Data Science and Big Data Projects

Important note:

We will check your projects by cheating tool. Please do your project with your efforts, each group is separate from the other group. When we realize that there is a case of cheating, this team will be taken zero in the project grade.

Team Members from3 to 4.

Project 1: Predicting Housing Price

Requirements from the student:

Use R language to build a model that predicts the price of a house based on its features (e.g., number of bedrooms, square footage, location).
Evaluate the accuracy of the model using metrics such as mean squared error (MSE) or root mean squared error (RMSE).
Dataset link:
Here's a sample dataset from Kaggle that can be used for this project:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Steps to complete the project:

- Download the dataset from the above link and import it into R using the read.csv() function.
- Clean the data by removing any unnecessary columns or rows, and fixing any errors or inconsistencies.
- Use R libraries to manipulate the data and create new features if necessary.

- Use R libraries \o build a regression model that predicts the housing prices based on the selected features.
- Evaluate the accuracy of the model using metrics such as mean squared error (MSE) or root mean squared error (RMSE).
- Experiment with different machine learning algorithms (e.g., random forest, support vector machines) to see which one performs best.
- Explore the impact of different features on the predicted housing prices by removing or adding variables to the model.

Present the findings of the project in a written report.

Project 2: Fake News Detection

Do you trust all the news you hear from social media? All news are not real, right? So we need to detect the news data to know the new if it is fake or real news.

Steps to complete the project:

- Download the dataset named “news” and import it into R.
- Clean the data by removing any unnecessary columns or rows, and fixing any errors or inconsistencies.
- Use R libraries to manipulate the data and create new features if necessary.
- Use R libraries \o build a classification model that detect the news based on the selected features.
- Evaluate the accuracy of the model .
- Experiment with different machine learning algorithms (more than one Classification algorithm) to see which one performs best.
- Explore the impact of different features on the detect news data by removing or adding variables to the model.

Present the findings of the project in a written report.

Project 3: Clustering Heart Disease Patients

Description

Doctors frequently study former cases to learn how to best treat their patients. A patient who has a similar health history or symptoms to a previous patient could benefit from undergoing the same treatment. This project investigates whether doctors might be able to group together patients to target treatments using common unsupervised learning techniques.

Steps to complete the project:

- Download the dataset named “heart disease patients” and import it into R
- Clean the data by removing any unnecessary columns or rows, and fixing any errors or inconsistencies.

- Use R libraries to manipulate the data and create new features if necessary.
- Use R libraries to split data set into groups based on the selected features.
- Experiment with different machine learning algorithms (more than one Clsutering algorithm) to see which one performs best.
- Evaluate the best number of clustering based on your data sets .
- Explore the impact of different features on the cluster your data by removing or adding variables to the model.

Present the findings of the project in a written report.