**PROJECT PLAN**

This assessment provides you with an opportunity to reflect on concepts in machine learning in the context of an open-ended research problem, and to strengthen your skills in data analysis and problem solving. The idea behind the project is for you to correctly implement general principles of statistical machine learning, while exploring data and algorithms of your interest. The goal of this project is not to obtain the best metric (e.g., accuracy) per se, but to perform different steps of machine learning in the proper way, according to what you have learnt in this subject. You should be clear about what is the research question in your project, what you plan to try, and what insights you might be planning to get. Then in terms of the results you get, you should discuss what worked or what did not work, and explain the possible reasons in light of what you learnt in class.

**By submitting this plan, you confirm that you have access to the required computational resources and tools to execute this plan. This plan (or any subsequent change) will be reviewed, to ensure that all plans have a similar level of complexity. If your plan is not reviewed, it might likely not be at the level of what we expect, which will end up impacting your grade. Thus, it is in your best interest to have your plan reviewed.**

1. Student names. (The project is to be done in groups of 3 students.)

Sibo Wang

Xuan Ji

Jiulong Xu

1. [Up to 5 lines] Definition of the problem and research question, possibly relevant to your interests.

Our study aims to use electronic health record (EHR) data from the MOVER database to explore how to use machine learning models, including logistic regression, support vector machine (SVM), and neural network, to build classification models to classify which patients need to transfer into the ICU after surgery. We will compare the prediction performance of these three models of different complexity, identify key preoperative features, and evaluate the potential application of machine learning in predicting postoperative monitoring needs to utilize overall patient outcomes.

1. [Up to 5 lines] Description of the dataset (or datasets) to be used, e.g., number of tables, rows, columns, type of data (discrete/categorical, continuous, sentences, images, etc.). Datasets should be already publicly available, since there is not enough time for you to collect data. Possible datasets are for instance: [ADHD 200 (Whole Brain Data)](http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline#Whole_Brain_Data), [Labeled Faces in the Wild](http://vis-www.cs.umass.edu/lfw), [Heritage Health Prize](https://www.kaggle.com/c/hhp), [Yahoo Bidding (A1)](http://webscope.sandbox.yahoo.com/catalog.php?datatype=a), [Yahoo Music (C15)](https://webscope.sandbox.yahoo.com/catalog.php?datatype=c). **You should choose a challenging dataset. Small datasets are not allowed. (Please also see question 5).**

The dataset name is Mover which stands for Medical Informatics Operating Room Vitals and Events Repository. This dataset is collected by UC Ivris which includes all adult-patients’ surgical details in Irvine Medical Center from 2015 to 2020. Due to the limiting of computing resources, we will only be using 9 of the total tables from the data section: EPIC. It has a total of 39,685 unique patients with 64354 unique surgical events along with more than 35 feature columns. This dataset mainly contains both discrete and continuous data such as use of medications and laboratory test results.

1. URL where the above dataset(s) is(are) available.

The Mover: https://mover.ics.uci.edu/index.html

1. [up to 5 lines] Which feature construction and preprocessing of the dataset will be performed, e.g., converting several tables to a single table, counting, summing, one-hot encoding, etc. **You are not allowed to just read a single data table and used it. Remember this is a Masters level subject and we require some level of complexity. You are allowed to either implement this from scratch or use third-party code.**

The unique identifiers for each patient and surgical events are MRN and LOG\_ID which are contained in every data table. We will be using references from research papers as supporting documentations to identify risk factors related to the ICU admissions, and using two identifiers as primary key to merge multiple tables into a single dataset for training the models. During this step, we will apply feature engineering to selected columns such as using one-hot encoding to transfer string data such as surgical types and gender into binary features.

1. [Up to 5 lines] Which 3 machine learning algorithms are going to be compared? You should list 3 different algorithms and with different model class complexity, i.e., simple, medium, complex. **You are allowed to either implement this from scratch or use third-party code. At least one of the 3 algorithms should be from a research paper in a conference or journal, e.g.,** [**NeurIPS**](https://proceedings.neurips.cc/)**,** [**ICML/UAI/AISTATS/JMLR**](https://proceedings.mlr.press/)**,** [**ICLR**](https://openreview.net/group?id=ICLR.cc)**,** [**TMLR**](https://openreview.net/group?id=TMLR)**, etc. (Since this research-paper algorithm might take you more time to figure out, you can let us know this later.)**

In this study, we will compare three machine learning algorithms of varying complexity: Logistic Regression (simple model), a linear classification model that is easy to interpret; Support Vector Machine (SVM) (medium complexity), which handles non-linear classification well and performs effectively on high-dimensional data; and Deep Neural Networks (DNN) (complex model), utilizing multiple non-linear layers to capture complex patterns, These algorithms will be evaluated for their ability to predict ICU admissions post-surgery.

1. [Up to 5 lines] Cross-validation technique, e.g., training/validation/testing, k-fold cross-validation, bootstrapping. **You MUST implement this from scratch. At least 20 independent repetitions should be run, so that means/variances can be computed for proper comparison between algorithms.**

We will implement 10-fold cross validation and run it 20 times independently to calculate the mean and variance to ensure robustness of model comparison. At the same time, we will use a training/validation/test data split with a ratio of 70% training, 15% validation, and 15% testing. This will ensure a fair comparison of different algorithms.

1. [Up to 10 lines] Which hyperparameter(s) is(are) going to be tuned for each of the 3 algorithms above, and what method is going to be used for the nested cross-validation. **You MUST implement this from scratch. Every algorithm should have at least one hyperparameter to be tuned and such hyperparameter should be expected to affect the results significatively.**

Logistic regression: The regularization strength parameter C is tuned to control the model's resistance to overfitting. The best C value is selected by grid search combined with nested 10-fold cross validation.

Support Vector Machine (SVM): Tune the kernel function type (such as linear, RBF) and the regularization parameter C, and select the optimal parameters through grid search plus 10-fold nested cross validation.

Neural Network (DNN): Tune the number of hidden layers and learning rate, and select the optimal combination through random search plus nested cross-validation to improve model performance.

1. [Up to 15 lines] Description of the experimental results, e.g., learning curves, ROC curves, plots of different datasets, etc. **You MUST implement this from scratch. Error bars should be computed across repetitions (at least 20 as described in question 7) and reported for proper comparison between algorithms.**

Given that the focus of this study is to predict whether a patient needs to be admitted to the ICU after surgery, we will prioritize recall (i.e. sensitivity). This is because in clinical practice, we would rather reduce the number of patients who actually need the ICU being incorrectly predicted as not needing it, even if this results in more false positives. Therefore, our experimental results will focus on the following:

1. Precision-Recall Curve: Each model generates a precision-recall curve that shows the trade-off between precision and recall. Since we want to increase recall, this curve helps evaluate the model's ability to recognize at high recall, even if precision is reduced.
2. Confusion Matrix: The confusion matrix for each model will be displayed, clearly showing the distribution of True Positives and False Negatives. The confusion matrix can visually show the performance of the model in terms of false positives and false negatives, especially the false negatives (not correctly identifying patients who need ICU) that we are concerned about.
3. F1 score and ROC curve: We will report the F1 score to balance precision and recall. In addition, the ROC curve will be used to evaluate the overall classification performance, and the area under the curve (AUC) provides a benchmark for comparison between models.
4. Error bars: All results are based on at least 20 independent replicates, and error bars are calculated to show the volatility of model performance to ensure the robustness of the results and to enable reliable comparisons of different algorithms.
5. Which programming language are you going to use? (Python, Jupyter, C++, MATLAB, Java are allowed.)

We will mainly use Python and Jupiter.