**Assignment 4: AutoML**

**Name: Tyler Conger Email: Tyler-Conger@Tyler-Conger.com**

Statement:

For this assignment's preparation, the author(s) have utilized Chat GPT-3.5, a language model created by OpenAI. Within this assignment, Chat GPT was used for purposes such as brainstorming, asking specific code error questions, and help with proof-reading”

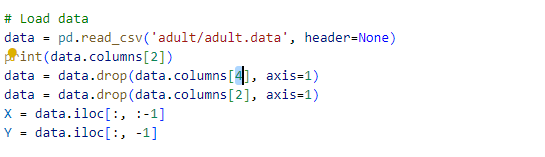
**Objective**: Use an AutoML tool or library to develop a machine learning model on a given dataset. Understand the strengths and limitations of using AutoML and compare results to traditional model development processes.

**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.

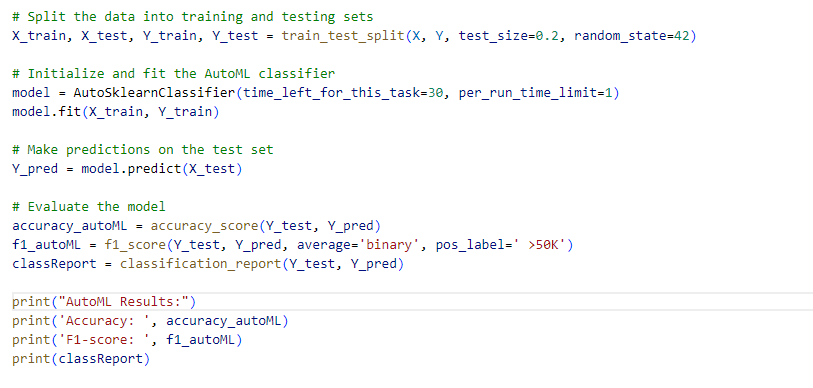
The UCI adult income census dataset was chosen for the AutoML task, introducing a departure from previous assignments, including the familiar MNIST dataset. The motivation behind this choice lies in the dataset's more unique characteristics, providing a fresh set of challenges and opportunities for exploration.

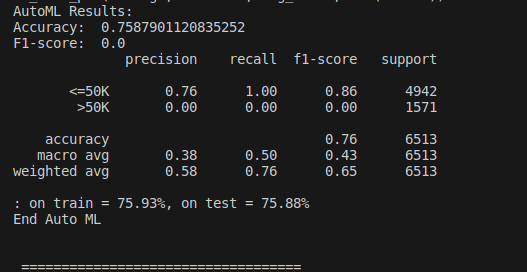
In preparation for AutoML, minimal preprocessing was required. A key step involved isolating specific columns and eliminating those deemed redundant or uninformative. Notably, the 3rd column containing fnlwgt and the education-num column were excluded, with the latter being redundant in representing education as categorical data. This preprocessing step aimed to streamline the dataset for more effective AutoML analysis. The specifics of this process are visually captured in the accompanying image.



1. **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).

In the initial phase of the lab, the AutoML framework auto-sklearn is the main library used. It's imperative to note that the application of auto-sklearn is currently restricted to Linux systems, excluding compatibility with Google Colab or Windows machines. Consequently, users are required to employ a Linux environment for package installation and execution.

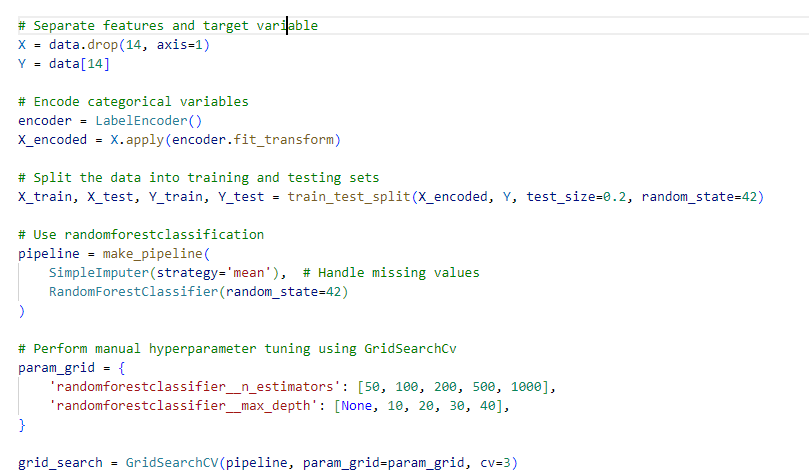


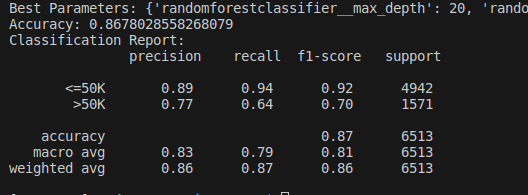


Note this image shows a dark background because it was run a separate Linux based machine running dark-mode vs my regular light mode.

1. **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.

Next we must implement this model using another method that is more traditional and not automatic. I am going to use random forest model from sklearn. The implementation can be seen in the images below and the analysis between the two can be seen in the next section. Initial accuracy report shows that accuracy is around .13 better than the auto-sklearn library used previously.





1. **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?

The benefits of AutoML are clear, it is so easy to use, with a limited amount of knowledge or understanding of machine learning or data science you could quickly start with AutoML. However, I was able to achieve almost 15% more accuracy with a more traditional model. However, it took a much more depth of knowledge of programming and ML to be able to create this model and tune it properly, which could be problematic to someone with limited knowledge. For someone who has limited programming experience it may be more useful to use AutoML because it is less barrier to entry, comparative to someone with more depth of knowledge and more time it would be better to use a more traditional model. It is also worth noting that between the two it was much more difficult to install and setup Auto-sklearn and thus would probably be something someone with limited knowledge was unable to do. I would suggest that auto-sklearn fix some issues that prohibit or cause conflict with google colab as that would help to broaden their audience.

**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.