PROJECT REPORT - Group 9

FINDING FREQUENT SETS OF SYMPTOMS & PREDICTION OF HEART DISEASE

Abstract:
Heart disease is a major public health concern globally, and early identification of at-risk individuals is crucial for prevention and intervention. However, the complex nature of the disease and its numerous contributing factors make it difficult to accurately estimate the risk of heart disease. To address this challenge, we explore the use of the Apriori algorithm, a data mining technique widely used for identifying frequent item sets in large datasets, to derive frequent itemsets from a dataset of heart disease patients. We then use these frequent itemsets as features to train a binary classification model that can predict the presence of heart disease in new patients.

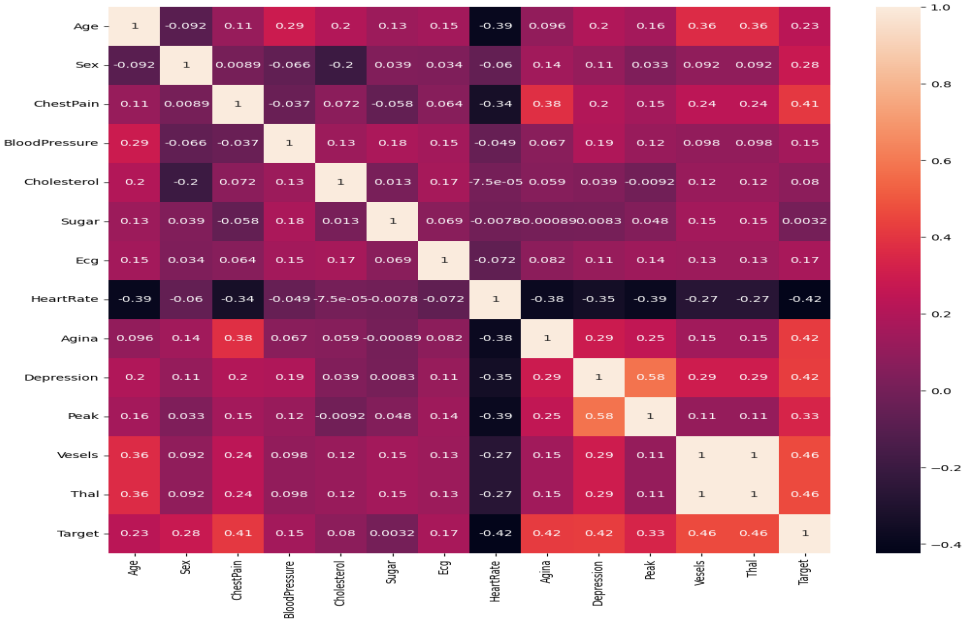
Introduction:
Early identification is essential for efficient treatment and prevention of heart disease, which is a major source of morbidity and mortality globally. By applying the Apriori algorithm to the dataset, we identified frequent itemsets (symptoms) and association rules that are indicative of the risk factors associated with heart disease. We then used the derived frequent itemsets as features to train a binary classification model to predict the presence or absence of heart disease in patients.

Related work:
Several studies have been conducted using the dataset that this project used using data mining methods to find hidden patterns in clinical datasets. Das et al. developed a neural network classifier for diagnosing valvular heart disease, which combines the predicted values from multiple models to achieve an accuracy of 97.4% on a dataset of 215 samples. Pandey et al. evaluated the performance of clustering algorithms for heart disease diagnosis and proposed a density-based clustering algorithm with a prediction accuracy of 85.8% as the most versatile. Karaolis and team developed a data mining system using association analysis based on the Apriori algorithm to assess heart-related risk factors. Different computational techniques can be used for pattern recognition in heart disease, including association rule algorithms, and various works have been done on different datasets. This project aims to add to the existing body of knowledge in this domain by attempting to find the most frequently appearing symptoms of heart disease based on our understanding of the Apriori algorithm and Decision tree method of classification. With this project, we will attempt to improve the accuracy of prediction of occurrences in comparison with existing body of knowledge.

The Dataset and Exploratory Data Analysis:

The data set used in this project is from the UCI repository and contains 76 attributes, for 4 different regions, Cleveland, Hungary,Switzerland and Long Beach. For the purpose of our study we used only the subset consisting of the Cleveland data, since it consisted of the most record and diversity suitable for our project goal upon performing some EDA. Some of the data exploratory analysis are described below.

- Heatmap showing a correlation matrix for all initial 14 columns



Notice that the 'Target' variable (categorical variable showing presence or absence of heart disease) shows high correlation with Thal, Vesels, Peak, depression, Angina and Chest pain. Considering 0.4 and above 'high'.

- Checking the 'Age' , 'blood pressure' and 'cholesterol' columns to see data distribution and detect the presence of outliers

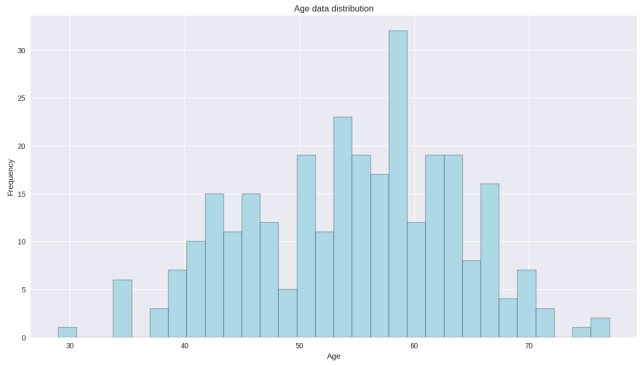


Figure: Visualization of the frequency of heart diseases at different age groups.

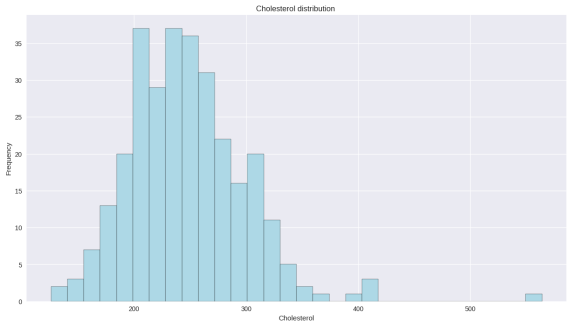


Fig:Impact of cholesterol on heart disease occurrences

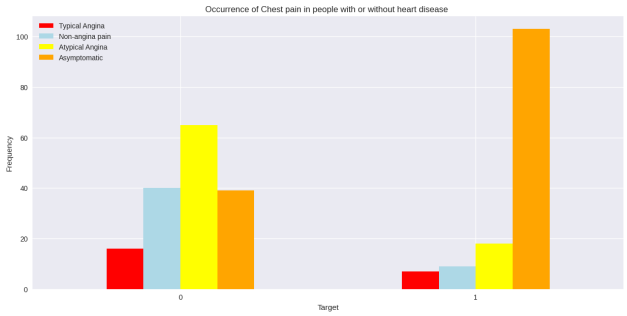


Fig:Coexistence of heart disease vs. types of angina

We see that Asymptomatic chest pain is clearly higher in patients with heart disease. And that not every type of chest pain indicates the presence of heart disease.

Methodology & Experimental Discussion:

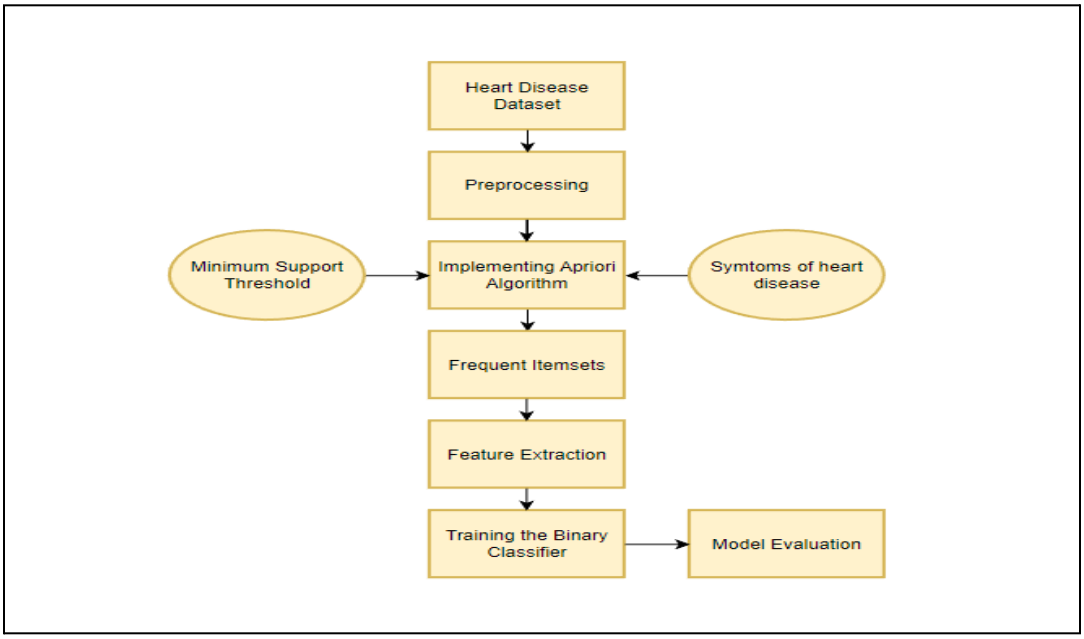


Figure: Workflow of the implementation

Part 1: Preprocessing and Exploratory data analysis:
We decided to work with the most popular dataset in the domain of heart disease prediction - the Heart Disease Data Set from the UCI Machine Learning Repository. This Dataset initially contained data collected from 4 locations- Cleveland, Long beach VA, Hungary and Switzerland over 76 attributes. After preliminary analysis, deleting rows with missing values, and checking other unrecognizable values present in the dataset, we were left with a total 1046 rows of data. During the preprocessing stage where we dropped the rows with values '?' in them, we realized that the data from 3 locations- Hungary, Long Beach and Switzerland had no rows remaining. Only the Cleveland dataset had 303 rows of usable data. This is the reason we were forced to use the data from only the Cleveland dataset. We considered imputing the values in the rows with missing values but the number of rows that had actual values were too little, so we risked losing the data integrity if we chose to impute the rows with missing values. Thus after making several more changes to the data, we were left with 297 rows of data from the Cleveland dataset to work with. We learned that this is also the reason why most of the previous research work in this area has only used the Cleveland dataset. Furthermore, we first chose the 14 attributes most relevant to our study out of 76 initial attributes. However after starting to work on the Apriori implementation, we realized it was feasible to work on only 4 columns - age, sex, sugar and angina.

Part 2: Finding the frequent sets of symptoms using Apriori :
Using a minimal support threshold of 0.1, we first tried to develop our method with all feasible features. However, we ran into a problem with zero output and narrowed our emphasis to just four characteristics: age, gender, angina, and sugar.
Step 1: We first found the frequency of singletons. And then a new RDD tuple has been formed using the lambda function which takes the first two values of each element.
Step 2: Using the map and reduceByKey transformations, we determined the frequency of various combinations of characteristics, such as age_angina and age_gender_sugar

Step 3: We applied the filter transformation to the RDD where the lambda function checks if the first value in the combination is greater than or equal to 2. The lambda function retains only those values and filters out the rest of the other tasks.
Step 4: In this step we extracted the key-value pairs from the combination of all 4 attributes using map and collectAsMap functions. Nextly, we have performed groupby to group the data using age and gender columns . The aggregate function counted the occurrences of rows with angina & sugar with value 1.
Step 5: The for loop iterates through the rows of the grouped dataframe, it extracts the age, gender, angina_count, sugar_count and calculates the support for the current age & gender combination using the get function of the age_gender_freq. At last, we calculated the confidence of angina & sugar using the calculated counts & support for the age-gender group and stored them in the confidences, confidences_agina and confidences_sugar dictionaries.

Part 3:Classification & prediction:
As discussed with the professor, we agreed to do this part in Python rather than in Spark.
Feature extraction:
We used the output of part 2 as the input dataset for binary classification and prediction. A dataset called 'datapred.csv' was created converting the spark dictionary with 'Angina' and 'Sugar' confidences into a csv file manually. Then a column 'Target' was added to this dataframe which represented the presence(1) or absence(0) of heart disease based on the confidence values of Angina and Sugar. A combination of Angina confidence > 0.3 and Sugar level confidence > 0.3 was decided to be indicating high potential of presence of heart disease. The decision of this range was based on literature review.

Classification methods:

1) **KNN with k=3**
   The classifier gave an Accuracy score of 0.75
   & ROC AUC score 0.6666666666666667
2) **Decision tree classification**
   The classifier gave an Accuracy score of 1
   & ROC AUC score 0.5

Model evaluation methods :

- Precision, Recall F1-score :
  Precision measures the fraction of true positives out of all positive predictions. Recall measures the fraction of true positives out of all actual positives and F1 score is the weighted average of precision and recall.
  KNN

	precision	recall	f1-score	support
No disease	0.88	1.00	0.93	7
Disease	0.00	0.00	0.00	1
accuracy			0.88	8
macro avg	0.44	0.50	0.47	8
weighted avg	0.77	0.88	0.82	8

DECISION TREE CLASSIFIER

	precision	recall	f1-score	support
No disease	1.00	1.00	1.00	7
Disease	1.00	1.00	1.00	1
accuracy			1.00	8
macro avg	1.00	1.00	1.00	8
weighted avg	1.00	1.00	1.00	8

- Confusion matrix :
  Confusion matrix is used to show the number of correct and incorrect predictions made by the model, comparing predicted labels to the true label. For KNN, we obtained 1 true positive, 2 false positives, no false negatives and 5 true negatives. From decision tree classifiers, we have 3 true positives, no false positives and false negatives and 5 true negatives.

Conclusion:
In the end, we were able to forecast the accuracy using KNN and decision trees after using the apriori approach. Our accuracy for KNN was 75%, and for the decision tree classifier it was 1.0. However, this accuracy rate does not efficiently represent the fit of the model because of the very small size of input data. Overall, the project's findings show how data mining methods like Apriori have the potential to increase the precision and effectiveness of heart disease prediction. For the purpose of creating more complex and precise cardiac disease prediction models, future research in this field may investigate the use of further data mining algorithms and machine learning approaches. Additionally, increasing the dataset's size to incorporate more patient information and a wider variety of symptom characteristics may help to shed light on other heart disease risk factors and enhance the precision of the model.

Contribution:

Name	**Task**
Anika Raisa Chowdhury	Tasks : 2,4
Fariha Danish	Tasks : 3,4
Pratiksha Gaikwad	Tasks: 2,3,4
Sayali Gaurav Agalave	Tasks: 2,4
Anubhuti Hiwase	Tasks: 1,3,4

Tasks:
1. Preprocessing and exploratory data analysis
2. Training and testing smaller samples from preprocessed data to find relevant samples (Apriori candidate pairs) for use cases of our algorithm .Design the logic for disease prediction - Apriori algorithm. Design and implement the algorithm on samples .
3. Prediction of heart disease using two classification methods. Model evaluation.
4. Report writing & Presentation

References:
[1] R. Das, I. Turkoglu, and A. Sengur, “Diagnosis of valvular heart disease through neural networks ensembles,” Elsevier, 2009.
[2] M. Karaolis, J. A. Moutiris, and C. S. Pattichis, “Association rule analysis for the assessment of the risk of coronary heart events,” Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009.
[3] A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, “Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method,” International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2277798, Vol 2, Issue10, October 2013
[4]Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Comput Intell Neurosci. 2021 Jul 1;2021:8387680. doi: 10.1155/2021/8387680. PMID: 34306056; PMCID: PMC8266441.
https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8266441/