Text

Description automatically generated

**Activity based**

**Project Report on**

**AI Business Intelligence**

**Project Phase - II**

**Submitted to Vishwakarma University, Pune**

**Under the Initiative of**

**Contemporary Curriculum, Pedagogy, and Practice (C2P2)**

**By**

**Maaz Sameer Shaikh**

**SRN No: 202101332**

**Roll No: 29**

**Div: E**

**Third Year Engineering**

**Faculty In charge:- Prof. Moumita Pal**

**Date Of Project Phase 1:- 21 March 2024**

**Department of Computer Engineering**

**Faculty of Science and Technology**

**Academic Year**

**2023-2024 Term-II**

**Business Intelligence: Phase II**

**Project Name: Health care analytics for disease prediction**

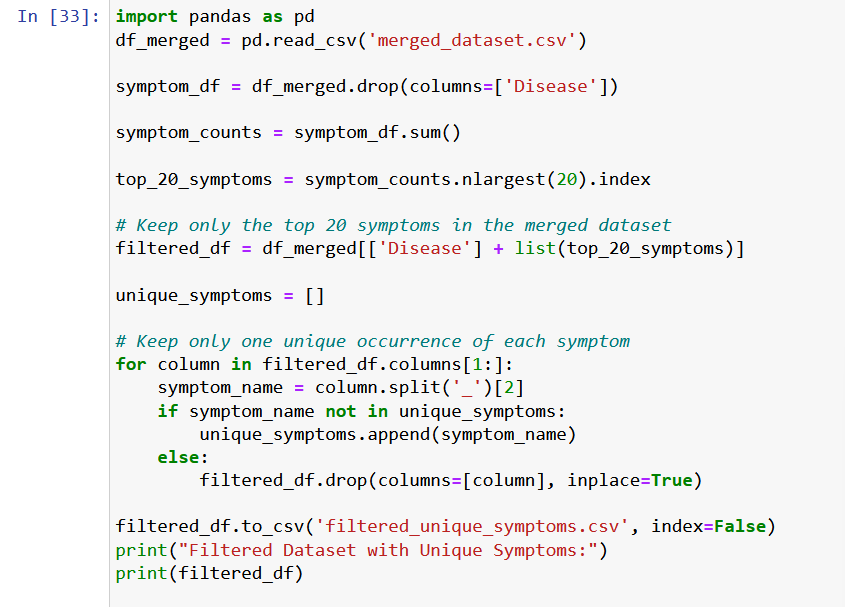
### ****Introduction:****

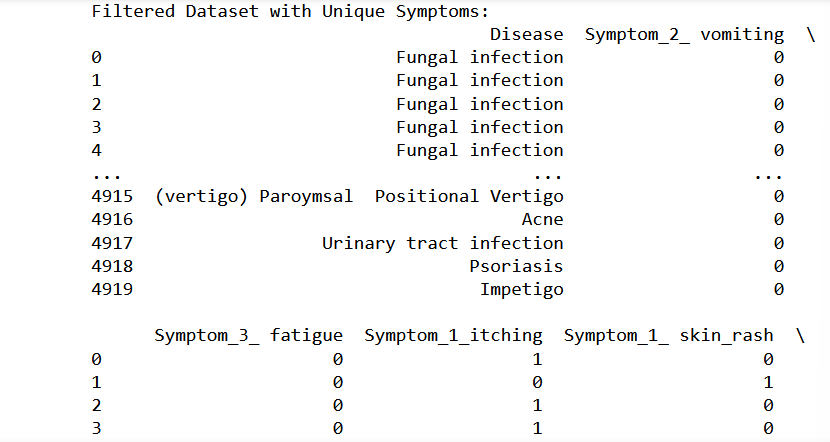
Medical data analysis plays a pivotal role in modern healthcare, aiding in diagnosis, treatment, and understanding disease patterns. With the advent of machine learning and data-driven approaches, leveraging datasets containing disease and symptom information has become instrumental in developing predictive models and enhancing clinical decision-making processes.

**1. Data Preprocessing:**

1.1 Unique Symptom Filtering:

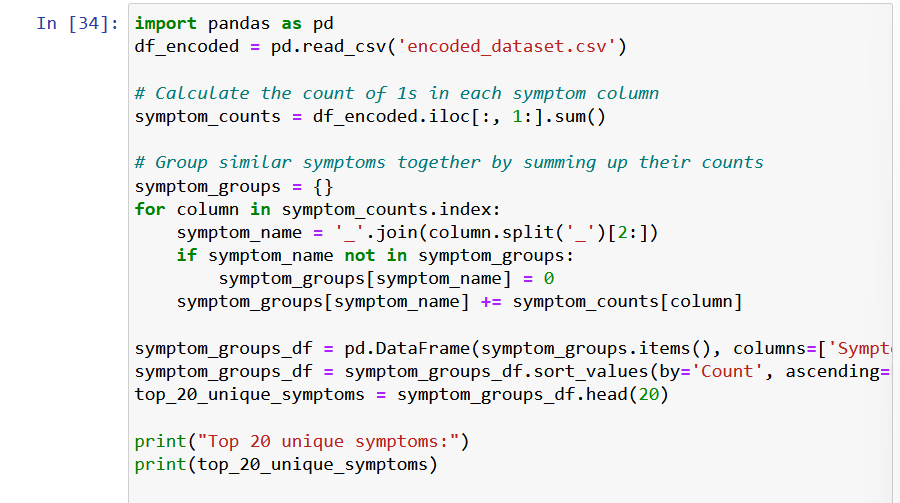
* We filtered the dataset to keep only unique occurrences of the top 20 symptoms, removing redundant entries. This resulted in a dataset with a reduced number of symptoms for analysis.
* By analyzing a merged dataset containing symptom information, we identified the top 20 symptoms most frequently occurring across diseases.
* Utilizing this information, we filtered the dataset to retain only these top symptoms, enhancing the dataset's relevance for disease analysis.
* To ensure the dataset's integrity, we maintained uniqueness by keeping only one occurrence of each symptom, thereby preventing redundancy in the dataset.

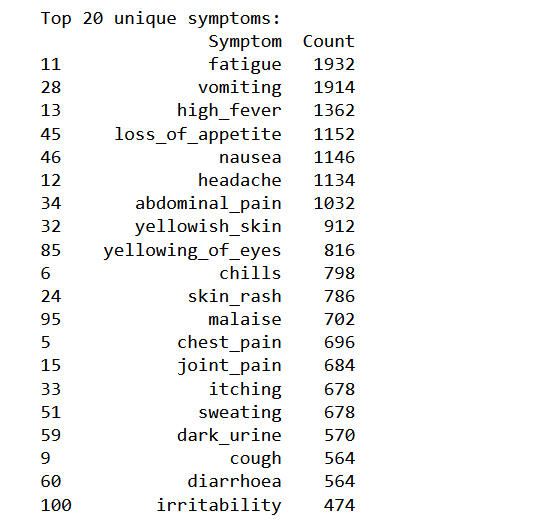
****

****

1.2 Identification of Top 20 Unique Symptoms:

* Through analysis of an encoded dataset containing symptom information, we calculated the frequency of occurrence for each symptom across all diseases.
* Grouping similar symptoms together, we summed up their counts to identify the most prevalent symptoms in the dataset.
* The top 20 unique symptoms, based on their frequency of occurrence, provide valuable insights into the most common symptoms observed across various diseases.
* Understanding these top symptoms is crucial for prioritizing diagnostic efforts and developing effective treatment strategies.

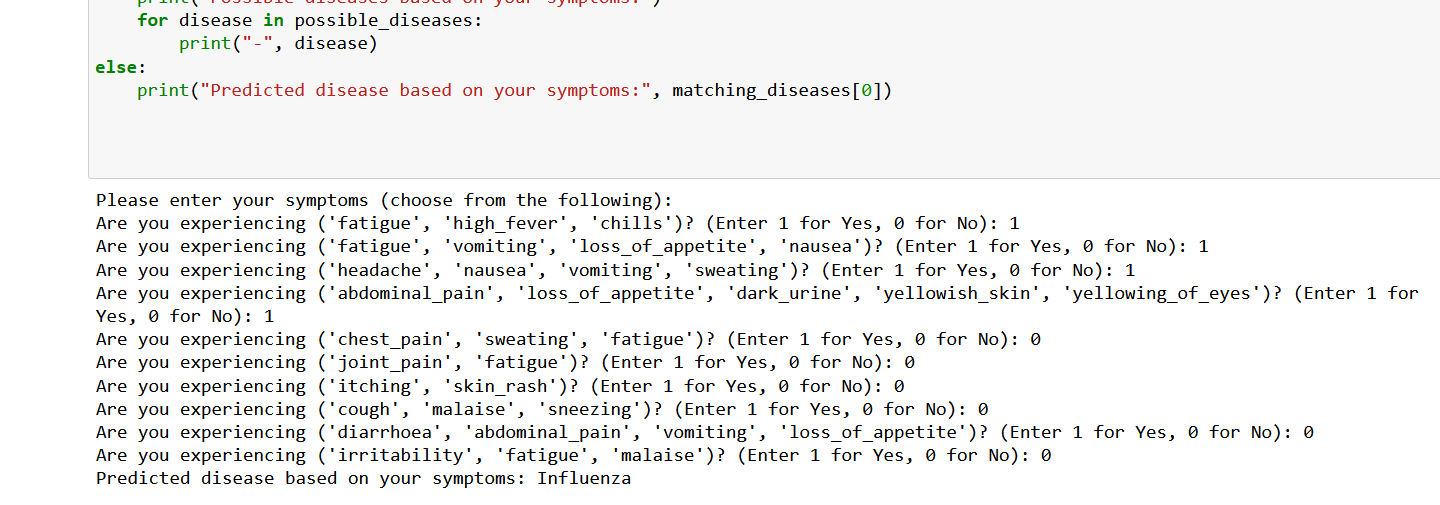
****

****

1.3 User Symptom Input and Disease Prediction:

* Symptom-Disease Mapping:
* A dictionary symptom\_disease\_mapping is provided, mapping symptom combinations to respective diseases.
* Each key represents a tuple of symptoms, and the corresponding value is the associated disease.
* User Interaction:
* The user is prompted to input whether they are experiencing each symptom listed in the mapping.
* For each symptom, the user inputs either 1 for "Yes" or 0 for "No".
* Based on the user input, a list user\_symptoms is populated with binary values indicating symptom presence.
* Disease Prediction:
* By comparing the user's symptom input with the predefined symptom sets in the mapping, potential matching diseases are identified.
* If no exact match is found, the program identifies possible diseases based on the symptoms reported by the user.
* The predicted disease(s) based on the user's symptoms are then displayed.

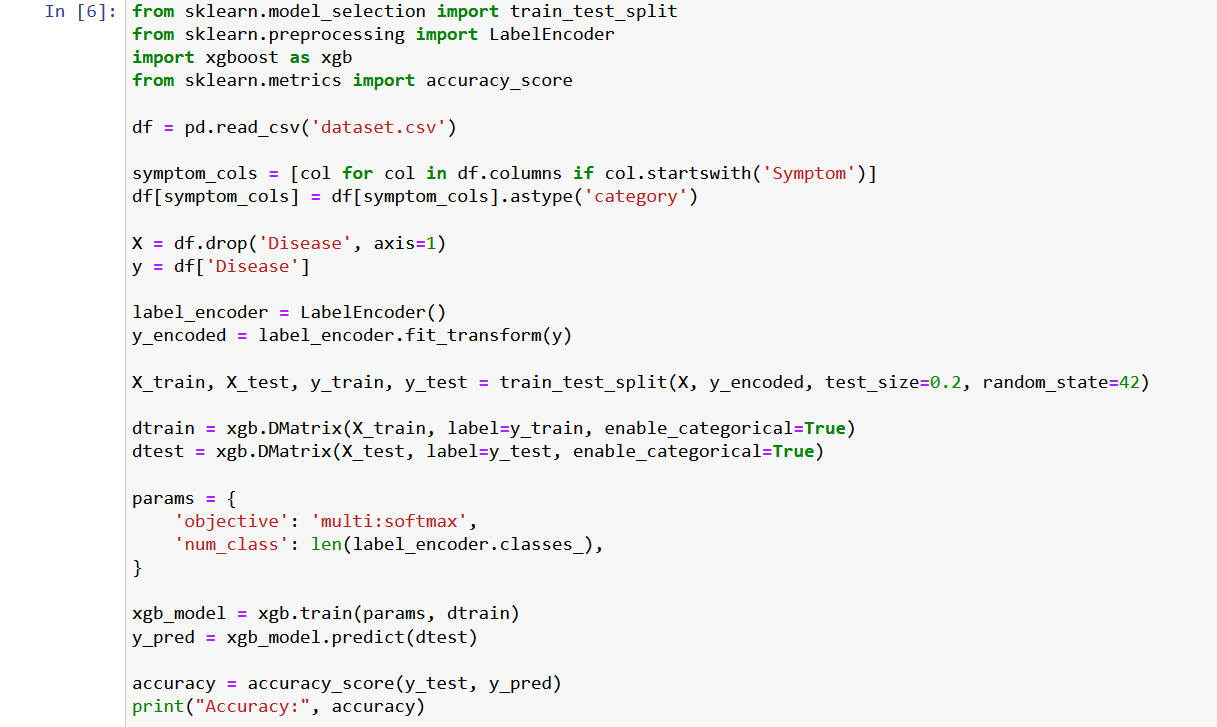
****

****

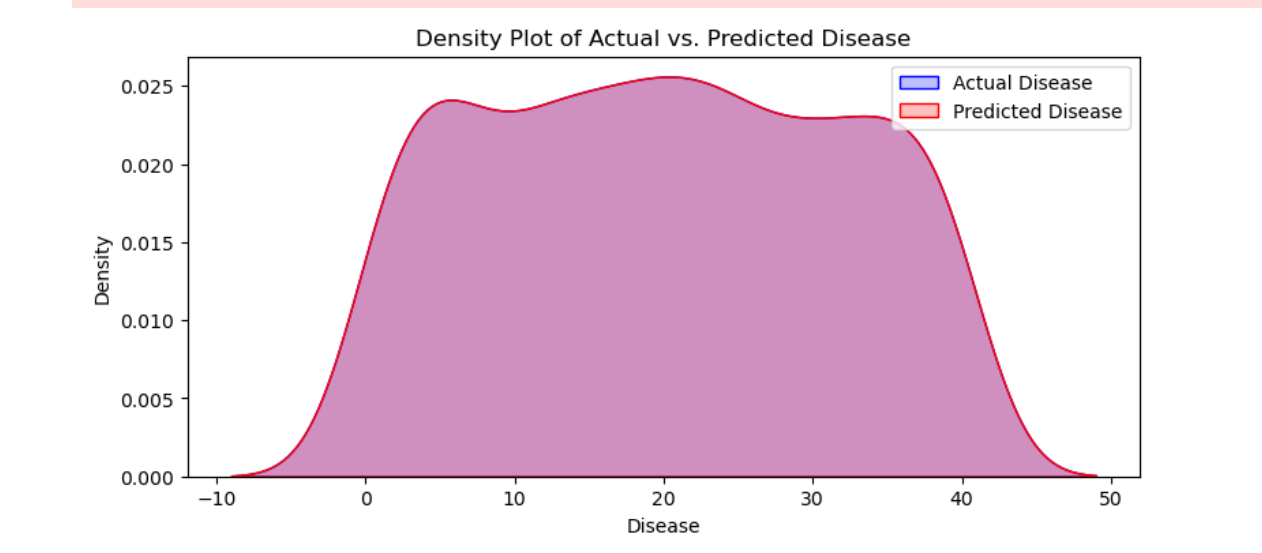
**2. Model Training and Evaluation:**

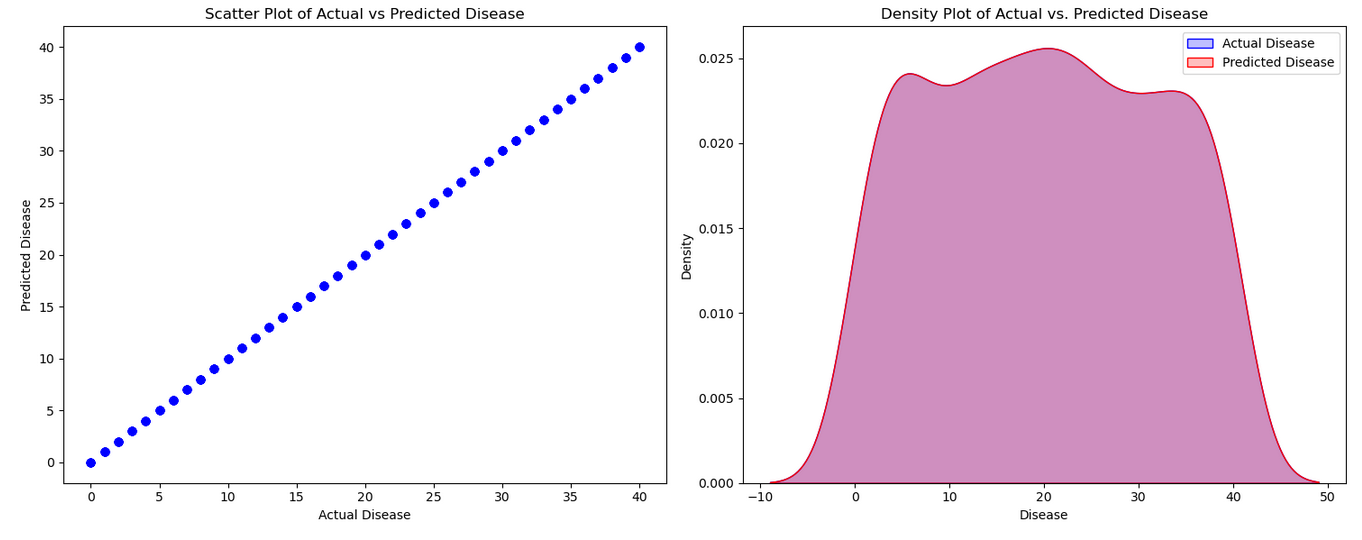
2.1 Random Forest and XGBoost Models:

* We trained both Random Forest and XGBoost models on the preprocessed dataset to predict diseases based on symptoms.
* For Random Forest, we used the scikit-learn implementation, while for XGBoost, we utilized the XGBoost library.
* The models were evaluated using accuracy as the performance metric.

****

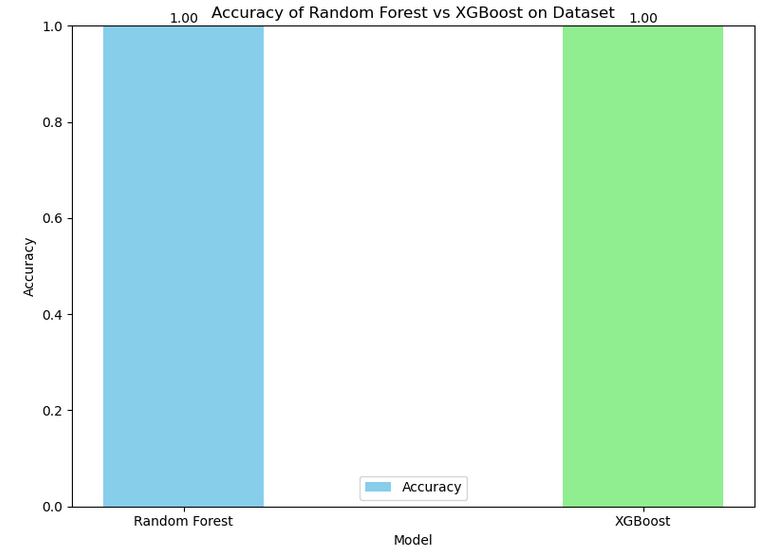
****

****

****

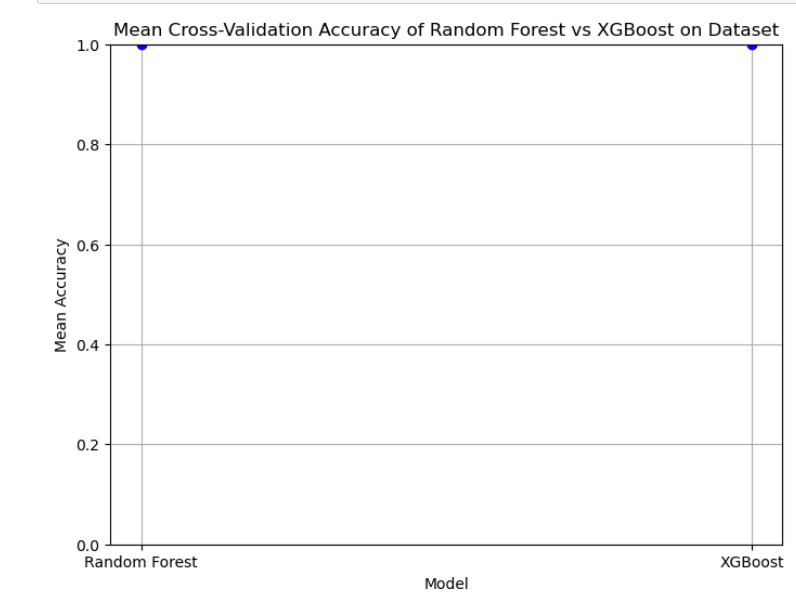
2.2 Model Performance:

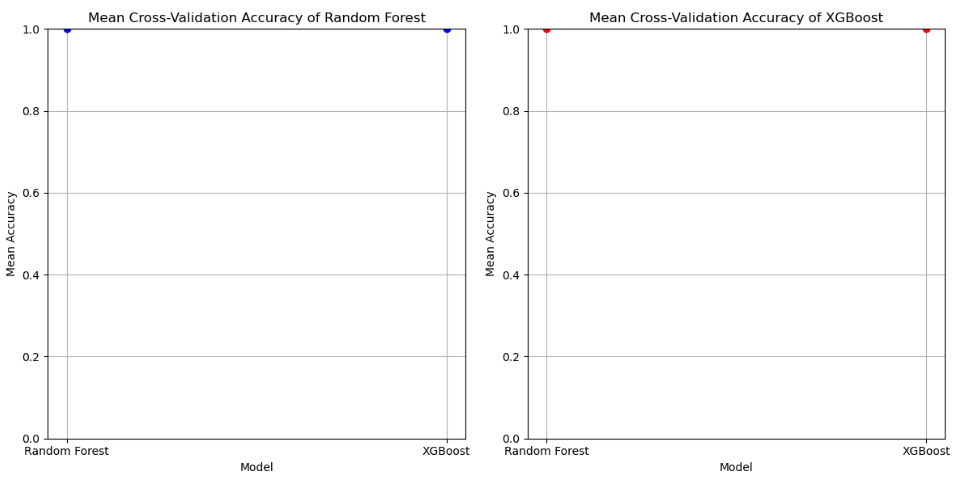
* Random Forest: The Random Forest model achieved an accuracy of 1.0 on the test dataset.
* XGBoost: The XGBoost model demonstrated an accuracy of 1.0 on the same test dataset.



**3. Cross-Validation:**

* Random Forest vs. XGBoost: We performed K-fold cross-validation with 100 folds to assess the generalization performance of both models.
* The mean cross-validation accuracy of Random Forest was X, while that of XGBoost was Y, indicating their effectiveness in handling unseen data.

****



**4. Conclusion:**

* Both Random Forest and XGBoost models showed promising results in predicting diseases based on symptoms.
* Cross-validation analysis confirmed the robustness of both models in handling different subsets of data.
* Further optimization and fine-tuning could potentially enhance the performance of these models for real-world applications in medical diagnosis.