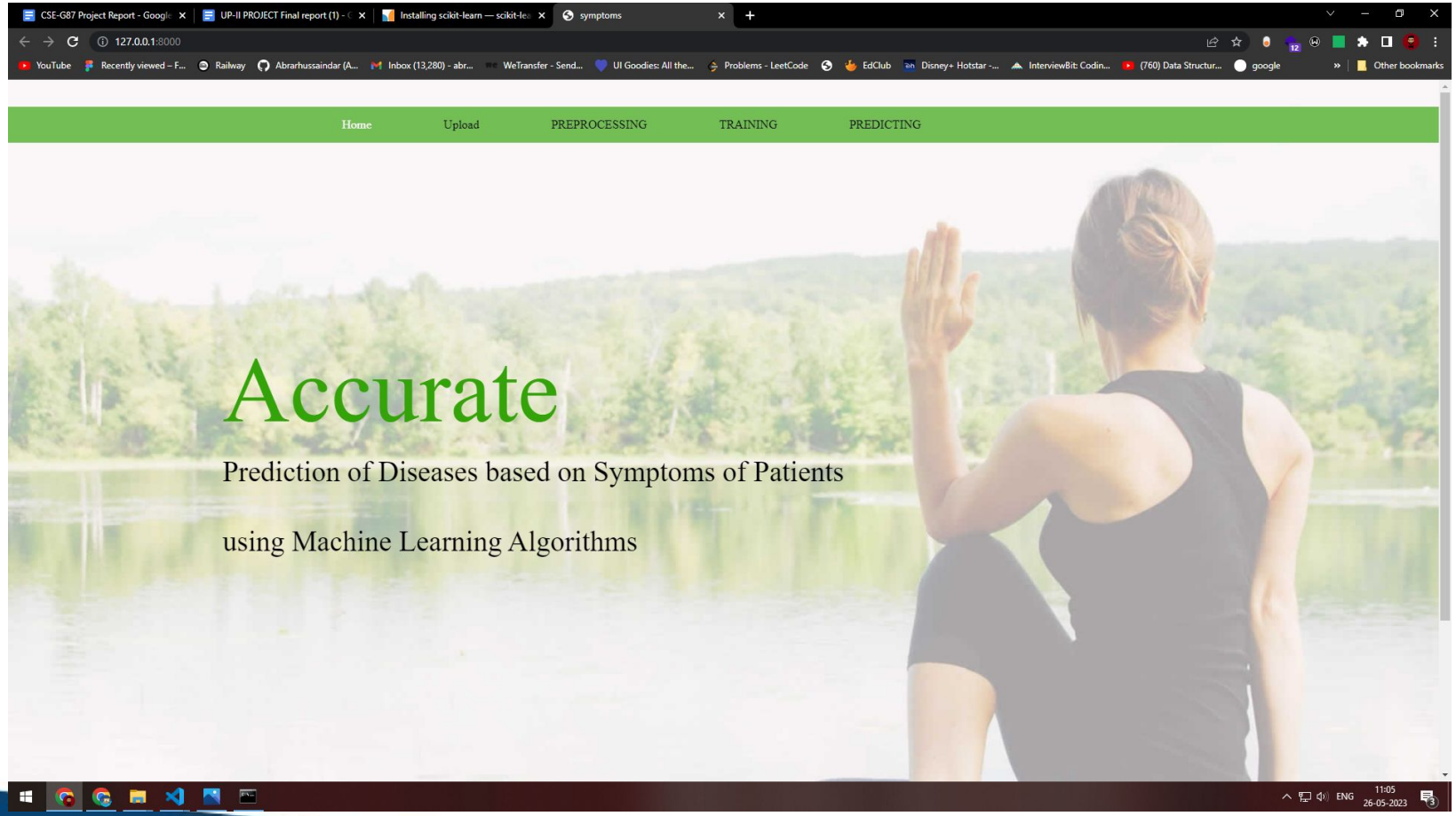
- **Data.csv**



Project Insight

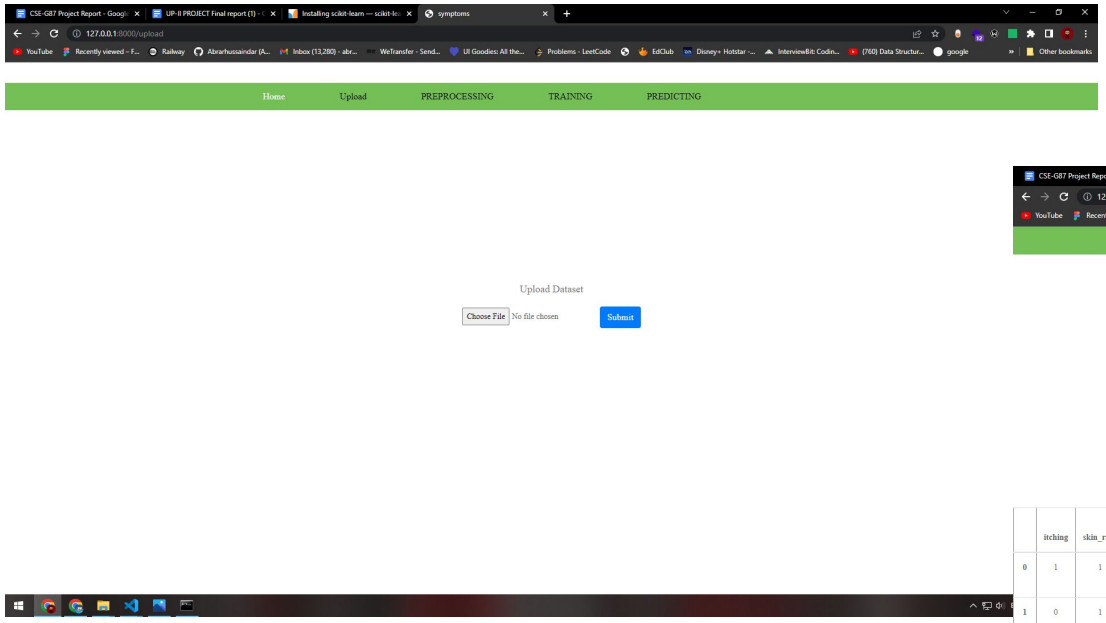
Project Insights:

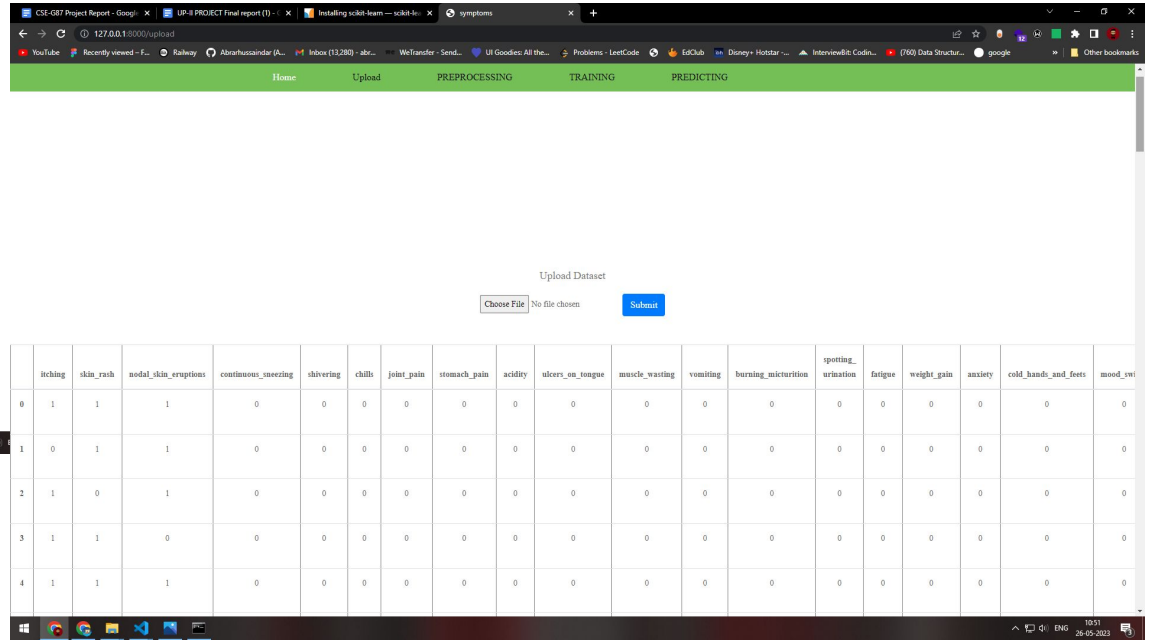
- **Home page**



Project Insight

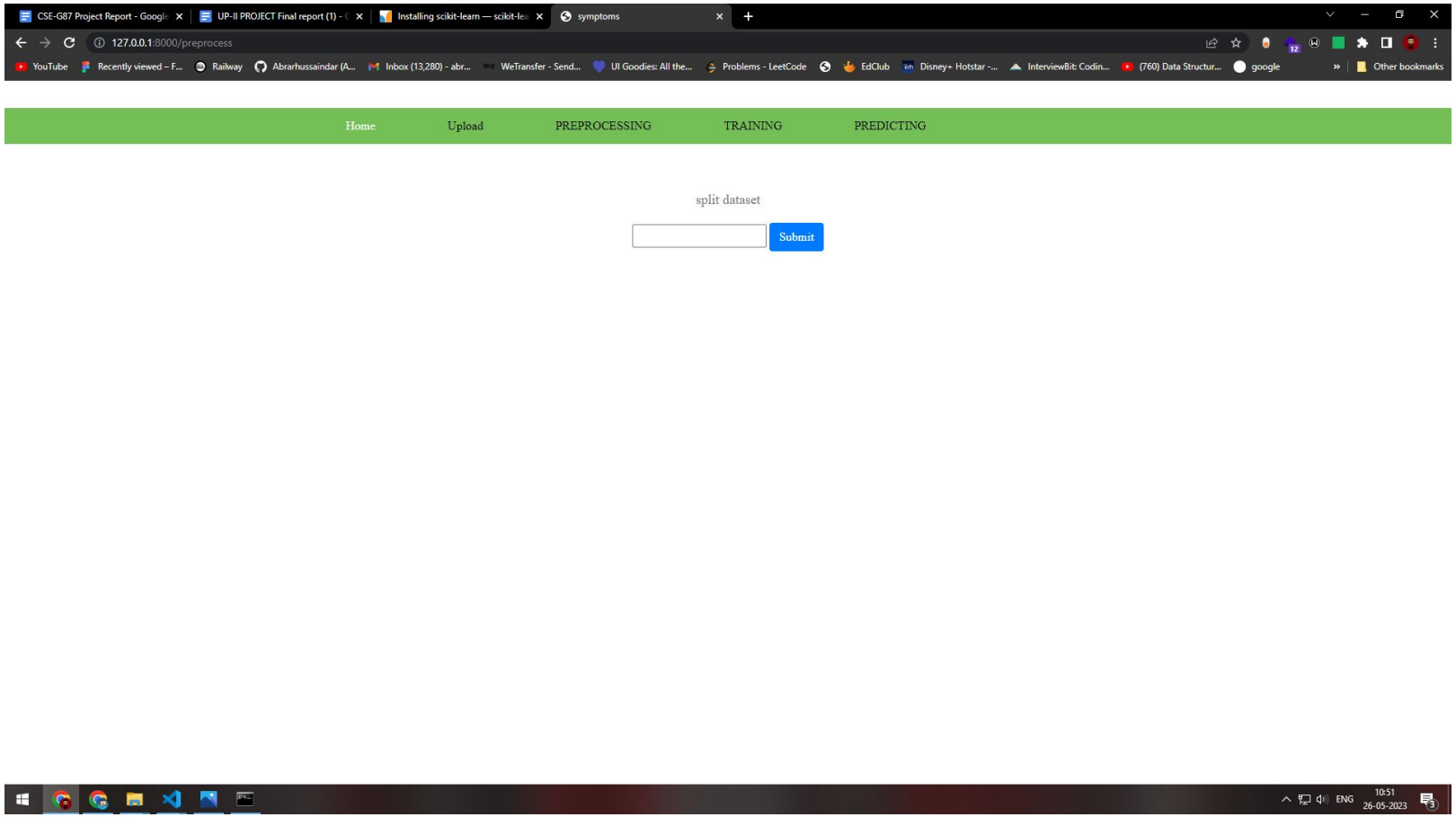
- **Upload Page**





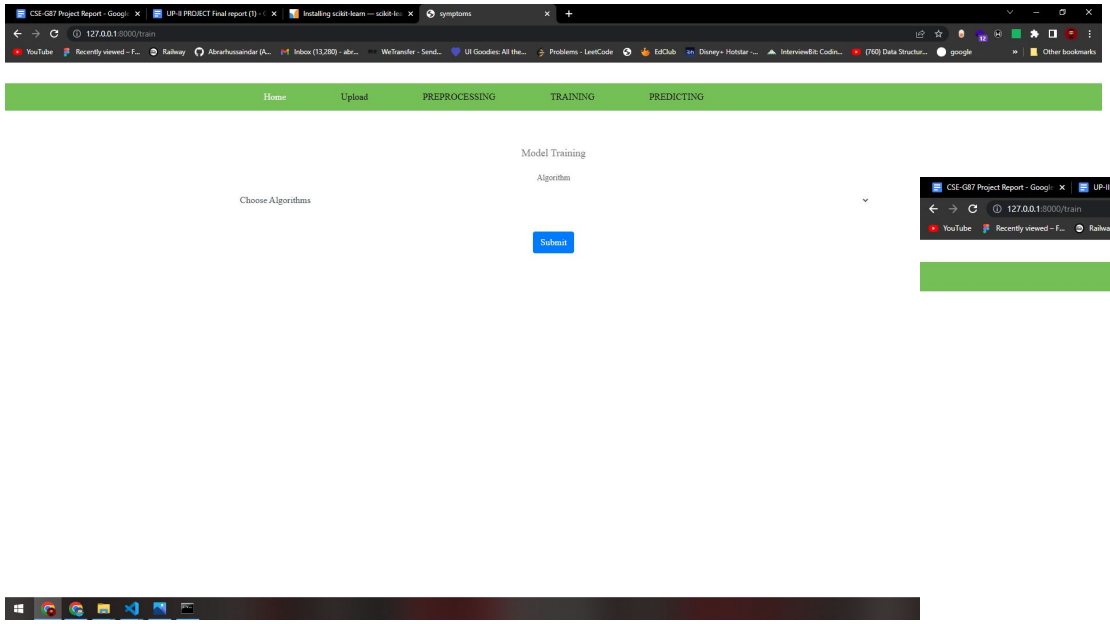
Project Insight

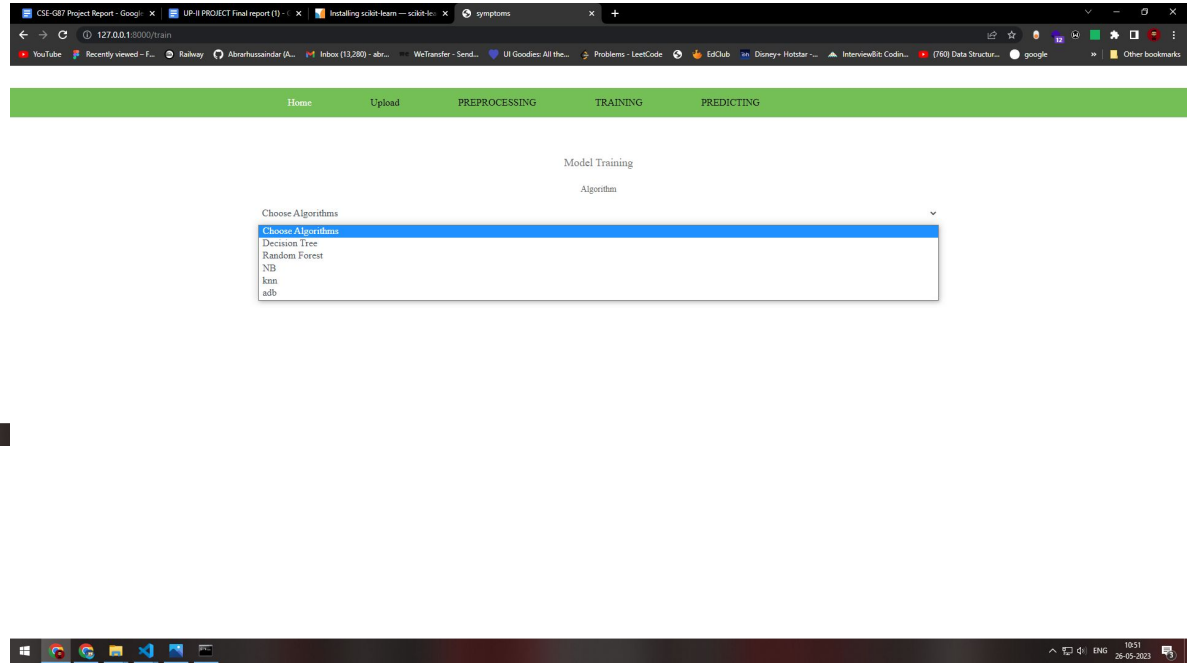
- **Processing Page**



Project Insight

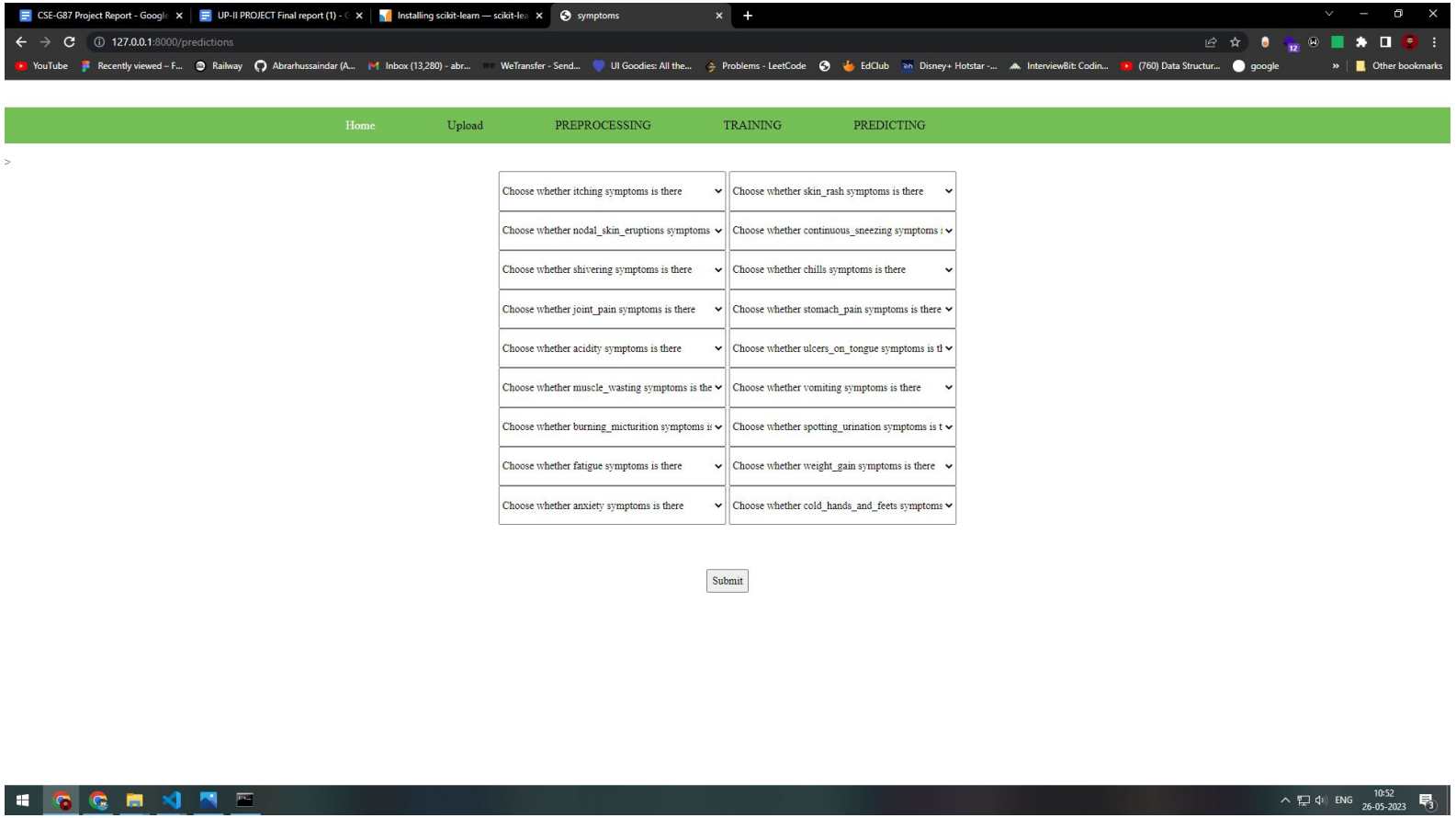
- **Algorithm Training Page**





Project Insight

- **Predicting Page**



Hardware and software specifications

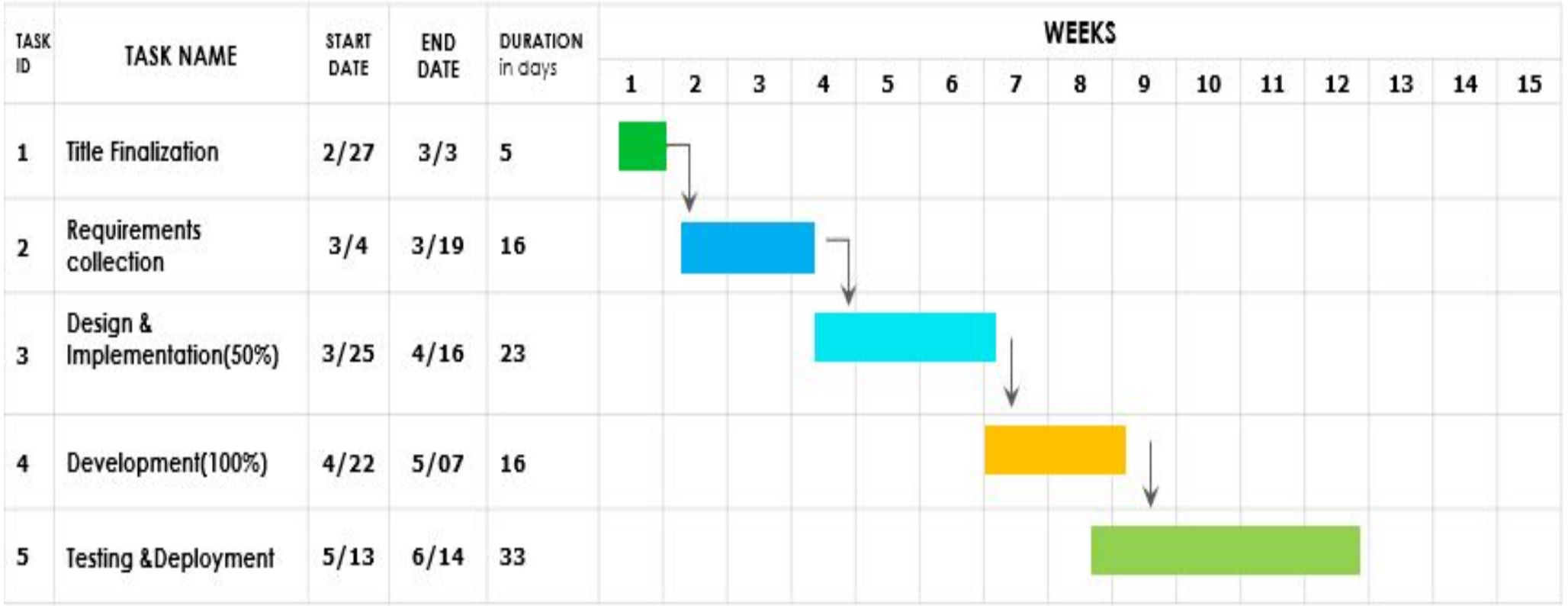
- Hardware requirements
  - Processor – intel Pentium 4(1.50GHz or above)
  - RAM – min 1GB
  - Hard disk – 128GB
- Software configuration
  - Operating system – Windows 7,10
  - IDE – Sublime text / Visual Studio Code

Hardware and software specifications

Software Requirements:

- **Python**: The primary programming language for your project will be Python
- **Python Libraries**: You will need to install several Python libraries for data manipulation, machine learning, and data visualisation. Some essential libraries include:
  - **NumPy**: For numerical computations and array manipulation.
  - **Pandas**: For data manipulation and analysis.
  - **scikit-learn:** For implementing machine learning algorithms and evaluation metrics.
  - **TensorFlow or PyTorch:** For deep learning models, if applicable.
  - **Matplotlib or Seaborn**: For data visualisation.
  - **Jupyter Notebook** or an integrated development environment **(IDE)** like **PyCharm** or **Anaconda Navigator**: To write and run your code.
  - **Machine Learning Frameworks**: Depending on the specific algorithms you choose, you may need to install additional machine learning frameworks. For example:
  - **scikit-learn**: Provides a wide range of machine learning algorithms and utilities.
  - **TensorFlow**: An open-source deep learning framework developed by Google.
  - **PyTorch**: Another popular deep learning framework with a strong focus on flexibility and usability.

Timeline of Project

- 

TASK ID	TASK NAME	START DATE	END DATE	DURATION in days
1	Title Finalization	2/27	3/3	5
2	Requirements collection	3/4	3/19	16
3	Design & Implementation(50%)	3/25	4/16	23
4	Development(100%)	4/22	5/07	16
5	Testing &Deployment	5/13	6/14	33

Outcomes

- The outcome of our project is a robust and reliable software system that can accurately predict diseases based on the symptoms exhibited by patients.
- The software utilises machine learning algorithms, such as decision trees, random forests, K-nearest Neighbors, Naïve Bayes (NB) Classifier and AdaBoost, to analyse the symptom data and generate predictions.
- The outcome of our project also includes a thorough literature review, where we explore existing research and approaches related to disease prediction using symptoms and machine learning algorithms.
- Overall, the outcome of our project is a valuable contribution to the field of healthcare, providing a powerful tool for accurate disease prediction based on symptoms.

Results AND Discussion

In this section, we present the results obtained from the implementation of the disease prediction software using machine learning algorithms based on symptoms. We discuss the performance of each algorithm and compare their accuracy in predicting diseases.

1. Performance of Machine Learning Algorithms

We evaluated the performance of several machine learning algorithms, including decision trees, random forests, support vector machines (SVM), neural networks, and AdaBoost. Each algorithm was trained on a dataset consisting of symptom data and corresponding disease labels.

2. Comparison of Algorithm Performance

All the tested machine learning algorithms achieved relatively high accuracy in disease prediction. However, there were some variations in their performance metrics.

3. Discussion of Findings

The results indicate that machine learning algorithms, particularly ensemble methods like Random Forests and AdaBoost, are effective in accurately predicting diseases based on symptoms. These algorithms show promising potential for real-world application in healthcare settings.

4. Limitations and Future Work

While the implemented software showed promising results, there are a few limitations to consider. Firstly, the accuracy of the predictions heavily relies on the quality and representativeness of the training data. Obtaining a diverse and comprehensive dataset with a large sample size could further enhance the performance of the algorithms.

Conclusion

- In conclusion, this project aimed to develop a software system for accurate disease prediction based on symptoms using machine learning algorithms.
- Through the implementation and evaluation of various algorithms, including decision trees, random forests, K-nearest Neighbours, Naïve Bayes (NB) Classifier and AdaBoost, we have achieved significant progress in the field of disease prediction.
- this project contributes to the field of healthcare by providing a reliable and accurate software system for disease prediction based on symptoms. The successful implementation of machine learning algorithms and the evaluation of their performance demonstrate the potential of this approach in supporting medical professionals in their decision-making process.
- Moving forward, further research can be conducted to explore advanced feature selection techniques, hyperparameter tuning, and the integration of additional data sources to improve the accuracy and robustness of the disease prediction models. With continued advancements in machine learning and healthcare technology, the field of disease prediction holds great promise for the future.

Overall, this project represents a significant step towards the goal of accurate disease prediction, contributing to the advancement of healthcare and ultimately benefiting patients worldwide.

References

1. S. M. a. M. K. Henry Barlow, “Predicting High-Risk Prostate Cancer Using Machine Learning Methods,” vol. 4, no. 3, p. 129, 2019.
2. M. Srivenkatesh, “Prediction of Prostate Cancer using Machine Learning Algorithms,” *Int. J. Recent Technol. Eng,* vol. 8, no. 5, pp. 5353-5362, 2020.
3. A. A. B. B. R. K. B. N. a. M. A. B. M. Tahmooresi, “Early Detection of Breast Cancer Using Machine Learning Techniques,” *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* vol. 10, no. 3-2, pp. 21-27, 2018.
4. R. Patra, “Prediction of Lung Cancer Using Machine Learning Classifier,” *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India,* vol. Revised Selected Papers 1, pp. 132-142, March 26-27 2020.
5. K. R. J. L. C. H. L. G. Y. C. L. a. X. X. Yijun Wu, “Machine Learning Algorithms for the Prediction of Central Lymph Node Metastasis in Patients With Papillary Thyroid Cancer,” *Frontiers in Endocrinology,* vol. 11, p. 577537, 2020.
6. B. S. a. C. O. Kubra Tunca, “Lung Cancer Incidence Prediction Using Machine Learning Algorithms,” *Journal of Advances in Information Technology,* vol. 11, no. 2, 2020.
7. T. D. J. Dr.B.Santhosh Kumar, “Breast Cancer Prediction Using Machine Learning Algorithms,” *International Journal of Advanced Science and Technology ,* vol. 29, no. 3, 2020.
8. S. L. S. K. W. N.-H. J. K. DongWook Kim, “Deep learning-based survival prediction of oral cancer patients,” *Scientific reports,* vol. 9, no. 1, pp. 1-10, 2019.
9. M. L. B. ,. M. C. M. I. P. E. M. I. C. N. C. V. Lerina Aversanoa, “Thyroid Disease Treatment prediction with machine learning approaches,” *Procedia Computer Science ,* vol. 192, pp. 1031-1040, 2021.
10. S. S. V. K. S. S. P. A. K. Kedar Pingale, “Disease Prediction using Machine Learning,” *International Research Journal of Engineering and Technology (IRJET) ,* vol. 6, pp. 831-833, 2019.
11. C. M. M. G. A. M. J. A. S.-S. J. R. L. C. C.-H. M. M.-Y. e. a. Chiesa-Estomba, “Machine learning algorithms as a computer-assisted decision tool for oral cancer prognosis and management decisions: a systematic review,” *ORL,* vol. 84, no. 4, pp. 278-288, 2022.

Thank You