University of Ottawa

School of Electrical Engineering and Computer Science

CSI4142 Fundamentals of Data Science

Project Phase 4: Data Mining

Due Date: April 5, 2024, 11:59pm

Instructions:

A. This is a team assignment. Use the ScikitLearn library to complete this assignment: https://scikit-learn.org/stable/index.html
B. Submit your documentation via BrightSpace using your team locker.
C. For your source code, you may either submit a zipped file or provide a link to a GitHub repository.

Part A. Data summarization, data preprocessing and feature selections:

1. An initial step of any data mining project involves exploring and summarizing the data to get a "feel" of the data. To this end, your team should conduct data summarization using techniques such as scatter plots, boxplots, and histograms to visualize and to explore attribute characteristics.
2. In addition, data pre-processing involves data transformation, including:
   - Handling missing values through e.g., imputation. (If not handled in the previous phase).
   - Handling categorical attributes through e.g., one-hot encoding or conversion to ordinal data,
   - Normalization of numeric attributes to ensure all attributes are of equal importance during learning, and
   - Feature selection to remove potentially redundant attributes.

Some relevant links:

https://scikit-learn.org/stable/index.html
https://www.postgresqltutorial.com/postgresql-python/connect/
https://scikit-learn.org/stable/modules/impute.html
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
https://scikit-learn.org/stable/modules/feature_selection.html

Deliverable Part A: Submit one page of summary to explain how you preprocessed the data. Your notes should detail any data transformation and data quality issues that you encountered.

Part B. Classification (Supervised Learning):

Note: This part does not apply to all projects; you need to consult your TA in case you are not sure that you need to do this part.

Next, conduct supervised learning using a label of your own choice. That is, you are required to identify your own classification task.

Complete the following steps:

1. Use the Decision Tree, Gradient Boosting and Random Forest algorithms to construct models against your data, following the so-called train-then-test, or holdout method.
2. Compare the results of the three learning algorithms, in terms of (i) accuracy, (ii) precision, (iii) recall and (iv) time to construct the models.
3. Submit a 200 to 300 words summary explaining the actionable knowledge nuggets your team discovered. That is, you should explain what insights you obtained about the data, when investigating the models produced by the three algorithms.

Some relevant links:

https://scikit-learn.org/stable/modules/tree.html (general discussion)
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html
https://scikit-learn.org/stable/modules/generated/sklearn.tree.export_text.html (useful to display the models in the form of rules)
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

Deliverables Part B:

1. Submit all your source code, either by uploading it to BrightSpace or providing us with a link to a GitHub repository.
2. Submit a PDF file for Part B.2 consisting of a table containing the (i) accuracy, (ii) precision, (iii) recall and (iv) time to construct of models,

constructed by the three algorithms and a 200 words summary explaining how, and motivating why, you would rank the quality of the models produced by the three algorithms.

3. Submit a PDF file containing your summary for Part B.3.

Part C. Detecting Outliers: (Bonus)

Complete the following steps:

1. Use the one-class SVM algorithm (or any algorithm of your choice) to identify global outliers in your data.
2. Write a 200 to 300 words summary detailing the outliers your team discovered. That is, you should describe how you identified the outliers and explain what insights you obtained from the data.

A relevant link: https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html

Deliverables for Part C:

1. Submit your source code either by uploading it to BrightSpace or providing us with a link to a GitHub repository.
2. Submit a PDF file containing your summary for Part C.2.