Kingdom of Saudi Arabia
Ministry of Education
UMM AL-QURA University

Data Mining (14016165-3)
First Assignment
First semester 1441

Course name	Data mining - 14016165-3 / Data mining - 14016313-3
Assignment tittle	Assignment 1 : Classification problem
due date	03-12-2020 (Week 14).
Assignment weight	20%

Please perform the following tasks:

Stage	Task
Downloading the Dataset	1. Download the cover type dataset from the following link: https://datahub.io/machine-learning/covertype 2. Describe the dataset and the classification task, more information about the dataset can be found in UCI repository. https://archive.ics.uci.edu/ml/datasets/Covertype
Data Exploration	3. Display the number of instances. 4. Display the number of attributes. 5. Display the number of classes. 6. For each class label, display the code of the class label and the name of that class. 7. Summarise the class distribution using a suitable graph. 8. Display a statistical summary for all the attributes.
Data Preprocessing	9. Check whether the selected dataset has any data quality issues and choose suitable strategies to deal with any issue (if exists). 10. Convert the multiclass classification problem into a binary classification problem. 11. Use a features selection technique to select those features in your data that contribute most to the prediction. 12. Divide your dataset into training, validation and testing datasets.
Classification	13. Build classification models. a. Use three different learning algorithms to generate three classification models. You should choose one learning algorithm from each of the following categories: i. {Decision Tree} ii. {Nearest Neighbor Classifier, Naïve Bayes Classifier, Support Vector Machine} iii. {Bagging, Boosting, Random Forest} 14. For each classification model: a. Try to find the most accurate classifier (avoid overfitting).
Evaluation	15. Evaluate your classification models on the validation and the testing datasets. a. For each classification model, print out a confusion matrix for the validation and testing datasets. b. Use the following evaluation measures to evaluate the performance of the generated classification models: i. Accuracy ii. Error rate iii. *F* -measure 16. Compare between the performances of all the classification models using suitable chart (The type of chart should be different from the type of the chart that is used in the data exploration stage).

Important notes:

- This an individual assignment.
- You need to use the Jupyter Notebook to perform the all the required tasks.
- The ipynd file name should be in the following format: (first name)_(last name)_classification.ipynb for example Majed_Farrash_classification.ipynb
- By the due date, you must submit your ipynb file using the blackboard.