CSCE 5300 Introduction to Big data and Data Science

ICE-8

Lesson Title: Machine Learning

Lesson Description: machine learning models (Decision Tree, Random Forest)

You can use google colab (https://colab.research.google.com) or Jupiter notebook (run spark on your own laptop) run different on the given dataset and explain the models or algorithms.

Source code given on Canvas.

1. Run decision tree model on the dataset. Change the hyperparameters:

A) Complete coding parts of evaluation methods.

B) Change the depth of the tree and report your results.

C) Explain how this model works.

D) With reference to your explanation, split the data into appropriate training and testing sets. Pick an optimal depth, State clearly how much (%) of data you have used in training and testing. Play around with these percentages and report the optimal set. Test at least 3 different scenarios and report your findings also explain why the model behaved this way in each case.

Use the below table to record your findings:

Train Test Split	**Depth of Tree**	**Findings**	**Why this happened**

E) Create a confusion matrix for any one of your results and calculate the Precision, Accuracy, Recall and F1-Score.

F) State and explain the 4-performance metrics used for evaluating classifiers i.e., Precision, Accuracy, Recall and F1-Score.

2. Run random forest model on the dataset. Change the hyperparameters:

A) The number of trees to improve the performance.

B) Complete coding parts of evaluation methods.

C) Explain how models work and the reason for improvement of performance.

D) Explain how we can improve the performance of this model apart from the above methods.

E) Create a confusion matrix for any one of your results and calculate the Precision, Accuracy, Recall and F1-Score.

3. Implement the Naïve Bayes model on the dataset.

A) Complete coding parts of evaluation methods.

B) Explain how models work and the reason for improvement of performance.

C) Explain how we can improve the performance of this model apart from the above methods.

D) Create a confusion matrix for any one of your results and calculate the Precision, Accuracy, Recall and F1-Score.

E) Explain the difference between supervised and unsupervised learning and explain the various methods associated with both learning methods.

ICE Submission Guidelines

1. ICE Submission is individual.

2. ICE code must be properly commented on.

3. The documentation should include screenshots of your code/queries and results.

4. Provide an explanation of the exercise for each question as per your understanding.

5. The similarity score for your document should be less than 15%.

6. Submit the source code (if any) properly commented and documentation (.pdf/.doc)

with explanation and screenshot of source code/queries having input logic and output.