695 WS2- Applied Machine Learning
Team 6
Final Project Proposal

Problem statement:

Predicting Annual Income of Individuals using Classification Techniques

Description of data set:

For this project, we are using a very well-known dataset named the ***Adult*** Dataset. It contains information on individuals from the US Census Bureau database, and includes demographic and socio-economic features such as age, education, occupation, work class, marital status, race, gender, and income.
(https://archive.ics.uci.edu/ml/datasets/Adult).

Features: The dataset provides 14 input variables that are a mixture of categorical and continuous data types. They are:

Categorical Features	Continuous Features
workclass	age
education	fnlwgt
marital-status	education-num
occupation	capital-gain
relationship	capital-loss
race	hours-per-week
sex:	
native-country	

Dataset Size: There are a total of 48,842 rows of data, and 3,620 with missing values, leaving 45,222 complete rows.

Classes: There are two class values '>50K' and '<=50K'.

Implementation plan:

We want to implement various classification techniques and compare the results.

1. Binary Classification: We will develop binary classification models that predict whether a person's income is above or below $50K per year. We will also experiment with different techniques for handling missing data, scaling features, and balancing the imbalanced classes.
    a. Logistic Regression
    b. Decision Tree
    c. Random Forest
    d. SVM
    e. Neural Network
2. Ensemble Model: We will develop ensemble models that combine the predictions of multiple classification algorithms to improve overall performance. We will try a variety of ensemble techniques, such as voting, bagging, boosting, or stacking, and compare their performance against individual models.

3. Hyperparameter Tuning: We will experiment with different hyperparameters for classification algorithms.
4. Feature Engineering: We will extract new features from the dataset and use these new features to train a classification model.

Team members & task allocation:

Darsh Patel:
- Decision Tree
- Random Forest
- SVM
- Hyperparameter Tuning

Janmejay Mohanty:
- Logistic Regression
- Ensemble Models
- Hyperparameter Tuning

Soumen Sikder Shuvo:
- Neural Network
- Ensemble Models
- Feature Engineering

Each member will contribute to the report writing of his assigned tasks. Also, it will be reviewed and edited by everyone in offline meetings.