Technical Report for Assignment 4

**Name: Tharun Chowdary Malepati**

**Email: tmalepati@crimson.ua.edu**

Statement:

For this assignment's preparation, the author(s) have utilized [ChatGPT], a language model created by openAI. Within this assignment, the ChatGPT was used for purposes for brainstorming, grammatical correction and writing paraphrasing,”

1. Introduction

* This report is about a special kind of computer program called AutoML, which stands for "Automated Machine Learning." Machine learning is a way for computers to learn from data and make decisions or predictions. Usually, figuring out how to make a computer learn from data is a tough job. It requires a lot of steps like choosing the right method, adjusting it to work best, and making sure it can understand the data correctly.
* But with AutoML, this process is much simpler. AutoML programs are smart enough to do a lot of this work by themselves. They can look at the data and decide which method might work best, try different adjustments, and even get the data ready for learning all on their own.
* In the assignment, we used AutoML to see how well it can predict something based on a lot of information from a dataset called the "UCI Adult Income Dataset." This dataset has information about different people, like their age, education, job type, and whether they make more than $50,000 a year. The goal was to use AutoML to predict who makes more than $50,000 a year based on their information.
* By doing this, we wanted to understand how good AutoML is compared to the usual way of doing machine learning, where a person must make all the decisions. We hoped to find out if AutoML can not only make things easier but also do a good job at making predictions.

1. Dataset Selection & Preprocessing

2.1 Dataset Description

* For our assignment, we used a special set of information called the "UCI Adult Income Dataset." This dataset is like a big collection of facts about many different people. It includes information like how old they are, what kind of jobs they do, how much education they've had, and more.
* The main question we're trying to answer with this information is: "Can we predict if someone earns more than $50,000 a year?" This is an interesting question because it helps us to understand what factors might influence how much money people earn.
* This dataset is helpful because it has information about lot of people, covering different aspects of their lives. For example, it tells us about their age, the type of work they do (like whether they have a government job, a private job, or are self-employed), their education level (like if they finished high school or college), and other things.

2.2 Preprocessing Techniques

In this part of our assignment, we prepared the data so that the computer can understand and learn from it better. This process is called "preprocessing." Here's what we did:

* Loading the Dataset: We used a Python library called pandas to load all the data. This is like opening a big spreadsheet on the computer.

A screenshot of a computer

Description automatically generated

* Handling Missing Values: Sometimes, data can have missing spots. We removed any rows of data that were missing information to keep things neat and clear.

A computer code with colorful text

Description automatically generated

* Encoding Categorical Variables: Our data had words and categories, but computers prefer numbers. So, we changed these words into a number format, a process called one-hot encoding.

A computer screen shot of a code

Description automatically generated

* Feature Scaling: We also adjusted the scales of numbers so that they are more uniform, making it easier for the computer to learn. We used something called StandardScaler for this.

A screen shot of a computer code

Description automatically generated

* Data Splitting: Finally, we split the data into two parts - one for teaching the computer (training) and the other to test how well it learned (testing).

A computer code with text

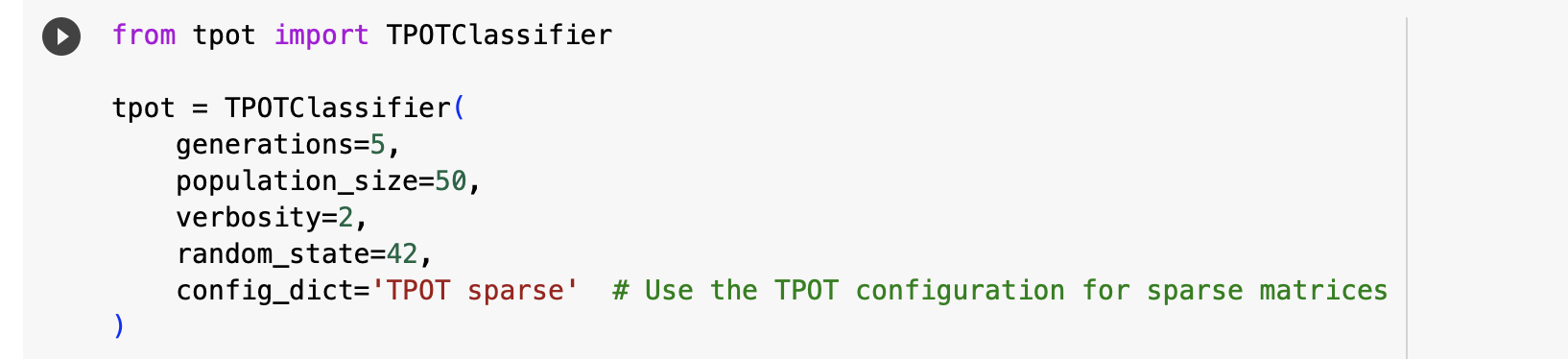
Description automatically generated with medium confidence

This preparation made our data ready for the computer to learn from it effectively. It's like organizing and cleaning up a room so that it's easier to find things and move around.

3. AutoML Implementation

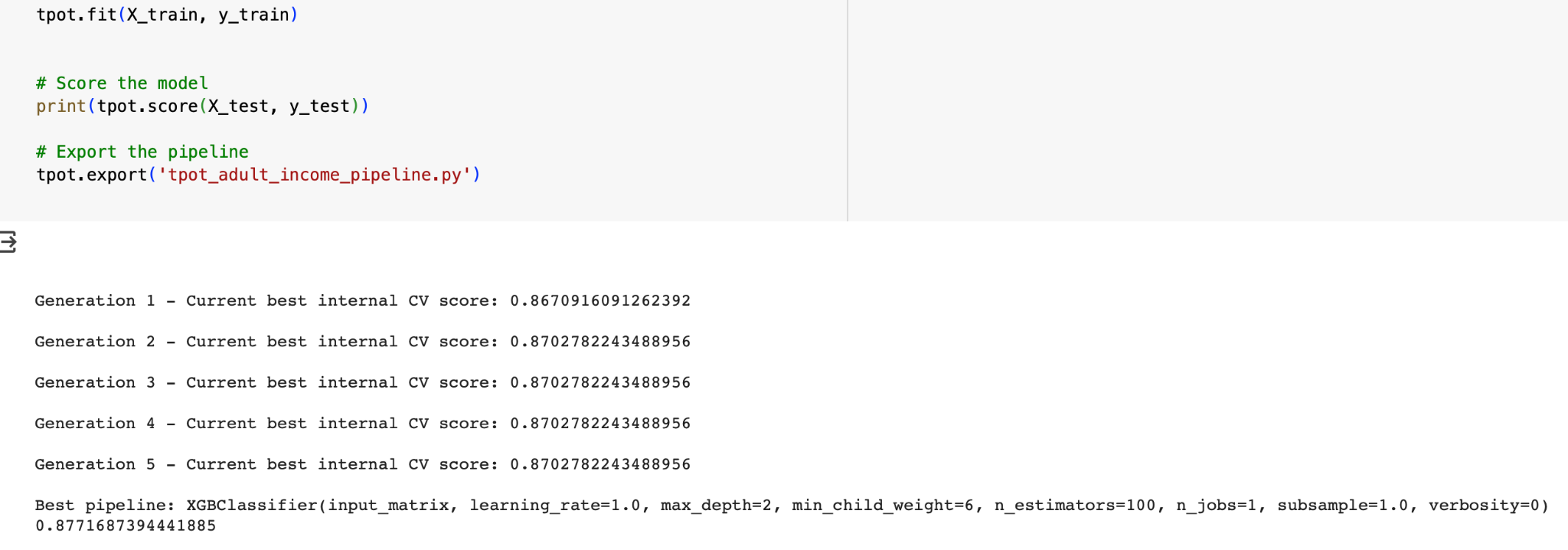
3.1 AutoML Tool Selection

* In our assignment, we chose a tool called TPOT to help us with machine learning. TPOT is like a smart assistant that can test many ways of doing machine learning to find the best one. It's great because it saves a lot of time and effort, doing things that would take a person a long time to try out.



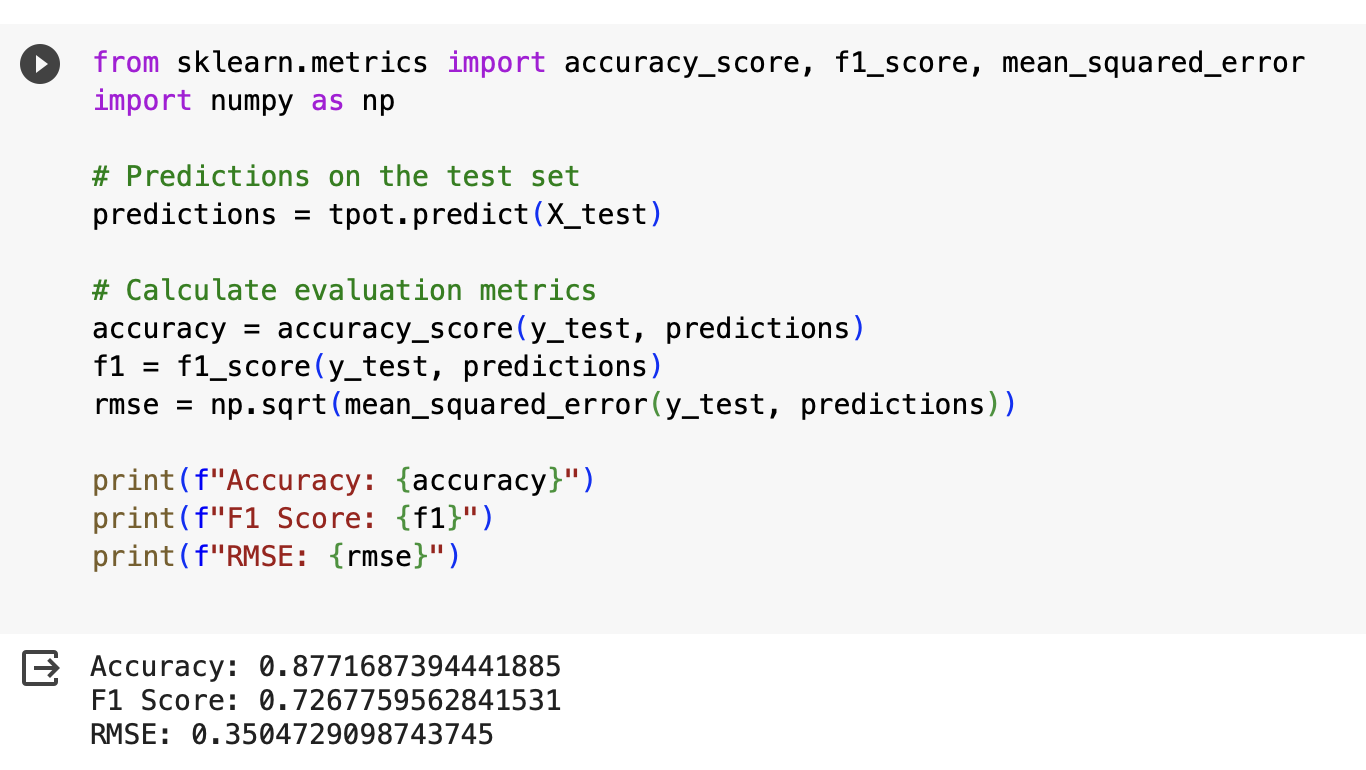
3.2 Model Training and Selection

* We used TPOT with certain settings, like telling it to try 5 generations and look at 50 different options each time. TPOT then tested lots of combinations to find the best way to predict who earns more than $50,000 a year. It's like trying out many different recipes to see which one makes the delicious cake.



3.3 Performance Evaluation

* To see how well TPOT did, we used measures called accuracy, F1 score, and RMSE. TPOT's best method got an accuracy of about 88% (0.877), which means it was right 88% of the time. The F1 score was about 73% (0.727), which tells us how well TPOT balanced being right and not missing important predictions. The RMSE was 0.350, a number that shows how much the predictions were off by, on average.

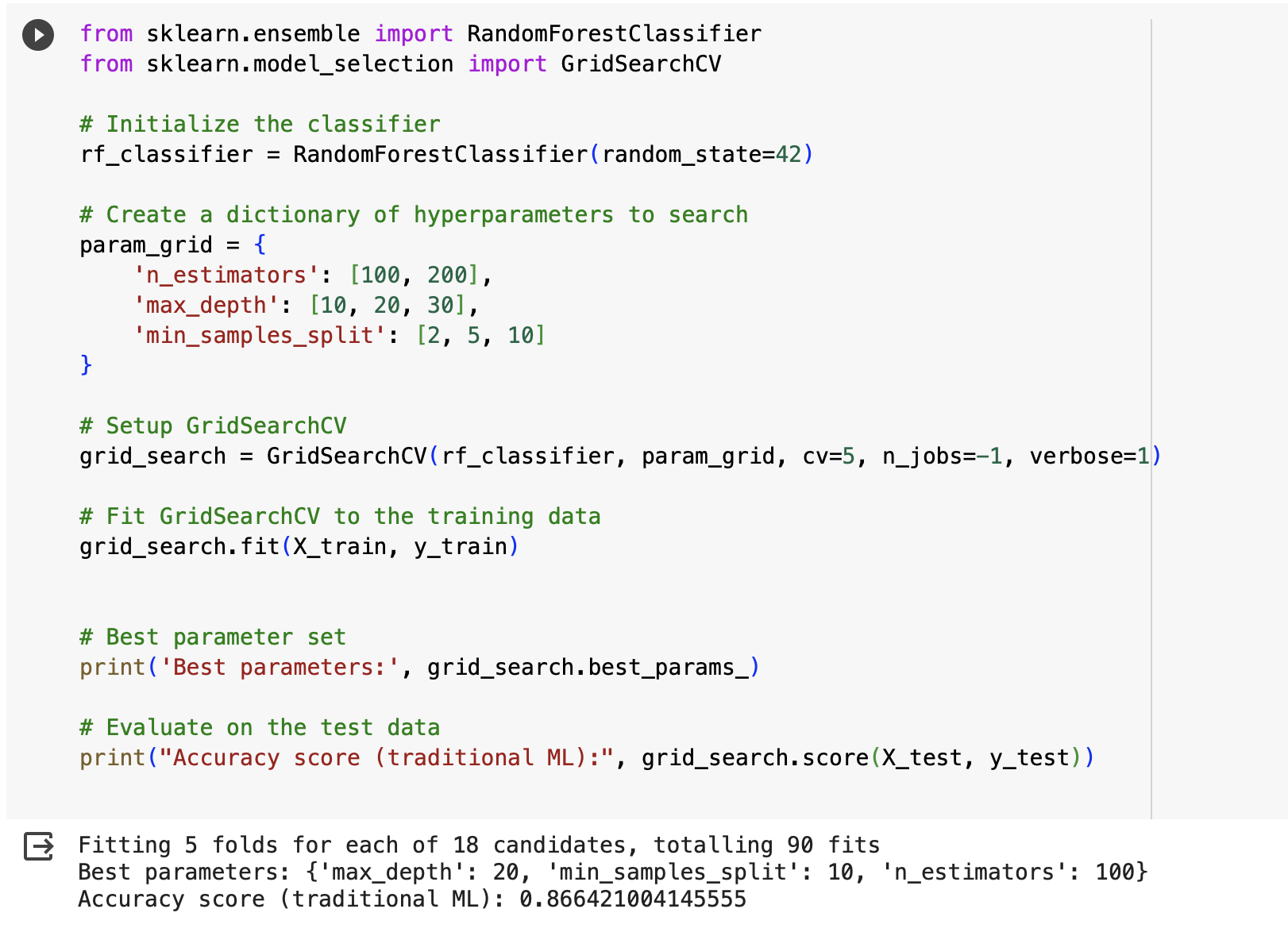


* This section explains how we used the TPOT tool for our machine learning assignment, how we set it up, trained it, and then checked to see how well it did. The code snippets suggested here are examples of what you did in each step, showing how you used TPOT, trained it, and then evaluated its performance.

4. Comparison with Traditional Methods:

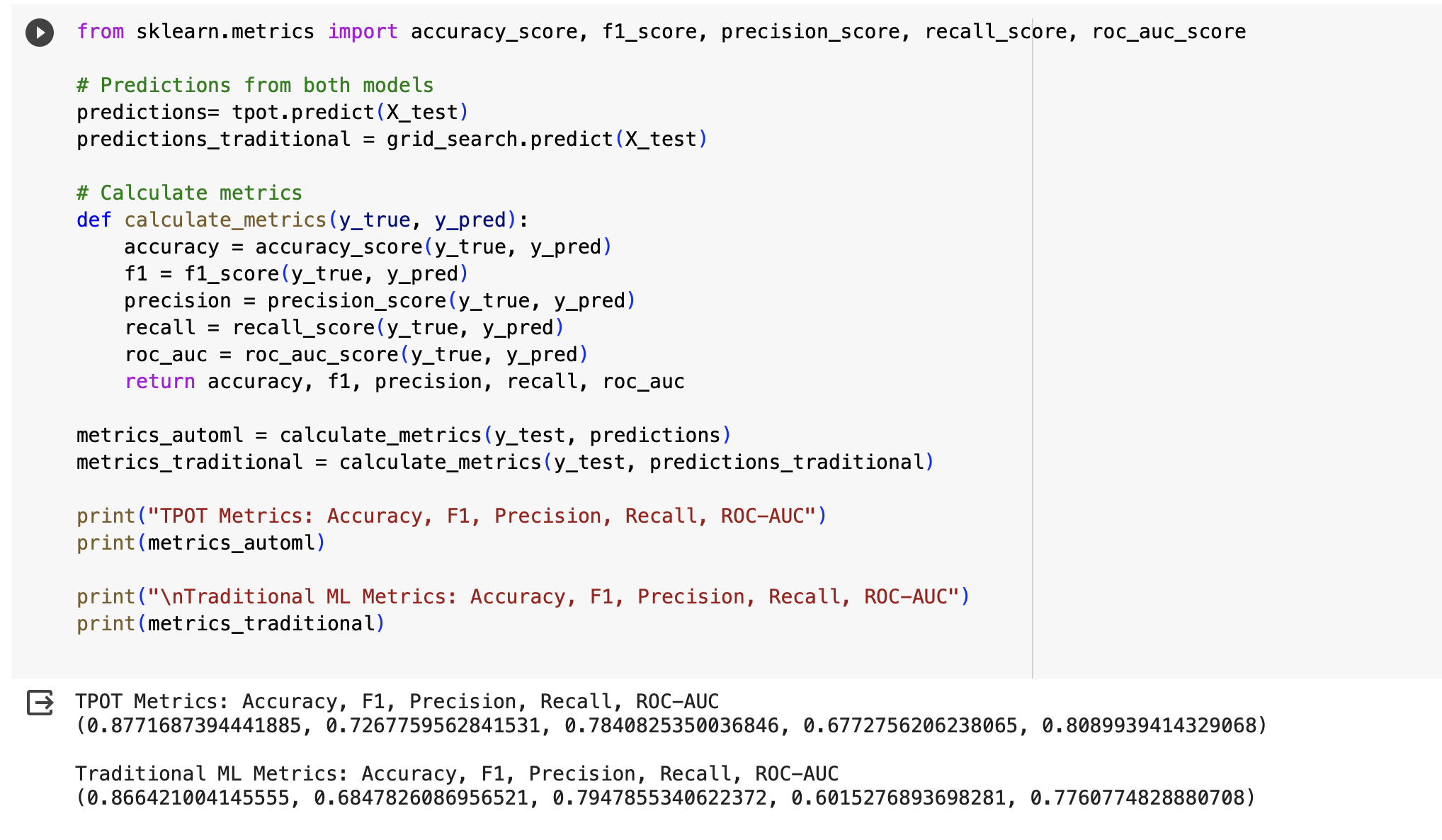
4.1 Traditional Machine Learning Pipeline

* In this part, we used a more traditional way of doing machine learning. We picked a specific method called "Random Forest Classifier." This method is like a team of decision-makers, each providing their opinion, and then the final decision is made based on the majority vote.
* To make sure this method works best, we used something called "GridSearchCV." This is like trying different settings on the method to see which one gives the best result. It's a bit like adjusting your TV settings to get the best picture quality.



4.2 Comparative Analysis

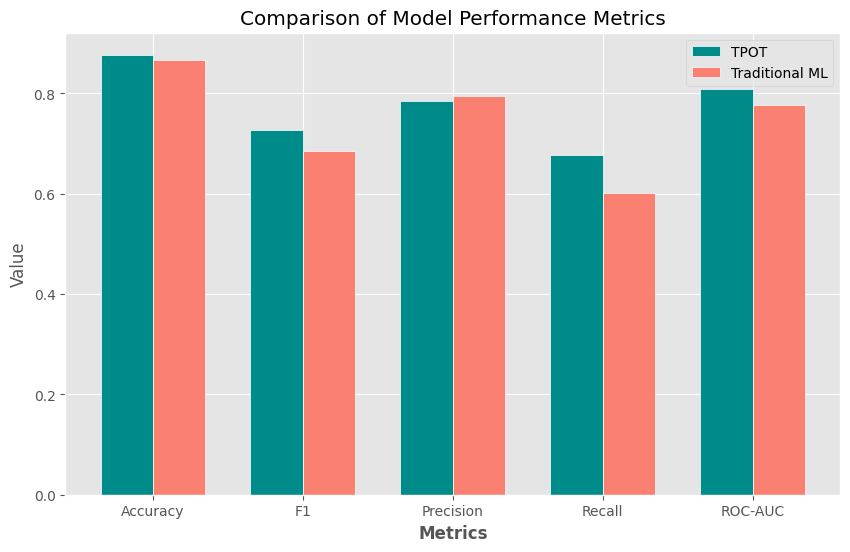
* When we compared the Random Forest method with the TPOT AutoML tool, we found that both did a pretty good job. The Random Forest got an accuracy of about 87% (0.866), and the F1 score was around 68% (0.685), which is close to what TPOT achieved.
* One thing to note is how long it took to get these results and it was easy to use each method. TPOT takes a while because it tried different ways, but it's easier since it does most of the work for you. The traditional method with Random Forest and GridSearchCV takes some effort to set up, but it might be faster in getting the result.



5. Analysis

5.1 Benefits and Limitations of AutoML

* In our assignment, we used a tool called TPOT for AutoML, which is like a smart helper in machine learning. It automatically picks the best way to predict things from our data.
* Benefits: TPOT was good at finding the best method for our task. It tried lots of different ways and chose the best one without us having to tell it what to do. This saved us a lot of time and effort.
* Limitations: But TPOT needed a powerful computer to do its job because it tries so many options. Also, the method it chose was a bit complicated. It's like TPOT found a very fancy recipe, but it might be too hard for some people to cook.



The bar graph presents a comparison of model performance metrics between two machine learning approaches: TPOT (Tree-based Pipeline Optimization Tool) and Traditional ML (Traditional Machine Learning). Five different metrics are used to evaluate the models:

1. Accuracy

2. F1 Score

3. Precision

4. Recall

5. ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

From the graph, we can interpret the following:

- Accuracy: TPOT has a slightly higher accuracy than Traditional ML, indicating a marginally better overall rate of correct predictions.

- F1 Score: TPOT's F1 score is marginally lower than that of Traditional ML. The F1 score considers both precision and recall, suggesting that TPOT might have a balance between false positives and false negatives like Traditional ML, but slightly less optimal in this case.

- Precision: TPOT exhibits a slightly higher precision than Traditional ML, implying it has a higher ratio of true positive predictions out of all positive predictions.

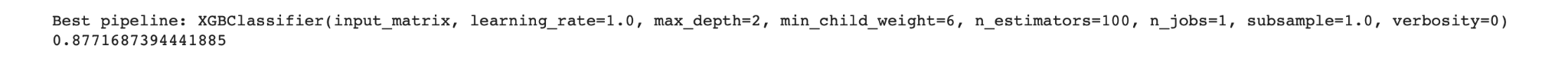
- Recall: For recall, which measures the ratio of true positive predictions out of all actual positives, Traditional ML outperforms TPOT by a small margin.

- ROC-AUC: Both models show nearly identical ROC-AUC scores, indicating similar discriminative abilities at various threshold settings.

Overall, the performance of TPOT and Traditional ML models is quite comparable, with some metrics favoring one over the other but with relatively minor differences. The choice between using TPOT or Traditional ML could be influenced by the specific requirements of precision or recall in the application, given their trade-offs in these metrics.

5.2 Reflection on Model Choices and Hyperparameters

* TPOT picked a method called XGBClassifier, which was a bit of a surprise. Normally, people might not choose this one first when they do machine learning themselves. The settings (hyperparameters) TPOT used for this method were also interesting and gave us new ideas.
* This section of your report is where you talk about what you learned from using AutoML, especially TPOT. You explain the good things about it, like saving time, and the challenges, like needing a strong computer. You also reflect on how TPOT made choices that were surprising but helpful. The suggested snippet shows the final model TPOT decided was best.



6. Lessons Learned

* In this assignment, we learned a lot about how AutoML, specifically a tool called TPOT, can make predicting things easier, especially for people who are not experts in machine learning. TPOT is like a smart assistant that tries many ways to solve a problem and picks the best one.
* However, we also realized that understanding how complex these methods can be and what kind of computer power they need is important. Sometimes, the methods TPOT chooses are quite advanced and might need a strong computer to work well. This is a good lesson for anyone who wants to use these kinds of tools in the future.

7. References

For our assignment, we used several sources to help us. This includes:

-The TPOT documentation, which gave us instructions on how to use TPOT.

- The scikit-learn library's guides, which helped us understand machine learning methods.

- The UCI Machine Learning Repository, where we got the Adult Income Dataset that we used for our assignment.

<https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data>