**Assignment 4: AutoML**

**Name: Email:**

Statement:

For this assignment's preparation, the author(s) did not use any generative AI tools.

For this assignment's preparation, the author(s) have utilized [Generative AI Tool Name], a language model created by [Generative AI Tool Provider]. Within this assignment, the [Generative AI Tool Name] was used for purposes such as [e.g., brainstorming, grammatical correction, writing paraphrasing, citation, specific sections of the assignment].”

**Objective**: Use an AutoML tool or library to develop a machine learning model on a given dataset. Understand the strengths and limitations of using AutoML and compare results to traditional model development processes.

**Tasks**:

1. **Dataset Selection & Preprocessing**:
   * Choose one of the datasets suggested below or any dataset of your interest.
   * Preprocess the dataset: handle missing values, normalize or standardize features, split the data into training and test sets.
2. **AutoML**:
   * Select an AutoML tool or library (e.g., Google Cloud AutoML, H2O.ai, TPOT, Auto-Sklearn).
   * Use the tool to automatically select a model, hyperparameters, and optionally, feature engineering techniques.
   * Train the model on the training data.
   * Evaluate the model's performance on the test data using appropriate metrics (e.g., accuracy, F1-score, RMSE).
3. **Comparison (Optional for CS451 students)**:
   * Implement a traditional machine learning pipeline (e.g., using scikit-learn) for the same dataset: select a model, perform manual hyperparameter tuning, etc.
   * Compare the results of your traditional pipeline with the AutoML results in terms of performance, time consumption, and other relevant metrics.
4. **Analysis**:
   * Discuss the benefits and limitations of using AutoML based on your experience.
   * Reflect on the model choices and hyperparameters the AutoML tool selected. Were there any surprises?

**Suggested Datasets**:

1. **Tabular Data**:
   * [UCI Adult Income Dataset](https://archive.ics.uci.edu/ml/datasets/adult): Predict whether income exceeds $50K/yr based on census data.
     1. <https://archive.ics.uci.edu/dataset/2/adult>
   * [Titanic Dataset](https://www.kaggle.com/c/titanic/data): Predict survival on the Titanic.
     1. <https://www.kaggle.com/c/titanic/data>
2. **Image Data**:
   * [Fashion MNIST](https://github.com/zalandoresearch/fashion-mnist): A more challenging version of MNIST with clothing items.
     1. <https://github.com/zalandoresearch/fashion-mnist>
   * [CIFAR-10](https://www.cs.toronto.edu/~kriz/cifar.html): 60,000 32x32 color images in 10 classes, with 6,000 images per class.
     1. <https://www.cs.toronto.edu/~kriz/cifar.html>
3. **Text Data**:
   * [Spam Classification Dataset](https://archive.ics.uci.edu/ml/datasets/sms+spam+collection): Classify SMS messages as spam or ham.
     1. <https://archive.ics.uci.edu/dataset/228/sms+spam+collection>
   * [20 Newsgroups](http://qwone.com/~jason/20Newsgroups/): Text classification across 20 different newsgroups.
     1. <http://qwone.com/~jason/20Newsgroups/>

**Assessment Criteria**:

1. **Data Preprocessing** (20%): Proper handling of missing values, outliers, and data splitting.
2. **Implementation of AutoML** (20%): Proper usage of the selected AutoML tool, and achieving reasonable performance on the test dataset.
3. **Comparison & Analysis** (20%): Depth and clarity of analysis on the AutoML process, model, and results. Quality of the traditional ML pipeline (if implemented).
4. **Report** (40%): Clear documentation on dataset preprocessing, AutoML process, results, and reflections.