1. Project Title

Heart Disease Prediction Using Data Mining Algorithm and Patient Segmentation

2. Names and Email Addresses Of Authors

Names: Ferial Najiantabriz, Subankar Chowdhury, Ujwala Vasireddy

Email Addresses: ferial@ou.edu, Subankar.Chowdhury-1@ou.edu, Ujwala.Vasireddy-1@ou.edu

3. Category and Objectives of the Project

We chose Category B because it allows us to thoroughly explore and compare multiple methods for predicting heart disease, which is essential given the complex nature of our dataset. Testing different methods is important as each may capture certain risk factors differently, helping us gain deeper insights into the data's impact on predictions. This evidence-based approach lets us assess performance through accuracy and reliability, ensuring we develop a robust model that is effective for medical decision-making.

Individual Objectives

- We will clean and preprocess the dataset, addressing missing values, normalizing data, and encoding categorical features.
- We will test and refine Logistic Regression, Decision Trees, and Random Forest to improve heart disease prediction.
- We will evaluate the models using accuracy, precision, and recall to identify the most effective one.
- We will develop an interactive visualization tool with Streamlit or Shiny to explore model performance and predictions.

4. The Significance of the Project

4.1 Application and its Significance

Application: The application developed in this project will predict heart disease risk based on medical features like age, cholesterol, and blood pressure, helping hospitals quickly assess and identify high-risk patients.

Significance: Heart disease is a major public health issue. The ability to predict the likelihood of heart disease based on patient data allows for timely interventions and treatments. With an accurate predictive model, healthcare providers can prioritize patients who need further medical attention, leading to better outcomes and potentially reducing mortality rates [Ali et al., 2021].

Data Mining Questions to be Answered

1. Can patient heart disease risk be accurately classified based on their medical attributes?
2. Which attributes contribute the most to predicting heart disease?
3. How do different classification algorithms compare in terms of predictive performance?
4. Can we predict whether a patient has heart disease based on demographic and medical features?

4.2 Why These Questions Require Data Mining

1. Can patient heart disease risk be accurately classified based on their medical attributes?

   Why it requires data mining: This task involves predicting heart disease based on medical attributes that require data mining techniques to build models that learn from past data and make accurate predictions for new patients. As discussed in [Tan et al., 2019], the principles of data mining provide the foundation for applying classification

algorithms to predict patient outcomes in heart disease.

Types of algorithms/tasks needed: We chose Logistic Regression for its widespread use in binary classification, making it ideal for predicting heart disease. It effectively models the relationship between patient attributes and disease likelihood.

2. Which attributes contribute the most to predicting heart disease?

   Why it requires data mining: To identify the most important features, we need to analyze the relationships between the input variables and the target outcome. Data mining helps us find which features have the biggest impact on prediction accuracy and remove those that aren't as useful.

   Types of algorithms/tasks needed: We choose Decision Tree algorithm because it can identify the most important attributes for predicting heart disease. It works by splitting the data into branches based on different features, showing which ones, like age, cholesterol, and chest pain type, have the biggest impact on heart disease risk. [Reddy et al., 2021][Oyeleye et al., 2022].

3. How do different classification algorithms compare in terms of predictive performance?

   Why it requires data mining: Comparing the performance of classification algorithms involves training each model on the same dataset and evaluating them using metrics like accuracy, precision, and recall [Ali et al., 2021][Almustafa, 2020]. Data mining is essential to experiment with different algorithms and identify the most effective one.

   Types of algorithms/tasks needed: We will implement and evaluate Logistic Regression, Decision Trees, and k-Nearest Neighbors (k-NN), comparing their performance using accuracy, precision, and recall [Al-Alshaikh et al., 2024] to select the most reliable model for predicting heart disease.The importance of algorithm comparison in healthcare decision support systems has been demonstrated, showing how data mining can improve the accuracy and reliability of medical predictions.[Dey & Rautaray, 2014]

4. Can we predict whether a patient has heart disease based on demographic and medical features?

   Why it requires data mining: This is another binary classification problem where the task is to predict whether a patient has heart disease based on data. Data mining methods are essential to train models that can identify patterns in this data to make accurate predictions.

   Types of algorithms/tasks needed: We chose Naive Bayes for heart disease prediction because it's simple, fast, handles categorical data well (like gender and lifestyle factors), and is highly effective for making quick predictions, especially when features are not strongly correlated [Almustafa, 2020] [Reddy et al., 2021]

4.3 Dataset Description

The dataset contains 1,025 records with 14 attributes related to heart health, totaling approximately 12 KB in CSV format. Each record is approximately 40 bytes, including CSV formatting overhead. A sample of records is provided for reference, and the dataset is sourced from the: Heart Disease Dataset on Kaggle.

S.N	Variable	Meaning	Description
1	age	Age of the patient	Age of the patient in years
2	sex	Gender of the patient	1 = male; 0 = female
3	cp	Chest pain type	0-3 indicating different types of chest pain
4	trestbps	Resting blood pressure	Resting blood pressure in mm Hg
5	chol	Serum cholesterol	Serum cholesterol in mg/dl
6	fbs	Fasting blood sugar	1 if fasting blood sugar is >120 mg/dl, else 0
7	restecg	Resting electrocardiographic results	0-2 indicating ECG results
8	thalach	Maximum heart rate achieved	Maximum heart rate
9	exang	Exercise-induced angina	1 = yes; 0 = no
10	oldpeak	ST depression	ST depression induced by exercise
11	slope	Slope of the ST segment	The slope of the peak exercise ST segment (0-2)
12	ca	Number of major vessels	Number of major vessels colored by fluoroscopy (0-3)
13	thal	Thalassemia	0 = normal; 1 = fixed defect; 2 = reversible defect
14	target	Heart disease presence	1 = heart disease; 0 = No heart disease

Figure 1: Dataset Table Of Format.

Age	Sex	Chest Pain Type	Resting BP	Cholesterol	Fasting Blood Sugar	Resting ECG	Max Heart Rate	Exercise Induced Angina	ST Depression	Slope of ST Segment	Major Vessels	Thalassemia	Target
52	1	0	125	212	0	1	168	0	1	2	2	3	0

Table 1: Sample Record from Heart Disease Dataset

Our target attribute is 'target', which indicates whether or not a patient has heart disease.

5. Implementation and Time Table

The heart disease prediction project will start by collecting and preprocessing data from Kaggle, followed by implementing and fine-tuning three machine learning algorithms: Logistic Regression, Decision Trees, and k-Nearest Neighbors. After evaluating model performance and conducting patient segmentation, an interactive visualization tool will be created using Streamlit or Shiny to display prediction results.

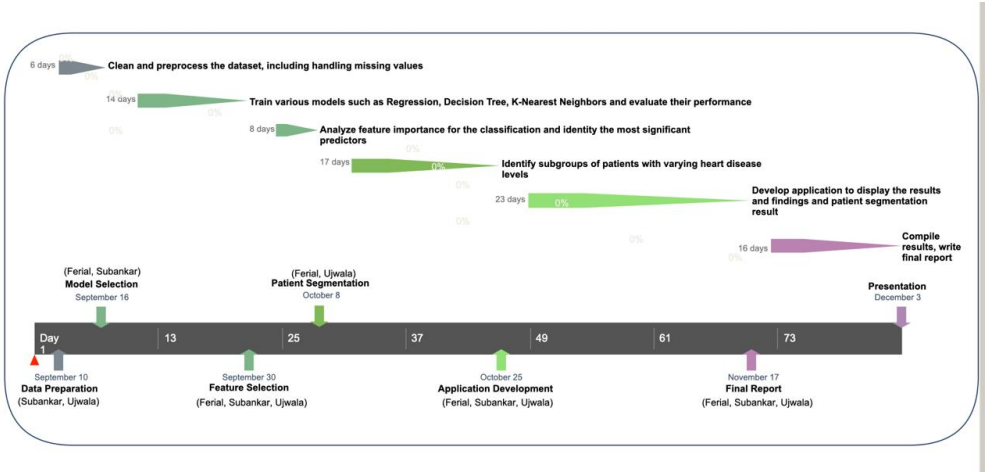

Figure 2: Time Table.

6. References

[1] Al-Alshaikh, H.A., P P, Poonia, R.C., Saudagar, A.K.J., Yadav, M., AlSagri, H.S., AlSanad, A.A. "Comprehensive evaluation and performance analysis of machine learn-

ing in heart disease prediction." *Scientific Reports*, 14(1):7819, April 2024. Available at: https://doi.org/10.1038/s41598-024-58489-7. Accessed on 09/07/2024.

[2] Almustafa, K.M. "Prediction of heart disease and classifiers' sensitivity analysis." *BMC Bioinformatics*, 21:278, 2020. Available at: https://doi.org/10.1186/s12859-020-03626-y. Accessed on 09/03/2024.

[3] Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M.W., Moni, M.A. "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison." *Computers in Biology and Medicine*, 136:104672, September 2021. Available at: https://doi.org/10.1016/j.compbiomed.2021.104672. Accessed on 09/07/2024.

[4] Dey, A., Rautaray, S. S., *Study and analysis of data mining algorithms for healthcare decision support system*, International Conference on Computer, Communication, Control and Information Technology (C3IT), IEEE, 2014, pp. 1–5. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d32ec14e005d9907603d2e46daa1b68a9b63d95b. Accessed on 09/01/2024.

[5] Oyeleye, M., Chen, T., Titarenko, S., Antoniou, G. "A Predictive Analysis of Heart Rates Using Machine Learning Techniques." *International Journal of Environmental Research and Public Health*, 19(4):2417, February 2022. Available at: https://doi.org/10.3390/ijerph19042417. Accessed on 09/01/2024.

[6] Reddy, K.V.V., Elamvazuthi, I., Aziz, A.A., Paramasivam, S., Chua, H.N., Pranavanand, S. "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators." *Applied Sciences*, 11(18):8352, September 2021. Available at: https://doi.org/10.3390/app11188352. Accessed on 09/01/2024.

[7] Tan, P.-N., Steinbach, M., Karpatne, A., Kumar, V. *Introduction to Data Mining*, 2nd Edition, Pearson Education, Inc., 2019.