

**University of Ottawa**  
**School of Electrical Engineering and Computer Science**  
**CSI4142 Fundamentals of Data Science**  
**Project Phase 4: Data Mining**

**Total Marks: 80 (+ 15 optional bonus marks)**

Instructions:

1. This is a team assignment. Use the ScikitLearn library to complete this assignment: <https://scikit-learn.org/stable/index.html>
2. Submit your work in BrightSpace using your team locker.
3. You may either submit a zipped file or provide a link to a GitHub repository.
4. Demonstrate your work during a Zoom meeting with the TA, in the timeslot allocated to you. Note that all team members are required to attend this demonstration and you will be asked to turn your cameras on.

**Project Description - Covid-19 Tracking and Lifestyle Trends Data Mart**

Data Science and Artificial Intelligence (AI) have been successfully used to study trends in our behaviours over time. However, our daily routines changed abruptly with the onset of the COVID-19 pandemic. In Canada, and many other countries, lockdown procedures were implemented, thus leaving citizens with little choice to adapt their lifestyles accordingly. For instance, people increasingly turned to online shopping, while participation in outdoor activities increased. Many non-essential businesses, notably in the hospitality sector, also adapted by offering only delivery or curbside pickup, leading to changes in consumer behaviour and traffic patterns.

During Phase 2 of the project, you created the database and staged the data, while you explored the data using OLAP queries in Phase 3. In this deliverable, you are required to explore the data using data mining techniques. Refer to the lecture slides, as well as the practical lecture as presented by the TA.

## **Part A. Data summarization, data preprocessing and feature selections: 30 marks**

1. (10 marks) 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 summarisation using techniques such as scatter plots, boxplots, and histograms to visualise and to explore attribute characteristics.
2. (20 marks) In addition, data preprocessing involves data transformation, including:
  - handling missing values through e.g., imputation,
  - handling categorical attributes through e.g., one-hot encoding or conversion to ordinal data,
  - normalisation of numeric attributes to ensure all attributes are of equal importance during learning,
  - feature selection to remove potentially redundant attributes, and
  - undersampling of the majority class(es) to balance the class distributions of the classification task your team selected in Part B.

Some relevant links:

<https://www.postgresqltutorial.com/postgresql-python/connect/>

<https://scikit-learn.org/stable/modules/impute.html>

[https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)

[learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)

[https://scikit-learn.org/stable/modules/feature\\_selection.html](https://scikit-learn.org/stable/modules/feature_selection.html)

<https://github.com/scikit-learn-contrib/imbalanced-learn> (separate repository)

**Deliverable Part A:** Submit one page of notes 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): 50 marks**

Next, conduct supervised learning using a label of your own choice. That is, you are required to identify your own classification task. For instance, you may consider using any one of the following attributes as your class (note that this list is not complete): Gender, Age-group, Month, Special-Measure-type, and so on.

Complete the following steps:

1. (15 marks) 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. (20 marks) 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. (15 marks) 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 Covid-19 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.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.plot_tree.html)

[https://scikit-learn.org/stable/modules/generated/sklearn.tree.export\\_text.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](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html)

### **Deliverables Part B:**

1. Submit all your source code for step 1.
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.

The following task is optional and may earn you up to 15 marks extra for this deliverable. That is, you will be able to earn up to 95/80 for deliverable 4.

### **Part C. Detecting Outliers: 15 marks (optional)**

Complete the following steps:

1. (10 marks) Use the one-class SVM algorithm to identify global outliers in your data.
2. (5 marks) 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.
2. Submit a PDF file containing your summary for Part C.2.