Project on Heart Disease Prediction

Healthcare is a major domain where data mining techniques are widely used. If you are curious about data mining projects in healthcare, you should explore the heart disease dataset.

Dataset: The dataset contains 75 particulars of 303 people. These particulars include parameters related to an individual's heart health like age, gender, serum cholesterol, blood sugar, etc.

The project implementation steps will go as follows:

- Understand the dataset attributes
- Apply the required data cleaning methods
- Implement different classification models to investigate the performance of each classifier on heart disease datasets.
- Mention your observations and study the parameters that play a vital role in determining the health condition of people's hearts (ex. Gender-based analysis, age-based analysis, … etc.)
  (What percentage of younger people are prone to be diagnosed with heart disease?
  Are women more prone to heart diseases, or is it the other way? …. etc.)

Project on Loan Prediction

The idea behind this ML project is to build a model that will classify Loan status for each customer who can take loan or not

It is based on the user's marital status, education, number of dependents, and employments and etc.

Dataset: The dataset "Train data" contain features of customers like marital status, education, number of dependents, and employments and etc. to detect the status of loan.

The project implementation steps will go as follows:

- Understand the dataset attributes
- Apply the required data preprocessing methods
- Implement different classification models to investigate the performance of each classifier on detecting the loan status for each customer
- Mention your observations and study the parameters (features) that play a vital role in accepting loan.
- After get the best classifier. Use "New Customer "data to predict the loan for each new customer
- From the new customer data, what is the percentage of married people in semiurban area that obtained the loan?

Project on a Fake News Detection

With the advent of the technological revolution, it is easier for users to have access to the internet which increases the probability of fake news to spread like a wildfire. In this project, you will learn how to classify news into Real or Fake.

Dataset: The dataset contains a list of 1000 text labeled with "fake" or “realnews”.

The project implementation steps will go as follows:

- Apply the required data cleaning methods
- Apply data preprocessing on text data
- Compute term frequency
- Implement different classification models to investigate the performance of each classifier for detect fake or real news.
- Mention the best model to detect the fake news

Project on Diabetes Prediction

Diabetes is one of the most common and hazardous diseases on the planet. It requires a lot of care and proper medication to keep the disease in control. If you are curious about data mining projects in healthcare, you should explore the diabetes dataset.

Dataset: The data consist of medical information, laboratory analysis… etc. The data that have been entered initially into the system are: No. of Patient, Sugar Level Blood, Age, Gender, Creatinine ratio(Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides(TG) and HDL Cholesterol , HBA1C, Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Predict Diabetic).

The project implementation steps will go as follows:

- Understand the dataset attributes
- Apply the required data cleaning methods
- Implement different classification models to investigate the performance of each classifier on diabetes datasets.
- Mention your observations and study the parameters (features) to determine the major factors affecting the onset of diabetes
- (What percentage of younger people are prone to be diagnosed with diabetes disease?

  Are women more prone to diabetes, or is it the other way? …. etc.)

Project on Customer Segmentation

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base.

“Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits”.

Dataset: The data contain some features about the customers of Mall.

The project implementation steps will go as follows:

- Understand the dataset attributes
- Apply the required data cleaning methods
- Implement unsupervised techniques to separate customers into several groups.
- Apply different techniques, mention the output for each technique and visualize it in graph

Project on Water Quality

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. Our goal is to classify the water potability

Dataset: contains water quality metrics for 3276 different water bodies with 9 factors affecting the potability of the water (which Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.)

The project implementation steps will go as follows:

- Understand the dataset attributes
- Apply the required data preprocessing methods
- Divide your data set to training and testing data
- Implement the classification model that suits the project best
- investigate the performance of your classifier on detecting the quality of water
- Mention your observations and study the parameters (features) that play a vital role in water potability.
- After building the classifier model. Use the testing data set to predict the water potability and calculate the error percentage

Project on Stroke Prediction

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

Dataset: is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

The project implementation steps will go as follows:

- Understand the dataset attributes
- Apply the required data preprocessing methods
- Divide your data set to training and testing data
- Implement the classification model that suits the project best
- investigate the performance of your classifier to predict whether a patient is likely to get stroke
- Mention your observations and study the parameters (features) that play a vital role in predict whether a patient is likely to get stroke
- After building the classifier model. Use the testing data set to predict whether a patient is likely to get stroke and calculate the error percentage

Project on College Scorecard

This project provides data to help students and families compare college costs and outcomes as they weigh the tradeoffs of different colleges, accounting for their own needs and educational goals.

The project implementation steps will go as follows:

- Understand the dataset attributes
- Apply the required data preprocessing methods (feature Selection and data cleaning ...etc.)
- Use cluster analysis to identify the groups of characteristically similar schools in the College Scorecard dataset. Mention your observations and study the parameters (features) that play a vital role in clustering.

Binary Classification Prediction

for type of Breast Cancer

Description: Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area. The key challenges against its detection are how to classify tumors into malignant (cancerous) or benign (non-cancerous).

Dataset: a dataset contains data from patients with breast tumors, either benign or malignant. The data contains many features and each patient has been labelled as 'B' (Benign) or 'M' (Malignant).

The project implementation steps will go as follows:

- Understand the dataset features (attributes)
- Apply the required data preprocessing methods
- Build classification models to predict whether the cancer type is Malignant or Benign and compare the evaluation metrics of various classification algorithms.
- Mention your observations and determine which factors are more prominent in deciding the type of Breast Cancer.

Mushroom Classification

Many people avoid eating mushrooms as they don't have an excellent idea of which mushrooms are poisonous and edible. It thus becomes essential to understand different types of mushrooms so that everyone can enjoy the taste of mushrooms without any worries.

Dataset: a dataset on mushrooms that contains interesting information about different types of mushrooms. The dataset mostly has physical features of the mushrooms like cap color, cap shape, gill color, gill shape, etc. Each mushroom has been labelled as 'e' (edible) or 'p' (poisonous).

The project implementation steps will go as follows:

- Understand the dataset features (attributes)
- Apply the required data preprocessing methods
- Implement different classification models to investigate the performance of each classifier to determine if the type of mushroom is edible (non-poisonous) or poisonous.
- Mention your observations and determine which factors are more prominent in deciding the nature of mushrooms.