Project Deliverables

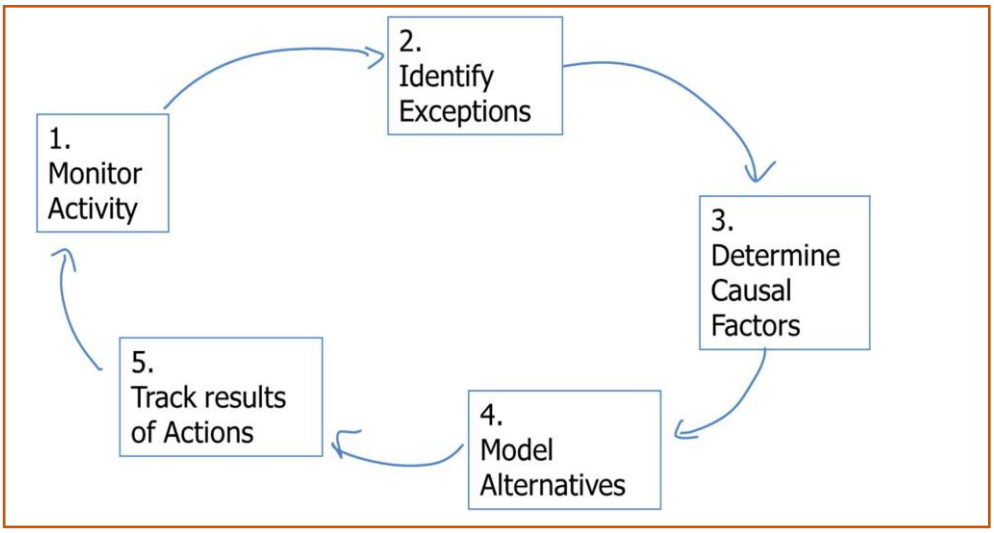
Phase 3: OLAP Queries, and BI Dashboard
Phase 4: Data Mining

Instructions:

A. This is a team assignment.
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. You are asked to submit the following details:
   i) Scripts to execute the SQL queries.
   ii) Screen shots of your Business Intelligence (BI) Dashboard that show the functionality.

Phase 3: Part A. OLAP queries

Write OLAP queries exploring the data to answer questions posed during the typical analytical lifecycle as covered in class and as shown below.



You should include a total of 8 queries, in the categories as shown below. (The examples are shown to illustrate the concepts. Teams are free to use their own examples.)

Part A.1. Standard OLAP operations – 8 queries in total

a. Drill down and roll up: by using concept hierarchies in your data mart, such as (name, region, continent) and (month, quarter, year, decade).

b. Slice, where only one dimension is selected.

c. Dice, where one creates a sub-cube.

d. Combining OLAP operations. In these queries, we combine the above-mentioned operations. For instance, we may explore the data characteristics i) during different time periods, ii) when certain events were taking place, iii) for different countries and regions, iv) while comparing age groups, or v) contrasting unemployment rates.

Part A.2. Explorative operation

Identify general trends using advanced SQL operations. Give *one query* from each one of these categories.

a. **Iceberg queries.** For instance, i) find the five years with the highest population growths, ii) find the five countries with the highest decreases in term of specific health conditions (e.g., tuberculosis) in subpopulations {children, male, female, total} when considering a particular decade.
b. **Windowing queries.** For instance, display the ranking of the countries in terms of the literacy rates, as reported per gender, over the last five years.
c. **Using the Window clause.** For instance, compare the number of hospital beds in Canada in 2019 to that of the previous and next years.

Note: Refer to the Module 4 Data Analytics lecture slides. The PostgreSQL syntax is available at:

https://www.postgresql.org/docs/current/queries-table-expressions.html
https://www.postgresql.org/docs/current/tutorial-window.html
https://www.postgresql.org/docs/current/sql-expressions.html#SYNTAX-WINDOW-FUNCTIONS

Part B. BI dashboard and Information Visualization

Create a dashboard that allows the users to explore the data and to visualize trends. Specifically, users should be able to traverse concept hierarchies, including the ability to roll up and drill down, slice and dice, as well as execute Top N or Bottom N queries. Your interface should include graphs and charts. You may use any dashboard tool of your choice.

Phase 4: Data Mining

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 preprocessing 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.

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.