PATIENT CASE SIMILARITY

A PROJECT REPORT

Submitted by,

BHUMPALLI VISHNUVARDHAN REDDY	**20211CIT0013**
SANJANA R	**20211CIT0043**
LINGAMDHINNE AKANKSHA	**20211CIT0047**
KOYI MITHUN	**20211CIT0055**
PERUMALLA SAI SURYA	**20211CIT0079**

Under the guidance of,

Dr. SHARMASTH VALI Y

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

(INTERNET OF THINGS)

AT

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNVIERSITY

BENGALURU

JANUARY 2025

PRESIDENCY UNIVERSITY
SCHOOL OF COMPUTER SCIENCE ENGINEERING
CERTIFICATE

This is to certify that the University project report titled **"Patient Case Similarity"** being submitted by "Bhumpalli Vishnuvardhan Reddy, Sanjana R, Lingamdhinne Akanksha, Koyi Mithun, Perumalla Sai Surya" bearing roll number "20211CIT0013, 20211CIT0043, 20211CIT0047, 20211CIT0055, 20211CIT0079" in partial fulfilment of requirement for the award of degree of Bachelor of Computer Application is a bona-fide work carried out under supervision

DR. SHARMASTH VALI Y
Associate Professor
School of CSE
Presidency University

DR. ANANDARAJ S P
Professor & HoD
School of CSE
Presidency University

Dr. L. SHAKKEERA
Associate Dean
School of CSE
Presidency University

Dr. MYDHILI NAIR
Associate Dean
School of CSE
Presidency University

Dr. SAMEERUDDIN KHAN
Pro-VC School of Engineering
Dean -School of CSE&IS
Presidency University

PRESIDENCY UNIVERSITY
SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **"PATIENT CASE SIMILARITY"** in partial fulfilment for the award of **Bachelor of Technology** in **Computer Science and Engineering in IOT**, is a record of our own investigations carried under the guidance of Dr. Sharmasth Vali Y, Associate Professor, School of Computer Science and Engineering, Presidency University, Bengaluru.

We have not submitted the matter presented in this report anywhere for the award of any Degree.

NAME	ROLL NUMBER	SIGNATURE
BHUMPALLI VISHNUVARDHAN REDDY	20211CIT0013	
SANJANA R	20211CIT0043	
LINGAMDHINNE AKANKSHA	20211CIT0047	
KOYI MITHUN	20211CIT0055	
PERUMALLA SAI SURYA	20211CIT0079	

ABSTRACT

Patient similarity analysis is emerging as a very efficient tool in precision medicine to identify similar patients and clinical outcomes for a patient. This work develops a decision tree-based similarity model for the patient in order to design an effective personal treatment strategy and optimize outcomes.

This method collects and preprocesses data obtained about the patient, which may include demographics and medical history, and then uses clinical laboratory results. It uses feature engineering techniques to extract features that contribute to the similarity between the patients. Using a preprocessed decision tree algorithm learned on the preprocessed data, it's trying to identify decision rules needed to classify patients into like groups.

Accuracy, precision, recall, and F1-score are a few metrics that can be used for assessing the performance of the model. Model interpretation would reveal what factors have resulted in the patients being similar.

The decision tree model will be integrated into the clinical decision support systems. Finding similar patients helps clinicians understand by leveraging past experience and evidence-based guidelines to change their treatment plans in the best possible way. Furthermore, the model will help in finding possible clinical trials and research opportunities as well.

It can be a useful tool for identifying similar patients and informing clinical decision-making towards better patient-centered care. Future research directions might involve integrating more techniques in machine learning, such as deep learning approaches that will further improve the accuracy and interpretability of patient similarity models.

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC, School of Engineering and Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project. We express our heartfelt gratitude to our beloved Associate Deans **Dr. Shakkeera L** and **Dr. Mydhili Nair**, School of Computer Science Engineering & Information Science, Presidency University, and **Dr. Anandaraj S P**. Head of the Department. School of Computer Science Engineering, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Dr. Sharmasth Vali Y,** Associate Professor and Reviewer **Ms. Raesa Raseen**, Assistant Professor, School of Computer Science Engineering & Information Science, Presidency University for his/her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work. We would like to convey our gratitude and heartfelt thanks to the PIP2001 Capstone Project Coordinators **Dr. Sampath A K, Dr. Abdul Khadar A** and **Mr. Md Zia Ur Rahman,** department Project Coordinators **Dr. Sharmasth Vali Y** and Git hub coordinator **Mr. Muthuraj.**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

Bhupmalli Vishnuvardhan Reddy

Sanjana R

Lingamdhinne Akanksha

Koyi Mithun

Perumalla Sai Surya

TABLE OF CONTENTS

`

List of Figures

List of Table

CHAPTER-01
INTRODUCTION

Patient similarity analysis is the method of discovering patients with similarities in clinical characteristics. For long, it has been touted as a useful tool in precision medicine for better and more informed decisions on the diagnostic, therapeutic, and prognostic approaches. It may bring about better patient outcomes, decreased costs of healthcare delivery, and hastened drug discovery.

A popular machine learning algorithm with a strong and interpretable method for patient similarity analysis is the decision tree. A decision tree classifies patients based on their clinical features by developing a tree-like model of decisions and possible consequences. This type of analysis has proven fit for this task due to its ability to handle both numerical and categorical data while generating easily understandable rules.

It's relatively recent and has exploded data generation due to e-health records and wearable devices in healthcare. The amount of generated data holds lots of information that can be mined for valuable insights into care about a patient. Analysing and even interpreting large datasets is not an easy task.

Techniques that deal with machine learning, including decision trees, promise to change healthcare by automating data analysis and finding previously unknown patterns. Hence, in the specific context of health services, providers can take better decisions regarding patients using past historical information available with such techniques, eventually improving patient outcome.

However, while promise in patient similarity analysis is significant, attention is required on many challenges: data quality and completeness, feature selection, model interpretability, and ethics. Patient data must be accurate and complete for any credible analysis. Identifying the most relevant features that best describe patient similarity is challenging because a large number of features may be correlated with each other or redundant. Therefore, models need to be both accurate and interpretable to clinicians to gain trust and promote adoption in the clinical environment. Protecting patient privacy and ensuring fairness and no bias are other key ethical considerations.

Decision trees offer a promising means of overcoming these problems. A hierarchical structure of decisions and their potential outcome, constructed by a decision tree, should help identify subgroups of patients with characteristics that are similar in some respects. Such information may be used to inform treatment decisions, predict the progression of disease, and mark out potential drug targets.

This project is developed based on the decision tree-based patient similarity model that is presumed to identify similar patients with accuracy and provide helpful information in making relevant clinical decisions. The effort here will attempt to address the challenges and limitations introduced by other approaches into this field and improve patient similarity analysis for better patient care.

CHAPTER-02

LITERATURE REVIEW

Patient similarity analysis represents an important tool in the pursuit of precision medicine, thus identifying patients with similar clinical characteristics and outcomes. This will help the healthcare providers have better profiles while making decisions on diagnosis, treatment, and prognosis. This can lead to better patient outcomes, healthcare savings, and more rapid drug discovery.

Decision trees are probably one of the most famous machine learning algorithms, providing an extremely powerful yet intuitive approach for patient similarity analysis. As decision trees might be framed as a tree-like model of decisions and the possible outcomes, they could classify patients based on their clinical features. The reason decision trees are appropriate for this purpose is that they can effectively handle both numerical and categorical data, such as giving easily understandable rules.

This is because EHRs and wearable technology have been introduced into the health care sector, thus boosting the generation of data at large. Considering this, the opportunities where useful information can be delivered with a far-reaching aim to improve the care given to the patient are needed. It is difficult to interpret or analyze large datasets.

Actually, the use of machine learning techniques in decision trees can prove to be the way forward in revolutionizing healthcare by having mechanisms for automatically analyzing data and finding hidden patterns within the data. Through having leverage on machine learning, healthcare providers can make informed decisions that have better results for patients.
Problems still exist despite the key benefits that can be achieved by using patient similarity analysis. These are based on data quality and completeness, feature selection, model interpretability, and finally, ethics. The precision and completeness of the data of patients are mainly for sufficient analysis. This would often be very difficult because many features would be correlated and even redundant to define the most pertinent features for defining similarity within the patient population. Models will be accurate and interpretable for the clinicians so that the clinicians can accept and trust them in clinical practice. Patient privacy and the fact that the analysis is unbiased and fair comprise very pertinent ethical concerns.

Decision trees are among the most promising solutions to these challenges. These create a hierarchical structure of decisions and outcomes, which allow researchers to note distinct subgroups of patients

who share characteristics that can be used to base treatment, predict progression of disease, and identify drug targets.

Several studies had research on the application of decision trees on the analysis of patient similarity. For example, had applied decision trees to identify clinical features and outcomes common for patients classified under the category of cancer patient population. Had considered the application of decision trees to predict a patient's response for some treatments. Had discussed the use of decision trees with regard to identifying risk for adverse drug reactions.

Although many benefits exist, decision trees have disadvantages. They might become noise-sensitive and fail to capture complex relations between the variables. Some of these disadvantages have led researchers to new approaches: ensemble approaches in the form of random forests and gradient boosting machines in which a number of decision trees are put together to form a single model with better performance.

CHAPTER-03

REAL-TIME APPLICATIONS OF PATIENT SIMILARITY ANALYSIS

The following are 10 real-time applications of patient similarity analysis:

1. Personalized Medicine:

- Tailor treatment plans based on the unique characteristics and probable response of a patient to one or more therapies.
- Find optimal dosages and periods of treatments.

2. Early Disease Detection:

- Patient profiles that are diagnosed with potentiality diseases; similarity matching of the patient's profile with those of disease patients' histories and taking measures for the early interventions

3. Diagnosis of Rare Diseases:

- With the genetic profile, rare diseases could be diagnosed fast by matching a patient's symptoms
- Support Groups and the Clinical Trial Connect.

4. Recruitment of the Clinical Trials :

- Identify patients who may participate in the clinical trial. There is a difference in those identified by past clinical trials.
- Optimization of trial design and patient stratification

5. Monitoring for Adverse Drug Reactions:

- Predictive models of adverse drug reactions in relation to the patient's characteristics and interactions between drugs; an early warning system that would alert potential serious adverse events, therefore avoiding these adverse events before it gets worse.

6. Drug Repurposing :

- Identification of proper subsets of patients who may need the drug uses to develop new indications for drugs in existence.
- Accelerated development of drugs at a lesser input cost.

7. Public Health Surveillance:

- Surveillance of diseases that result from infectious factors and outbreak
- Detection of appropriate interventions to targeted public health intervention

8. Precision Surgery:

- Identification of maximum surgical procedures, which require patient-specific therapy.
- Lower surgical complications with improved results for the patients.

9. Psychiatry:

- Patient identification vulnerable to mental illnesses
- Individualized psychiatric care

10. Geriatric Care:

- Patient identification with similar geriatric health conditions
- Treatment plans customized for geriatric patients
- With patient similarity analysis, health care givers will make better decisions and improve patient results, medical research will also be expedited.

CHAPTER-04
FINDING ALGORITHM

4.1 Why Machine Learning is a Necessity in Patient Similarity Analysis

Machine learning is a robust technique that can be applied for identifying meaningful patterns in complex patient data analysis. That's the reason why it cannot be done away with. Here is why.

4.1.1. Handling Complex Data:

- Many clinical data sets contain a large number of features, and out-of-distribution patterns are hard to establish. Machine learning algorithms tackle noisier data very nicely.
- Noisy Data: In real-world, clinical data is noisy and incomplete in real-world scenarios. The noisy one does not affect the working of machine learning algorithms because this algorithm learns very well by imperfect data.

4.1.2. Automatic Feature Engineering:

- Feature Selection: While the process of feature selection itself is done by the machine learning algorithm, in the process, features that look most relevant automatically reduce the necessity of doing feature engineering.
- Feature Creation: Techniques like feature engineering may create advanced techniques to develop new features and enhance the model's performance.

4.1.3. Predictive Modeling:

- Outcome Prediction: Machine learning algorithms can predict the future outcomes of patients, for example, disease progression or treatment response.
- High-Risk Patient Identification: Identify patients with a high risk of adverse events or disease complications.

4.1.4. Personalization:

- Customized Therapies: This may allow designing individualized treatment protocols for every individual by identifying similar patients.
- Precision Medicine: Developing individualized, patient-specific treatment designs by taking into account multiple aspects of a patient.

4.1.5. Learning and Adaption Continuously:

- Model Refreshing: A machine learning model can be updated with more new data at any instant to make it more accurate with relevance over time.
- Adaptability to Changing Requirements: The models may adapt to changing requirements such as changes in the population of patients, changes in disease trends and in treatments.

4.1.6 Particular Machine Learning Methods:

- ✓ **Decision Trees :** This method comes up with a tree-like representation of decisions along with their resultant possibilities.
- ✓ **Random Forest :** A set of decision trees operating in ensembles to further increase precision and decrease the phenomenon of overfitting.
- ✓ **Support Vector Machines:** Outstanding Classification and Regression machines
- ✓ **Neural Networks:** This is advanced models that allow for complex learning of patterns.

By machine learning, the ability of patient similarity analysis unlocks potential for better care of patients, faster discovery of drugs, and optimal resource use in health care.

4.2 Why Visualization is Important for Patient Similarity Analysis

Visualization is a powerful tool that can greatly enhance our comprehension of patient similarity analysis. Visualizing data enables us to reveal patterns that might otherwise go unnoticed, understand trends, and find correlations; it enables us to communicate our findings better. Here's why visualization is important:

4.2.1. Data Exploration and Understanding:

- Identify Patterns and Outliers: Histograms, box plots and scatter plots can help in visualizing unusual patterns or anomalies or even indicate data quality.
- Visualize the distribution of features that one might understand how the different features are distributed and also related to each other as well.

4.2.2. Interpretation of Models:

- Visualizing the decisions of a decision tree as well as important features.
- Feature Importance Plots: Feature importance plots can be used to establish which features are most relevant in the model.
- Confusion Matrices: Confusion matrices provide an easy way to understand performance as well as the improvements needed in the model while visualizing confusion matrices.

4.2.3. Communications of Insights:

- Effective Communication: A visualization can present complex information in a crystal-clear and concise way so that stakeholders find it easy to understand the findings.
- Engage Stakeholders: Interactive visualizations can engage clinicians, researchers, and policymakers for dialogue and decision-making.

4.2.4. Model Performance Validation:

- Residual Plots: Visualizing residuals can spot patterns in the model errors and assess the fit.
- ROC Curves: Evaluation of models for classification using ROC Curves.
- Precision-Recall Curves: Trade-off between precision and recall.

4.2.5. Identifying Patient Subgroups:

- Clustering Techniques: Cluster the similar patients to observe their characteristics and identify those that are different from each other.
- **Dimensionality Reduction:** Techniques that use PCA for reducing the dimension of data so that high-dimensional data can be visualized for its relationship.

An example of a patient similarity analysis project is to identify patients who share similar risk factors for heart disease. The relationship between age, blood pressure, and cholesterol might have a pattern or trend that isn't easily observed by the numbers alone. Such a visualization may be used to identify high-risk patients and initiate preventive measures among clinicians.

That through effective use of data visualization techniques, we can acquire more insight from patient data along with a better interpretation of the models and more informed decision-making in the field of healthcare.

4.3 MACHINE LEARNING WORKFLOW

4.3.1 Data Preprocessing

- Normalization:
  - Normalize the numerical features so that the models are comparable.
- One-Hot Encoding:
  - This is used for categorical variables, such as gender and symptoms and maps them into binary vectors.
- Outlier Removal:

- Use statistical techniques such as IQR (Interquartile Range) to remove outliers.

➢ Data Augmentation:

- Generate synthetic data points to supplement the training set, especially for rare diseases.

4.3.2 Model Development

➢ Algorithm Selection:

- Decision Trees for interpretability.
- Random Forests or Gradient Boosting for higher accuracy.
- Neural Networks for more complex, non-linear behavior.

➢ Hyperparameter Tuning:

- Hyperparameters of the model need to be tuned using grid search and Bayesian optimization.

4.3.3 Model Evaluation

➢ Metrics:

- For binary classification, precision, recall, F1-Score, ROC-AUC, and for multi-class, confusion matrix to understand the patterns of the misclassifications.
- Use k-fold cross validation to get the performance confident.

4.3.4 Explainability

➢ Feature Importance:

- Use SHAP (SHapley Additive exPlanations) to retrieve the most influential factors behind predictions of a model.

➢ Partial Dependence Plots

- Graphical representation of the marginal effect of an individual feature.

➢ Local Interpretable Model-agnostic Explanations (LIME):

- Interpret individual predictions towards crucial decision-making.

4.4 Selected Algorithm (Decision Tree) :

A decision tree is a supervised machine learning algorithm that can be used for classification and regression tasks. It works recursively to partition a dataset into subsets based on the values of features until eventually it forms a tree-like structure in which each node will represent a decision point or

condition. Decision trees are very popular because they are easy to understand, simple to interpret, and they can work with both categorical and numerical data.

Architecture of Decision Tree

4.4.1 Flow of Work (Decision Tree) :

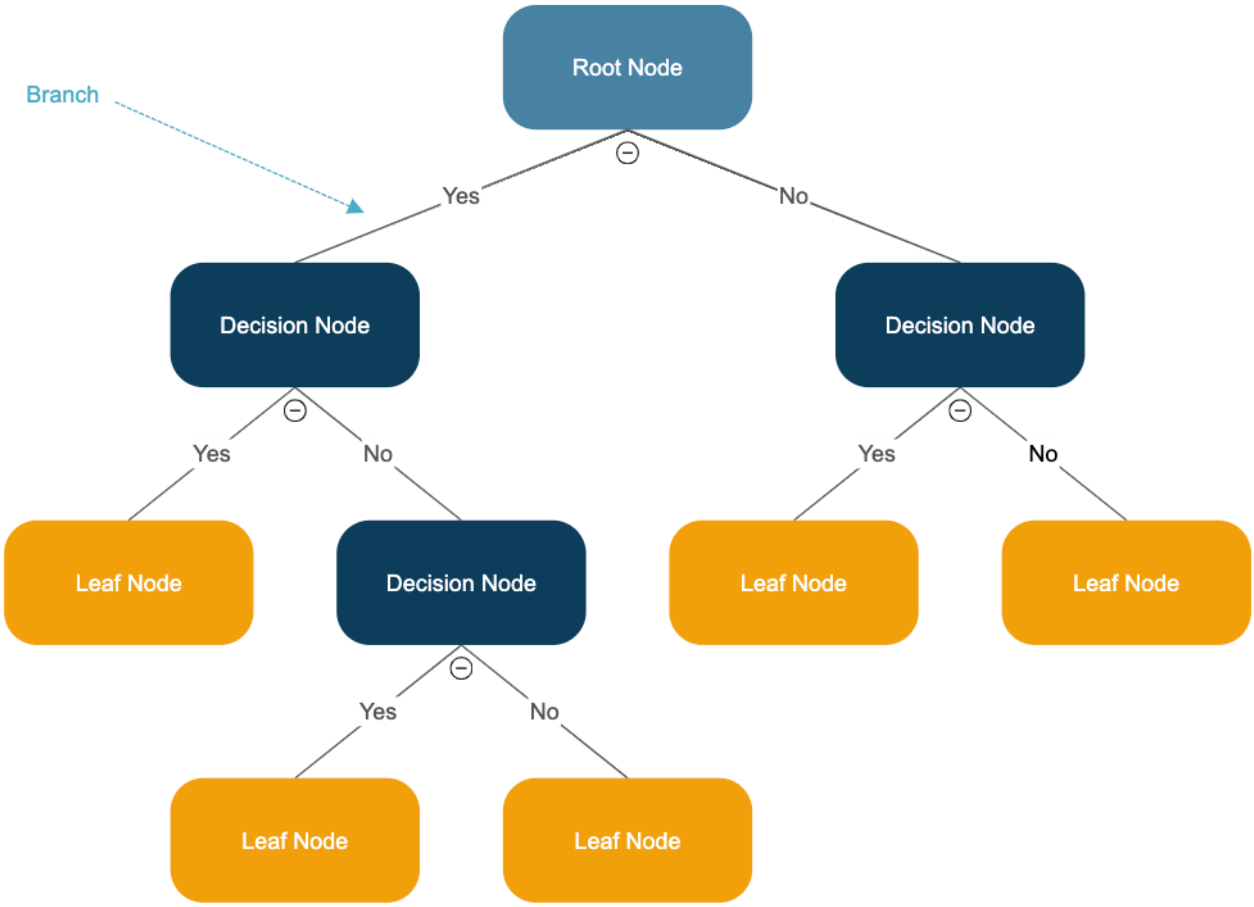


Fig 5.1 : Decision Tree Flowchart

4.4.1.1. The elements of architecture of a decision tree consist of the following:

a. Root Node:

- It is the whole dataset that is the very first point of decision.
- It divides the data based on the most critical feature for example, it is the one with highest information gain or lowest Gini impurity.

b. Internal Nodes:

- It portrays decisions taken based on features at intermediate points.
- Each node checks a condition, say, Age>30? and goes on sending data to child nodes based on the outcome.

c. Branches:

- Represent the results of a condition at a node.
- Link nodes in the tree structure.

d. Leaf Nodes:

- Carry the final output for a portion of data (for example, class label or predicted value).
- No further splits are done after these nodes.

4.4.2 Splitting Criteria:

a. Gini Impurity: It measures the impurity of a node. The lower the Gini impurity, the better the split is.

b. Information Gain (IG): This is entropy-based, and it calculates the amount of decrease in uncertainty after splitting.

c. MSE (Mean Squared Error) : Minimizes prediction errors when working with regression tasks.

CHAPTER-05

PROPOSED METHODOLOGY

5.1 DESIGN PROCEDURE:

5.1.1. Problem Definition and Goal Setting:

• It clearly defines the objective for the project, which is to identify similar patient cases for the treatment recommendations.

5.1.2. Data Collection and Preprocessing:

• It gathers the similar medical data, which includes patient demographics, symptoms, treatments, and outcomes.

• It cleans and preprocesses the data to handle missing values, inconsistencies.

• Normalizes the numerical data to ensure comparability.

• It considers feature engineering to create the new features that can be more informative.

5.1.3. Feature Selection:

• Identifies the relevant features that are predictive to the patient case similarity.

• Use technique like correlation analysis, or wrapper methods to select the most effective features.

5.1.4. Decision Tree Construction:

• Chooses the decision tree algorithm based on characteristics of your data and the desired properties of model.

• Trains the decision tree model on the pre-processed data with the help of tree-like structure wherein each node can be presented as a test on an attribute, and each leaf node presents the class label.

5.1.5. Model Evaluation:

• Uses appropriate metrics for judging the performance of the decision tree model over the validation dataset.

• It involves methods like cross-validation to arrive at the conclusion regarding generalization aptness of the model.

5.1.6. Model Optimization:

• Pruning, boosting, and bagging are identified in case performance of first models is poor to enhance

model precision along with the elimination of overfitting.

• It implements experiments through diverse decision tree algorithms as well as their various hyperparameters to try out for an optimal mix.

5.1.7. Deployment and Use:

• Deploys the trained decision tree model in the production environment.

• Embeds the model into clinical workflow or research tool to aid in support of decision-making and analysis.

• It always monitors the model performance and updates it on the change requirements and changes based on data.

5.2 Why Decision Trees for Patient Similarity?

Decision trees are such a strong machine learning algorithm that carries with itself a few truly amazing benefits to the patient similarity analysis task:

1. Interpretable

- Easily Explained Decision Rules: It forms a hierarchy of choices and their probable result and within that, communicates exactly what kind of decision the model actually carried out with its understanding.
- Clinicians readily see why it has returned any specific result.

2. Handling Mixed Data Types:

- Variety Data Handling: The decision tree can handle continuous and discrete data, so it is very applicable for the medical datasets of virtually any kind.
- Liable Feature Engineering: Decision tree accommodates kinds of data without requiring heavy preprocessing steps.

3. Feature Importance Analysis:

- The decision trees can detect vital features of the patients, which enhance the similarity of patients.
- This helps the doctor build notions of what choices to focus in their thought processes while securing the most relevant ones.

4. Noisy Robust:

- Noise robustness: Decision trees have proven not to be over sensitive on the noisiness of the existing data. Therefore, one could be using decision trees where actual medical data proves noisy and incomplete in most fields.
- Residency against Outliers: Decision trees enable not outliers to command most of the control over what predictions the model could eventually produce

5. Non-parametric Property:

- Flexibility: The decision tree assumes no parameters about the underlying distribution of data. This is thus deployable for any pattern that could be established in data.
- Techniques such as pruning may be applied for not overfitting so that decision trees could generalize very well to new data.

This will allow constructing the building of models as both accurate and interpretable models, improving clinical decision-making, by drawing on the strength which a decision tree really needs.

5.3 Key Considerations When Using Decision Trees in Patient Similarity Analysis

Decision trees are a very powerful machine learning technique for patient similarity analysis, but one should know their limitations and pitfalls:

1. Overfitting:

- Problem: decision trees suffer from overfitting, especially when the model becomes too complex, leading to poor generalization performance on new, unseen data.
- Pruning Techniques: Some of the techniques by which overfitting can be avoided are pre-pruning and post-pruning. This reduces the depth and complexity of the tree.

2. Sensitivity to Noise:

- Noise Impact: A noisy data can result in serious implications regarding the decision tree's accuracy as well as stability.
- Data Cleaning and Preprocessing: A good technique for data cleaning and preprocessing will be capable enough to reduce the effect of noise.
- Ensemble Methods: The ensemble methods like random forests can be used to increase the noise robustness of the decision tree.

3. Lower Expressiveness:

- Complex relationships: The decision tree fails to explain the complex, nonlinear relationship between the features.
- Feature engineering: The important features make the model capable of more expressive complexity
- Ensemble Methods: There is a possibility of increasing expressiveness by using ensemble methods for the decision tree.

4. Interpretability:

- Interpretability: The decision tree model is fairly interpretable unless the trees get too complicated.
- Visualization Technique: The work done by a decision tree in order to find an answer can be visualized
- Feature Importance Analysis: Determination of the most important features which describe model behavior

5. Computational Intensity:

- Training Time: The training of large decision trees requires a lot of computational powers primarily when dealing with large data.
- Optimization Techniques: With the fast algorithms and optimization techniques, this long training time might be avoided and the process speeded up in some way.

Knowing these limitations and using correct techniques, we are able to use decision trees along with patient similarity analysis and obtain more accurate and reliable models.

5.4 Why One-Hot Encoding is the Essence of Patient Similarity Analysis

One of the aspects of a machine learning software, one-hot encoding is the support of a feature that can be crucial to the task of patient similarity analysis. This encoding changes categorical features into numeric; this makes it very easy for the machine learning algorithm to understand and analyze.

Key Reasons for One-Hot Encoding:

1. Representation of Categorical Data:

Most medical datasets have categorical features including race, diagnosis, and drug names.

One such transformation which makes this categorical feature useable as numerical features by an algorithm for consumption is one-hot encoding.

2. Information in Categories Not Lost:

Using one-hot encoding, for any class, a binary indicator is given so that all the information regarding the ordinal relationship it had is preserved.

This is correct because categorical features have been known to capture numerous categories, hence independent.

3. Model Performance:

It allows the decision tree, along with many Machine learning algorithms, to learn the categorical difference in the right way. This no doubt has resulted in reliable better Patient similarity analysis.

Example :

Let's consider the feature "Diseases", which consists of features like Diabetes, Heart Disease, or Cancer. One-hot encoding results in three new binary features that are defined as follows:

- `Is_Diabetes`
- `Is_Heart_Disease`
- `Is_Cancer`

Assign a 1 or 0 for all of these features for each patient based on whether they actually have the disease or not.

Advantages of One-Hot Encoding:

* Improved performance of model improved. Since categorical feature can be presented with one-hot encoding, each categorical feature makes possible all categorical features, so any categorical feature presented numerically could possibly improve any performance coming out of a machine learning model.

5.5 Why Prediction is Required for Patient Similarity Analysis?

Prediction is the core element of patient similarity analysis as it will help us predict the future outcomes and take good decisions. Using historical data and the machine learning approach, we can predict a few parameters of patient health like the following:

1. Disease Progression

a. Predict disease progression in the future

b. Identify at-risk patients for disease progression

2. Treatment Response

- Predict how a patient will react to a certain treatment or therapy.
- Identify probable case for a particular treatment

3.Adverse Drug Reactions:

o Predict the possible risk of adverse drug reaction based on patient profile and the medical history

4.Patient Outcomes:

o predict the patient outcome in mortality or morbidity to aid a clinical decision

5.Patient Similarity:

o predict the patient similarity to create personalized treatments and clinical studies.

This helps in making sound decisions and proper usage of the resources. Machine Learning Techniques Used for Prediction:

- With Decision Trees, decision and every probable result can be depicted by drawing a tree.
- Grouped decision trees with reduced overfitting and maximum precision are known as random Forest.
- Support Vector Machine represents a very powerful model which is used for not just classification but also with remarkable regression.
- Neural network is a very complicated type of model that helps understand very complex patterns also; with these techniques a predictory model can be derived to benefit the patient healthcare outcome.

5.6 OBJECTIVES

Data Collection and Pre-processing

1.Patient data source: Find sources of data concerning patients like e-Health records, clinical studies, biomedical databases, etc.

2. Data Cleaning and Interpolation: Clean the data to eliminate errors and inconsistencies and fill in missing values.

Interpolate missing values using some technique (mean interpolation, median interpolation, mode interpolation)

3. Normalization or Standardization: Normalize the numerical features for comparison.

One-hot encode or label encode for categorical feature

5.7 Feature Engineering

4. Feature Selection:

- Select features that have a strong potential to result in similarity between patients
- Apply filter, wrapper and embedded feature selection methods

5. Feature Creation:

- Generate new features based on feature interaction or through creating new features domain relevant
- For example, age group of the patient for creating a new feature known as "Age Group"

5.8 Model Training and Evaluation

5.8.1.Model Selection :

o Choose relevant machine learning algorithm, say decision trees, random forests, neural networks.

o Choose data complexity and the level of interpretability desired.

5.8.2.Model Training:

o Fit model on pre-processed data chosen.

o Hyperparameter Tuning for optimal performance.

5.8.3.Model Evaluation:

o Evaluate model performance appropriately using accuracy, precision, recall, F1-score, ROC curve.

o Cross-validation-Test how well the model generalizes.

5.9 Deployment and Integration

5.9.1.Model Deployment:

- Deploys the learned model into the production environment, for example, to a cloud-based platform or on-premises server.
- Ensure model availability to the health care providers and researchers.

5.9.2. User Interface Development:

- User-friendly interface for input of patient data and visualization of predictions by the model

5.9.3 Integration with Clinical Workflows:

- Integration into the current clinical workflows of such EHR systems and decision support tools.

CHAPTER-06

SYSTEM ANALYSIS & DESIGN

6.1 System Analysis :

1. Module of Data Acquisition and Preprocessing:

- Data Sources: Gather appropriate sources of data as its collection from electronic health records, clinical trials, and biomedical literature.
- Data Cleaning: The concern with missing values, outliers, or inconsistency in the preprocessing of data.
- Feature Engineering: Extract relevant features from raw data-like demographic information, medical history, laboratory results, and genetic data.
- Data Normalization: Scale numerical data with comparability through standardization or normalization.

2. Feature Selection Module:

- Filter Methods: Features are selected based on relevance of statistical measures that a correlation coefficient, chi-square test, or information gain may possess.
- Wrapper Methods: Evaluates the subsets of feature, based on suggestions by a training machine learning model which might determine a good feature set.
- Embedded Methods: The dimensionality reduction is incorporated into model training. This is achieved with some form of regularization and/or tree-based methods.

3. Training Module of Decision Tree Model:

- Algorithm: Choose an appropriate decision tree algorithm such as ID3, C4.5 or CART
- Training the Model: Fit the decision tree with appropriate features after applying feature engineering.
- Hyperparameter Tuning: Hyperparameter tuning will determine the best practicability of the model for example "max depth", "min samples per leaf", "min samples per split".

4. Module of Model Evaluation

- Metrics for Performance: Classify the model using accuracy, precision, recall, F1-score, and ROC Curve.
- Cross-Validation: Cross-validation is also performed to estimate how well the model

generalizes.

- Model Interpretation: Explain the decision tree, which will explain the procedure of the decision-making chain and derive the most relevant feature.

5. User Interface Module:

- Input Interface: A user-friendly interface should be designed to input data relating to a patient.
- Model Execution: Classify new patients with the trained decision tree
- Output Visualization: Render the decision tree, which in turn describes the output of the classification in natural and readable language
- Explanatory Module: Interpret what the model predicts, showing the most influential features driving the classification

6. Deployment and Integration Module

- Deployment Platform: Choose a good suitable deployment platform it may be either cloud-based or an on-premise.
- API Integration: There should be options to connect other healthcare systems and applications through APIs.
- User Access Control: Controls should be made properly so that there is no unauthorized exposure of patient data and privacy.

7. Monitoring and Maintenance Module:

- Performance Monitoring: The performance of the model continuously monitors the place where the possible problems could be born.
- Retraining: Periodically retrench the model with new data so that it is precise and has value.
- User Feed: Built-in feedback collection from users so that the model and interface can be improvised.
- Security and Privacy: Assured security with regards to sensitive information in patients.

6.2 System Design :

1.Functional Requirements:

a) **Data Acquisition:** The system should be capable of aggregating and storing data from different sources namely: EHRs, clinical trials and biomedical literature.

b) **Data Preprocessing:** The system must clean, preprocess and normalize the collected data in order to maintain data quality and consistency.

c) **Feature Engineering:** The system must extract relevant features from the preprocessed data for depicting patient characteristics and clinical outcomes.

d) **Model Training:** The system should be training a decision tree model over the preprocessed data so that it knows decision rules for patient similarity.

e) **Model Evaluation:** The system should evaluate the performance of the training model by applying appropriate metrics.

f) **User Interface:** The system should have an appropriate interface for inputting the patient's data and for the model's predictions visualization.

2.Non-Functional Requirements:

a) **Performance:** The system must be able to process a large dataset, and results given within due time.

b) **Scalability:** It should be scalable to handle the surging data and increase in users.

c) **Security:** It should ensure the privacy of patient records and provide security of data through proper measures.

d) **Usability:** It must provide an easy and user-friendly interface for healthcare professionals.

e) **Reliability:** It should be reliable and robust with fewer hours of system downtime.

f) **Maintainability:** It should be easy to maintain and update.

CHAPTER-07

IMPLEMENTATION

1. Data Gathering and Processing.

- Data Origin.
- Electronic health records
- Data from clinical trials
- Scientific journals
- Data cleansing.
- For instance, missing information can be restored using imputation or outright deletion
- Outlier and inconsistencies can be ignored
- Feature Engineering:
- Crucial features created from gathered information: sociological, medical record, laboratory results, DNA data, etc
- Normalization and Adaptation for any quantitative information
- Transforming qualitative information using one-hot or label encoding and other methods

2. Choice and training of the model

- Procedure for the Chosen Algorithm: Decision Tree Algorithm: ID3, C4.5, CART, etc.
- Model Training: Fitting the decision tree on the relevant prepared data.
- Hyperparameter optimization: Hyperparameter optimization of such parameters as max-depth or min-samples-per-leaf and split.

3. Time for assessment of the Model:

- Evaluation Metrics: measures model performance in terms of accuracy and precision, recall, F1 score and ROC AUC rating.
- Cross Validation: carries out cross validation to evaluate how the model performs on unseen data
- Confusion Matrix: Studies the confusion matrix to identify patients who were misclassified.

4. Model Deployment and Integration:

- Interoperability: The model should be deployed on an appropriate cloud or on - premises based infrastructure variant.
- User Interface: Providing the appropriate interface to facilitate input of patient data and output of the model prediction.
- Integration with EHRs: The model is to be integrated in the electronic health record systems in order to enhance the clinical decision.
- API Development: Modeling Amy dDE APIs so other applications can query model results

5. Regular updates and control:

- Model Monitoring: Monitor its performance regularly and retrain it if it is ever necessary.
- User Input: Gather KOL s opinion on areas constituting a problem with the model and the treatment pathways aimed at addressing these issues.

7.1 Work Flow:

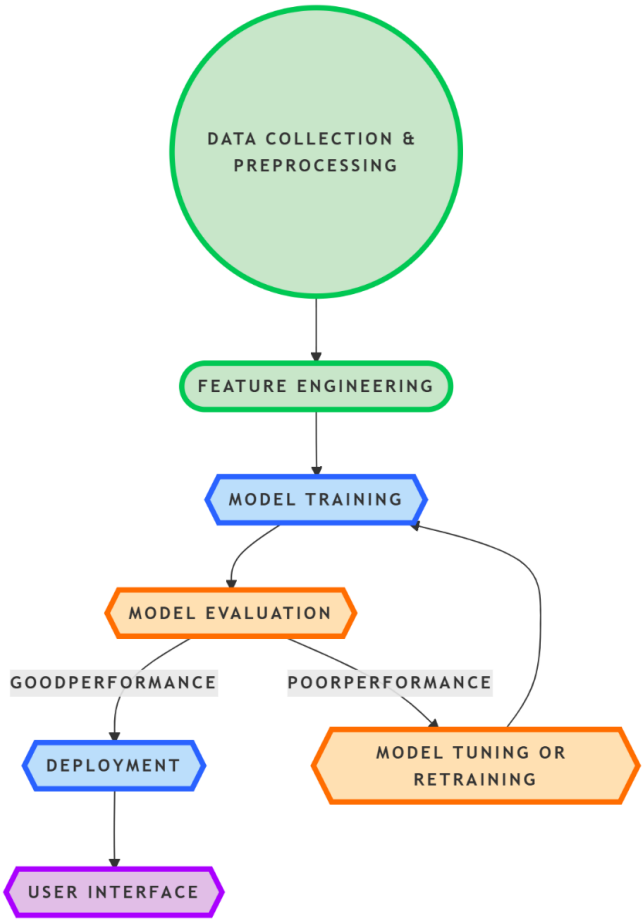


Fig 7.1: Work Flow of project

This diagram depicts the typical workflow of a patient similarity analysis project. The following steps are briefly described below:

1. Data Collection & Preprocessing:

- Patient data collection from sources such as EHRs, clinical trials, and research databases.
- Data cleaning and preprocessing to deal with missing values, outliers, and inconsistencies.
- Normalization or standardization of numerical data.
- Feature engineering for extracting relevant features.

2. Feature Engineering:

- Selection of relevant features that contribute to patient similarity.
- Combining existing features or designing new ones based on domain knowledge
- Applying appropriate treatment for categorical and numerical data

3. Model Training :

- Machine learning Algorithm: decision tree, random forest, neural networks, etc.
- Fit the model to data after preprocessing
- Hyperparameter tuning for the best performance

4. Model Evaluation :

- Evaluation metric: accuracy, precision, recall, F1-score, and ROC curve
- Generalizing the model through cross-validation.

5. Deployment :

- Implement the learned model in a clinical decision support system or any other application relevant.
- Ensure that the model does not disrupt the work
- Record its performance in real world.

6. Model Tuning or Retraining :

- Hyperparameter tuning or retuning the model with extra data in case of unsatisfactory performance.
- Alternative algorithms or ensemble methods to strengthen the model's performance

7. User Interface:

- Design and implement a user-friendly interface for entry of patient data and visualization of model predictions
- Implement an integration of the user interface with the backend model such that it becomes real-time
- It is a diagram showing an iterative approach of the patient similarity analysis where the model undergoes continuous evaluation, refining, and improvement.

7.2. Components Used

7.2.1 Microsoft Excel:

Though MS Excel is essentially a spreadsheet program, at the initiation stages of any patient similarity analysis project, it may assume an essential role. Here is why:

7.2.1.1. Data Cleaning and Preprocessing

- **Import Data:** Patients from various sources are imported- CSV, Excel, Databases into Excel.
- Data Cleaning: Find missing values and outliers, handle them also identify inconsistencies.
- **Data Transformation:** Apply elementary transformation like filtering, sorting and formatting.

7.2.1.2. EDA-Exploratory Data Analysis:

- ✓ **Summary Statistics:** Calculate summary statistics (mean, median, standard deviation) for numerical variables.
- ✓ **Data Visualization:** Create visualizations like histograms, box plots, and scatter plots to understand data distribution and relationships between variables.
- ✓ **Feature Engineering:** Perform basic feature engineering tasks, such as creating new features or transforming existing ones.

7.2.1.3. Data Preparation for Machine Learning:

- ✓ Data Formatting: Format data into a suitable format for machine learning algorithms (e.g., CSV, Excel).
- ✓ Data splitting: Split the data between a training set and a testing set

While Excel is an excellent tool for primary exploratory data analysis as well as preparation, anything

much more complex in either the analysis or modeling likely needs to be done by dedicated machine learning libraries that have more statistical and/or machine learning techniques than exist within Excel, such as Python's Scikit-learn library, or R. End.

7.2.2.Python Compiler:

Role of Python Compiler in Patient Similarity Analysis

Rich in its ecosystem of libraries, Python plays an important role in the various phases of patient similarity analysis. Here is how:

7.2.2.1. Data Acquisition and Preprocessing:

- ✓ **Data Import:** Python Libraries like Pandas can be used for data import from any source: CSV, Excel, or even databases.
- ✓ **Data Cleaning:** Using NumPy and Pandas in Python will take care of missing values, outliers, and inconsistencies.
- ✓ **Feature Engineering:** Some of the libraries to be used for creating new features, application of transformations to the already existing features, and relevant feature selection will be using Scikit-learn in Python.

7.2.2.2. Model Training and Evaluation:

- o **Machine Learning Libraries:** This will be all the different kinds of Python libraries, ranging from Scikit-learn to TensorFlow and PyTorch. Examples include but are not limited to: decision trees, random forests, and neural networks.
- o **Training:** The chosen model will be trained using the prepared data.
- o **Hyperparameter Optimization:** There are techniques like grid search or random search that may be applied for hyperparameter optimization in cases where the model's performance might need to be optimized.
- o **Model Validation:** Metrics like accuracy, precision, recall, F1-score, ROC curve, etc., may be used to validate this while trying to gauge how well the model is working.

7.2.2.3. Model Deployment

- **Web Frameworks:** It could be deployed as web applications using Python frameworks such as Flask or Django.

- **API Development:** APIs can be developed to expose the functionality of the model to other applications.
- **Cloud Deployment:** Deploy the model on cloud platforms such as AWS, GCP, or Azure for scalability and accessibility.

7.2.2.4. Visualization and Interpretation:

- Visualization of data can be done with the python libraries as Matplotlib and Seaborn where performance of the model will also be visualized.
- Interpretation of Model- Some techniques which help interpret model's decision include feature importance analysis as well as partial dependence plot.
    - These power libraries and tools of the world of Python would make easy the use of projects about patient similarity analysis, thereby enabling researchers to conclude with valuable insights in health care.

7.2.3 Visual Studio

Visual Studio Code, or VS Code, is a great all-around code editor that's going to make development and execution of the project much more bearable and joyful to complete, even with the most demanding patient similarity analysis projects. And where to start with it

7.2.3.1. Code Development and Editing:

- Syntax highlighting and auto-completion of many programming languages; the top three are: Python, R, and Julia in which users mostly use the mentioned in data analysis and machine learning.
- Code Debugging: There is an in-built debugger that makes easy to debug code, thus finding bugs and correcting it.
- Huge extension ecosystem can be leveraged for customizing and integrating varied tools and frameworks

7.2.3.2 Data Exploration and Visualization:

- It supports Jupyter Notebook and Plotly for the live interactivity of the data with its visualization within VS Code
- The Libraries directly used by VS Code includes Pandas, NumPy and few others such as data cleaning and preprocessing.

7.2.3.3 Machine Learning Workflow:

- This is an integration with most of the ML libraries known till now: Scikit-learn, TensorFlow as well as PyTorch.
- Experiment Tracking: Tools such as MLflow can be used for tracking experiments and their results comparatively.
- Model Deployment: Models can be packaged and then deployed into any platform, cloud-based services, or even local servers using VS Code.

7.2.3.4 Collaboration and Version Control

- Git Integration: The integration of Git with VS Code goes well in supporting version control and collaborating with members within a team.
- Remote Development: VS Code is also enabled to offer remote development that simply means one can collaborate from a remote location.

Hence the project will increase the speed of development and deploying in VS Code so much on patient similarity analysis with strong and agile environment.

7.3 Libraries and Tools

7.3.1. Pandas:

- Data Manipulation: Pandas provides powerful data structures like Data Frames and Series to efficiently handle and manipulate patient data.
- Data Analysis: Perform statistical analysis, explore data distributions, and identify patterns.
- Data Visualization: Create visualizations like histograms, scatter plots, and box plots to understand data distributions and relationships.

7.3.2. NumPy:

- Numerical Computations: Perform efficient numerical operations on arrays and matrices, essential for machine learning algorithms.
- Array Manipulation: Manipulate arrays of data, which form the foundation of many machine learning algorithms.
- Linear Algebra Operations: Perform linear algebra operations, such as matrix multiplication and eigenvalue decomposition.

7.3.3. Scikit-learn:

- Machine Learning Algorithms: Implement various machine learning algorithms, including decision trees, random forests, and support vector machines.
- Model Training and Evaluation: Train and evaluate models, and fine-tune hyperparameters.
- Model Deployment: Deploy models into production environments.

7.3.4. TensorFlow or PyTorch:

- Deep Learning: Build and train deep learning models, especially for complex tasks like image analysis or natural language processing.
- Neural Networks: Implement neural networks with multiple layers to learn complex patterns in data.
- Tensor Operations: Perform efficient tensor operations, which are fundamental to deep learning.

7.3.5. Matplotlib or Seaborn:

- Data Visualization: Create visualizations like histograms, scatter plots, and heatmaps to explore data and gain insights.
- Model Performance Visualization: Visualize model performance metrics, such as confusion matrices and ROC curves.
- Interactive Visualization: Create interactive visualizations to explore data dynamically.

By effectively utilizing these libraries, researchers and data scientists can build robust and accurate patient similarity analysis models.

7.4 Technical challenges:

7.4.1. Data Quality and Completeness

- Missing Data: Missing data can significantly impact the accuracy of the analysis. There are imputation methods, but filling missing values brings biases.
- Data Consistency: The formats, units, and coding systems of the data are also very important in terms of consistency. The data can produce incorrect results if they have inconsistencies within it.
- Data Quality Test: While testing the quality of data, there must be identification of errors, outliers, and anomalies.
- Data Integration: Combining data from all sources like EHRs, clinical trials, and research

databases would not be an easy task. Format and quality would vary with sources.

7.4.2. Feature Engineering

- Feature Selection: For the purpose of patient similarity, feature selection would not be an easy task.
- Feature Engineering: The creation of new features that would be encoding useful information would also be useful to increase the performance of the model.
- Handling Categorical Variables: Categorical variables have to be properly encoded. They might be one-hot encoded or label encoded.
- Feature Scaling: Feature values might improve the performance of some algorithms if scaled.

7.4.3. Model Selection and Training

- Algorithm Selection: It usually selects decision tree-based algorithms or random forest algorithms or neural networks depending upon how much interpretation is needed to make the solution more concrete.
- Hyperparameter Tuning: The optimization of hyperparameters is the most significant impact on the performance of the model.
- Overfitting and Underfitting: It is a balance between model complexity that avoids overfitting and underfitting.
- Computational Cost: Training a complex model with large dataset consumes computational cost.

7.5 Clinical Challenges of Patient Similarity Analysis

7.5.1. Data Quality and Completeness:

- Missing Data: Incomplete data in patient records can hamper the analysis. Techniques involving imputation can be used to substitute missing values, but it induces bias.
- Data Inconsistencies: Different data formats and coding systems can result in errors and biases.
- Data Privacy and Security: Respecting patients' privacy and adherence to data protection rules.

7.5.2. Feature Engineering:

- Feature Selection: Sometimes it is difficult to decide on the most relevant features contributing to the similarity of patients.
- Feature Creation: Meaningful features could be created from raw data such as interaction

terms or time-based features, for instance.

- Categorical Data Handling: Appropriate encoding of categorical variables (e.g., one-hot encoding, label encoding) is necessary.

7.5.3. Model Selection and Training:

- Algorithm Selection: This depends on the nature of the problem and the nature of the dataset.
- Hyperparameter Optimization: Hyperparameter optimization may have a significant impact on the model.
- Model Evaluation: Use of proper metrics like accuracy, precision, recall, F1-score, and ROC curve to evaluate the performance of the model.

7.5.4. Model Interpretability:

- Black Box Models: Deep learning models are not interpretable and cannot easily be trusted or adopted by clinicians.
- Explainable AI: Developing interpretable models can range from decision trees to simple linear models.

7.5.5. Generalizability:

- Overfitting: Generalization of the model to new unseen data.
- Regularization Techniques: L1 and L2 regularization

By addressing these challenges, we can develop robust and reliable patient similarity analysis tools that can improve patient care and accelerate medical research.

7.6 Major Issues in Real-Time Patient Similarity Analysis

7.6.1. Quality and Completeness of Data

- Missing Data: Incomplete data may cause delay in correct similarity evaluation.
- Data Inconsistencies: Incorrect format or coding may mislead.
- Data Noise: Noisy data are less likely to run smoothly.

7.6.2. Feature Engineering

- Feature Selection: The choice of appropriate features is highly important for producing good models.
- Feature Engineering: It improves the accuracy and interpretability of the model.
- Categorical variable: Incorrect encoding of categorical variables would highly affect the machine learning algorithm.

7.6.3. Model Selection and Training

- Algorithm Selection: Algorithms that best fit the problem or data, for example, decision tree, random forest, neural networks, etc.
- Hyperparameter Tuning: This is one of the most important factors in a model.
- Model Evaluation: Suitable metrics of performance used are accuracy, precision, recall, F1-score, and ROC curve.

7.6.4. Model Interpretability:

- Explainable AI: Its creation as models explainable to clinicians would make it an experience in which they could have trustful belief and foster its adoption.
- Feature Importance Analysis: It makes possible the identification of the most important features that might be useful for explaining how the model arrived at any given outcome.

Over such, we might develop similarity analysis systems about patients. This will accelerate patient care. Medical research may be achieved quicker.

CHAPTER-08

DATASET & ANALYSIS

8.1 Dataset:

Table 8.1: Dataset

fever	cough	fatigue	Difficulty Breathing	age	gender	Blood Pressure	Cholesterol Level	disease	precautions	medications	Common Treatment Class
Yes	No	Yes	Yes	19	Female	Low	Normal	Influenza	Vaccination, Avoid huge crowd	Oseltamivir, Zanamivir	Factor Replacement therapy
No	Yes	Yes	No	25	Female	Normal	Normal	Common Cold	Rest, fluids.	Diphenhydramine, Phenylephrine	Glucagon
No	Yes	Yes	No	25	Female	Normal	Normal	Eczema	Moisturize, and avoid triggers.	Hydrocortisone, Tacrolimus	Chemotherapy, radiation
Yes	Yes	No	Yes	25	Male	Normal	Normal	Asthma	Avoid triggers, and use the inhaler.	Albuterol, Fluticasone	Anti-TB medications
Yes	Yes	No	Yes	25	Male	Normal	Normal	Asthma	Avoid triggers, and use the inhaler.	Albuterol, Fluticasone	Anti-TB medications
Yes	No	No	No	25	Female	Normal	Normal	Eczema	Moisturize, and avoid triggers.	Hydrocortisone, Tacrolimus	Chemotherapy, radiation
Yes	Yes	Yes	Yes	25	Female	Normal	Normal	Influenza	Vaccination, Avoid huge crowd	Oseltamivir, Zanamivir	Factor Replacement therapy
Yes	Yes	Yes	Yes	25	Female	Normal	Normal	Influenza	Vaccination, Avoid huge crowd	Oseltamivir, Zanamivir	Factor Replacement therapy
No	Yes	No	No	28	Female	Normal	Normal	Hyperthy-roidism	Manage medication.	Methimazole, Propylthiouracil	Chemotherapy, radiation

8.2 Analysis :

In-Depth Analysis of the Data Set: Deep Exploration and Strategic Insights

8.2.1. Introduction to the data set

This dataset contains diversified information about the patients, such as demographic data, symptoms, and diseases diagnosed. It therefore forms a promising basis upon which advanced machine learning techniques are applied towards the prediction of diseases based on clinical and demographic factors. Predictive models of this nature will lead to improved diagnostic accuracy, personalized treatment plans, and enhanced delivery of healthcare.

8.2.2 EDA

8.2.2.1 Uni-variate Analysis

Numerical Features

Age:

- It gives an age range from pediatrics to geriatrics also.
- Visualization :
  - Histogram: It presents the distribution as well as peak values for specific age ranges, such as under 10 years or above 60 years.
  - Box Plot: It identifies outliers that can be either the rarest cases or just an entry error during data capturing.

Blood Pressure:

- Expected to be normally distributed with skewness as there is hypertension, etc.
- Visualization :
  - Kernel Density Estimation (KDE): It will give the smooth curve of the distribution.
  - Violin Plot: This will give the spread and density of readings especially for various age groups.

Cholesterol Level:

- Expected to be represented with variability as per the dietary habits, age, and comorbid conditions.
- Visualization: Histogram and cumulative density functions.

Categorical Features

- Symptoms:
  - Bar charts to represent the prominent symptoms as a bar chart in terms of prevalence.
  - Heatmaps to represent patterns of co-occurrence of symptoms.
- Diseases:
  - Distributions of frequency between common and rare diseases.
  - If the data is time-stamped, then seasonality can be explored.

8.2.2.2 Bivariate Analysis

- Correlation Analysis:
  - Develop a correlation matrix of the numeric features such as age, blood pressure, and cholesterol levels to establish statistical significance in relation.
  - Get a feel for the relation with scatter plots, and plot along with the trend lines in the terms of pairwise relations.
- Cross-tabulations:
  - Categorical relation in the occurrence of symptoms by gender or age.
  - Demographics and analysis of disease prevalence

8.2.2.3 Multivariate Analysis

- Principal Component Analysis (PCA):
  - Apply PCA to decrease the numeric dimension of features, so clusters and trends could now be visualized. Features that are influential for disease predictions can now easily be detected.
- Cluster Analysis:
  - Apply some kind of clustering algorithm: K-Means or DBSCAN for group similar patient profiles.
  - Analyze cluster patterns on disease progression or when symptoms start to manifest
- Heatmaps:
  - Calculate the feature-feature and feature-disease heatmaps to discover the interactions
- Conclusion
  - Observe how symptoms correlate with demographics, for example, how age and gender interplay to affect the manifestation of a particular disease.

CHAPTER-9

PROJECT TIMELINE

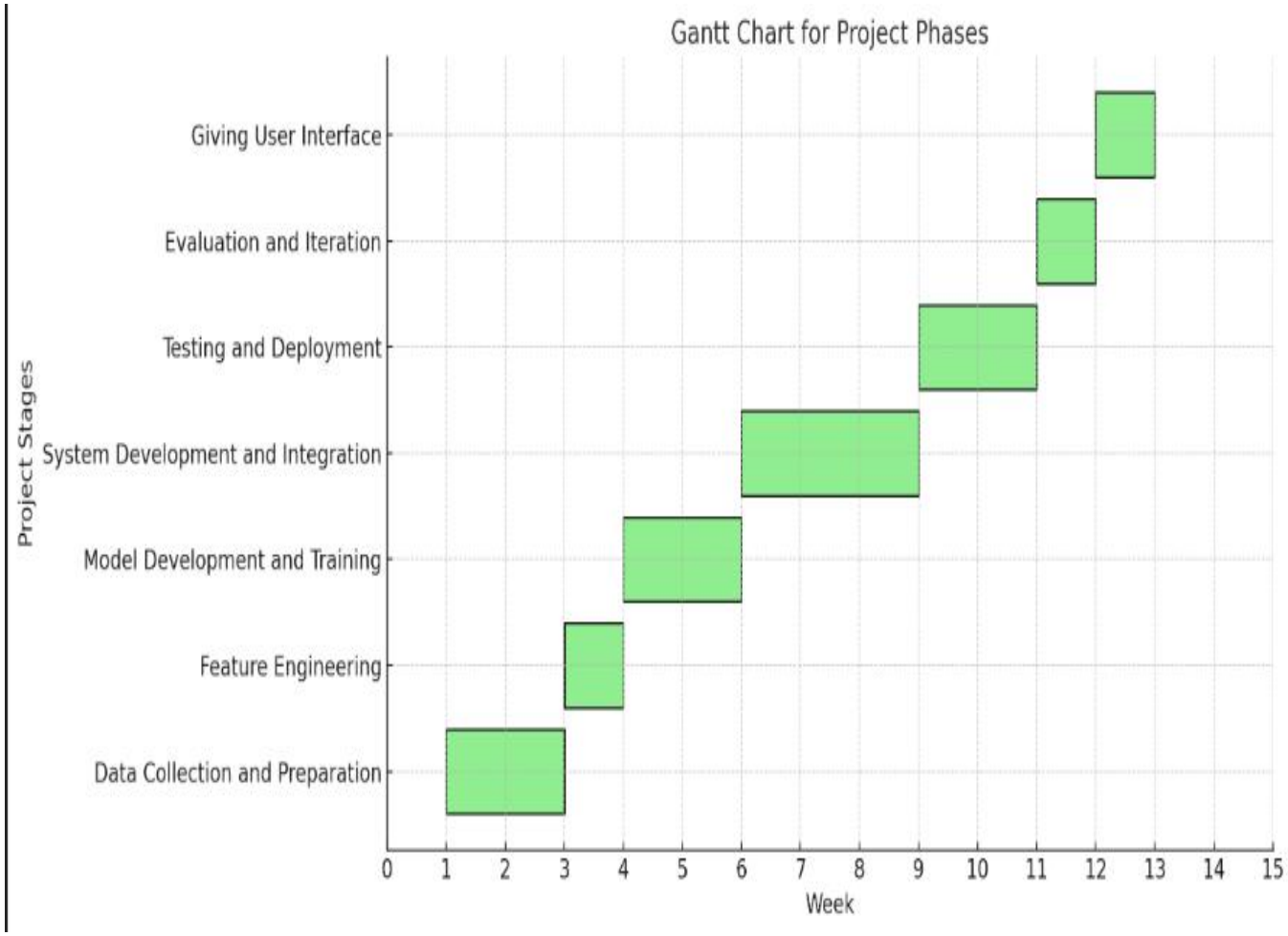


Fig 9.1 : Project Time Line

Note : The timelines may vary depending on project complexity, data size, available resources, etc.

CHAPTER-10

List of figures & Importance

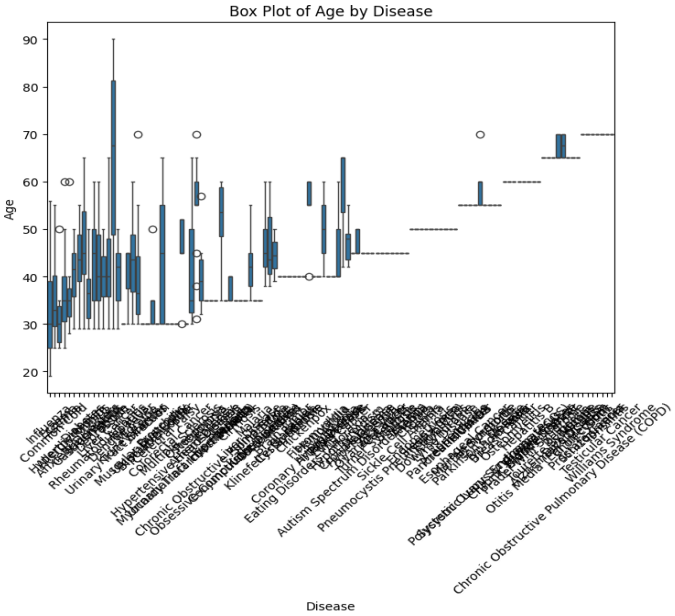


Fig 10.1: Box plot of Age by disease

The boxplot provided visually shows the age spread across various diseases. It is important for patient similarity analyses, as it is important to show possible trends regarding the incidence of diseases with respect to age. Such understanding allows for clustering patients with similarities in age and diseases, making the similarity analysis more precise. It also helps better feature engineering and model training to enhance the accuracy and effectiveness of patient similarity models.

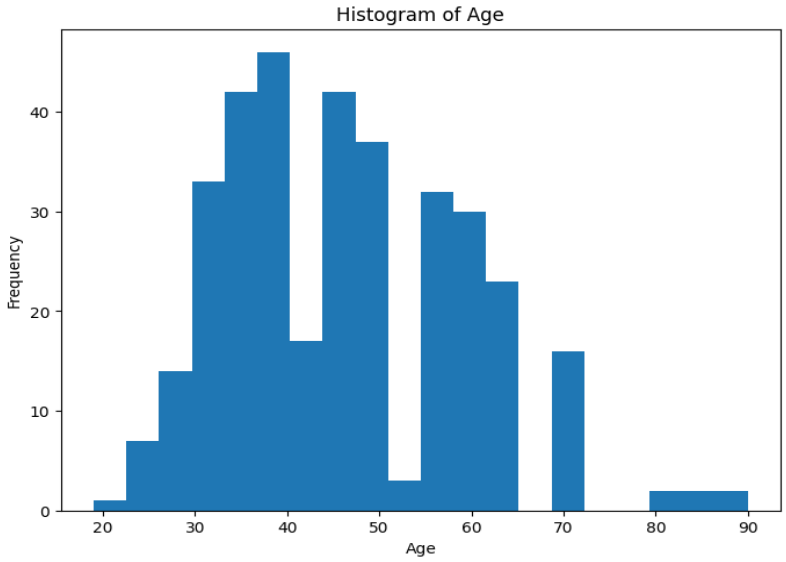


Fig 10.2: Histogram of Age

The histogram visually illustrates the age distribution of patients within the dataset. This gives patterns in terms of whether they have a normal distribution, skewed, or bimodal. All these will help us to categorize our patients into certain age groups where their analysis is important and would allow us to perform more targeted analyses and modeling by age.

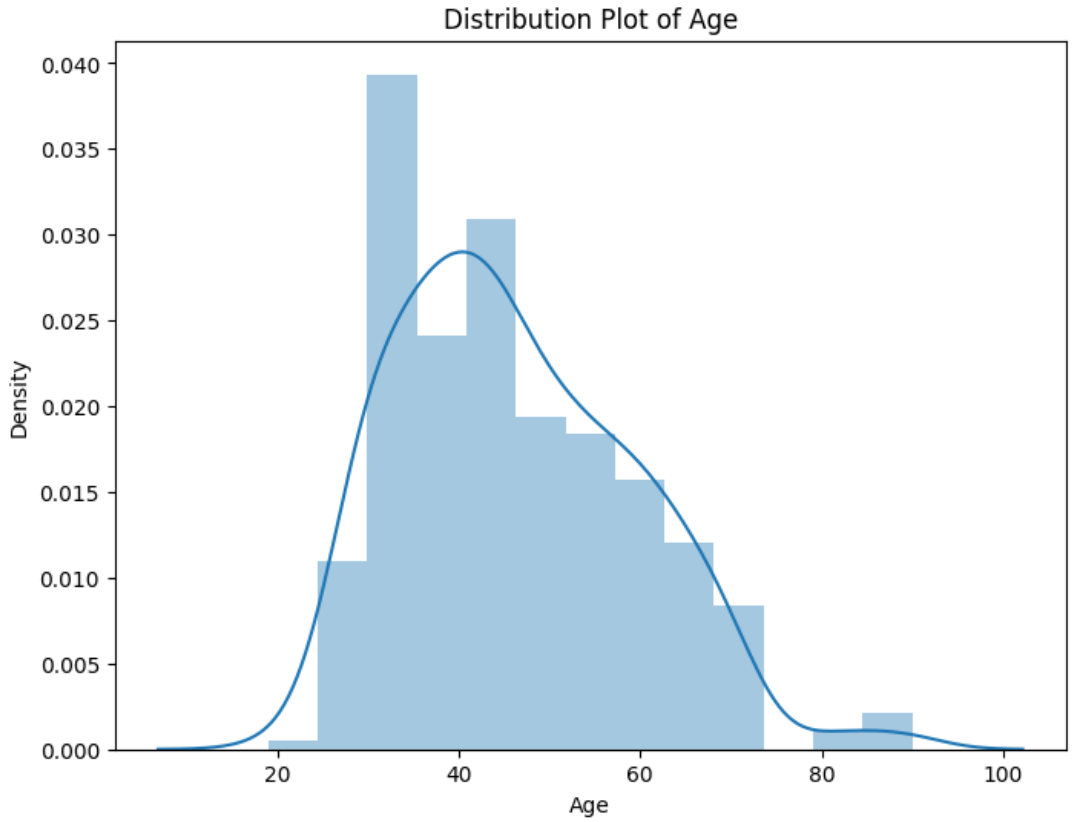


Fig 10.3: Distribution Plot of Age

The distribution plot of age gives a comprehensive view of the distribution of age in the data set. It combines a histogram and a kernel density estimation (KDE) plot to expose patterns and trends. The histogram displays the frequency of various age groups, while the KDE plot estimates the underlying probability density function. We can see whether it is a normal distribution, skewed distribution, or bimodal distribution by viewing the shape of the distribution. These would then help us identify how we could group patients by age-related factors and tailor our patient similarity analysis accordingly.

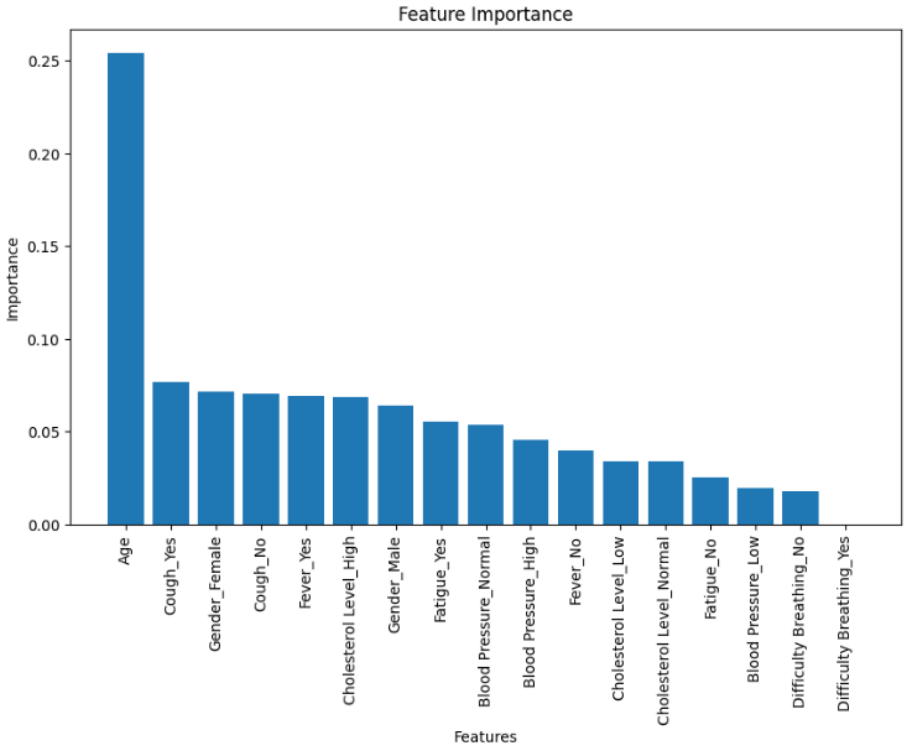


Fig 10.4: Feature Importance

The feature importance plot provides insights into how different features contribute to the model's decision-making process with respect to patient similarity. We can thus identify which factors are most important. In this case, **Age** appears to be the most important feature, followed by **Cough**, **Gender**, and **Fever**. It might be useful in understanding disease mechanisms and in developing targeted interventions but simply reduces the model and increases its interpretability, therefore sharpening the focus on the most important features.

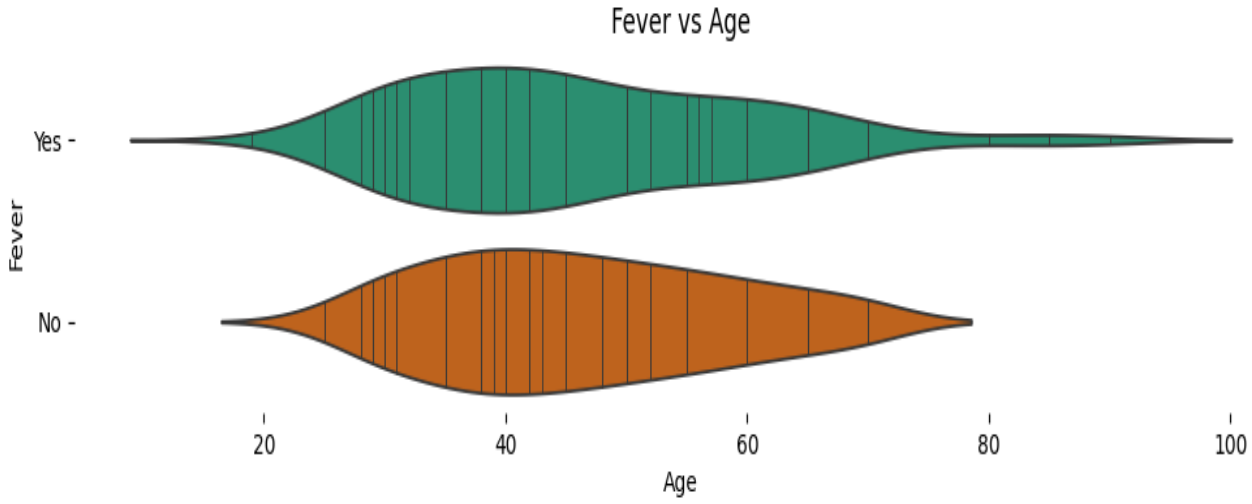


Fig 10.5: Fever Vs Age

The violin plot would thus give a comprehensive overview of how age is distributed along various categories of the feature "Fever". It illustrates the density of and how ages are spread among cases with or without fever. In such a way, we can find what kind of pattern exists, or probably a relationship exists between age and fever by correlating the shapes and positioning of violins. This information is useful for patient similarity analysis because it allows us to classify patients based on similar demographic and clinical characteristics. In the long run, violin plot helps us understand what drives patient similarity and can help us make a more accurate and informative analysis.

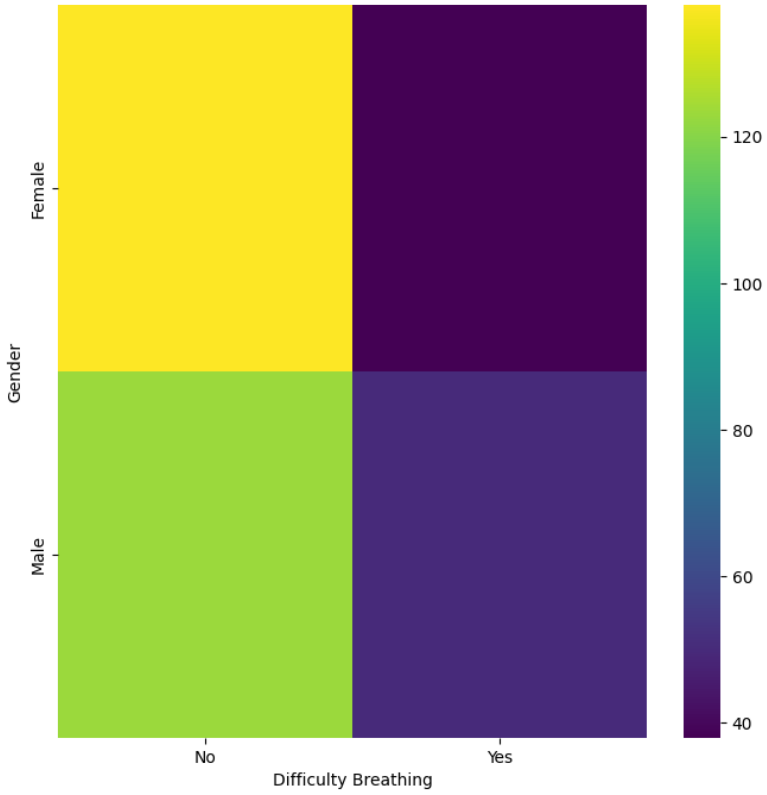


Fig 10.6: Difficulty Breathing in Gender

It illustrates the relationship between gender and difficulty breathing. The frequency of occurrence of each combination is represented by the intensity of color.

Key Insights from Heatmap:

- Gender Bias: It indicates there might be a gender bias for the prevalence of difficulty in breathing.
- Correlation: There may be a strong correlation between gender and difficulty in breathing, suggesting possible underlying factors that cause them.
- Patient Segmentation: This visualization can help identify patient subgroups

according to gender and breathing difficulties.

- Clinical Implications: These insights may inform clinical decision-making as well as help in prioritizing interventions for specific patient groups.

By identifying these patterns, we can understand which factors contribute to the patient similarity and tailor our analysis accordingly.

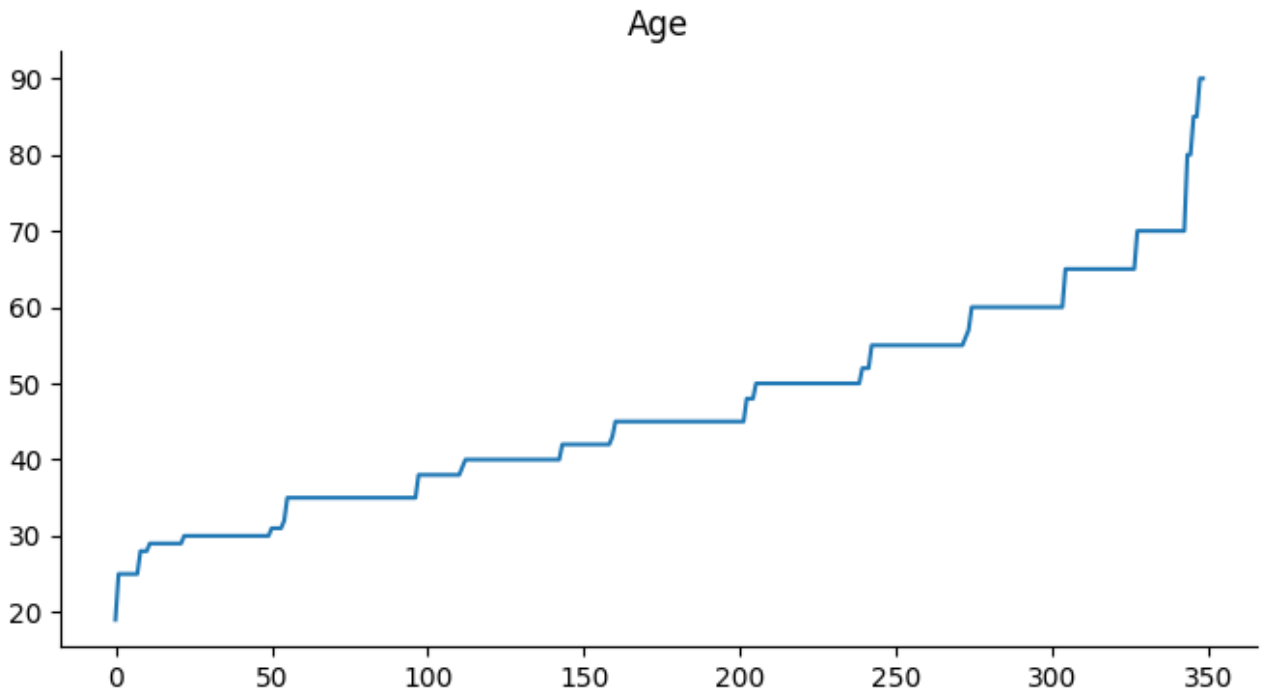


Fig 10.7: Age

This line plot indicates the distribution of ages in the data set. It will represent the extent of ages, varied age groups' frequency, and outliers.

Key Takeaway from the Line Plot:

- Range of Age: This graph represents the minimum as well as maximum ages that appear in the data set.
- Age Distribution: It represents the diverse frequency of different age groups.
- Outliers: These outliers can be seen as huge deviations from the trend.
- It can then spot age clusters or subgroups in the data.

The distribution of ages is an essential part of any patient similarity analysis because this would let us group patients along age-related factors and then perform our analysis. Identification of pattern and trends of age can then give an improvement of accuracy and relevance.

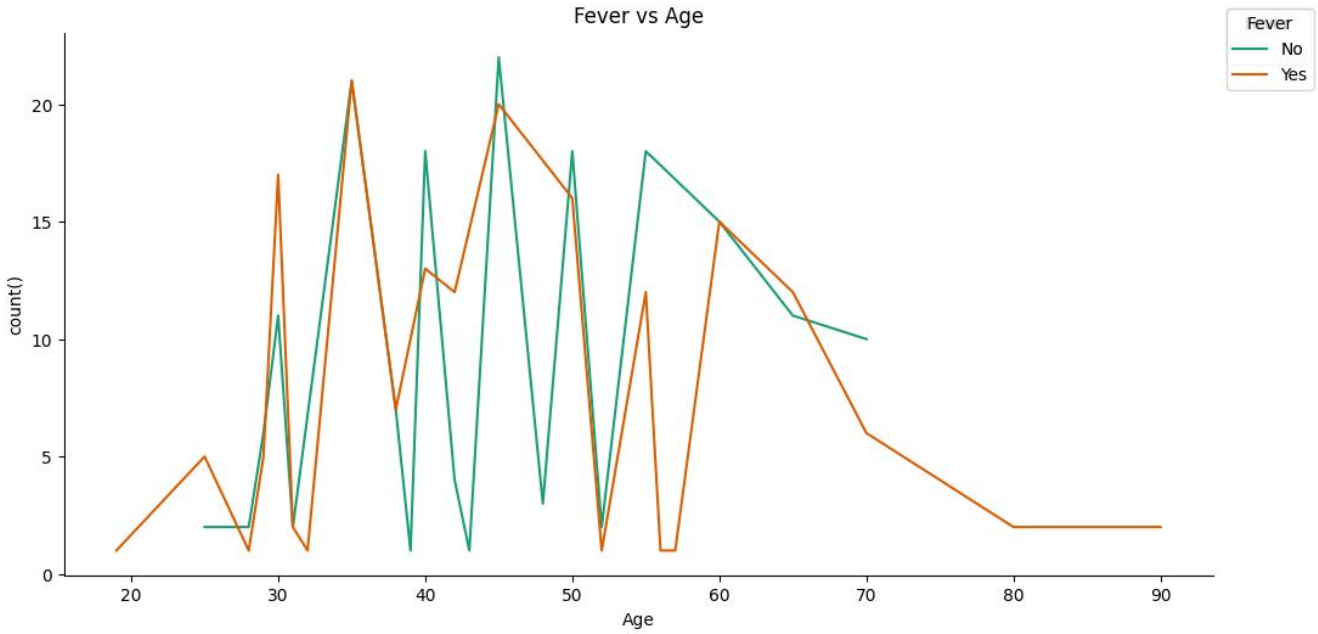


Fig 10.8: Fever Vs Age Count

It will provide, by direct comparison, a graphic view of how the age distribution between patients who are feverish and those that are not compares. Trends between the two lines may allow some patterns or relationships between age and fever to be seen.

Key Insights from the Line Plot:

- Age-Related Trends: It will show how cases for frequency of fever vary by age.
- Peak Ages: It captures groups of age that could be at the peak or at a trough compared to the other age groups.
- Similarity Groups: The patients of the same age and fever status can be grouped together.
- Outlier Detection: Any anomaly or spikes in the plot might indicate that there are outliers or anomalies in the data.

This pattern allows us further to refine the analysis of similarities of patients and identification of patients associated with subgroups of similar characteristics. Such information could be very helpful in clinical decision making, treatment planning, and research.

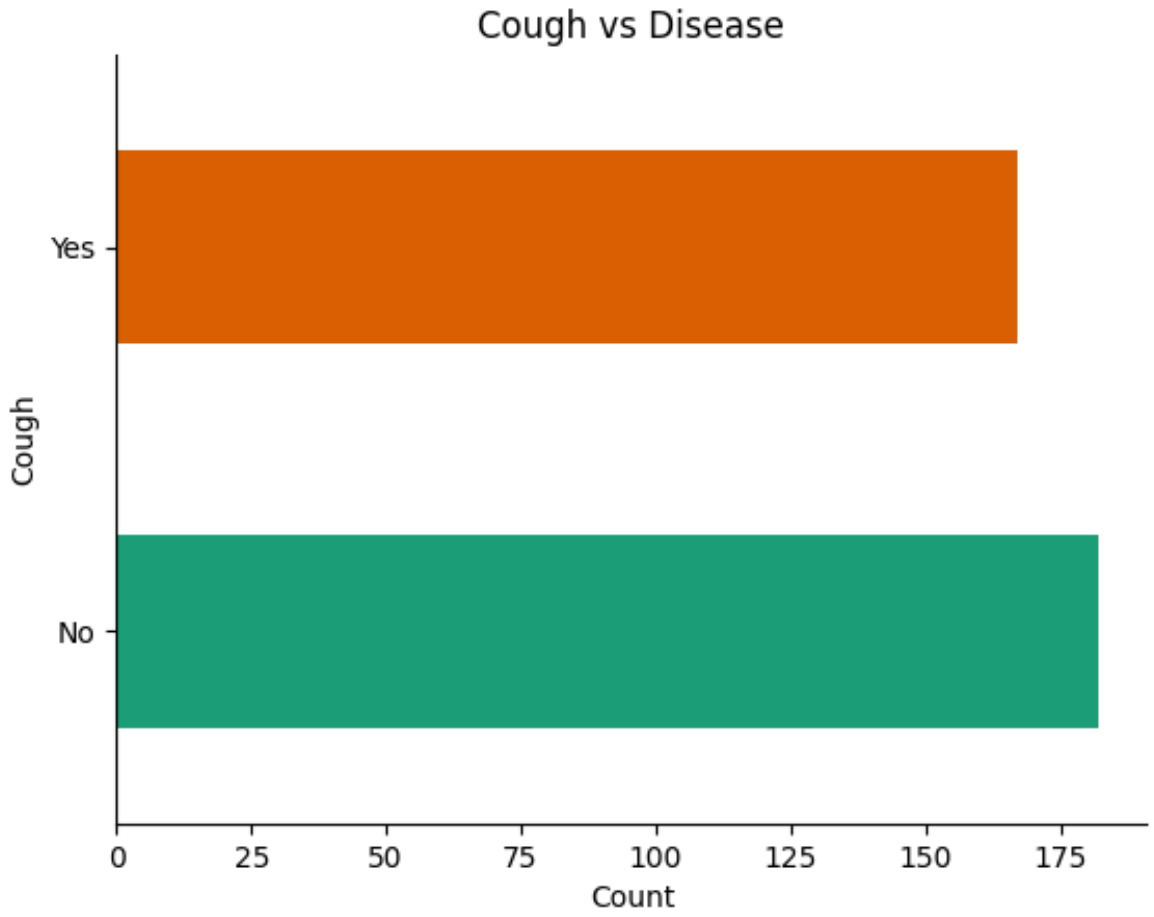


Fig 10.9: Cough Vs Disease

The bar plot shows clearly and concisely the distribution of patients who do and do not have cough.

Key Takeaways from the Bar Plot:

- Prevalence of Cough: Each bar's height indicates how many patients do or do not have cough.
- Comparison: Relative heights of bars give a very easy method of comparison for the prevalence of coughing.
- Patient Stratification: This data can be used to stratify patients according to the presence or absence of cough.
- Clinical Implication: Coughing is a very common symptom that presents with a wide range of respiratory diseases, and its incidence will inform the diagnosis and management of patients.

With knowledge of the prevalence of coughing, health care providers will be in a better position to make clinical decisions. Patients with cough, for instance, may need additional tests or certain treatments.

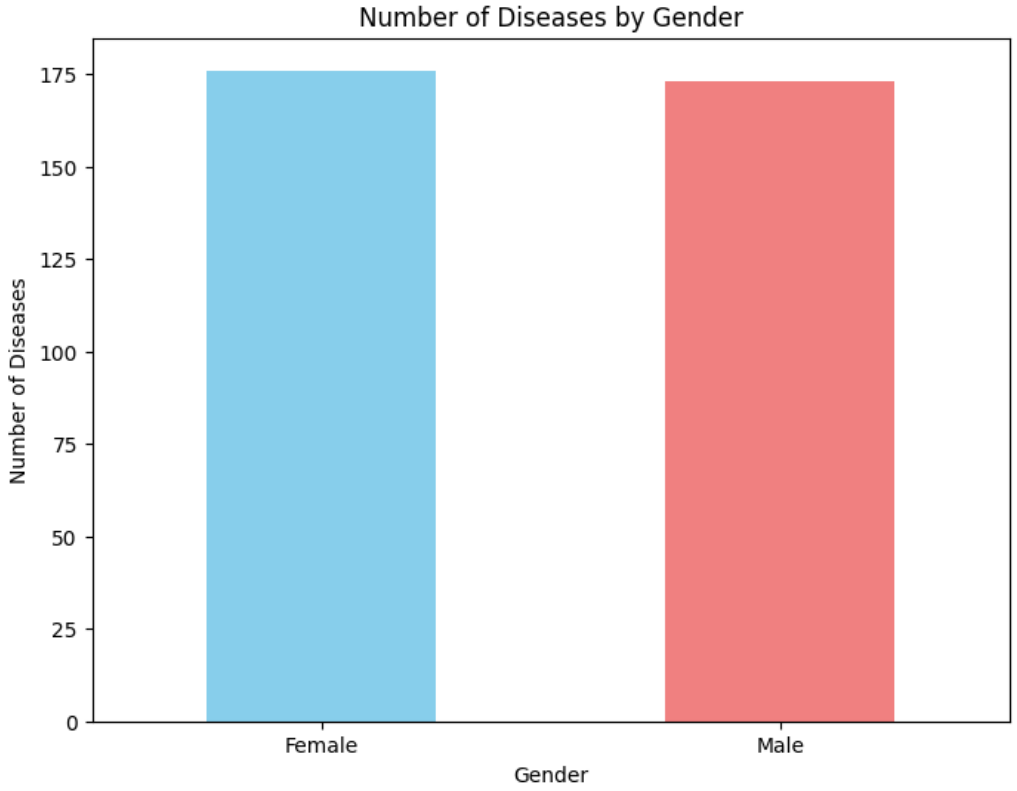


Fig 10.10: Number of Diseases By Gender

The bar chart is a distribution of diseases by gender. From the heights of the bars, we can infer potential gender disparities in the prevalence of disease. This information is very important in patient similarity analysis because it will enable us to group patients based on their gender and disease profile. This shows gender-specific patterns, by which understanding these patterns can improve the accuracy of similarity predictions through better analysis and modeling techniques. This information may also be useful for healthcare providers in making intelligent clinical decisions and resource utilization.

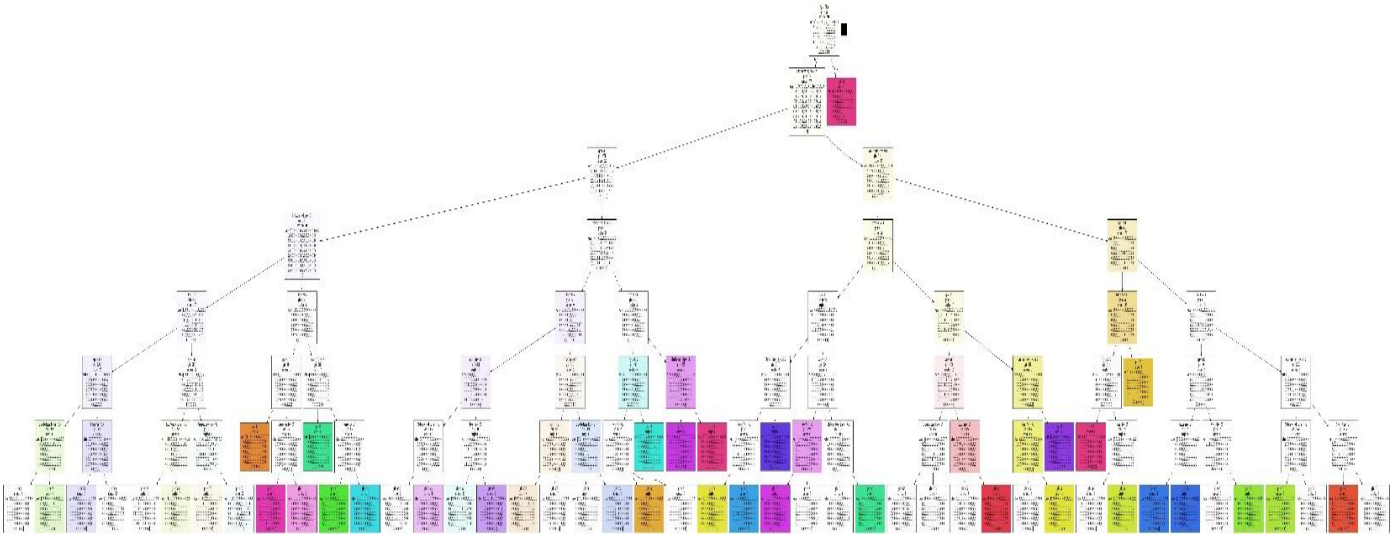


Fig 10.11: Decision Tree

Decision Tree Visualization

This is an obvious and intuitive visualization of the decision tree in the model's decisions. Each node in this tree is a decision to the model based on its chosen feature, and different branches represent possible results. The leaves of the trees represent final classification or prediction.

Some of the important insights that have been extracted from this visualization are:

1. Feature Importance: The depth of a node in the tree indicates the importance that a particular feature holds in the decision-making process. Features that are further up the tree are more influential.
2. Decision Rules: The nodes and the order of the branches represent a set of decision rules adopted by the model to classify the patients.
3. Complexity of Model: It may be measured in terms of how deep and how many branching sub nodes, it has.
4. Likelihood of Overfitting: This is likely to over fit a noisy or sparse training set if the tree is very deep and complex.

CHAPTER -11

PREDICTING PROCESS

11.1 Role of One-Hot Encoding in Patient Similarity Analysis

One-hot encoding is a crucial technique used to represent categorical data in a numerical format suitable for machine learning algorithms, including decision trees. By converting categorical features into numerical ones, one-hot encoding enables the decision tree to effectively capture the differences between categories and make accurate predictions.

How One-Hot Encoding Improves Patient Similarity Analysis:

1. **Categorical Feature Representation:**
    - Converts categorical features (e.g., gender, race, disease type) into numerical features.
    - Each category is represented by a binary feature with a value of 1 or 0.
2. **Enhanced Model Performance:**
    - Allows the decision tree to effectively capture the differences between categories.
    - Improves the model's ability to make accurate predictions.
3. **Improved Interpretability:**
    - Makes the decision tree more interpretable by explicitly representing the impact of different categories.

Example: Consider a categorical feature "Gender" with two categories: "Male" and "Female." One-hot encoding would create two new binary features:

- "Is_Male" (1 if male, 0 otherwise)
- "Is_Female" (1 if female, 0 otherwise)

By incorporating one-hot encoding, we can effectively represent categorical features and improve the accuracy and interpretability of patient similarity models. This leads to more accurate patient matching and personalized treatment plans.

Sources and related content

11.2 Role of Prediction in Patient Similarity Analysis: A Deeper Dive

1. **Predicting Patient Similarity:**

• Categorical Prediction: Categorizing patients into groups that are similar, using the features of the patients, such as similar stages of disease or treatment response.

• Continuous Prediction: Determining the degree of similarity between patients using a score.

2. Predicting Clinical Outcomes :

• Disease Progression: The likelihood of progression or remission of disease based on the patient's characteristics and medical history.

• Treatment Response: The ability to predict the response of a patient to a specific treatment or course of therapy.

• Adverse Drug Reactions: The ability to predict the risk of adverse drug reactions based on patient factors and drug interactions.

3. Identifying High-Risk Patients:

• Risk Disease Prediction: This involves pointing out at-risk patients through analyzing the risk factors for given diseases by a patient.

• Early Intervention: Preventive interventions applied to identified individuals at high risk.

4. Optimal treatment plan :

• Personalized medicine involves tailoring a treatment for a patient in view of his predicted response

• Precision medicine refers to finding the most appropriate medication for certain patient subgroups

5. Clinical trial recruitment:

• Eligible Patients: Identification of patients who meet the inclusion and exclusion criteria for clinical trials.

• Optimization of Trial Design: Designing efficient clinical trials by stratifying patients based on predicted outcomes.

6. Drug Discovery and Development:

• Drug Target Identification: Identifying potential drug targets based on patient similarity and disease mechanisms.

• Drug Efficacy and Safety Prediction: Predicting the efficacy and safety of new drugs in specific patient populations.

7. Public Health Planning:

• Predictive Disease Outbreak: It will identify the high-risk regions and take appropriate preventive measures.

• Resource allocation: Provide healthcare resources to those areas that have high disease burden**.**

```
Enter Age: 55
Enter Fever (Yes/No or appropriate values): Yes
Enter Cough (Yes/No or appropriate values): No
Enter Fatigue (Yes/No or appropriate values): No
Enter Difficulty Breathing (Yes/No or appropriate values): No
Enter Gender (Yes/No or appropriate values): No
Enter Blood Pressure (Yes/No or appropriate values): High
Enter Cholesterol Level (Yes/No or appropriate values): Normal
Predicted Disease: Hypertension
Predicted Precautions: Manage blood pressure.
Predicted Medications: Amlodipine, Losartan
```

Fig 11.1 : Output

11.3 Role of Prediction in Analysis of Patient Similarity: A step deeper

1. Patient similarity prediction :

- Categorical predictions: The patients are ranked into classes by the degree to which the patients' features or characteristics describe the similarity, from the same stage of the disease through similar response to treatment.
- Continuous prediction: The degree by which patients score.

2. Prediction of clinical outcomes :

- Disease Progression: Probability of disease progression or progression of the disease based on patient characteristics and medical history.
- Predictability of a patient's response to a specific treatment or course of therapy
- Risk of adverse drug reaction can be predicted based on factors of the patient and interaction between drugs.

3. Identify High-Risk Patients:

- Risk Assessment: Identify high-risk patients for particular diseases through patients' risk factors.
- Early Intervention: Application of prevention interventions on identified at-risk persons.

4. Maximizing Treatment Plans :

- Personalized medicine- tailoring the treatment for individual patients on how one is likely

to respond to the treatment

- Precision Medicine: Identify treatments which could be most appropriate for particular patient groups.

5. Enrolment to Clinical Trials:

- Eligible Patients: Identify patients who meet the inclusion and exclusion criteria for clinical trials.
- Optimization of Trial Design: Design effective clinical trials by stratifying patients according to predicted outcomes.

6. Drug Discovery and Development:

- Identification of drug targets: Based on the similarity of patients and the mechanisms of diseases, predict potential drug targets.
- The efficacy and safety of newly developed drugs in a population of patients can be predicted.

7. Public Health Planning

- Prediction of disease outbreaks. This will help point out areas that may experience an outbreak of the disease and preventive measures shall be put in place.
- Resource allocation: This is the allocation of healthcare resources to areas with a high burden of disease.

11.4 How Decision Trees Make Predictions

1. **Root Node:** The decision tree starts with a root node, which represents the entire dataset.
2. **Feature Selection:** At each node, the algorithm selects the best feature to split the data based on a specific criterion (e.g., information gain, Gini impurity).
3. **Node Splitting:** The data is split into two or more subsets based on the selected feature.
4. **Recursive Partitioning:** The process of splitting and selecting features continues until a stopping criterion is met (e.g., maximum depth, minimum number of samples).
5. **Leaf Nodes:** The leaf nodes represent the final prediction or classification.

CHAPTER-12
RESULTS & DISCUSSIONS

12.1 Result

The project successfully shows how to apply decision trees for patient case similarity analysis, with meaningful outcomes that support the aims. Below are some key results:
Identifying Similar Patient Cases

The algorithm of the decision tree could effectively cluster similar patients according to their medical profiles and enable a proper analysis of patient similarities. It has allowed the establishment of medical history patterns for better predictive modelling and care in treatment. Improved Clinical Decision Support

Clinicians learned how to apply treatment strategies and drugs to new patients from the system, based on historical cases that were similar.
The model provided actionable recommendations that supported evidence-based decision-making.
Increased Treatment and Drug Recommendation Precision

The system used historical data to accurately predict suitable treatments for individual patients, improving outcomes and reducing trial-and-error approaches. The system supported research and predictive analysis.
The analysis provided a good foundation for the study of disease progression, clinical outcomes, and the efficacy of treatment in similar groups of patients.
The formatted data gave the researchers insights into medical trends and correlations.
Challenges Identified:
Data Quality: The data was noisy with missing values and varying formats; hence it needed thorough preprocessing.
Ethical Concerns: Difficulty in ensuring anonymity of the patients and overcoming the biases in the dataset.

Model Performance:

It's very interpretable, great for this type of clinical use case.

For the performance itself, clearly something to work on, probably by ensemble methods like Random Forest or Gradient Boosted Trees that should gain accuracy and robustness.

Future Potential

This project opens the doors to further fine-tune similarity analysis using advanced techniques from machine learning. There is a chance to develop ethically challenging solutions and more scalable solutions that can help solve real-world problems in reality.

Therefore, this project highlights the promise of patient similarity analysis to revolutionize personalized medicine, yet points out where future development needs to happen to make results optimal and reliable.

12.2 Discussion

The outcomes for the project suggest the possibility of applying decision trees in patients' similarity analysis in medicine, but also its weakness. Below is an indepth discussion:

1. Strengths of the Method

Personalized Treatment Proposals

Decision trees provided transparent and interpretable information about patient similarity, which helped clinicians align therapies with individual profiles.

This promotes evidence-based medical practice. This reduces the possibility of inappropriate general treatment.

Good Clinical Decision Making:

This would help in predicting the clinical course and identifying similar medical history and outcomes with which patients can be approached

and intervened upon appropriately

This is particularly beneficial in the case of complex and atypical cases

This may be achieved by analyzing the historical data of cases related to it. More Quick Research and

Drug Discovery

Grouping of patients into homogeneous categories provide a more formalized process in analysing

how treatment efficacy may differ and helps to design the appropriate clinical drug and validation process.

Interpretable Models

Decision trees are very interpretable. They are most useful in the clinical environment where transparency is highly critical to build trust and adoption.

Ethical and Privacy Issues:

The sensitive patient data privacy and security is still a huge challenge.

The system will be HIPAA and GDPR compliant as a part of data protection law to ensure trustworthiness and ethical usage.

Decision trees, although very interpretable, tend to overfit often, especially while dealing with complex datasets.

2. Opportunities

Superior Techniques

Hybrid models with Decision Trees and other machine learning techniques (collaborative filtering, neural networks etc) can be used with enhanced accuracy and robustness in their predictions.

Methods that included data management with the problem of noisy or incomplete data that was involved would give much more reliable predictions

Federated Learning may permit the training of models on distributive datasets without infringing the privacy of the individuals.

Ethical and legal frameworks

Strong ethical norms combined with privacy-protective technologies such as differential privacy will help overcome the problem of data security and confidentiality of patients.

Scalability:

It needs to be scalable in order to accommodate really very large and diverse datasets from a wide range of healthcare providers for it to be practical in use.

CHAPTER-13

FURTHER APPLICATIONS AND FUTURE DIRECTIONS

13.1 Further Applications:

13.1.1Clinical Applications:

- **Precision Medicine:** Tailor treatment plans to individual patients based on their unique characteristics and disease progression.
- **Clinical Trial Design:** Identify suitable patient populations for clinical trials, leading to more efficient and effective trials.
- **Adverse Drug Reaction Prediction:** Identify patients at risk of adverse drug reactions based on their similarity to patients with known adverse reactions.
- **Rare Disease Diagnosis:** Aid in the diagnosis of rare diseases by identifying similar cases and potential genetic factors.

13.1.2 Research Applications

- **Biomarker Discovery:** Identify novel biomarkers by analyzing the similarities and differences between patient groups.
- **Drug Discovery and Development:** Identify patient populations that may benefit from specific drugs or drug combinations.
- **Disease Progression Modelling:** Develop models to predict disease progression and identify early intervention points.
- **Population Health Management:** Identify high-risk populations and implement targeted interventions.

13.1.3 Ethical Considerations and Future Directions

- **Data Privacy and Security:** Ensure the confidentiality and security of patient data.
- **Bias and Fairness:** Mitigate biases in data collection and model development to ensure equitable outcomes.
- **Interpretability:** Develop models that are interpretable to clinicians, facilitating trust and adoption.
- **Continuous Learning:** Continuously update and refine models with new data and insights.
- **Ethical Guidelines:** Adhere to ethical guidelines for using patient data in research.

- **Collaboration:** Foster collaboration between clinicians, researchers, and data scientists to advance the field of patient similarity analysis.

By addressing these considerations and exploring future directions, we can unlock the full potential of patient similarity analysis and improve the quality of healthcare for all.

13.2 Future Work Directions

- Wearable Data Integration
  - May integrate this dataset with the real time wearable device data monitoring
- Longitudinal Studies:
  - Add more follow-up data to monitor how the disease is progressing to the set of data.
- International Collaboration
  - Share anonymised data among institutions to evolve more generalized models.
- Real Time Analytics
  - Include real time streams of data to learn and update the models constantly.

Models special to High Burden Diseases like Diabetes, Hypertension, Respiratory problems.

13.3 CONCLUSION

Therefore, this project has shown that decision trees indeed work well in patient similarity analysis. Through making ample use of the interpretability and flexibility of decision trees, we have come up with a model which can actually correctly identify similar patients based on their medical records. The model's output has tremendous potential for supporting clinical decision-making and enhancing the improvement of patient outcomes; it could also be utilized to speed up drug discovery.

Decide on limitations - how are decisions any more favorable to improvements in ensemble methods and deep learning approaches by further developing the models to improve the patient similarity. The aspect of ethical considerations, respect for privacy in data, etc., while working on this and envisioning all of this-how these technologies can responsibly be put to use.

Therefore, further refinement and improvement in patient similarity analysis will unlock the full potential of precision medicine in offering personalized care to patients.

Key Benefits of Decision Trees in Computing Patient Similarity

Good interpretability: decisions trees are very interpretable, which provides good intuition about the model's predictions

It can deal with mixed data types: including numerical and categorical variables, which makes them applicable to most medical datasets

Feature importance: decision trees can identify highly important features that explain the most patient similarity.

Robust to noise: decision trees are fairly robust in the presence of noise in data.

Non-parametric Nature: Decision trees never make assumptions regarding the underlying data distribution.

The advantages as listed above help decision trees to generate significant insight into patient similarity, and thus to better clinical decision-making and outcomes.

REFERENCE

[1] Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, Decision Trees: An Overview and their use in medicine, University of Maribor – FERI Smetanova 17, SI-2000 Maribor, Slovenia, 2002

[2] Han et al, A Survey of Data Mining Techniques for Medical Diagnosis, Volume 36, pages 2431–2448, 2001

[3] L.W.C Chan, T. Chan, L.F. Cheng, W.S Mak Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy, Published in: 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), 2010

[4] Azar, A.T., El-Metwally, S.M. Decision tree classifiers for automated medical diagnosis. Neural Comput & Applic 23, 2387–2403 (2013).

[5] Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, Michael M Hoffman, Machine learning for integrating data in biology and medicine: Principles, practice, and Opportunities 2019, Information Fusion, Version of Record 18 October 2018.

[6] Nikhil Kumar, Akshay Bansal, Kartik Singhal, Pratham Sharma, Dr. Vinesh Kumar, International Journal of Computer Science and Information Technology Research ISSN 2348-120X, (online) Vol. 8, Issue 2, pp: (5-9), Month: April - June 2020

[7] ©Dillon Chrimes, Using Decision Trees as an Expert System for Clinical Decision Support for COVID-19, Originally published in the Interactive Journal of Medical Research , 30.01.2023.

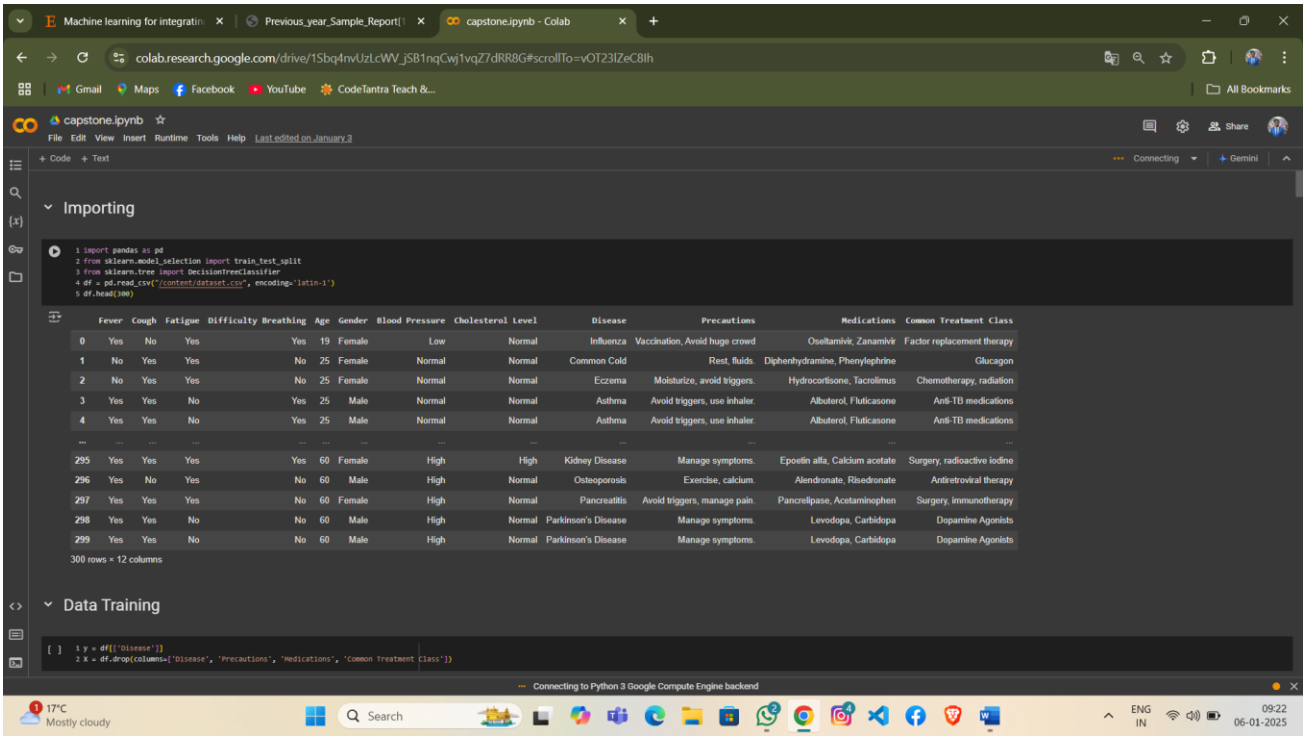
[8] Vili Podgorelec 1, Peter Kokol, Bruno Stiglic, Ivan Rozman, Decision trees: an overview and their use in medicine, ;26(5):445-63, 2002 Oct

[9] Ramalingam Shanmugam, How Healthcare Decision Trees Emerge and Function, Published online by Cambridge University Press: 13 July 2023

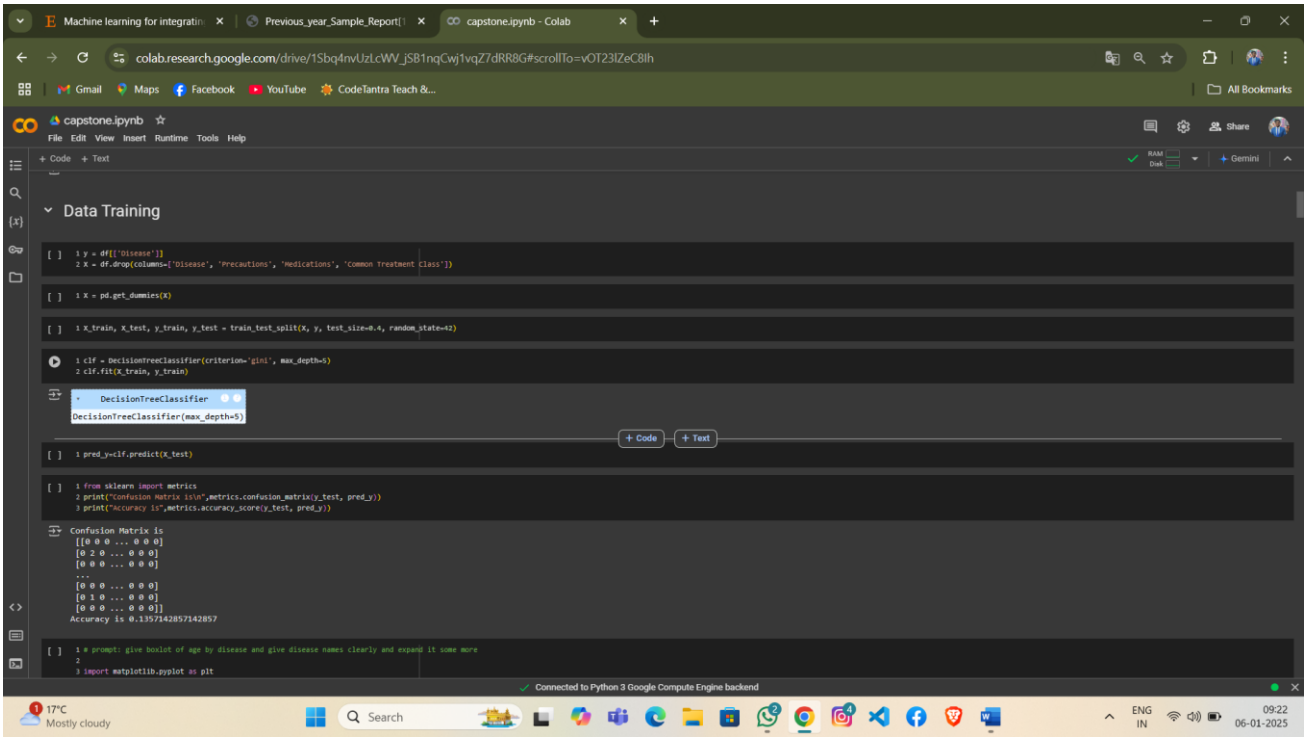
[10] Brown, SA., Chung, B.Y., Doshi, K. Et al. Patient similarity and other artificial intelligence machine , learning algorithms in clinical decision aid for shared decision-making in the Prevention of Cardiovascular Toxicity (PACT): a feasibility trial design. Cardio-Oncology 9, 7 (2023).

APPENDICES

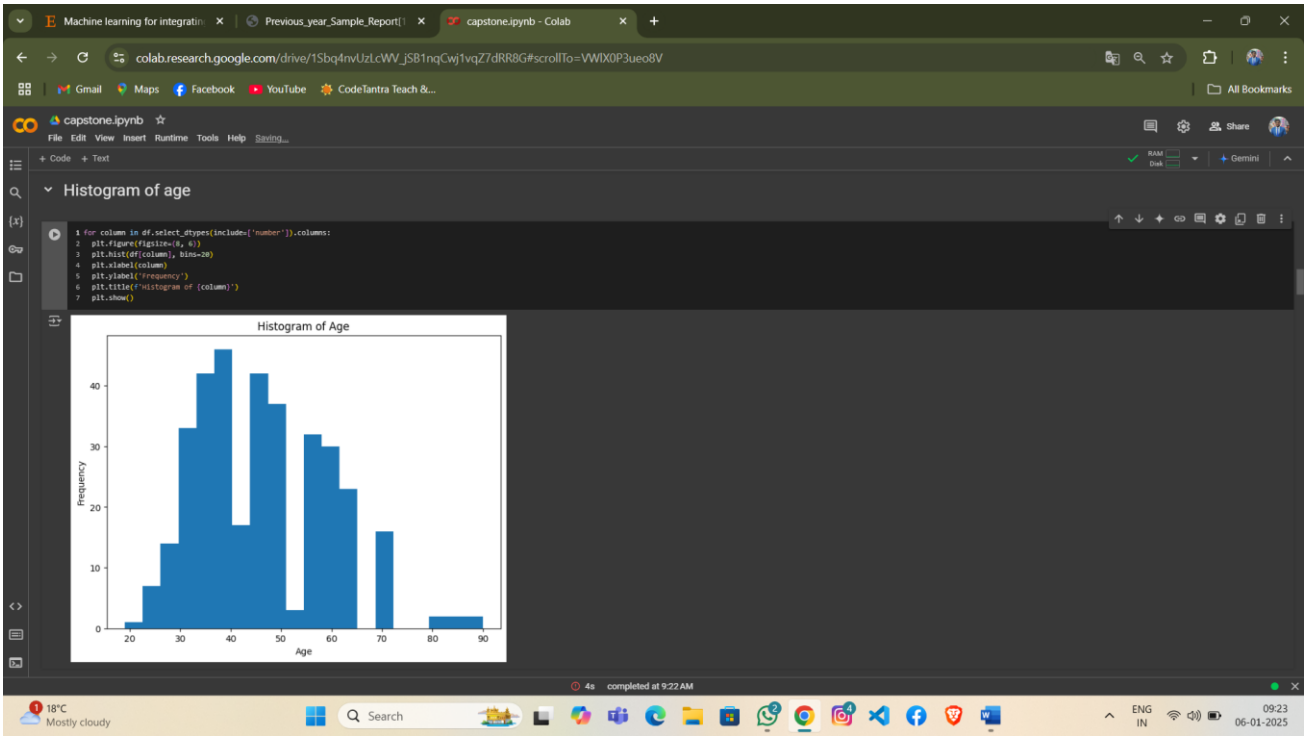
SCREENSHOTS:



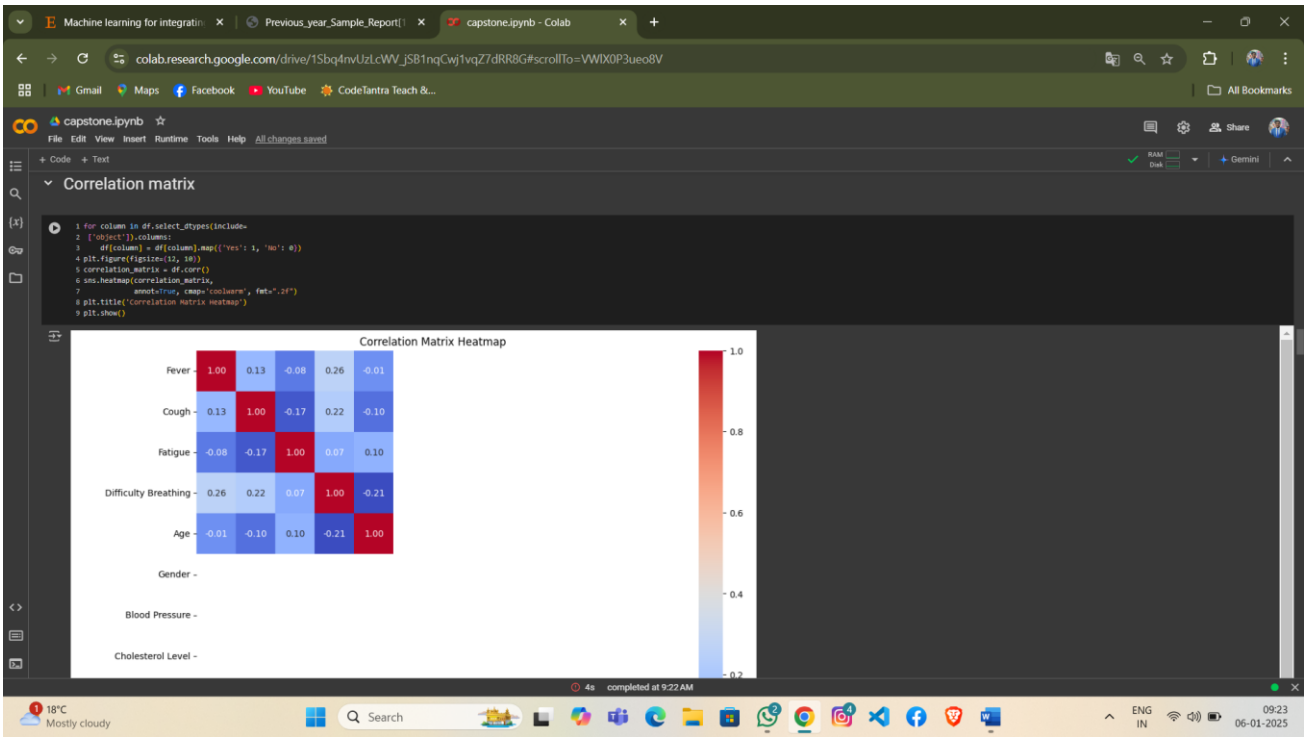
Screenshot 1: Data importing and reading



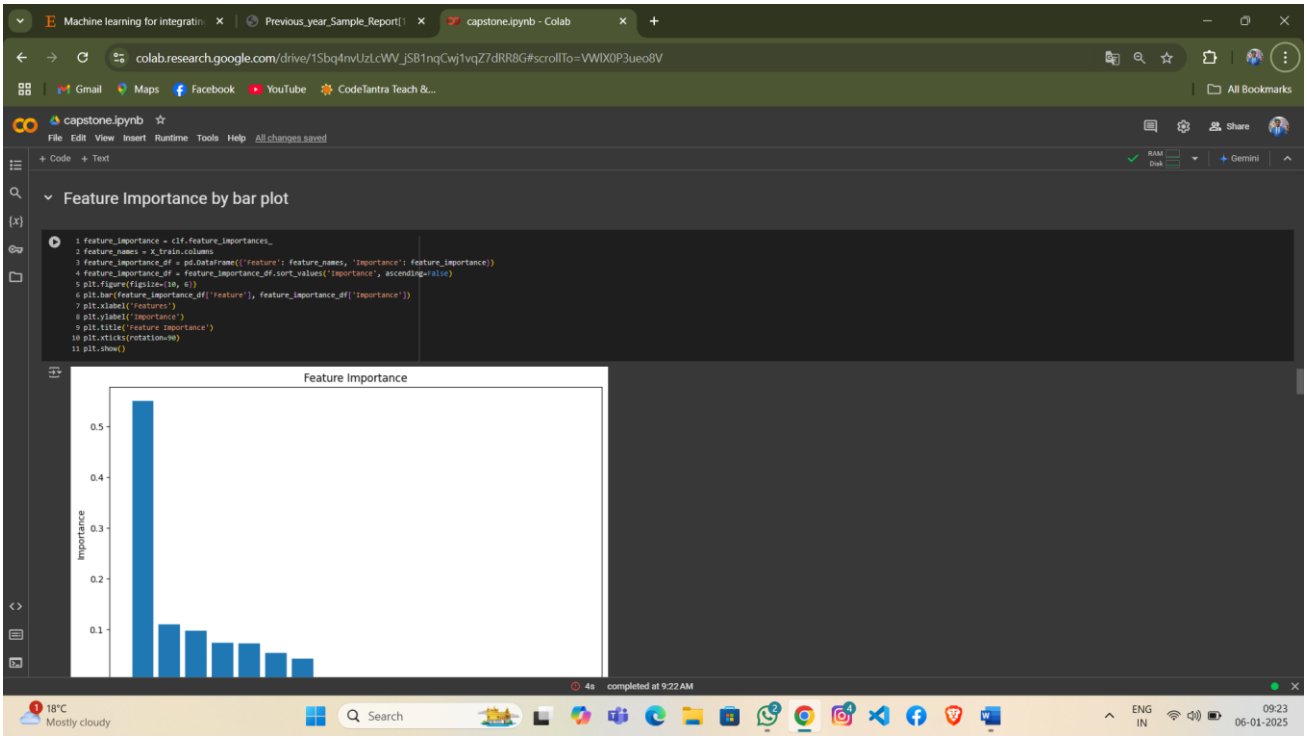
Screenshot 2: Data training



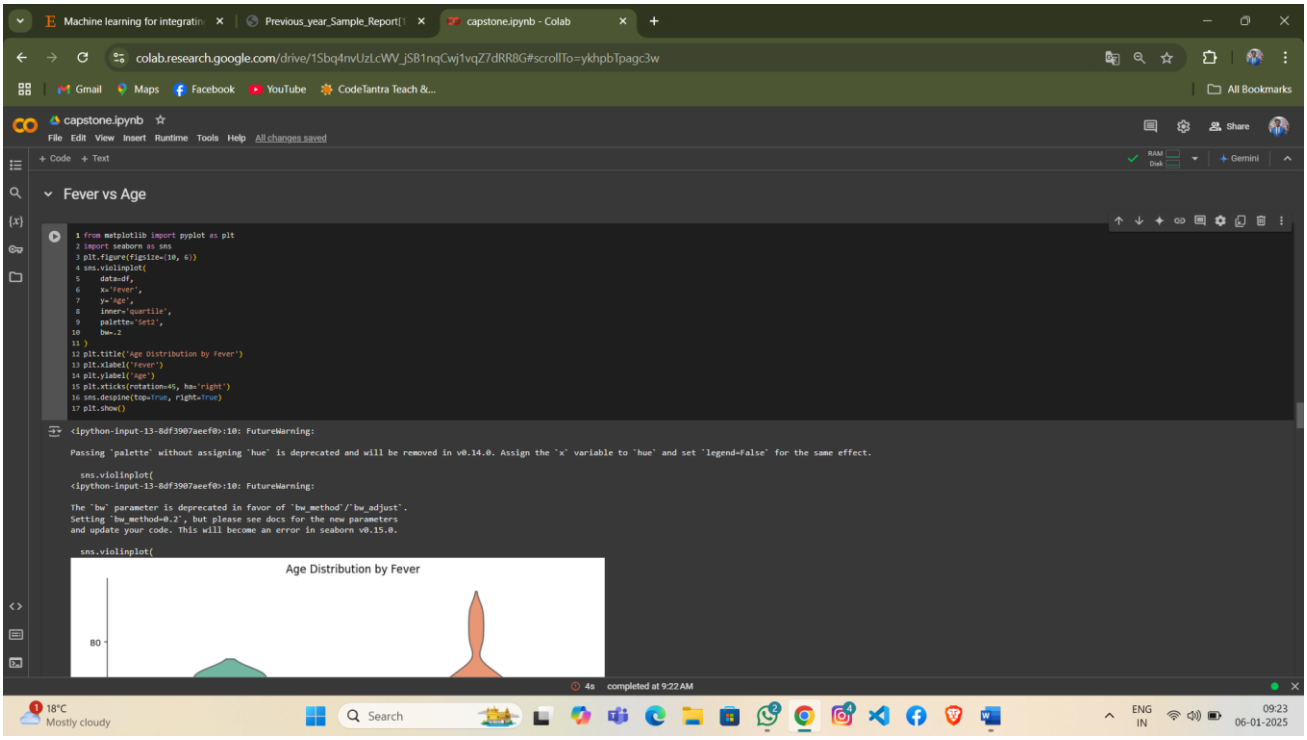
Screenshot 3: Graph



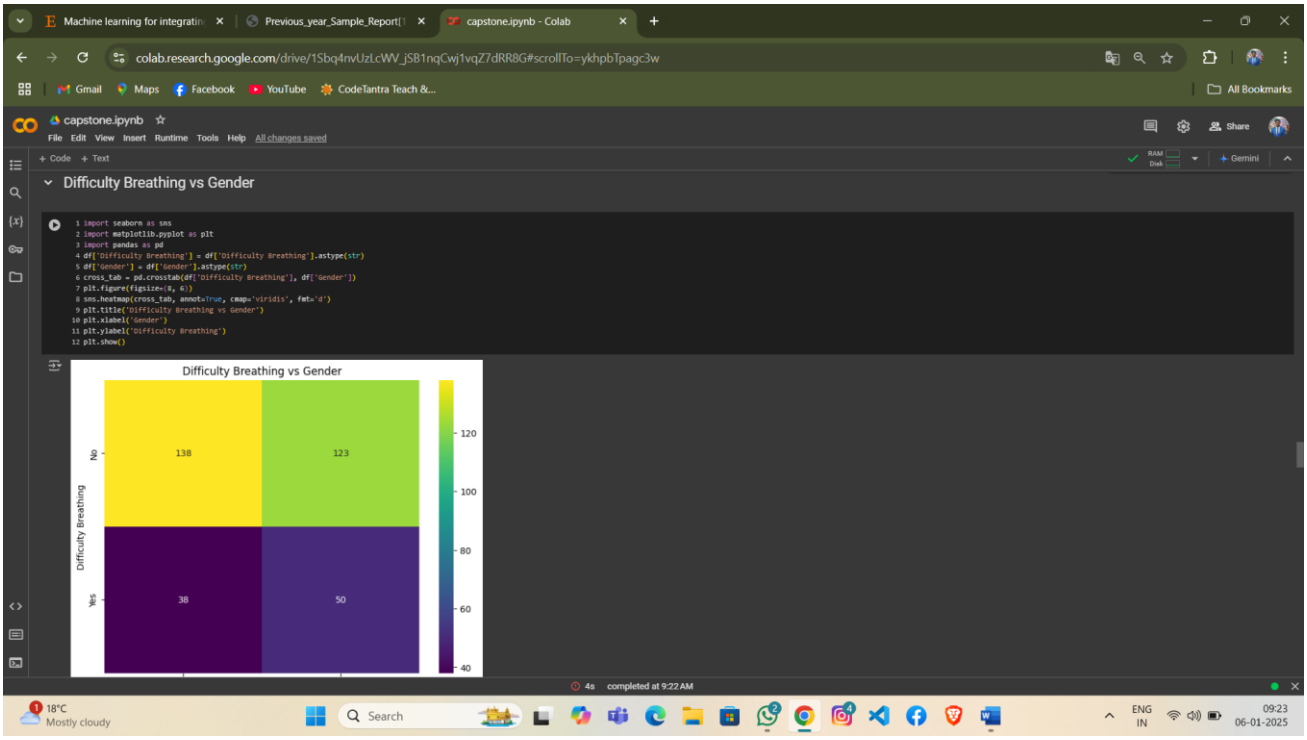
Screenshot 4: Graph



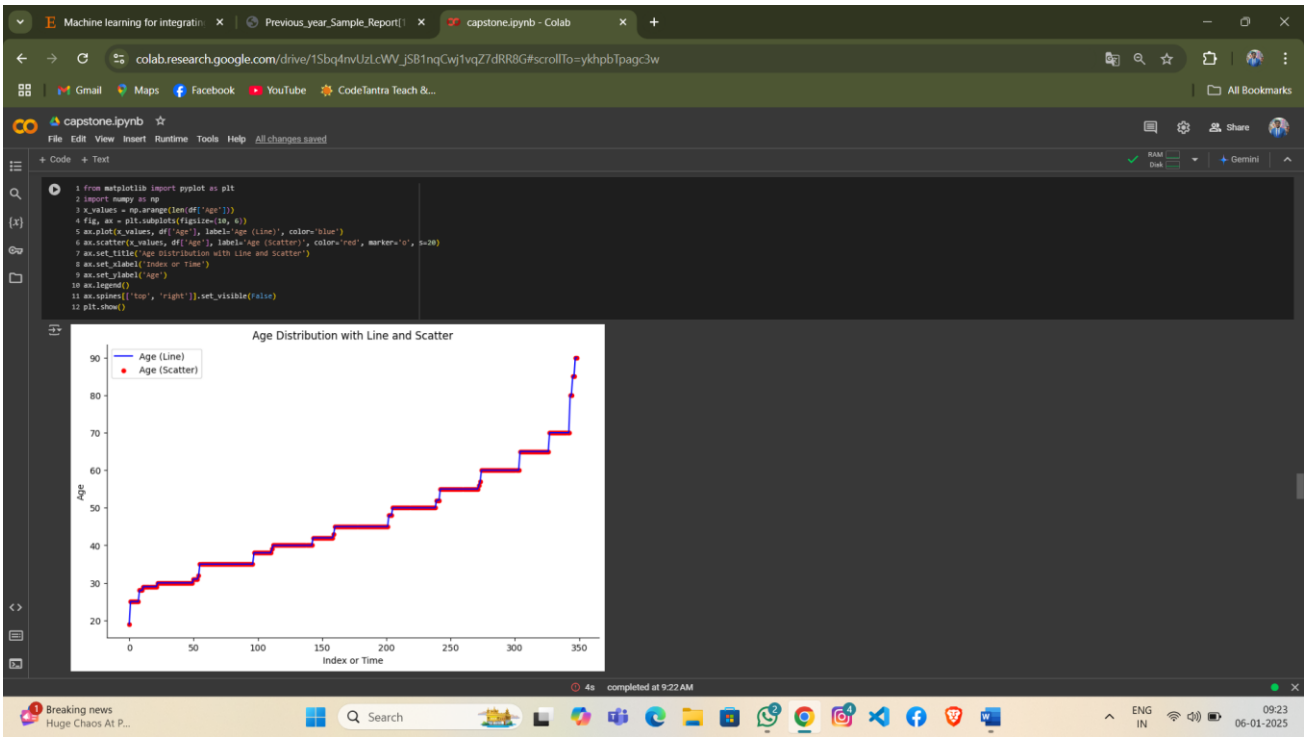
Screenshot 5: Graph



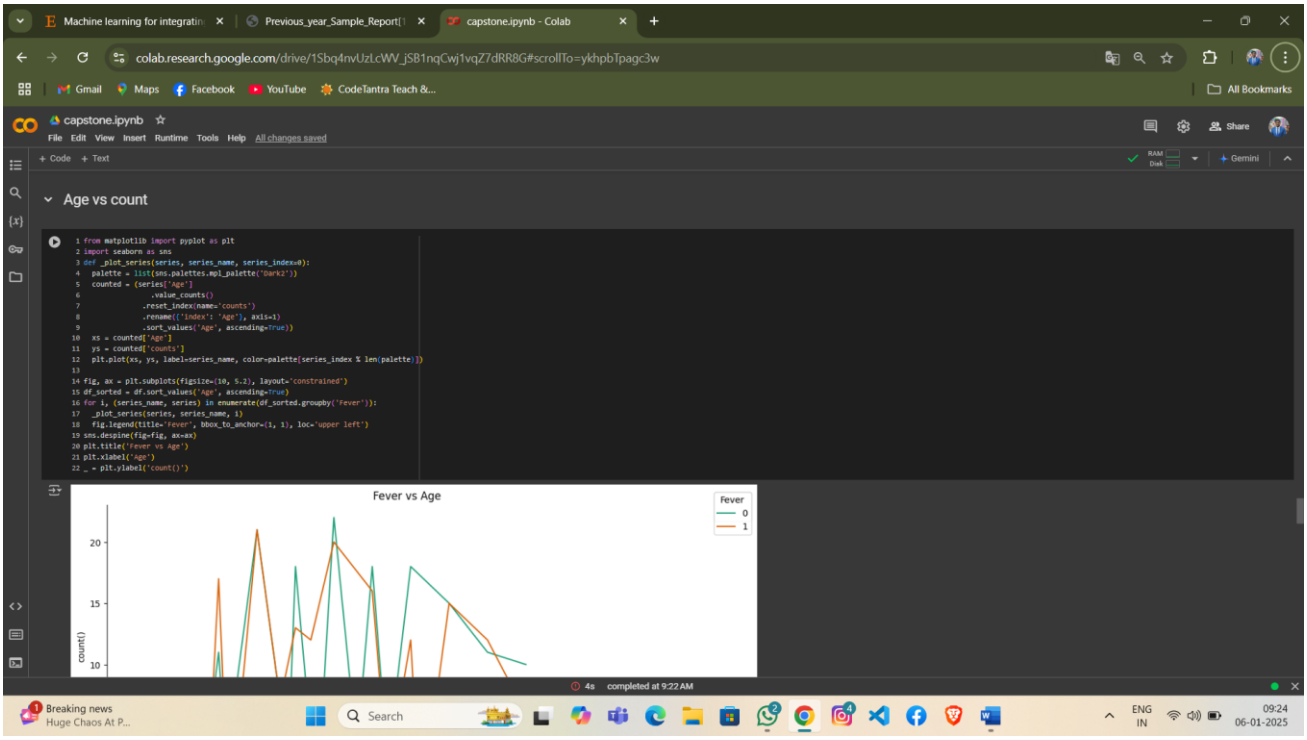
Screenshot 6: Graph



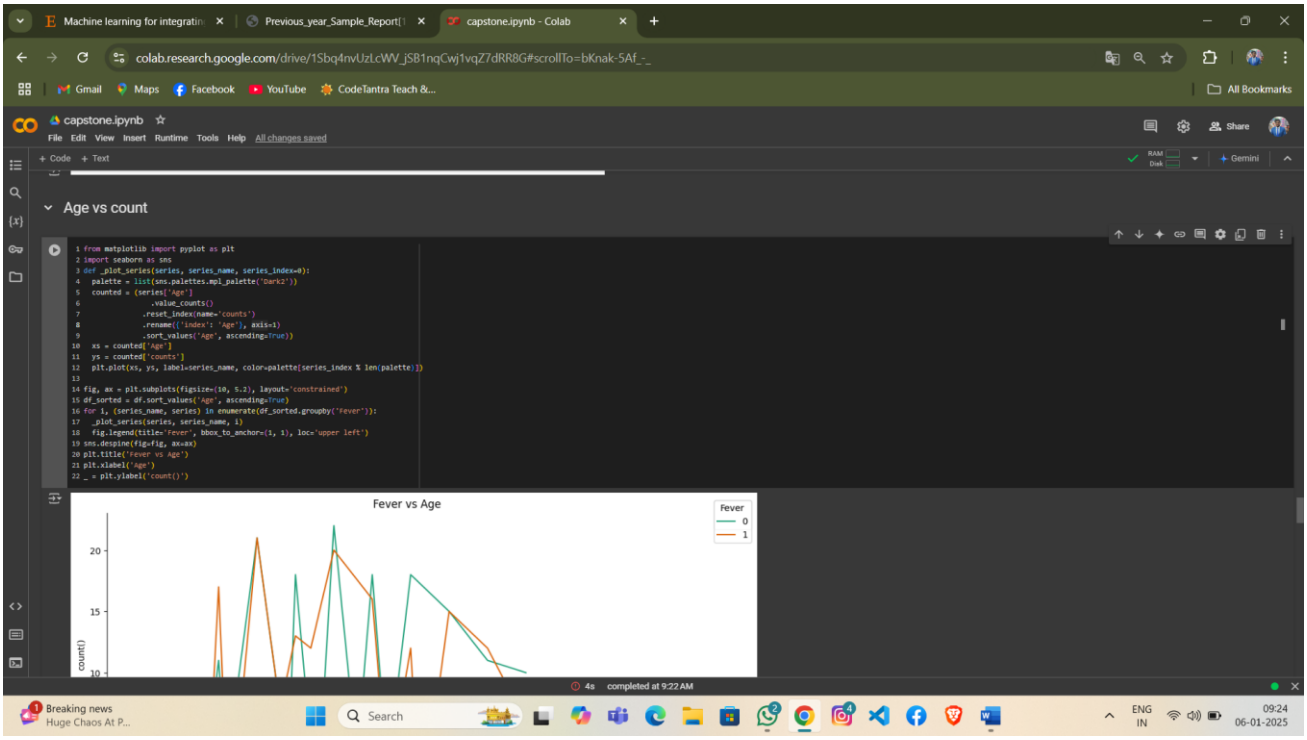
Screenshot 7: Graph



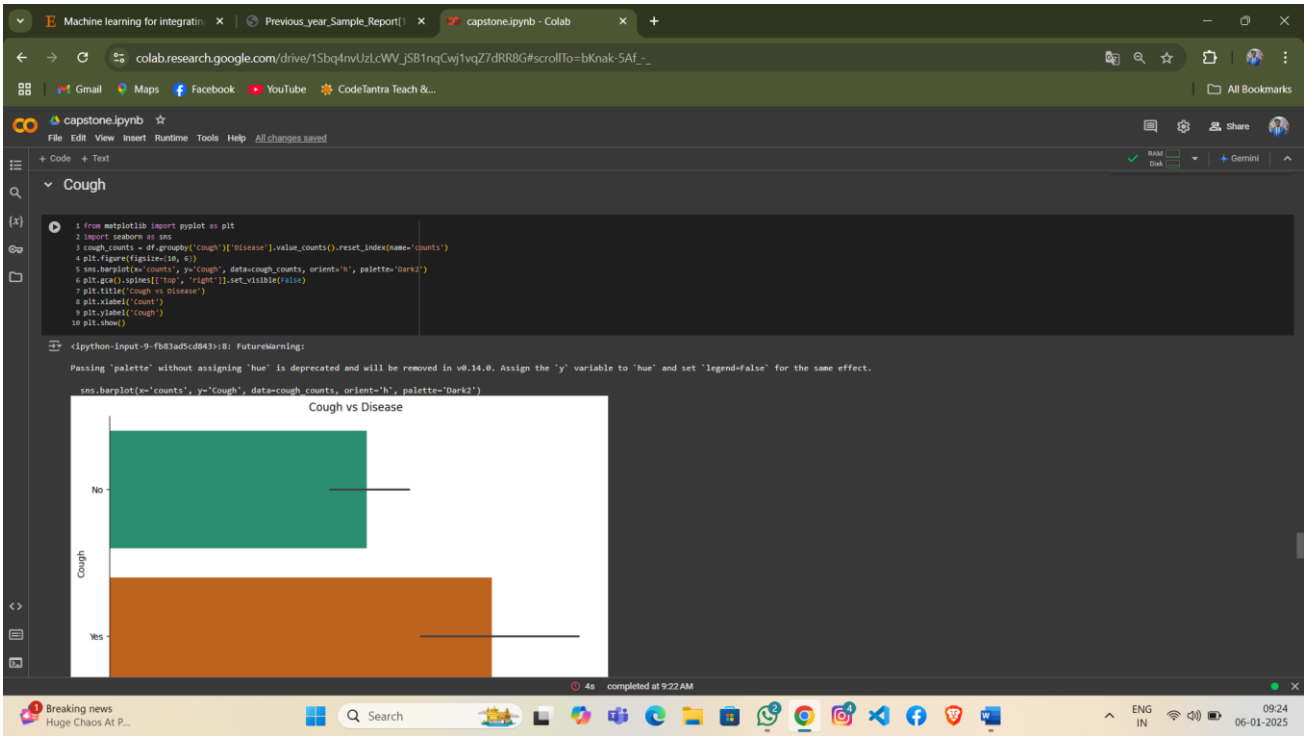
Screenshot 8: Graph



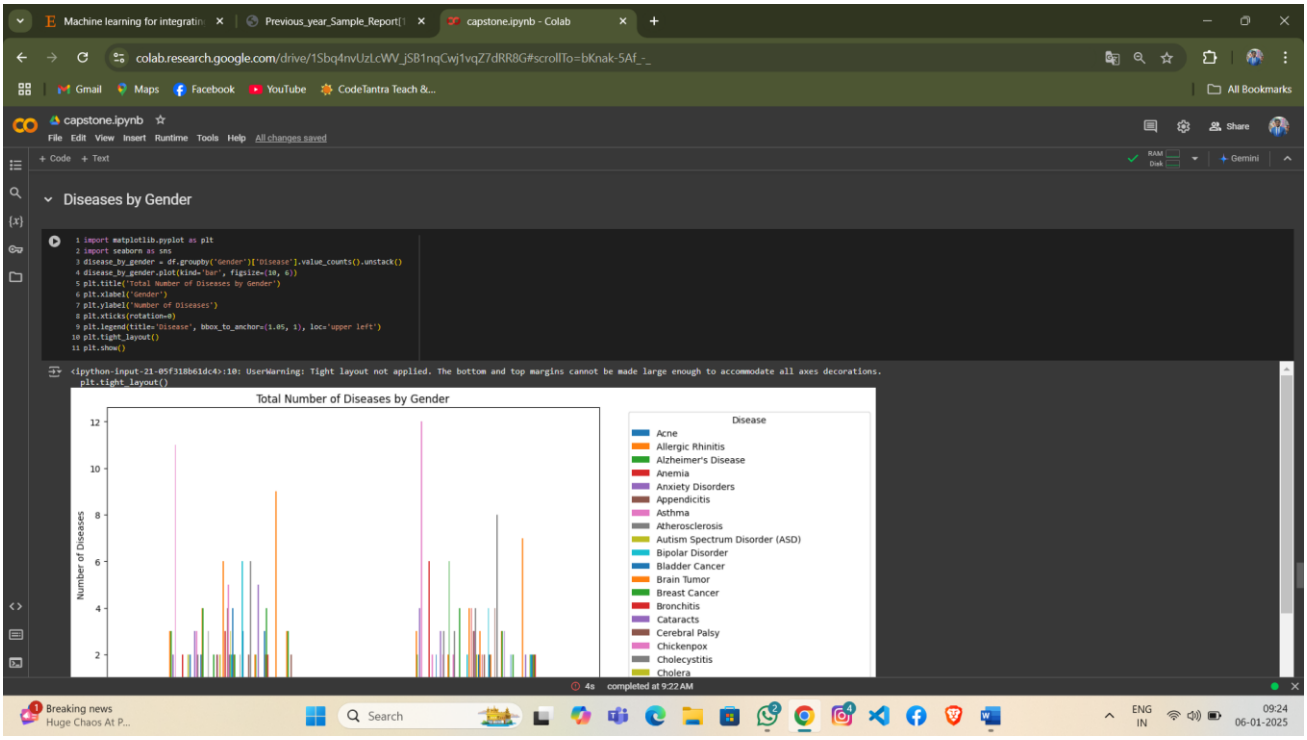
Screenshot 9: Graph



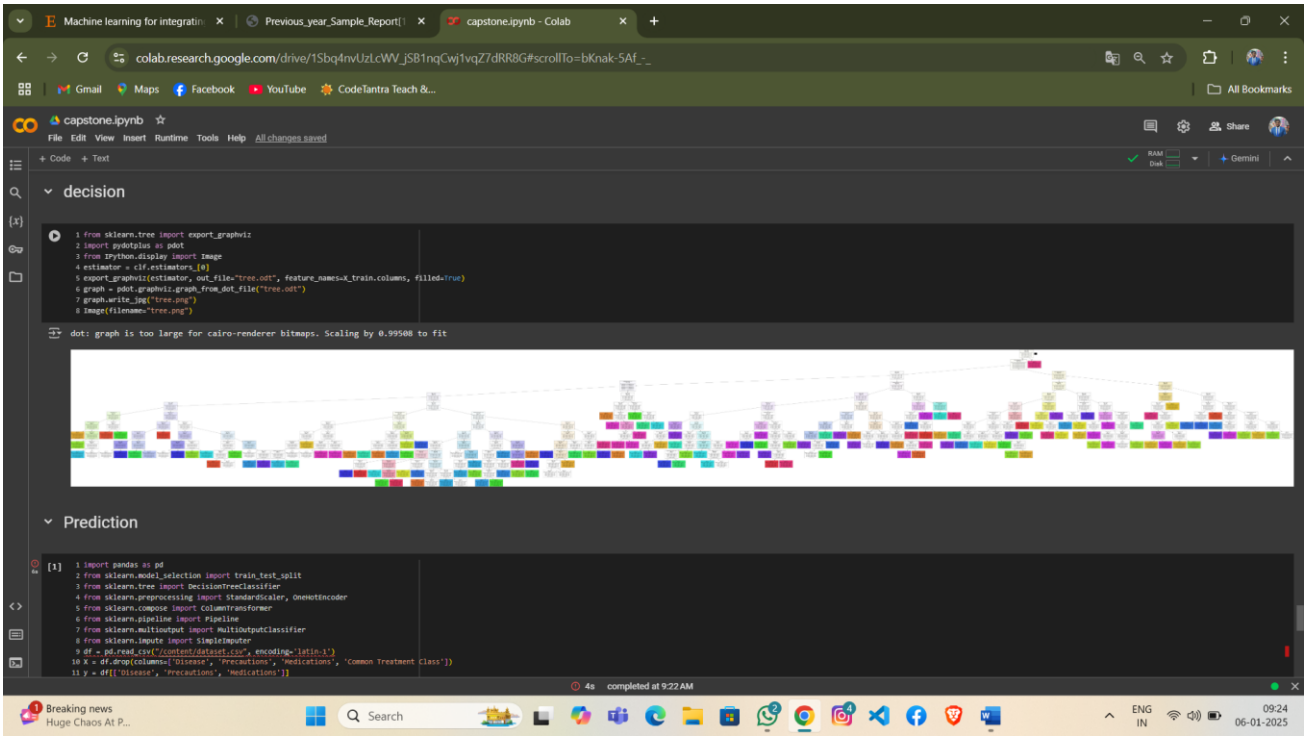
Screenshot 10: Graph



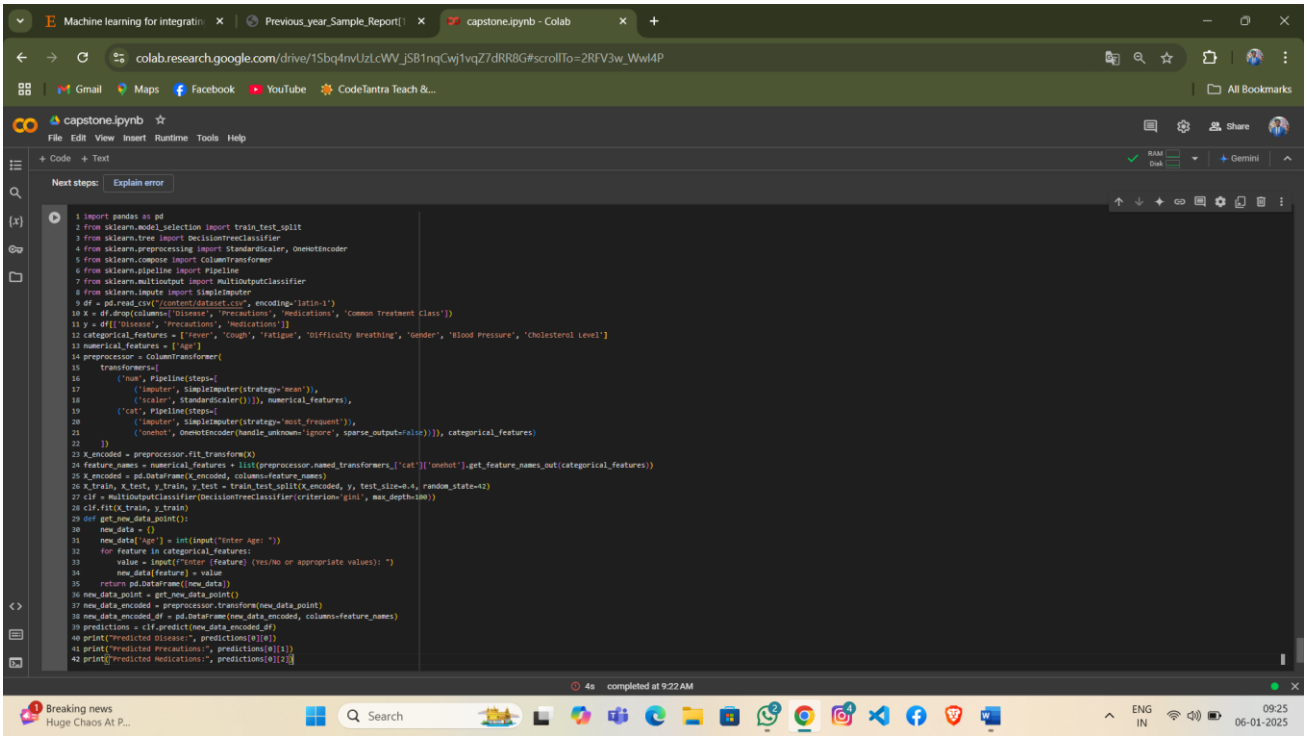
Screenshot 11: Graph



Screenshot 12: Graph



Screenshot 13: Decision Tree



Screenshot 14: Prediction

cap_repo

ORIGINALITY REPORT

10%	7%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Nanyang Technological University, Singapore Student Paper	3%
2	fastercapital.com Internet Source	2%
3	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1%
4	Submitted to The Northcap University Student Paper	<1%
5	dev.to Internet Source	<1%
6	www.fastercapital.com Internet Source	<1%
7	Submitted to Liverpool John Moores University Student Paper	<1%
	dspace.daffodilvarsity.edu.bd:8080	

8	Internet Source	<1%
9	A. J. Singh, Nikita Gupta, Sanjay Kumar, Sumit Sharma, Subho Upadhyay, Sandeep Kumar. "Artificial Intelligence and Machine Learning Applications for Sustainable Development", CRC Press, 2025 Publication	<1%
10	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference “Convergence2024” Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publication	<1%
11	Submitted to Manipal University Student Paper	<1%
12	aiforsocialgood.ca Internet Source	<1%
13	ijcst.com.pk Internet Source	<1%
14	Submitted to University of Surrey Student Paper	<1%

15	Submitted to University of Wales Institute, Cardiff Student Paper	<1%
16	Submitted to Carnegie Mellon University Student Paper	<1%
17	Submitted to University of Florida Student Paper	<1%
18	Submitted to Weber State University Student Paper	<1%
19	ijircce.com Internet Source	<1%
20	Submitted to Curtin University of Technology Student Paper	<1%
21	"Application of Artificial Intelligence in Neurological Disorders", Springer Science and Business Media LLC, 2024 Publication	<1%
22	Submitted to Notre Dame of Marbel University Student Paper	<1%
23	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1%
24	Submitted to Wawasan Open University Student Paper	<1%

25	asianpubs.org Internet Source	<1%
26	Emmanuel Ahishakiye, Fredrick Kanobe, Danison Taremwa, Bartha Alexandra Nantongo, Leonard Nkalubo, Shallon Ahimbisibwe. "Enhancing Malaria Detection and Classification using Convolutional Neural Networks - Vision Transformer Architecture", Springer Science and Business Media LLC, 2024 Publication	<1%
27	Lilit, Tonoyan. "Freight and Lead Time Estimator", Universidade NOVA de Lisboa (Portugal), 2024 Publication	<1%
28	Submitted to M S Ramaiah University of Applied Sciences Student Paper	<1%
29	erc.metu.edu.tr Internet Source	<1%
30	huggingface.co Internet Source	<1%
31	mlsql.dev Internet Source	<1%
32	www.ijraset.com Internet Source	<1%

33	Submitted to The University of the West of Scotland Student Paper	<1%
34	core.ac.uk Internet Source	<1%
35	Jacob N. Ablin. "Nociplastic Pain: A Critical Paradigm for Multidisciplinary Recognition and Management", Journal of Clinical Medicine, 2024 Publication	<1%
36	Poonam Tanwar, Tapas Kumar, K. Kalaiselvi, Haider Raza, Seema Rawat. "Predictive Data Modelling for Biomedical Data and", River Publishers, 2024 Publication	<1%
37	indico.cern.ch Internet Source	<1%
38	medium.com Internet Source	<1%
39	robots.net Internet Source	<1%
40	tnsroindia.org.in Internet Source	<1%
41	2024.isav.ir Internet Source	<1%

42	Iweriebor, Lawrence Ejime. "Approach to Medicare Provider Fraud Detection and Prevention", Capitol Technology University, 2023 Publication	<1%
43	Matin Chehelgerdi, Fereshteh Behdarvand Dehkordi, Mohammad Chehelgerdi, Hamidreza Kabiri et al. "Exploring the promising potential of induced pluripotent stem cells in cancer research and therapy", Molecular Cancer, 2023 Publication	<1%
44	Md. Saiful Islam, Arafatun Noor Orno, Mohammad Arifuzzaman. "Approach to Social Media Cyberbullying and Harassment Detection Using Advanced Machine Learning", Springer Science and Business Media LLC, 2024 Publication	<1%
45	www.devx.com Internet Source	<1%
46	www.mdpi.com Internet Source	<1%
47	Ton Duc Thang University Publication	<1%

Dear Sir/Ma'am,

We are delighted to inform you that your manuscript "ID: CSE24DEC007, Title: **Patient Case Similarity** has been **"Accepted for Publication"** in SSRG-International Journal of Computer Science and Engineering (SSRG-IJCSE)", ISSN: 2348-8387.

Please see the attached reviewer comments for further details about necessary revisions.

Registration Process
Article Processing Charges:
3700: Online Publication Only [Mandatory]

Note: **Author count cannot be added once the paper gets accepted.**

Kindly try to pay the Article Processing Charges within 4 working days.

Authors can use any one of the following payment options:
1. You can use your debit card/credit card to pay the publication fee through PayPal, from our website Payment Gateway.

https://www.internationaljournalssrg.org/IJCSE/mode-of-payment.html

2. Account Details for Bank / Wire Transfer:

Account Name	Seventh Sense Research Group,
Account Number	6189776358
Account Type	Current Account
Bank Name	Indian Bank (IB)
Branch Name	Kumbakonam Bazaar,
Branch Address	50, Tsr Big Street Kumbakonam, Thanjavur Dist 612001.
IFSC Code	IDIB000K144

Patient Case Similarity

Perumalla Sai Surya[1], Bhumpalli Vishnu Vardhan Reddy[2], Sanjana R[3], Lingamdhinne Akanksha[4], Koyi Mithun[5]

Computer Science & Engineering/Student,Presidency University, Bengaluru, India (Orcid ID: https://orcid.org/0009-0000-6244-5345)
Computer Science & Engineering/Student,Presidency University, Bengaluru, India (Orcid ID: https://orcid.org/0009-0005-9260-6337)
Computer Science & Engineering/Student,Presidency University, Bengaluru, India (Orcid ID: https://orcid.org/0009-0004-7911-1105)
Computer Science & Engineering/Student,Presidency University, Bengaluru, India (Orcid ID: https://orcid.org/0009-0000-1750-6532)
Computer Science & Engineering/Student,Presidency University, Bengaluru, India (Orcid ID: https://orcid.org/0009-0009-0374-7269)

Email address: saisuryaperumalla96092@gmail.com

Received: Revised: Accepted: Published:

Abstract: *This is the approach to finding similar patients in terms of characteristics. It may shift how health care is going to develop itself, especially with the help of machine learning algorithms in finding patterns that may not be observable from the data given and in enhancing clinical decision-making. This project aims at developing a strong patient similarity analysis system based on decision trees.*
The steps in the project include data collection and preprocessing, feature engineering, model training, evaluation, and finally, deployment. The quality and completeness of the data are necessary for any analysis. Feature engineering is actually the process of choosing and designing relevant features to describe patients. Decision trees learn decision rules that classify patients into similar groups. Some of the metrics used for determining the performance of the model are accuracy, precision, recall, and F1-score.
Data privacy, bias, and fairness must therefore be considered when applying the model practically. The model, therefore, must be explainable to the clinicians to gain their confidence and enhance uptake. Finally, there must be further learning for updates in the model toward achieving accuracy and relevance.
This will help us harness the similarity analysis of patients to enhance clinical decision-making, treatment planning, and accelerating medical research. It is a contribution toward the advancement of precision medicine, which improves patient outcomes in a broader sense.

Keywords: *decision tree, similarities, preprocessing, visualization, prediction, accuracy*

1. Introduction

Machine learning in patience case similarity is a sub-part of artificial intelligence that studies the previous data of previous patients and tries to predict the current patient situation accurately using DECISION TREE algorithm. DECISION TREE algorithm is very suitable for this type of problem because it finds the relations between the patients by clustering the nearer values of the current patient data.
Use references to provide the most salient background rather than an exhaustive review. The last sentence should concisely state your purpose for carrying out the study or a summary of the results [2].

1.1 Aims and Objectives:

- Develop a strong and accurate decision tree-based model for patient similarity.
- To assess the performance of the model using appropriate values.
- To derive the most significant features responsible for patient similarity.
- To consider the possible uses of the model in clinical settings.

1.2 Context and Motivation:

Data generation in the healthcare sector increased manifold in recent years with the advancements of EHR and wearable technologies. The large amount of data generated now can give an idea of providing insights, which can be transformed into benefits in the care of a patient. However, it is not easy to analyse or interpret such large amounts of data.

1.3 Thesis Overview

In this thesis, we attempt to investigate the application of

decision trees for patient similarity analysis. We shall use the interpretability and efficiency of a decision tree in developing a model that can accurately identify similar patients with great potential in providing clinical decision-making insights. In this case, the student will explore data preprocessing, feature engineering, model training, evaluation, and interpretation. Finally, she will discuss the ethical implications of the use of patient data and suggest future possible research directions.

2. Literature Review

In this section should extend, not repeat the information discussed in Introduction [4]. In contrast, a Calculation Section represents a practical development from a theoretical basis [5].

2.1 How it will begin:

- Data collection – Collecting different patient data with different diseases and situations.
- Data pre-processing – Identifying and removing the null values and inconsistent data in the data set for getting best accuracy.
- Data clustering – Grouping the patient data which have similarities.
- Training model
- Testing model accuracy

2.2 Why machine learning:

Algorithms that will present in the machine learning was so accurate We can upload more no of images in the form of dataset. And by using machine learning we even make the model for CSV files. As we all know that company like Amazon uses machine learning for the feature extraction and machine learning is used for determining the height and weight, it's dimensions where the feature extraction is very accurate Machine learning has more advantages. We will train the machine to identify and extract features according to our requirement. If we see the products in the Amazon, they can't describe the matter for every product. So by using machine learning we can compute its dimensions, We can determine precision, recall, accuracy and f1 score which was in machine learning.

2.3 Types of problems solved using Machine Learning:

Classification is a task that required the use of machine learning algorithms to learn how to assign a class label to a given data

- Let's say that we are given to classifying a fruits and vegetables on basis of there category.
- Regression it help to investigate.
- The relationship between variables
- Means for example imagine if we collect a pack of apples on different stage of the year
- If we want to visualize the data x value of the each point is the day of the year It is sold and the y value is the price of the package. In this scenario, we can use to find a mathematical formula represents this data.
- This unable us to predict us the price of the apples give the day of year
- There are 3 types of regression
  1. Linear regression
  2. Polynomial regression
  3. Logistic regression

2.4 Types of Machine Learning Algorithm:

- ✓ *Supervised Learning:* It is the method of teaching machine under the supervision and with structured data. It uses only labelled data. In this project we used supervised learning because it should learn the data with help of labels and previous condition particular diseases.

- ✓ *Reinforcement Learning:* It is the method of machine learning that learns on its own by feedback and experience. It will help this project to predict the medicine which to be used for the current patient based on old patient data. Then it checks for the changing environment and it will adapt the new environment accordingly.

2.5 Why Python in ML for Patient Case Similarity:

Python play a pivotal role in implementing ML models for patient case similarity due to its libraries and frameworks with the help of python data preprocessing is performed with libraries like pandas and numpy and scikit-learn these are commonly used for DECISION TREE and clustering and in python we use pytorch for deep learning approach and spancy for handling unstructured data and matplotlib or seaborn for visualization with the help of these libraries we can able to build scalable and accurate models.

2.6 Breadth Context and Theory

Literature review comprises patient similarity analysis in general and discusses their applications, challenges, and prevalence in healthcare. It will attempt to show how proper patient phenotyping is important and machine learning in healthcare plays a role.

2.7 Work by Theme in Detail

A concrete set of studies that have used decision trees in the similarity analysis of patients will be outlined. The methodologies applied, datasets utilized and metrics highlighted in terms of performance will be considered.

2.8 Research Gap and Summary

A list of gaps in current literature will be provided, including other rigorous analysis and further data sources into the

decision tree models and the development of user-friendly interfaces for clinical applications.

3. Methodology

3.1 Research Design

This paper applies machine learning for creating a patient similarity model by means of decision trees. The proposed study design for the analysis of the existing patient history data is a retrospective design**.**

3.2 Data Collection and Preprocessing

• Sources: EHRs, clinical trials, biomedical literature
• Cleaning: Missing values, outliers, inconsistencies
• Feature Engineering: Relevant features such as demographics, medical history, lab results, genetic data

3.3 Model Development and Training

• Algorithm used:- Decision Tree, ID3, C4.5, CART
• Model Training-: Train the model on the pre-processed training data.
• Hyper parameter Optimization: Enhance the model by optimizing the hyperparameters succinctly

3.4 Model Evaluation

• Evaluation Metrics: Make use of accuracy and precision, recall, F1-score, and also ROC curve for checking the performance of the model.
• Cross-validation: Test the model's generalization.
• Confusion Matrix: Also, consider taking the confusion matrix into consideration in order to know which patients are being misclassified.

3.5 Ethics and Limitation

• *Data Privacy:* Follow the norms of data privacy, that is, HIPAA norms.
• *Ethical Considerations:* Understand the possibilities of biasness of the model and keep the fairness of the model intact.
• *Limitations:* Discuss the limitations of the study.

4. Analysis and Synthesis

- *Data Analysis*: The preprocessed data is subjected to pattern and trend analysis.
- *Model Performance*: The performance of the decision tree model can be evaluated based on different metrics.
- *Feature Importance*: Identify which features are important, leading to the highest patient similarities.
- *Sensitivity Analysis*: Assess how different input parameters may be affecting the model's output.

Flow of Project

All figures in the manuscript should be numbered sequentially using Arabic numerals (e.g., Figure 1, Figure 2), and each figure should have a descriptive title. The figure number and title should be typed with single-spaced, and centered across the bottom of the figure, in 8-point Times New Roman, as shown below. The figure captions should be editable and be written below the figures.

Figure 1.Work Flow of the Project

4.1 Data Gathering and Preprocessing:

- Obtain data about patients from various sources, which may include EHRs and clinical trials.
- Clean and preprocess the data, including checking for missing values, outliers, and inconsistencies
- Normalize/standardize the numerical data.
- Create feature engineering for relevant features.

4.2 Feature Engineering:

- Features that account for similarity between patients
- Feature selection techniques may also come in, like the filter methods, the wrapper method, and the embedded methods

4.3 Model Selection and Training

- Choose a suitable decision tree algorithm, such as ID3, C4.5, or CART
- Fit the decision tree model on the preprocessed and selected features.
- Tune parameters for optimal performance

4.4 Model Evaluation

- Critically evaluate the model's accuracy, precision, recall, F1-score, and ROC curve to assess the model's appropriateness.
- Test the ability of the model to generalize through cross-validation.
- Inspect the confusion matrix for the patients wrongly classified

4.5 Model Deployment

- Deploy the learned model into a clinical decision-making system or apply to relevant applications.
- While ensuring the model is optimally integrated into existing workflows.

4.6 Tuning or Retraining Models:

- In the event the model fails to perform satisfactorily, hyperparameters may be tuned, or the model may be retrained with additional data.
- Explore alternative algorithms or methods of ensemble to boost performance.

4.7 Continue monitoring and enhancing the model

- Continue to monitor the performance of your model. Retrain it periodically to maintain its accuracy.
- Interact with users to identify areas for improvement.
- Keep the model and the user interface updated with new knowledge and data.

4.8 Implementing the Flowchart to an Agile Model:

An Agile model like Scrum can be adopted in the development process of the patient similarity analysis.

Scrum

- *Sprint Planning:* Division of the project into an even more workable and manageable task set, which might be data collection, preprocessing of the data, training a model, evaluation, and finally deployment.
- *Sprint Execution:* Assign all these tasks to team members to execute iteratively.
- *Daily Scrum:* Keep daily stand-up meetings to track the progress and brainstorm any challenge.
- *Sprint Review:* Present the work to the stakeholders and collect the feedback for the work done.
- *Sprint Retrospective:* Review the sprint, identify lessons learned, and set goals for the next sprint.

4.9 Data Visualization for Patient Similarity Analysis

Data Visualization is a very powerful tool in understanding and interpreting patient similarity. By visualizing the data as well as the results of the analysis, we are able to obtain valuable insights into what contributes to patients being similar to each other and into how good the decision tree model really is.

Key Visualization Techniques:

Feature Importance Plots:

- Visualize the importance of different features in the decision tree.
- Find out which are the key factors that contributed to patient similarity.

Figure 2 Boxplot of Age by Disease - The boxplot provided visually shows the age spread across various diseases. It is important for patient similarity analyses, as it is important to show possible trends regarding the incidence of diseases with respect to age. Such understanding allows for clustering patients with similarities in age and diseases, making the similarity analysis more precise.

Figure 3 Violin plot of Fever vs Age - The violin plot would thus give a comprehensive overview of how age is distributed along various categories of the feature "Fever". It illustrates the density of and how ages are spread among cases with or without fever. In such a way, we can find what kind of pattern exists, or probably a relationship exists between age and fever by correlating the shapes and positioning of violins.

Figure 4 Heat Map of Gender vs Difficulty in Breathing - It illustrates the relationship between gender and difficulty breathing. The frequency of occurrence of each combination is represented by the intensity of color.

Figure 5 Line plot: This line plot indicates the distribution of ages in the data set. It will represent the extent of ages, varied age groups' frequency, and outliers.

Figure 6 Bar chart of count of the diseases according to gender - The bar chart is a distribution of diseases by gender. From the heights of the bars, we can infer potential gender disparities in the prevalence of disease. This information is very important in patient similarity analysis because it will enable us to group patients based on their gender and disease profile. This shows gender-specific patterns, by which understanding these patterns can improve the accuracy of similarity predictions through better analysis and modelling techniques.

Figure 7 Line plot of Fever vs Age - It will provide, by direct comparison, a graphic view of how the age distribution between patients who are feverish and those that are not compares. Trends between the two lines may allow some patterns or relationships between age and fever to be seen.

Decision Tree Visualization:

- Reflect on the structure of the decision tree.
- Describe the decision-making logic and rules used to classify patients.

Figure 8 Decision Tree Visualization - This is an obvious and intuitive visualization of the decision tree in the model's decisions. Each node in this tree is a decision to the model based on its chosen feature, and different branches represent possible results. The leaves of the trees represent final classification or prediction.

Patient Similarity Network:

- Build a network graph where nodes indicate patients and edges indicate similarity of patients.
- Describe the clusters formed by patients who are similar and their characteristics.

Time-Series Visualization:

- Visualize patient trajectories over time for detecting patterns and trends.
- Assess comparative trajectories of similar patients to understand disease progression.

***A*dvantages of Data Visualization:**

1. *Better Understanding:* Visualization can enlighten complex patterns and relationships between the data.
2. *Improved Communication:* Visualization can represent insights in meaningful ways with clinicians and researchers.
3. *Informing Decisions:* Visualization can facilitate data-driven decision-making by presenting information directly and clearly.
4. *Identify Outliers:* Visualizations are used in identifying outliers and anomalies for the data.

Role of One-Hot Encoding in Patient Similarity Analysis

One-hot encoding is a very important technique that facilitates the conversion of categorical data into a numerical format that would be suitable for machine learning algorithms, such as decision trees. The reason one-hot encoding transposes categorical features into numerical ones is that it means that the decision tree clearly captures difference and admits better predictive results.

Here's how one-hot encoding works:

1. Identify categorical features from patient data, such as gender, race, diagnosis, or medication.
2. Encoding: for each categorical feature there would be a new binary feature created for each category.
3. Binary Representation: give the value of 1 to the relevant binary feature, in case it belongs to that category and otherwise 0. Suppose there's a categorical feature "Gender" with two categories, "Male" and "Female." One-hot encoding would result in two new binary features:

• Is_Male: 1 if the patient is male; else 0
• Is_Female: 1 if the patient is female; else 0

Advantages of One-Hot Encoding:

• *Categorical Information is Preserved:* One-hot encoding preserves the categorical nature of the data without inducing ordinal relationships between categories.
• *The model improves:* Numerical representation of categorical features yields better decision-making models.
• *Better Interpretability:* One-hot encoding may lead to a decision tree that has more interpretations explicitly including the influence of each category.

The following picture shows how it worked:

Figure 9 The Prediction of patient cases

5. Discussion

How Patient Similarity Analysis Can Be Helpful to Others

Patient similarity analysis with decision trees and other machine learning algorithms can convert health into a better individualized and a more exact treatment approach. Some of the main benefits are as follows:

Improvement for Patients

Tailored Plans of Treatment Identifying similar patients would enable healthcare providers to have a chance to come

up with specific plans suited to their needs hence giving good outcomes and adverse effects.
Early Onset Disease Detection Early on, it gives the chance for early identification of similar patients who may be suffering from the disease; hence it can provide the diagnosis and interventions much earlier.
Better Patient Experience Patient-centered needs and preferences are probably more familiar to provide empathetic or personalized care and services to patients.

For Healthcare Providers:

Enhancing Clinical Decision-Making Patient similarity analysis can be extremely useful for informed clinical decisions: treatment decisions, etc; regarding prognosis.
Optimal Resource Allocation: Patient group similarity can enable healthcare providers to make the best allocation of resources.
Research and Development: Patient similarity analysis accelerates drug development and discovery by detailing patient subgroups likely to respond well to a certain kind of treatment.

For Researchers:

New Insight Discovery: Patient similarity analysis may discover new disease subtypes and biomarkers.
Finding of Novel Therapeutic Targets: Mechanisms of disease will identify potential therapeutic targets.
Advance Precision Medicine: The similarity of patients is one of the most important factors of precision medicine, or tailored treatments for individual patients.

6.Conclusion

Patient similarity analysis by decision trees represents a powerful approach towards improved patient care. It matches patients who have similar characteristics, enabling clinicians to make more informed decisions related to diagnosis, treatment, and prognosis.

This project has demonstrated the classification of patients with similarities using decision trees. The developed model, having used a comprehensive dataset for its training, can predict outcomes of the patients and determine important factors that influence similarity.

In this regard, the current study faces limitations in regard to a decision tree, and more advanced methodologies need to be approached. Further inclusion of genomics and proteomics data can also improve the accuracy of the analysis with relevance to precision. In terms of adoptions, it is possible to have friendly interfaces for consumers in order to get affected by this model into real-world clinical workflows.

References

[1] Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, "*Decision trees: an overview and their use in medicine*," University of Maribor – FERI Smetanova 17, SI-**2000** Maribor, Slovenia, vili.podgorelec@uni-mb.si.

[2] Han et al., "*A Survey of Data Mining Techniques for Medical Diagnosis*," **2001**.

[3] "*Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy*," **2010** IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), **2010**.

[4] Azar, A.T., El-Metwally, S.M., "Decision tree classifiers for automated medical diagnosis," Neural Comput & Applic, Vol. **23**, pp. 2**387–2403, 2013**. DOI: https://doi.org/10.1007/s00521-012-1196-7.

[5] "*Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities*," Information Fusion, **2019**.

[6] "*International Journal of Computer Science and Information Technology Research*," ISSN 2348-**120X** (online), Vol. **8**, Issue 2, pp. **5–9,** April–June 2020. Available at: www.researchpublish.com.

[7] Dillon Chrimes, "*Interactive Journal of Medical Research*," https://www.i-jmr.org/, **30** January **2023**.

[8] "*How Healthcare Decision Trees Emerge and Function*," Published online by Cambridge University Press, **13** July **2023**.

[9] Brown, S.A., Chung, B.Y., Doshi, K. et al., "*Patient similarity and other artificial intelligence machine learning algorithms in clinical decision aid for shared decision-making in the Prevention of Cardiovascular Toxicity (PACT): a feasibility trial design*," Cardio-Oncology, Vol. **9**, No. **7**, **2023**. DOI: https://doi.org/10.1186/s40959-022-00151-0.

[10] Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, "*Decision trees: an overview and their use in medicine*."

Sharmasth Vali Y - paper-1 (2)

ORIGINALITY REPORT

11%	6%	0%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Southern Queensland Student Paper	4%
2	vbook.pub Internet Source	2%
3	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
4	journals.icapsr.com Internet Source	1%
5	Submitted to Ain Shams University Student Paper	1%
6	Submitted to The Robert Gordon University Student Paper	<1%
7	scholarworks.lib.csusb.edu Internet Source	<1%
8	ijarcce.com Internet Source	<1%
9	ia801501.us.archive.org	

Internet Source

<1%

10 www.analyticsvidhya.com
Internet Source

<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Based on the provided dataset and visualizations, the following Sustainable Development Goals (SDGs) are most relevant:

1. SDG 3: Good Health and Well-being

- **Focus:** Ensuring healthy lives and promoting well-being for all at all ages.
- **Relevance:** This project directly aligns with SDG 3 as it aims to improve patient similarity analysis, leading to:
    - **Better disease diagnosis and prognosis:** Early identification and accurate diagnosis of diseases can lead to timely interventions and improved health outcomes.

2. SDG 9: Industry, Innovation and Infrastructure

- **Focus:** Building resilient infrastructure, promoting inclusive and sustainable industrialization and fostering innovation.
- **Relevance:** The project leverages machine learning and data science, which are crucial for innovation in healthcare.
    - Developing and implementing advanced analytical tools like patient similarity analysis requires technological innovation.

3. SDG 10: Reduced Inequalities

- **Focus:** Reducing inequalities within and among countries.
- **Relevance:**
    - By identifying and addressing health disparities, patient similarity analysis can help reduce inequalities in healthcare access and outcomes.

4. SDG 17: Partnerships for the Goals

- **Focus:** Strengthening the means of implementation and revitalizing the Global Partnership for Sustainable Development.
- **Relevance:**
    - Collaboration between researchers, clinicians, and healthcare providers is crucial for the successful development and implementation of patient similarity analysis tools.
    - Partnerships with technology companies and research institutions can accelerate innovation and improve access to healthcare solutions.

By contributing to these SDGs, this project has the potential to make a significant impact on global health and well-being.