STA314H Winter 2026: Final Project

Team Sign-up Start: Jan 19th   Kaggle-Start: Feb 2nd, 10 AM   Due: Mar 22nd, 11:59 PM

1 Project description

The project includes the following stages. The in-class synchronous competition will be hosted on Kaggle as a private competition. All times mentioned in the project guidelines are in EST (Eastern Standard Time).

- **Stage 1: Team sign-up (Jan 19 – Feb 1, 23:59 PM)**: Please form a group of size 2 to 4 and register your team on both the poll form and Quercus (under the "People" module). Each group needs to list all group members and the team name that will be used for the Kaggle competition. The form is editable until the sign-up deadline. During this period, we will provide a small sample of each dataset (see Section 4) so you can get familiar with the goal and setting. The sign-up form will be closed on Feb 1st, 23:59 PM so please fill out the form promptly!

- **Stage 2: Kaggle sign-up and enrollment (Feb 2, 10:00 AM – Feb 6, 23:59 PM)**: We will send an invitation link (that expires on Feb 6th 23:59 PM) for each dataset's Kaggle competition. Each group may register for *one* competition of their choice, using the exact group name created in Stage 1. Any unrecognized group will be removed from the competition.

- **Stage 3: Kaggle competition (ends on Mar 20, 23:59 PM)**: Registered groups will validate their models and submit predictions through Kaggle by Mar 20 at 23:59 PM. The competition will be closed after this time.

- **Stage 4: Report submission (due on Mar 22, 23:59 PM)**: The project report is due on Mar 22 at 23:59 PM. Please refer to Section 2 and Section 3 for the report format. We will also post additional guidance later in the course with a more detailed report structure.

- **Stage 5: Selected presentations (after final results and report screening)**: Top-performing teams and teams with exceptionally well-written reports will be invited to give a 10–15 minute in-class presentation to showcase their approach. The teaching team will reach out in advance so you have enough time to prepare.

Note: The final delivery of the project should be one written report and one final prediction submission from each group (see details in Section 2). Your final score will not depend on group size. It is your responsibility to make sure that you have the correct teammates on poll-form and Quercus, by Feb 1st, 11:59 PM.

For the chosen dataset, you should familiarize yourselves with the context and meaning of each measurement. The first thing you and your teammates need to decide is:

- Formalize the goal(s) of your study, and think about the motivation.

You are expected to conduct statistical analysis which include, but are not limited to,

- exploratory analysis of the data,
- prediction of the specified target, (you may also predict other targets if you want)

- feature selection,
- statistical inference.

For each dataset, prediction of the specified target is the only requirement. However, you may choose to conduct whatever other statistical analyses that you think meaningful and aligned well with the goal(s) of your study. You may refer to Section 3 for the evaluation criterion.

In terms of statistical methods you might use, there are no restrictions, and you are allowed to employ any statistical models or methods, including those not covered in this class. The goal of this project is to give you concrete applications to apply various machine learning algorithms you have learned in this course, as well as those we were unable to cover. You can also use this opportunity to explore more sophisticated algorithms that you and your teammates find interesting and wish to learn. There are no strict rules for this project, so feel free to explore, experiment, and enjoy the learning and application process!

2 Submission

There are two required submissions per team:

1. *Prediction of your chosen data set.* This is submitted via Kaggle (see the link of each data set in Section 4) and you can find detailed instruction of the submission therein. The competition is closed by Mar 20th, at 11:59pm, and you will not be able to update your prediction after that.

2. *Final report.* Your final report should be submitted by Mar 22nd, at 11:59pm via Quercus. Note that we DO NOT accept any late report. A score of 0 will be assigned if we do not receive your report on time.

   In terms of format, your report should be in PDF format and must not exceed 8 pages on A4 paper, with a font size of 11, single line spacing, and margins set as follows: top = 2 cm, left = 2 cm, right = 2 cm, bottom = 2.5 cm.

Content of the final report. There are no specific rules regarding the content of your report. However, it should be a complete report; for instance, it needs to address all the following aspects:

- Clearly state the problems you studied and explain your motivation.

- Clearly describe the statistical analysis you conducted and explain your results.

- For prediction performance, your report must include your team name on Kaggle, the final prediction accuracy of your model, and your final ranking.

- Statistical analysis can be done in either R or Python. However, all the code used in your data analysis must be included at the end of the report (code does not count toward the 8-page limit).

3 Evaluation and grading policy

The evaluation of the final report will depend on the statistical problems you have addressed. For instance, if you focus solely on prediction, your report will be evaluated based only on the predictive performance of your final model (e.g., your ranking in Kaggle) and the efforts you made to improve it. If you also consider, for example, selecting a subset of features and interpreting your model, your final report will be evaluated based on both prediction and the additional aspects you include.

The final report should be well-written and coherent, containing all necessary elements of a scientific report. For example, a top-ranked predictive model accompanied by a poorly written report will not result in a high grade. Conversely, even if your final model does not rank among the top, a coherent statement of your thought process and an exposition of the efforts you made to improve it can still contribute to achieving a high grade.

Since this is a group project, please make sure you and your chosen teammates are correctly registered in poll-form and assigned in the same group on Quercus, by Feb 1st, 11:59pm. All group members are expected to contribute equally in the project and the same grade will be assigned to all group members.

4 Datasets

There are four datasets provided below. The sample dataset shared through the link is a small subset of the training data, so teams can get familiar with the data. When the Kaggle competition opens on February 2nd, we will release the full training dataset and the submission guidelines.

Please note that each team needs to choose and work on ONE dataset.

- *Pet Face Images dataset.* You can access a sample of the data set via this link. Here are a few background on this dataset.

  Pets can display a wide range of facial expressions that may reflect their emotional state or intent. This dataset contains face images of various pets, including dogs, cats, rabbits, hamsters, sheep, horses, and birds. The images capture diversity in both species and facial expressions (e.g., “happy”, “sad”, “angry”).

  You may treat this as a facial-expression classification problem (predict an emotion/expression label from an image). Note that animal expressions can be subtle, and labels (if provided) may be subjective or noisy. Your final report should briefly discuss potential sources of label noise and bias (e.g., differences in lighting/background, species imbalance, or ambiguous expressions), and how you attempted to handle them.

- *Dataset of three Cancer Diseases.* You can access a sample of the data set via this link. Here are a few background on this dataset.

  The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset, describing tumor tissue and matched normal tissues from more than 11,000 patients, is publically available and has been used widely by the research community. The data have contributed to more than a thousand studies of cancer by independent researchers and to the TCGA research network publications.

  In this data set, we focus on three cancer types: the glioblastoma multiforme (GBM), the lung squamous cell carcinoma (LUSC) and ovarian cancer (OV). The features contain gene expression data using the Affymetrix HT Human GenomeU133a microarray platform by the Broad Institute of MIT and Harvard University cancer genomic characterization center. Data are in log space. Genes are mapped onto the human genome coordinates using UCSC xena HUGO probeMap. Both data consist of 12,043 identifiers. We have 886 samples in total. To be specific, the GBM dataset has 376 samples, LUSC dataset has 90 samples and OV dataset has 420 samples.

- *Mental Health Text Classification Dataset.* You can access a sample of this data set via this link. Here are a few background on this dataset.

  Mental health concerns such as suicidal ideation, depression, and anxiety can be reflected in individuals’ written posts on online platforms. This dataset contains labeled mental-health–related text posts for **4-class** classification: `Suicidal`, `Depression`, `Anxiety`, and `Normal`. The goal is to build a model that predicts the mental-health status label from the text content.

  The dataset is a derived and preprocessed combination of several public sources (e.g., mental-health Reddit datasets and Kaggle collections). The provided files are cleaned, and the labels are *not* clinical diagnoses. The sample file contains an **unbalanced** class distribution (reflecting real-world class skew).

- *COVID-19 Patient Symptoms & Diagnosis dataset.* You can access a sample of the data set at this link. Here are a few background on this dataset.

  COVID-19 is an infectious respiratory disease that can range from mild symptoms to severe illness. Because symptoms may overlap with other respiratory infections, and because early identification can support timely isolation and care, it is useful to study how reported symptoms, basic clinical indicators, and exposure history relate to diagnostic outcomes.

  This dataset contains synthetic patient-level healthcare data. It includes records for a few thousand patients with 18 columns in CSV format. The dataset provides demographic information (e.g., age and gender), common COVID-19 symptoms (e.g., fever, dry cough, fatigue, shortness of breath, loss of smell/taste, chest pain), basic clinical indicators (oxygen level and body temperature), and medical history / exposure variables (comorbidity, travel history, and contact with a COVID-19 patient). The target variable is `covid_result`, a binary indicator of whether the patient tested positive or negative for COVID-19.

  Note on data quality. For this project, we manually included a small amount of data issues to better reflect real-world settings. For example, we added some errors and flipped a subset of labels in the training data. You may consider strategies to identify and handle potentially mislabeled observations in your final project. If you do so, please clearly describe what you did and why.