Table of Contents

[Assignment 1 1](#_Toc188859122)

[Finding Data 1](#_Toc188859123)

[Report Submission Instructions: 1](#_Toc188859124)

[1. Problem Definition (5 marks) 1](#_Toc188859125)

[2. Exploratory Data Analysis (15 marks) 1](#_Toc188859126)

[3. Model Development and Tuning (10 marks) 2](#_Toc188859127)

[4. Model Evaluation (10 marks) 2](#_Toc188859128)

[5. Code Appendix (10 marks) 2](#_Toc188859129)

[Restrictions 3](#_Toc188859130)

# Assignment 1

Develop machine learning predictive models for any data set that interests you provided that the data set has not been covered in the Predictive Analytics option.

## Finding Data

You can find data at:

* <https://data.world/search>
* <https://www.kaggle.com/datasets>
* Web scraping
* API’s

## Report Submission Instructions:

Please submit a Word document. Do not use pdf please. Also, please include your dataset with the submission (or provide a hyperlink to your data set in a separate \*.txt file).

Please submit a report which contains the following:

### Problem Definition (5 marks)

Explain the problem you are trying to solve.

### Exploratory Data Analysis (15 marks)

Please provide a thorough exploratory data analysis with visualizations (not just a heatmap and histograms or a listing of variables and data types). At the start of the exploration, and at the start of the report, identify the best model features in your pool and be sure to focus on the relevant important features in your exploration and visualizations. Show the histograms for all features of the best model. Make sure the report is informative and rich with information for a non-technical audience. Use visualizations to how the best model features are related to the target. Do not only show a heat map.

### Model Development and Tuning (10 marks)

Use the following automated feature selection routines to generate a pool of potential features. Experiment with different feature combinations and record the results in a reader-friendly table. Use this section to explain how you created and tuned your model. You are required to build the following types of models:

* A bagged model using either:
  + BaggingClassifier

or

* + BaggingRegressor
* A voting or mlextend ensemble model that uses several models of different types using either.
* A stacked model. Please use the technique that was demonstrated either in example 6 or example 7 of lesson week 3, COMP4948.
* Linear regression or logistic regression as appropriate.
* Any other model that you are interested in.

### Model Evaluation (10 marks)

Use cross fold validation and remember to split the data into three data sets: Train, Validation and Test. Test should be used as unseen data to evaluate the model after the cross-fold loop. The cross fold validation loop for training must use this format:

|  |
| --- |
| from sklearn.model\_selection import KFold  kfold = KFold(n\_splits=NUM\_SPLITS, shuffle=True)  for train\_index, test\_index in kfold.split(X): |

Present statistics and summaries that compare your models in a professional user-friendly manner for a non-technical audience to explain how your model was evaluated and selected.

Please compare the models from the previous step in a table. In this table, show all features and a robust set of evaluation metrics for each model. Remember to show averages and standard deviations in the same table. Make it easy for readers to compare models in one single table.

Clearly identify your preferred model and why. Please also highlight your selected model so it is easy for the reader to know which model is recommended.

### Code Appendix (10 marks)

Please place your code in the appendix. The code must:

* Prepare the data.
* Build and evaluate the stacked model as shown in example 6 or example 7 of lesson week 3, COMP4948. Use cross fold validation and appropriate data treatments.

Please keep the code neat and tidy. Use a mono-spaced font. The code needs to run from start to end without error.

### Restrictions

Do not use data sets that are covered in the class. Please do not use the same type of model that you use for any of your other assignments in COMP3948, COMP4948 and COMP4949.