PROJECT PROPOSAL

Nagavardhan Bitragunta (Artificial Intelligence)

Omkar Datta Sowri Vullaganti (Artificial Intelligence)

Manasa Tallapaka (Artificial Intelligence)

Jagruth Reddy Palle (Artificial Intelligence)

Predictive Modeling and Machine Learning - 02 (Fall 2024)

Noor Al-Hammadi

Saint Louis University

Dataset for Classification problem:

1.Source of the data:

The classification analysis dataset that I recommend is related to heart issues is the Heart Disease UCI dataset, which is available on the UCI Machine Learning Repository.

Source: UCI Machine Learning Repository, Heart Disease dataset.

Link : Heart Disease - UCI Machine Learning Repository

This dataset contains information about patients, with the target variable being whether they have heart disease. It's a binary classification problem, making it ideal for this project.

Permission to share the data:

Yes, the heart disease dataset is publicly available for academic and research purposes. It can be shared and used freely in the UCI Machine Learning Repository.

Motivation for choosing this dataset:

The motivation behind selecting this dataset is because it's relevant to the healthcare domain and has the potential for developing predictive models that can help us in the early detection of heart disease. Heart disease is one of the leading causes of death globally, and this dataset provides valuable clinical variables that can be used to create models aimed at diagnosing whether a patient has heart disease.

This dataset has a good balance of categorical and continuous features, and its real-world importance makes me to choose this classification task.

Variables Table

Variable Name	Role	Type	Demographic	Description	Units
age	Feature	Integer	Age		years
sex	Feature	Categorical	Sex		
cp	Feature	Categorical			
trestbps	Feature	Integer		resting blood pressure (on admission to the hospital)	mm Hg
chol	Feature	Integer		serum cholestoral	mg/dl
fbs	Feature	Categorical		fasting blood sugar > 120 mg/dl	
restecg	Feature	Categorical			
thalach	Feature	Integer		maximum heart rate achieved	
exang	Feature	Categorical		exercise induced angina	
oldpeak	Feature	Integer		ST depression induced by exercise relative to rest	

Image from UCI Website.

The dataset consists of both continuous and categorical variables, satisfying the project requirements.

For example: age, trestbps, chol, thalach are continuous, while sex, cp,fbs,restecg,exang are categorical.

5. What will be the goals of your analyses?

The primary goal of this analysis is to build a machine learning model that predicts whether a patient has heart disease based on clinical and demographic variables. Specifically, this analysis will aim to address the following questions:

1. Which factors are most predictive of heart disease? (For example, does chest pain type play a bigger role in determining heart disease risk?)
2. How accurately can machine learning models such as logistic regression, decision trees, or random forests classify whether a patient has heart disease?
3. Can we improve classification accuracy by using techniques like feature selection or dimensionality reduction?
4. How do categorical features such as thalassemia and chest pain type interact with continuous variables like resting blood pressure to impact the model's predictions?
5. What is the overall performance of the model (in terms of accuracy, precision, recall, F1-score) when predicting heart disease?

This analysis plays a crucial role in medical diagnosis and helps identify high-risk patients early on, enabling timely intervention and potentially saving lives. Through this project, we will be able to analyze important medical data and contribute to a solution for a significant healthcare problem.

This heart disease classification task will use machine learning techniques to derive insights that could have practical applications in the medical field.

Dataset 1 for Regression problem:

1. **Source of the data:** The regression analysis dataset that I recommend for the study is the ‘‘Boston Housing’’ dataset. This is an open-source dataset which is accessible from different sources such as UCI Machine Learning Repository and it can be instated in MASS package in R.

2. **Permission to share:** Yes, this dataset is publicly available and can be shared but, it is kindly requested to acknowledge from this source. The original source should be cited as: Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. J. Environ. Economics & Management, 5, 81-102.

3. **Motivation for choosing this dataset:** I selected the Boston Housing dataset as it aligns with the following criteria or requirements: It is relevant for regression analysis, widely utilized in machine learning education and learning, and available for broader use in the field of study. It gives a realistic idea of understanding and predicting the house prices based on its features which is very useful in real world.

4. **Details of the variables:**

Variable Name	Measurement Type	Role
CRIM	Continuous	Predictor
ZN	Continuous	Predictor
INDUS	Continuous	Predictor
CHAS	Categorical (Binary)	Predictor
NOX	Continuous	Predictor
RM	Continuous	Predictor
AGE	Continuous	Predictor
DIS	Continuous	Predictor
RAD	Categorical (Ordinal)	Predictor
TAX	Continuous	Predictor
PTRATIO	Continuous	Predictor
B	Continuous	Predictor
LSTAT	Continuous	Predictor
MEDV	Continuous	Output

5. **Goals of the analysis:** Our target variable will be 'MEDV' or, in other words, the median value of owner-occupied houses for which other features need to be predicted. Specific questions to address include:

    1. Of which features are the most powerful predictors of housing prices?
    2. What sort of accuracy of the result is possible when using these features for forecasting housing prices?
    3. Are there any associations which are not proportional between the predictors and the target variable?
    4. Are there any clusters or some pattern in the housing market which can be detected by unsupervised learning techniques?