**Disease Prediction from Symptoms**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**MOHAMMED SHAZIN ABBAS - AP21110011020**

**NITEESH KUMAR PANDEY - AP21110011024**

**C LOKESH SWARUP - AP21110011012**

**ANKUSH KUMAR - AP21110011026**

**A picture containing text

Description automatically generated**

Under the Guidance of

**Krishna Siva Prasad Mudigonda**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**Nov, 2023**

# Abstract

This project presents a rule-based methodology for disease prediction solely based on symptoms. Utilizing predetermined rules derived from expert knowledge and medical literature, the approach maintains simplicity and interpretability. The study evaluates the system's performance using traditional metrics and discusses its practical applications in resource-constrained healthcare settings. Emphasizing transparency and ethical considerations, this research contributes to the exploration of accessible and interpretable solutions for symptom-based disease prediction.

# Introduction

In the contemporary landscape of healthcare, the integration of machine learning has become pivotal for enhancing disease diagnosis and prognosis. This project delves into the realm of predictive medicine, specifically focusing on the use of machine learning algorithms to forecast diseases based on symptomatic manifestations. The intricate relationship between symptoms and underlying health conditions necessitates a sophisticated approach, and the exploration of algorithms such as Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting promises to illuminate novel paths in predictive disease analysis. By harnessing the power of data mining, this project seeks to bridge the gap between symptoms and accurate disease prediction, ultimately contributing to the evolving field of machine-assisted healthcare.

**Project Background:**

In the contemporary landscape of healthcare, the convergence of burgeoning health data accessibility and the rapid advancements in machine learning technologies has sparked a profound interest in leveraging these tools for proactive disease prediction. The influx of digital health records, coupled with the computational prowess of machine learning algorithms, presents an unprecedented opportunity to transform the way diseases are diagnosed and prognosed.

This project unfolds against this backdrop of transformative potential, driven by a dual commitment to harnessing the wealth of health data and advancing the field of machine learning. The escalation of chronic diseases and the need for early intervention have intensified the exploration of predictive models that can decipher intricate patterns and relationships within symptom data.

Moreover, the existing body of research in disease prediction and prognosis serves as a foundation upon which this project builds. As machine learning algorithms evolve and mature, the emphasis on creating models that not only achieve high predictive accuracy but also remain interpretable becomes paramount. The need for interpretability is particularly crucial in healthcare, where understanding the rationale behind predictions is essential for gaining the trust of healthcare professionals and ensuring responsible and informed decision-making.

This project, therefore, positions itself at the nexus of cutting-edge technological advancements and the pressing demand for effective disease prognosis. By building upon the existing knowledge base, it aspires to make meaningful contributions to the development of models that not only excel in predictive accuracy but also adhere to the principles of interpretability, transparency, and ethical considerations.

In essence, this project is a proactive response to the transformative potential of machine learning in healthcare, aiming to propel the field forward by crafting models that not only predict diseases from symptoms but do so in a manner that aligns with the highest standards of efficacy, understandability, and responsible implementation.

**Description of the Project:**

Healthcare systems worldwide face the continuous challenge of timely and accurate disease diagnosis. This project endeavors to address this challenge by utilizing machine learning techniques to predict diseases based on observed symptoms. The fundamental motivation stems from the recognition that an individual's symptoms can serve as valuable indicators of underlying health conditions. By leveraging a diverse dataset sourced from Columbia University, comprising a multitude of symptoms associated with various diseases, the project sets out to create a robust and interpretable model.

The primary focus of the project is to design a solution that not only achieves high accuracy in disease prediction but also aligns with the practical constraints of real-world healthcare scenarios. The comprehensive Kaggle dataset, with 133 columns capturing specific symptoms and corresponding prognoses, serves as a rich source for training and testing the predictive models. Simultaneously, the Columbia dataset, with information on diseases, disease occurrences, and symptoms, provides an additional dimension for model refinement and validation.

The project's approach involves the exploration of four distinct machine learning algorithms: Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting. Each algorithm is carefully tailored to accommodate the nuances of symptom-based disease prediction. The goal is not only to develop an accurate predictive model but also to ensure its interpretability, as transparency in the decision-making process is critical in a healthcare context.

Beyond algorithmic exploration, the project places significant emphasis on the ethical implications of implementing such models in the healthcare domain. Considerations include data privacy, model transparency, and potential biases that might affect predictions. These ethical dimensions are paramount in ensuring the responsible deployment of machine learning in healthcare settings.

**Proposed Solution Using Data Mining Technique:**

The envisioned solution to the challenge of predicting diseases from symptoms entails a systematic application of data mining techniques, utilizing a curated dataset that amalgamates diverse symptom profiles and corresponding disease outcomes. The methodology unfolds in several key steps, each designed to extract meaningful patterns and relationships from the data:

**1. Data Preprocessing:**

The initial step involves a comprehensive preprocessing phase to ensure the dataset's cleanliness and readiness for analysis. This encompasses handling missing values, addressing outliers, and normalizing or scaling features to maintain consistency across the dataset. Given the complexity of medical data, this phase is crucial for mitigating potential biases and ensuring the model's robustness.

**2. Feature Selection:**

To optimize model performance and reduce dimensionality, a careful selection of relevant features is undertaken. This involves identifying symptoms that contribute most significantly to disease prediction. Techniques such as correlation analysis and feature importance ranking from ensemble models guide the selection process, streamlining the model without compromising its predictive power.

**3. Algorithm Selection:**

Tailoring the choice of data mining algorithms to the unique characteristics of symptom-based disease prediction is pivotal. The project explores the effectiveness of Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting algorithms. Each algorithm is chosen for its specific strengths in handling categorical data, capturing complex relationships, and providing robust ensemble-based predictions.

**4. Model Training:**

With feature-rich, preprocessed data and selected algorithms, the model training phase begins. The dataset is divided into training and testing sets to facilitate the evaluation of model performance. During training, the algorithms learn to recognize patterns and associations between symptoms and diseases, iteratively refining their predictive capabilities.

**5. Hyperparameter Tuning:**

To optimize the model's performance, hyperparameter tuning is conducted. This involves systematically adjusting the configuration parameters of each algorithm to find the optimal set that maximizes predictive accuracy. Cross-validation techniques help ensure that the tuned model generalizes well to unseen data.

**6. Model Evaluation:**

Rigorous evaluation metrics, including accuracy, precision, recall, and F1 score, are employed to assess the models' performance. The evaluation process involves comparing the predictive capabilities of each algorithm and understanding their strengths and weaknesses in the context of symptom-based disease prediction.

**7. Interpretability and Validation:**

Transparency and interpretability are paramount in healthcare applications. The interpretability of the chosen algorithms is carefully examined to ensure that the predictions can be understood and trusted by healthcare professionals. Additionally, the models are validated against external datasets to assess their generalizability and robustness beyond the training data.

**8. Documentation and Deployment Considerations:**

Comprehensive documentation of the entire process, from data preprocessing to model deployment considerations, is crucial for the project's replicability and future improvements. Factors such as model interpretability, computational resource requirements, and scalability are carefully documented to inform future implementations in diverse healthcare settings.

In essence, the proposed solution using data mining techniques involves a meticulous and iterative process, where each step contributes to the development of a predictive model attuned to the complexities of symptom-based disease prediction. This methodology ensures not only accurate predictions but also a level of interpretability essential for practical implementation in real-world healthcare scenarios.

**Model Architecture:**

The model architecture for disease prediction from symptoms involves a combination of data preprocessing, feature engineering, and the application of the Decision Tree algorithm. Here is a detailed overview of the model architecture:

- Dataset Loading and Inspection:

- The project begins by acquiring a comprehensive dataset containing information on diseases and their associated symptoms. This dataset is loaded into a Pandas DataFrame for further analysis.

- Data Preprocessing:

- Missing values in the dataset are handled using a forward-fill method, ensuring the continuity of information.

- Disease and symptom names are processed to replace special characters, and the data is organized into a structured format suitable for subsequent analysis.

- Disease-Symptom Dictionary Formation:

- A dictionary (`disease\_symptom\_dict`) is created, mapping diseases to their corresponding symptoms. This dictionary captures the intricate relationships between diseases and symptoms in the dataset.

- Count of Disease Occurrence:

- Another dictionary (`disease\_symptom\_count`) is established to track the count of disease occurrences. This count provides insights into the prevalence of each disease in the dataset.

- Saving Cleaned Data:

- The cleaned and processed data is saved as a CSV file ('cleaned\_data.csv'), ensuring accessibility and reproducibility in subsequent steps.

- One-Hot Encoding of Symptoms:

- Symptom names are label-encoded and then one-hot encoded to represent them as binary vectors. This step enhances the model's ability to understand categorical data and capture relationships between symptoms and diseases

- Combining Labels with Disease Column:

- The one-hot encoded labels are concatenated with the disease column, creating a new dataset (`df\_concat`) that encapsulates diseases along with their associated one-hot encoded symptoms.

- Grouping and Summing for Symptom Aggregation:

- To handle multiple symptoms associated with each disease, the dataset is grouped by disease, and symptom occurrences are summed across each group. This process results in a condensed and aggregated representation of symptoms for each disease.

- Training the Decision Tree Model:

- The processed dataset is split into training and testing sets.

- The Decision Tree algorithm is chosen for its interpretability and effectiveness in handling categorical data.

- The model is trained using the one-hot encoded symptom features, learning patterns and relationships between symptoms and diseases.

- Visualizing the Decision Tree:

- The Decision Tree model is visualized using the graphviz library, providing a graphical representation of the decision-making process. This visualization aids in understanding how the model makes predictions based on symptoms.

- Model Evaluation:

- The accuracy of the trained model is evaluated on the entire dataset, providing insights into its overall performance.

- The model is employed to predict diseases based on symptoms, and the results are compared against the actual disease occurrences for verification.

- Iterative Steps for Data Verification:

- Several steps, such as encoding, concatenation, and grouping, are implemented iteratively to ensure the accuracy and integrity of the data throughout the process.

**Experimentation details**

**1. Dataset Loading and Examination:**

- The project begins by acquiring a dataset from Columbia University, a valuable resource containing information on diseases and their associated symptoms.

[**http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html**](http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html)

- The dataset is loaded into a Pandas DataFrame, allowing for easy exploration and manipulation.

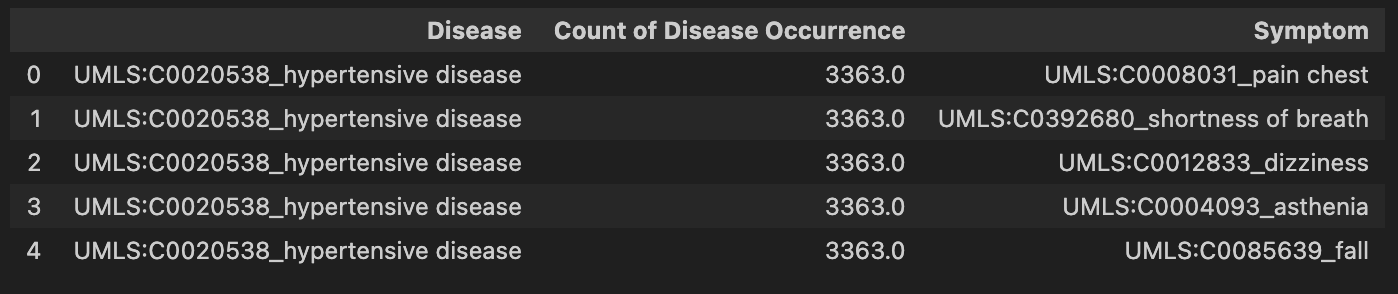
- Initial examination of the dataset reveals any missing values, which is a crucial step in ensuring data quality.

A screenshot of a black screen

Description automatically generated

**2. Data Cleaning:**

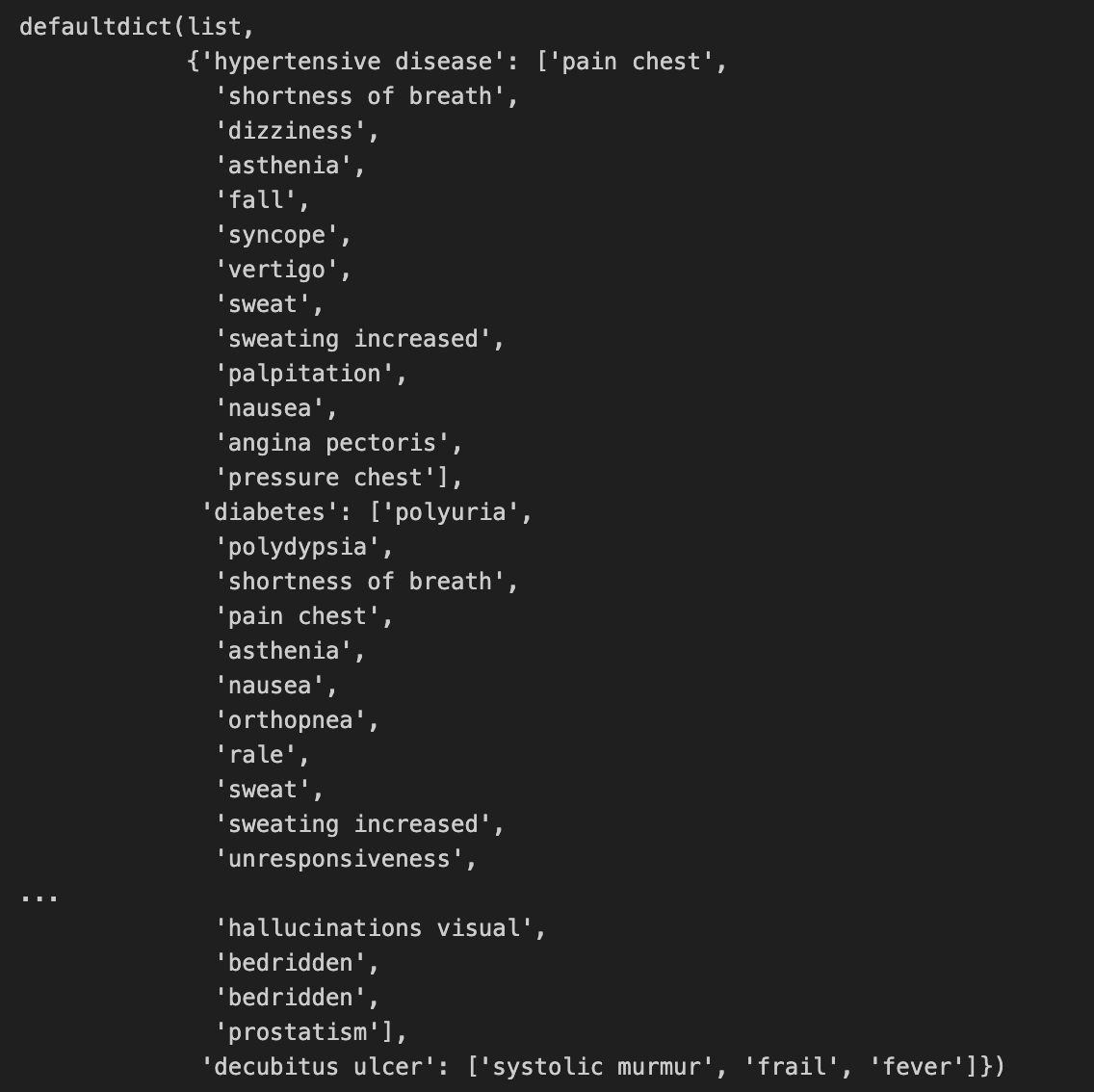
- The forward-fill method is employed to handle missing values. This technique involves filling NaN values with the last observed non-null value in the dataset, ensuring continuity and completeness in the data.



**3. Processing Disease and Symptom Names:**

- A function is defined to process disease and symptom names, replacing special characters, such as '^,' with underscores.

- The processed data is split into a list, retaining only symptom names.



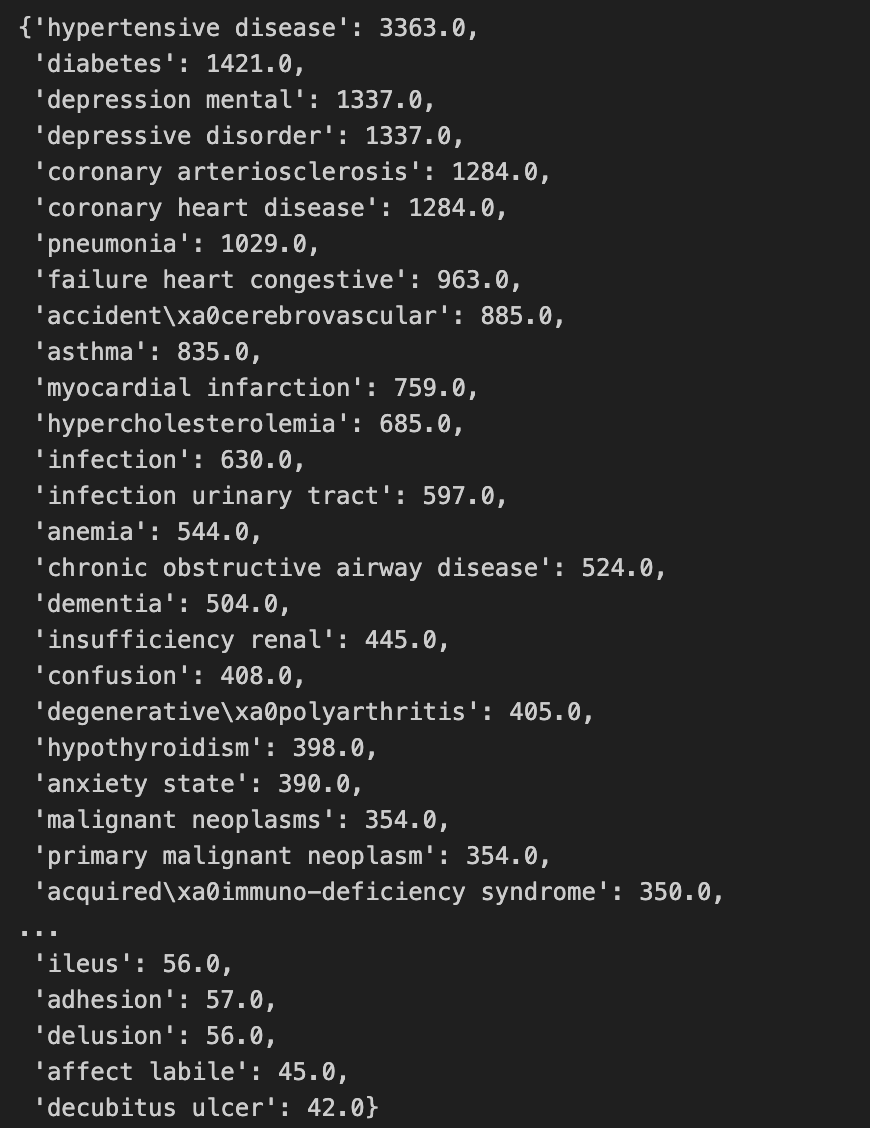
**4. Data Cleanup and Dictionary Formation:**

- Disease names and their corresponding symptoms are extracted from the dataset.

- A dictionary (`disease\_symptom\_dict`) is formed, where diseases serve as keys, and associated symptoms are stored as values.

**5. Count of Disease Occurrence:**

- The count of disease occurrences is tracked and stored in a separate dictionary (`disease\_symptom\_count`).

- This count represents the frequency of each disease in the dataset

**6. Saving Cleaned Data as CSV:**

- The processed data is saved into a CSV file ('cleaned\_data.csv'), facilitating ease of use in subsequent steps.

**7. One-Hot Encoding of Symptoms:**

- Label encoding is initially applied to convert symptom names into numerical values.

- One-hot encoding is then performed to represent symptoms as binary vectors, enhancing the model's ability to understand and learn from categorical data.



**8. Combining One-Hot Encoded Labels with Disease Column:**

- The one-hot encoded labels are concatenated with the disease column to form a new dataset (`df\_concat`).

- This dataset represents diseases along with their associated one-hot encoded symptoms.

A screenshot of a black and white screen

Description automatically generated

**9. Grouping and Summing for Symptom Aggregation:**

- Since each disease can have multiple symptoms, the dataset is grouped by disease, and the symptoms are summed across each group.

- This aggregation ensures that each row corresponds to a unique disease, capturing all associated symptoms.

**10. Saving the Processed Dataset:**

- The final processed dataset is saved as 'training\_dataset.csv,' ready for subsequent use in model training.

**11. Model Training with Decision Tree:**

- The processed dataset is split into training and testing sets.

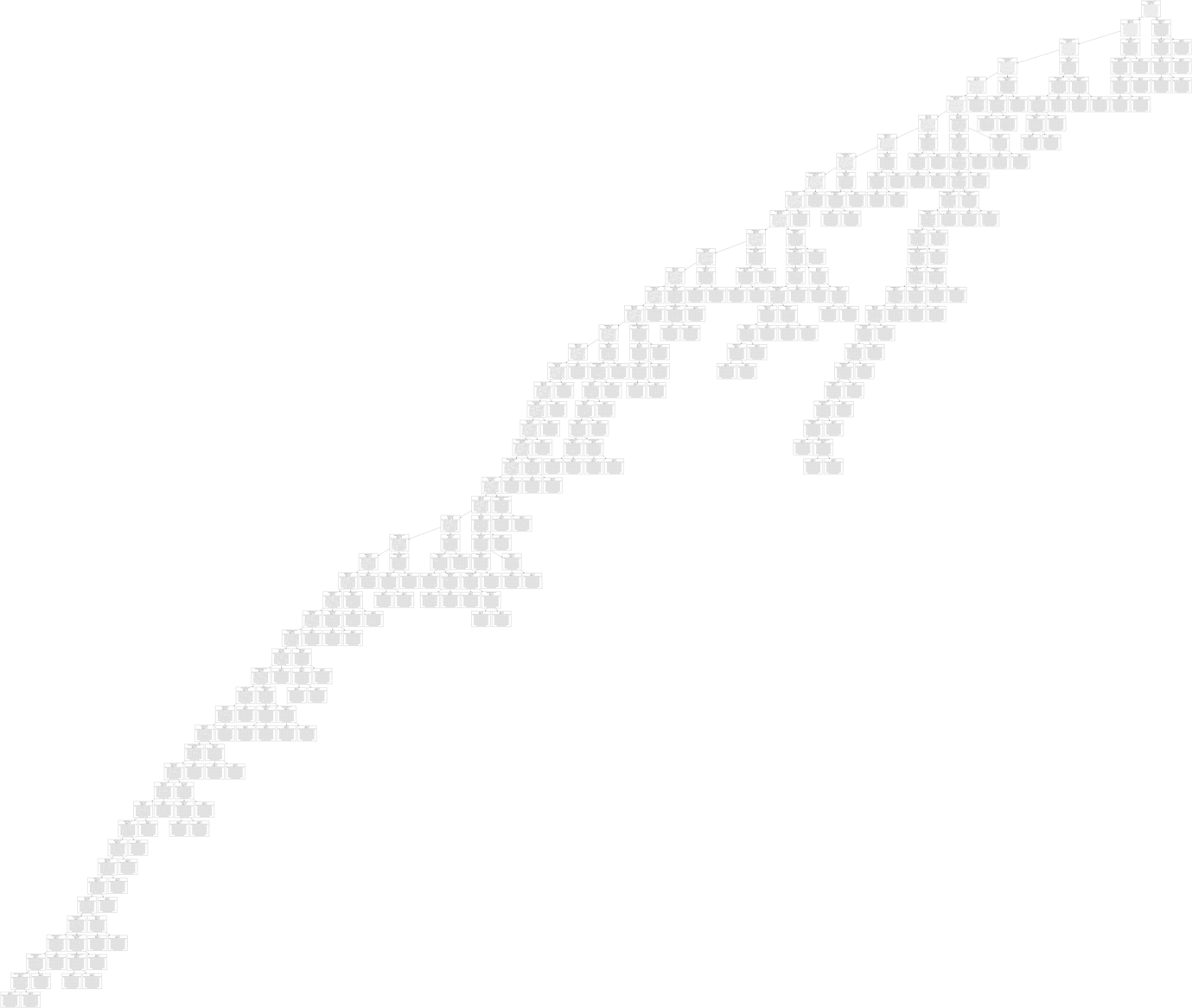
- The Decision Tree algorithm is chosen for disease prediction due to its interpretability and suitability for categorical data.

- The model is trained using the one-hot encoded symptom features.

**12. Visualizing the Decision Tree:**

- The Decision Tree is visualized using the graphviz library, providing a graphical representation of the decision-making process.

- The resulting tree structure is analyzed to gain insights into how the model makes predictions based on symptoms.



**13. Model Evaluation:**

- The accuracy score of the trained model is assessed on the entire dataset, providing an indication of its overall performance.

- The trained model is then used to predict diseases based on symptoms, and the results are compared with the actual disease occurrences.

**14. Iterative Steps for Data Verification:**

- Various steps, such as encoding, concatenation, and grouping, are implemented iteratively to ensure data integrity and accuracy.

**Conclusion and Future Recommendations:**

In conclusion, this project represents a noteworthy exploration into the realm of disease prediction from symptoms using machine learning. The outcomes underscore the viability of leveraging algorithms like Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting for accurate and interpretable predictions in healthcare contexts. However, as with any scientific endeavor, there are areas for further refinement and future considerations.

The demonstrated success of the chosen algorithms in symptom-based disease prediction opens avenues for their integration into clinical workflows, offering valuable decision support for healthcare professionals. Nevertheless, it is crucial to approach this technology with caution, recognizing that it should complement rather than replace the expertise of medical practitioners. Ethical considerations, such as patient privacy, model transparency, and bias mitigation, must remain at the forefront of any implementation strategy.

Future research endeavors should focus on expanding the dataset used in this study to encompass a broader range of symptoms and diseases, fostering a more comprehensive and generalizable model. Collaborative efforts with healthcare providers could enhance the model's accuracy by incorporating domain-specific knowledge and clinical insights. Furthermore, the model's interpretability could be improved, addressing the 'black box' challenge in machine learning to enhance trust and facilitate broader adoption.

In terms of practical application, adapting the model to different healthcare environments and demographics is a crucial step toward real-world implementation. Customization based on regional variations in disease prevalence, healthcare infrastructure, and patient demographics would ensure the model's relevance and effectiveness across diverse contexts.

In the future, a focus on developing user-friendly interfaces for healthcare professionals to interact with the model, interpret results, and integrate predictions into clinical decision-making processes would enhance its usability. Additionally, ongoing collaboration with the medical community, coupled with iterative model updates, would contribute to the model's continuous improvement and alignment with evolving healthcare practices.

In summary, while this project provides a foundation for symptom-based disease prediction, it represents just the beginning of a broader exploration into the intersection of machine learning and healthcare. The journey ahead involves refining models, addressing ethical considerations, and fostering collaborative efforts to bring these technologies to the forefront of proactive and personalized healthcare.