CS777 – Term Project Proposal Submission Template

**Student: Natasya Liew**

1. Data set description: Provide a detailed description of the public data set you have selected, including its source, format, and any relevant details about the data.

|  |
| --- |
| *For my project, I will use previous season NFL SuperBowl data to analyze features contributing to a successful tackle. The original dataset contains 13 csv files containing information of the different games within the season, along with game highlights, players information, and plays tracking for each week of the first 9 weeks within the season. The original data can be found at:* [*https://www.kaggle.com/competitions/nfl-big-data-bowl-2024/overview*](https://www.kaggle.com/competitions/nfl-big-data-bowl-2024/overview)*.*  *Unfortunately, the data will require some level of cleaning and joining between the multiple datasets. More information about the games can also be seen in the visualization in the NFL website: https://www.nfl.com/scores/2022/REG1.* |

1. Research question: Clearly define your research question and explain why studying is important. What do you want to learn from the data?

|  |
| --- |
| *As mentioned from the previous section, I want to analyze the different features for feature selections and compare the selected features with three different ML models to analyze the accuracy of predicting successful tackles.* |

1. Machine Learning model: Specify the type of machine learning model you plan to use, such as classification or clustering, and explain why you have chosen this model.

|  |
| --- |
| *I intend to use k-means, logistic regression, and random forest as the models to classify successful tackles. If possible, I’d like to also explore the use of gradient boosting and/or regularization on the three models as well for parameter tuning.* |

1. Expected outcomes: What do you expect to achieve after implementing your learning model? What do you hope to learn or discover from your data analysis?

|  |
| --- |
| *Based on the type of regularization model, feature removal/selection, and data cleaning, I would imagine that logistic regression will give us the best explainability in terms of correlation (or weights) of the features against prediction of successful tackles. However, random forest will probably provide the highest accuracy in terms of predicting successful tackles.* |

1. Evaluation plan: Explain how you plan to evaluate your project and assess the correctness of your model. What metrics or methods will you use to evaluate the effectiveness of your learning model? How well do you expect the model to work, and how will you measure its performance?

|  |
| --- |
| *In this study, I plan to use the confusion matric along with the precision, TPR, and TNR to analyze the accuracy/errors in the models’ predictive ability. It would also be interesting if there is enough time to add the F1, recall, along with the AUC/ROC curve as well to further explain the accuracy of the model.* |