**MACHINE LEARNING**

**CREDIT APPROVAL PREDICTION**

**A Project Report submitted in partial fulfillment of the requirements**

**for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

**Submitted by**

**Guntreddi Rakesh Naidu (121810305023)**

**V V Mani Chandra Puppala (121810305032)**

**H N V R Shashidhar Kappera (121810305043)**

**Kokkirala Vishnu (121810305008)**

**Under the esteemed guidance of**

## Dr.J.Hyma

## **Associate Professor**



**DEPARTMENT OFCOMPUTER SCIENCE & ENGINEERING**

**GITAM**

**(Deemed to be University)**

**VISAKHAPATNAM**

**OCTOBER**

**2021**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GITAM INSTITUTE OF TECHNOLOGY**

**GITAM**

**(Deemed to be University)**



**DECLARATION**

I/We, hereby declare that the project report entitled “Credit Approval Prediction” is an original work done in the Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University) submitted in partial fulfillment of the requirements for the award of the degree of B.Tech. in Computer Science and Engineering. The work has not been submitted to any other college or University for the award of any degree or diploma.

Date: 3-11-2021

**Registration No(s). Name(s) Signature(s)**

121810305023 Rakesh Naidu Rakesh

121810305032 Mani Chandra Mani

121810305043 Shashidhar Shashidhar

121810305008 Vishnu Vishnu

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GITAM INSTITUTE OF TECHNOLOGY**

**GITAM**

**(Deemed to be University)**

**CERTIFICATE**

This is to certify that the project report entitled “**CREDIT APPROVAL PREDICTION**” is a bonafide record of work carried out by **Guntreddi Rakesh Naidu (121810305023), V V Mani Chandra Puppala (121810305032), H N V R Shashidhar Kappera (121810305043), Kokkirala Vishnu (121810305008)** students submitted in partial fulfillment of requirement for the award of degree of Bachelors of Technology in Computer Science and Engineering.

**Project Guide Head of the Department**

**Dr.J.Hyma Dr. R.Sireesha**

## **Associate Professor Professor**

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Abstract | | 5 |
| 2 | Introduction | | 5 |
| 3 | Literature Review | | 8 |
| 4 | Problem Identification & Objectives | |  |
| 5 | System Methodology | | 14 |
| 6 | Overview of Technologies | | 15 |
| 7 | Implementation | | 18 |
|  | 7.1 | Coding | 18 |
| 7.2 | Testing | 22 |
| 8 | Results & Discussions | | 26 |
| 9 | Conclusion and Future scope | | 31 |
| 10 | References | | 32 |

**Description**

1. **Abstract**

Credit Card Approval Prediction is a machine learning solution to automate the process of credit card approval to an account. The model is trained with multiple attributes and the accuracy is measured with both train and test data.

1. **Introduction**

The accurate assessment of consumer credit risk is of uttermost importance for lending organizations. Credit scoring is a widely used technique that helps financial institutions evaluates the likelihood for a credit applicant to default on the financial obligation and decide whether to grant credit or not.

The goal here is to build an end to end automated Machine Learning solution where a user will be able to predict whether a bank customer should be approved for attaining the credit card or not. The user is only need to give the value of feature variables and the model will able to predict the binary outcome (Approve/ Not Approve). The model will be able take care of all intermediate functionalities like cross validation, hyper parameter tuning, algorithm selection etc.

This project shall be delivered in two phases:

Phase 1: All the functionalities with PyPi packages.

Phase 2: Integration of UI to all the functionalities.

Note: All the code will be written in python version 3.6

1. **Literature Review**

Credit card has evolved to a great level in banking industry. Each banking system consists of an enormous number of datasets to carry customer's transactions of their credit cards. So, banks would be in need of customer profiling. Customer Profiling in banks cognizes the issuer's decisions about whom to give banking facilities and what credit limit to be provided. In previous researches, profiling mainly depended on transaction data or demographic data, but in this research, both transaction and demographic data are merged in order to get more accurate results and minimize the possibility of risk occurrence

By using the best techniques, it leads to improvement in accuracy and helps banks to have high profitability through customer satisfaction by focusing on the valuable customer (companies) which are considered as the main engine in the bank's profitability. This study used k-mean, improved k-mean, fuzzy c-means and neural networks. The used dataset is labeled and for neural network classification creating a new label as a target becomes the main aspect of this study, which helps to reduce the execution time of clustering process and provide the best results with accuracy. Finally, by comparing the accuracy ratio the results show that the neural network is the best clustering technique which could give an accuracy percentage of about 98.08%.

The logistic regression (Wiginton, 1980) allows overcoming these deficiencies and became a common credit scoring tool of practitioners in financial institutions. Non-parametric techniques applied to credit scoring include the k-nearest neighbor (Henley and Hand, 1996), decision trees (Frydman et al., 1985; Davis et al., 1992), artificial neural networks (Jensen, 1992), genetic programming (Ong et al., 2005) and support vector machines (Baesens et al., 2003). More recently, research on hybrid data mining approaches has shown promising results (Lee et al., 2002; Hsieh, 2005; Lee and Chen, 2002). While the pursuit of better classifiers for credit scoring applications is a crucial research effort, improved accuracies can be easily achieved by aggregating scores predicted by an ensemble of individual classifiers. West et al. (2005) found that the accuracy of an ensemble of neural networks is superior to that of a single neural network in credit scoring and bankruptcy prediction applications. Thanks to all the researchers who published their studies.

1. **Problem Identification & Objectives**

The proposed project is built end to end. Starting from Data Preprocessing to Deployment. This project includes the features like:

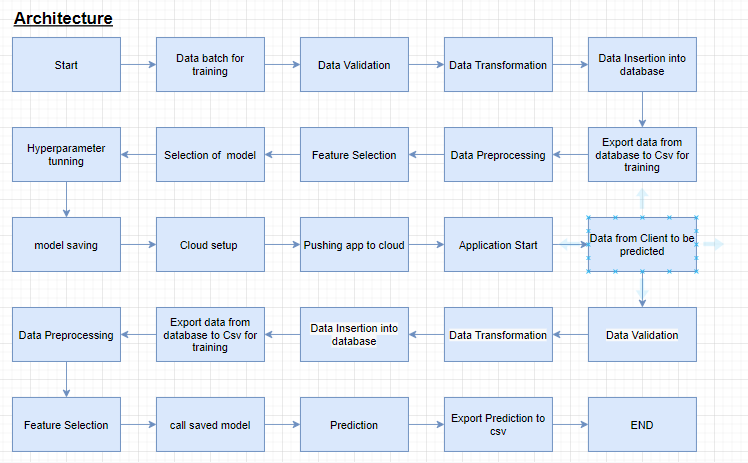
1. Statistical analysis
2. Hyper parameter tuning
3. Best algorithm selection
4. Deployment in Heroku using flask.

The main objectives of the proposed project are to:

1. Increase the accuracy
2. Do Exploratory data analysis
3. Test the model with different algorithms
4. Try different model selection criteria
5. Do Hyperparameter tuning
6. Deploy the project for easy use

Different technologies or libraries used in the project are:

1. numpy, pandas
2. matplotlib, seaborn
3. scipy, scikit learn
4. xgboost
5. html, css
6. flask, gunicorn
7. **System Methodology**



1. **Overview of Technologies**
   1. **Numpy**

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

* 1. **Pandas**

Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like ActiveState’s ActivePython.

* 1. **Matplotlib**

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB. Developers can also use matplotlib’s APIs (Application Programming Interfaces) to embed plots in GUI applications.

* 1. **Seaborn**

Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily. Below are a few benefits of Data Visualization.

* 1. **Scipy**

SciPy is a scientific computation library that uses NumPy underneath. SciPy stands for Scientific Python. It provides more utility functions for optimization, stats and signal processing. Like NumPy, SciPy is open source so we can use it freely. SciPy was created by NumPy's creator Travis Olliphant.

* 1. **Scikit learn**

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It’s built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib!

The functionality that scikit-learn provides include:

* **Regression**, including Linear and Logistic Regression
* **Classification**, including K-Nearest Neighbors
* **Clustering**, including K-Means and K-Means++
* **Model selection**
* **Preprocessing**, including Min-Max Normalization
  1. **Xgboost**

XGBoost is a tree based ensemble machine learning algorithm which is a scalable machine learning system for tree boosting. XGBoost stands for Extreme Gradient Boosting. It uses more accurate approximations to find the best tree model. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

* 1. **HTML, CSS**

HTML (the Hypertext Markup Language) and CSS (Cascading Style Sheets) are two of the core technologies for building Web pages. HTML provides the structure of the page, CSS the (visual and aural) layout, for a variety of devices. Along with graphics and scripting, HTML and CSS are the basis of building Web pages and Web Applications. Learn more below about:

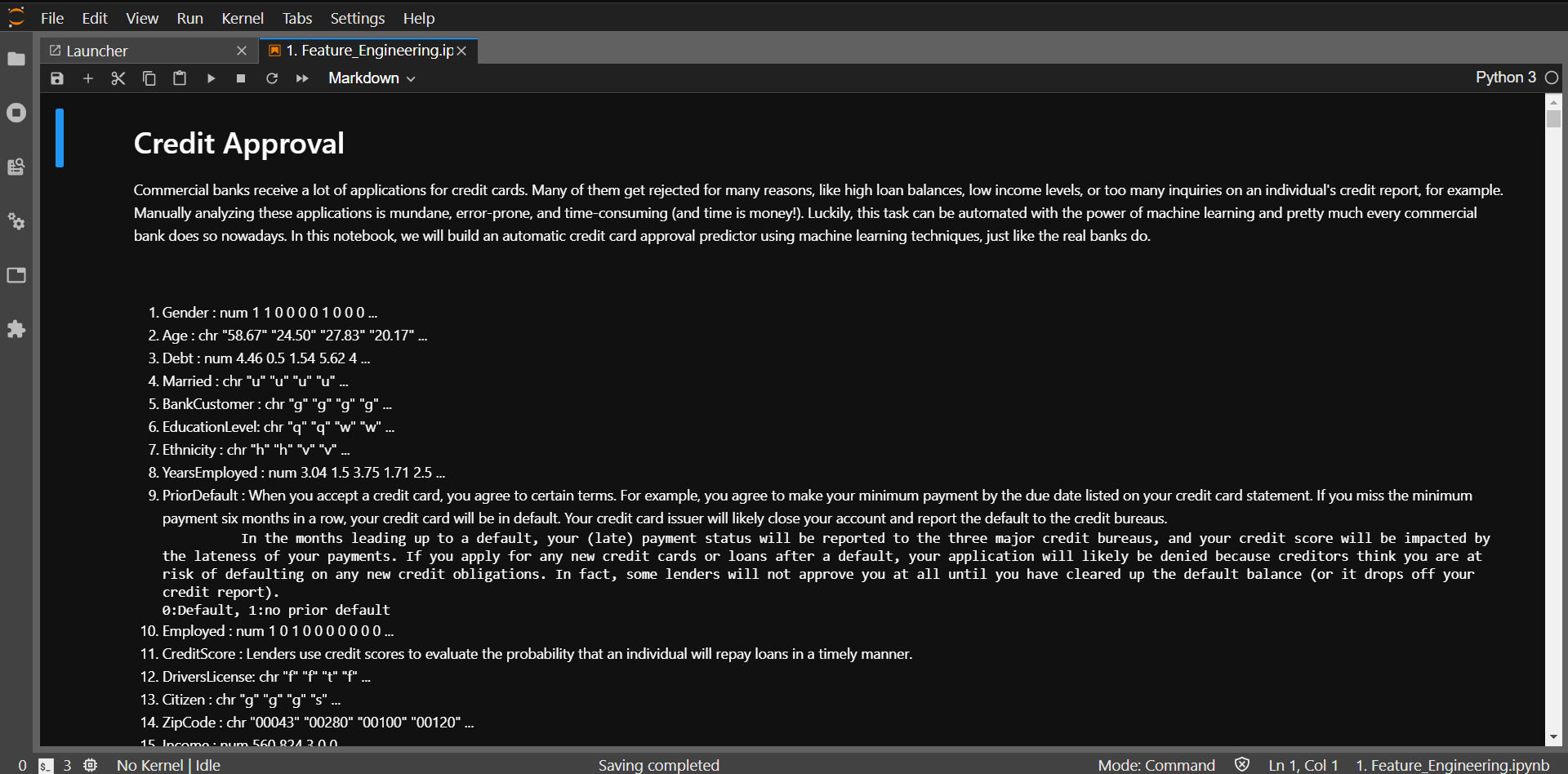
* 1. **Flask, gunicorn**

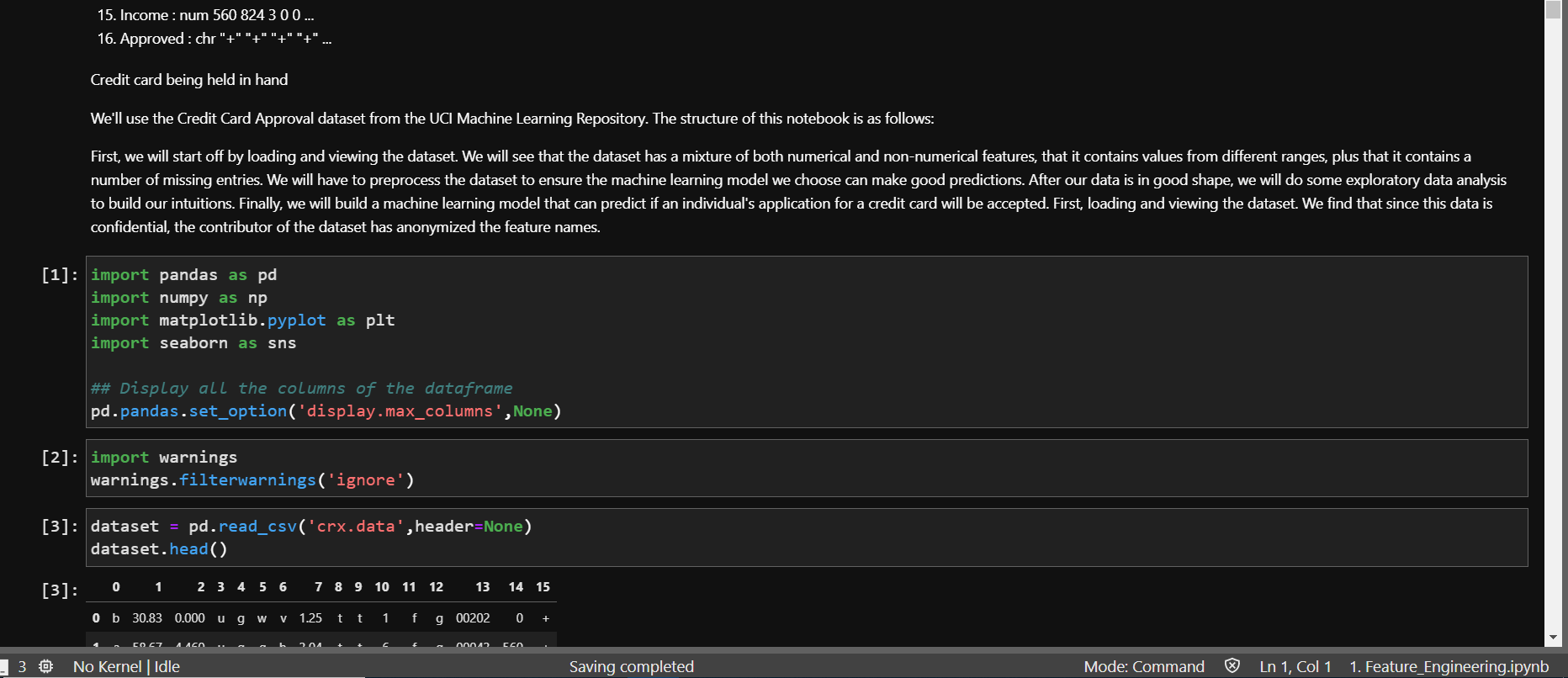
Flask is a Python-based microframework that is popular with web developers, given its lightweight nature and ease of use. Gunicorn is a Python WSGI HTTP Server that uses a pre-fork worker model. By using gunicorn, we will be able to serve your Flask application on more than one thread.

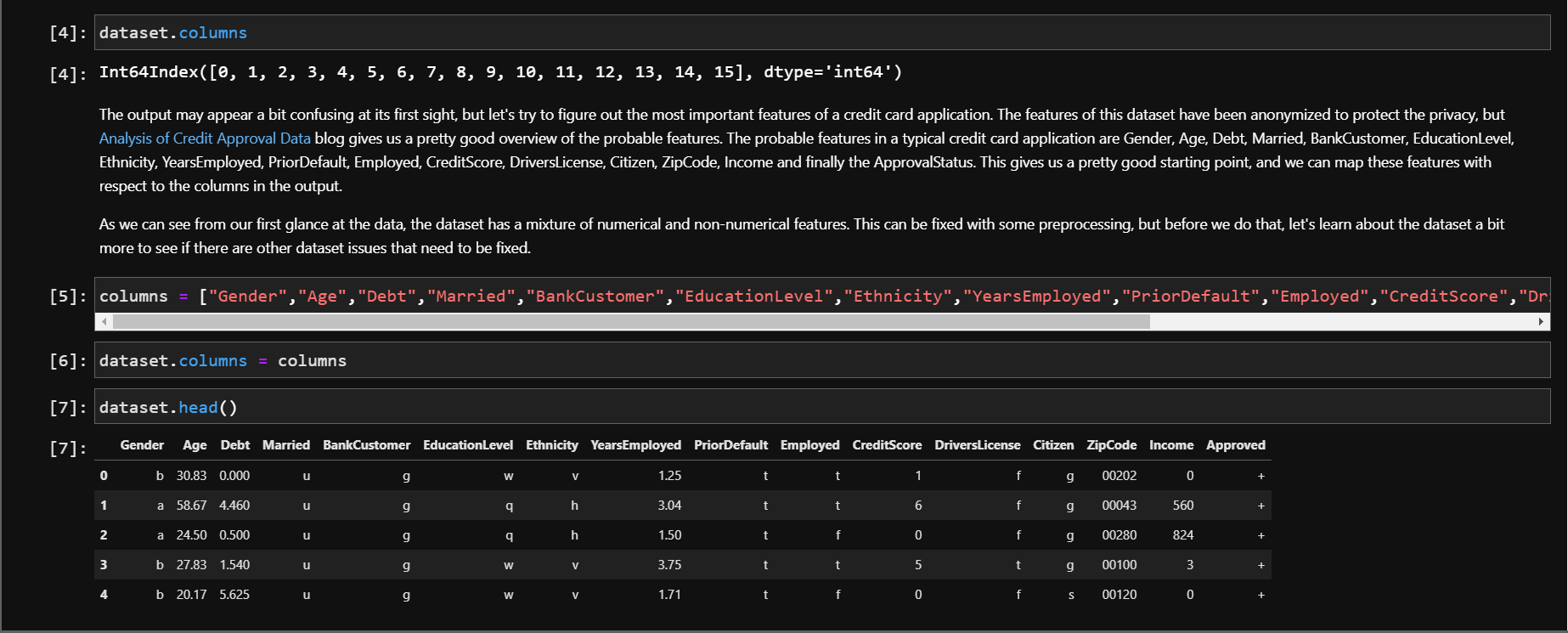
1. **Implementation**

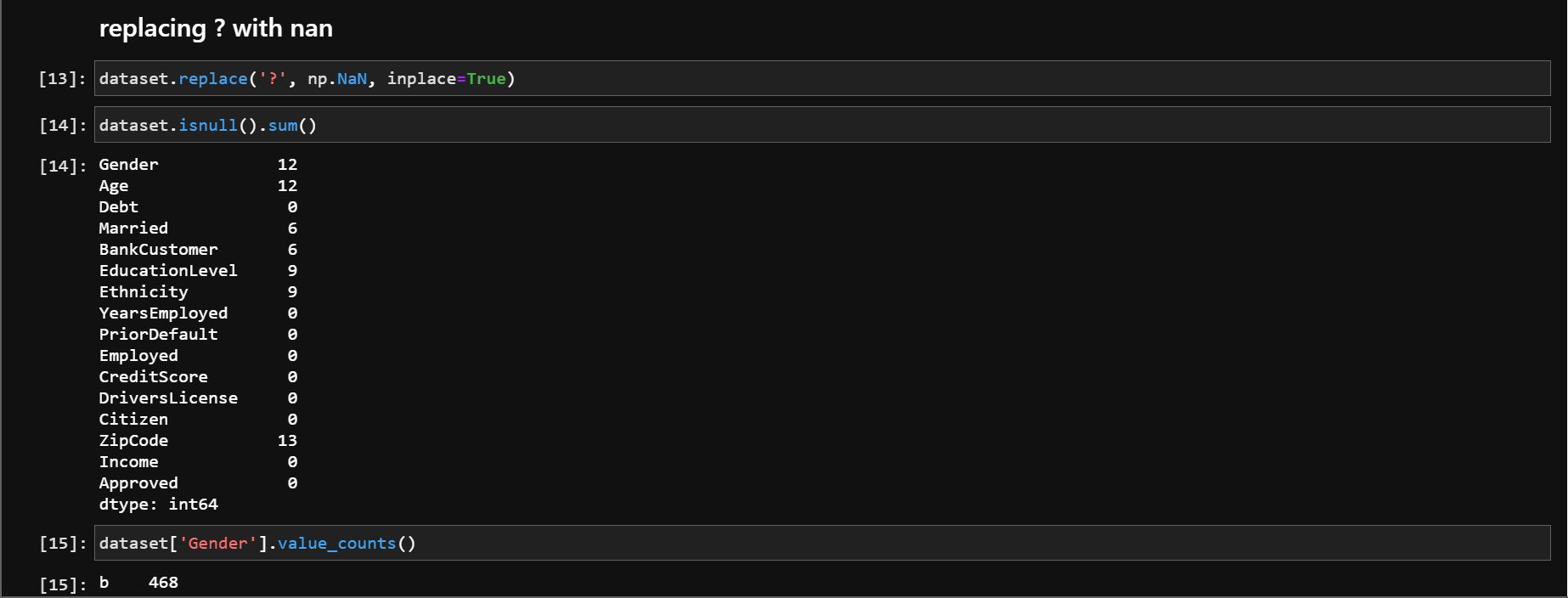
**7.1 Coding**

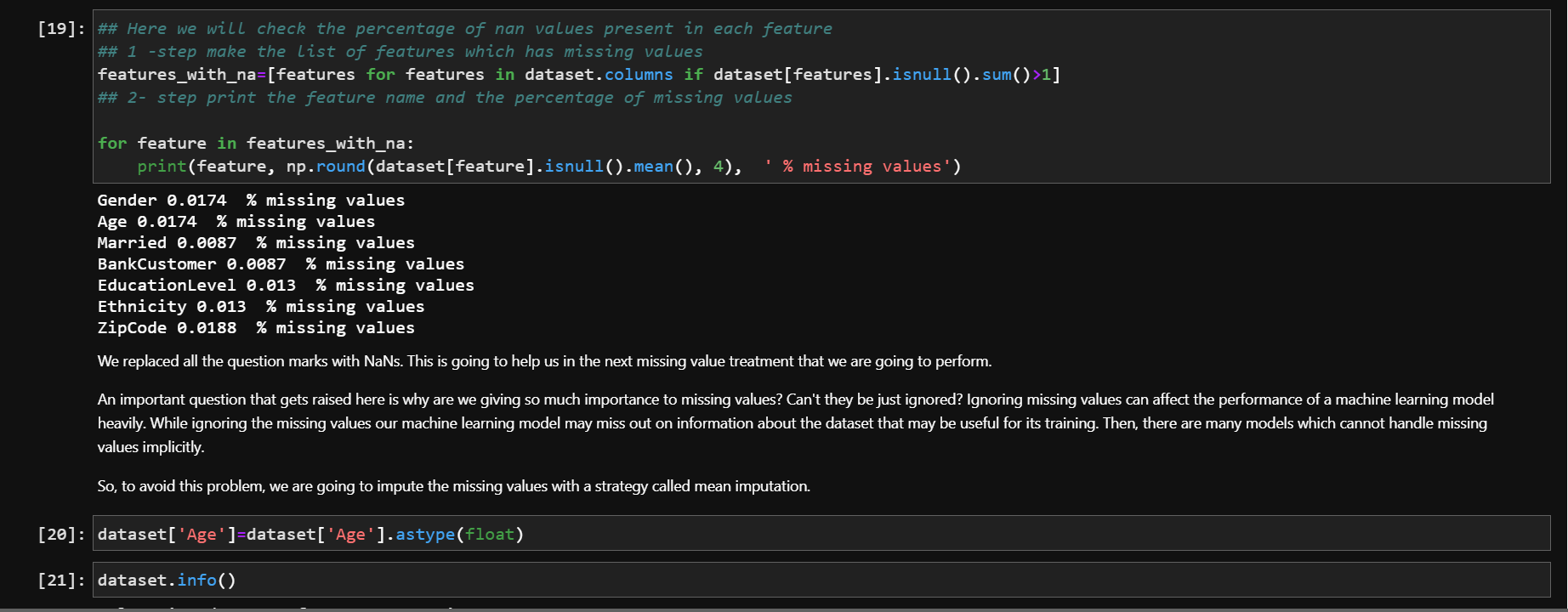
**Feature Engineering**

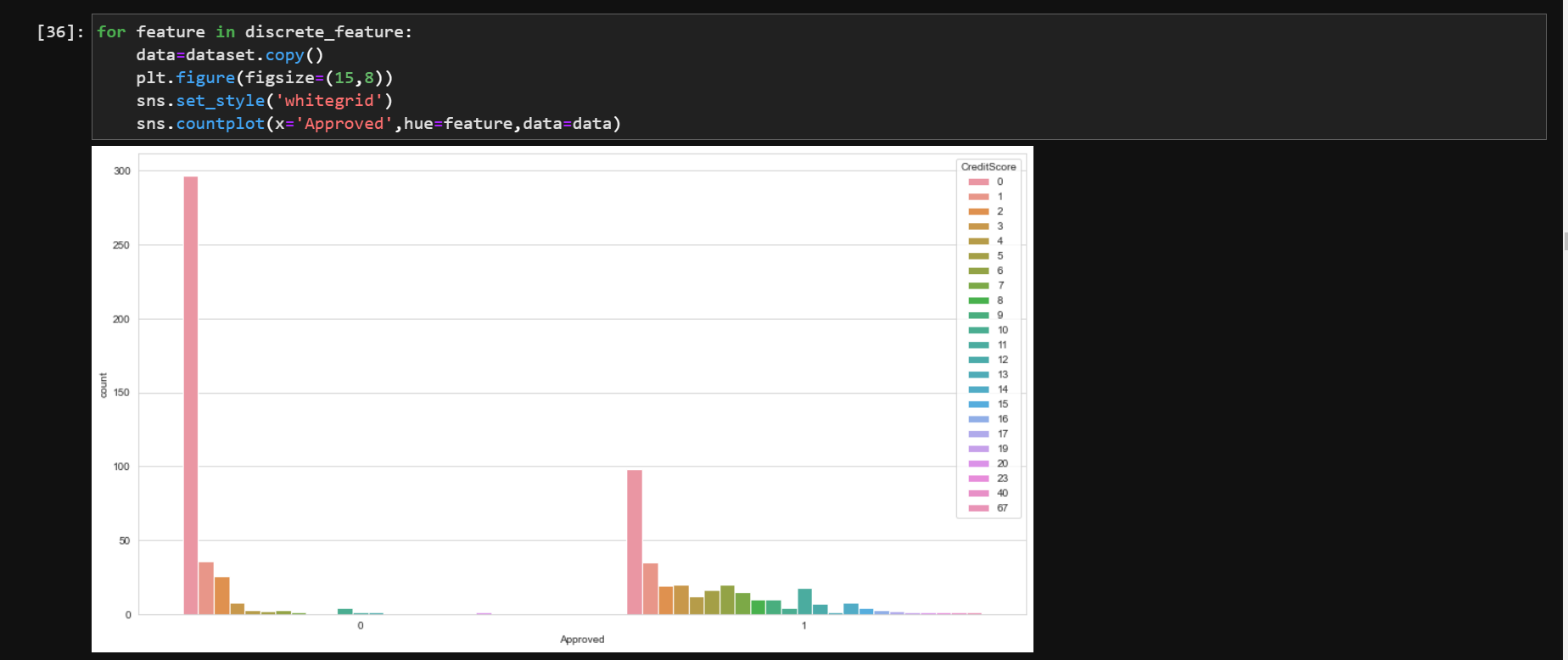


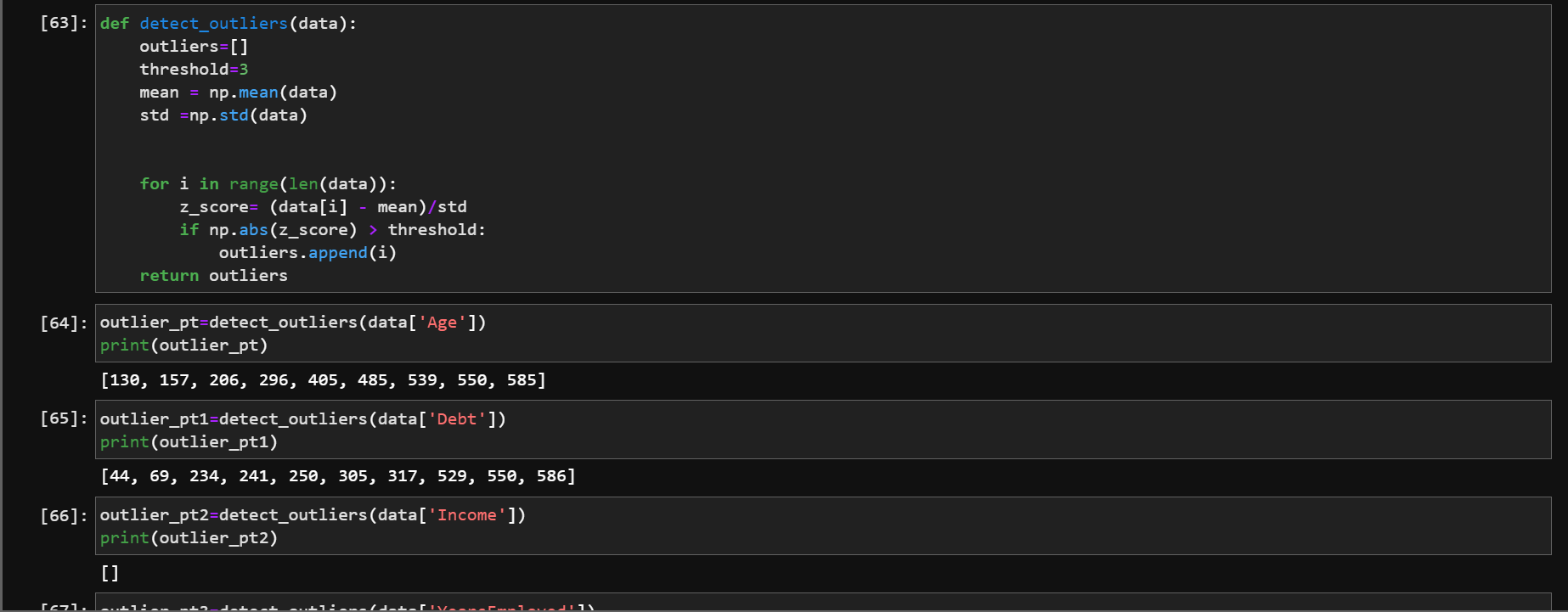






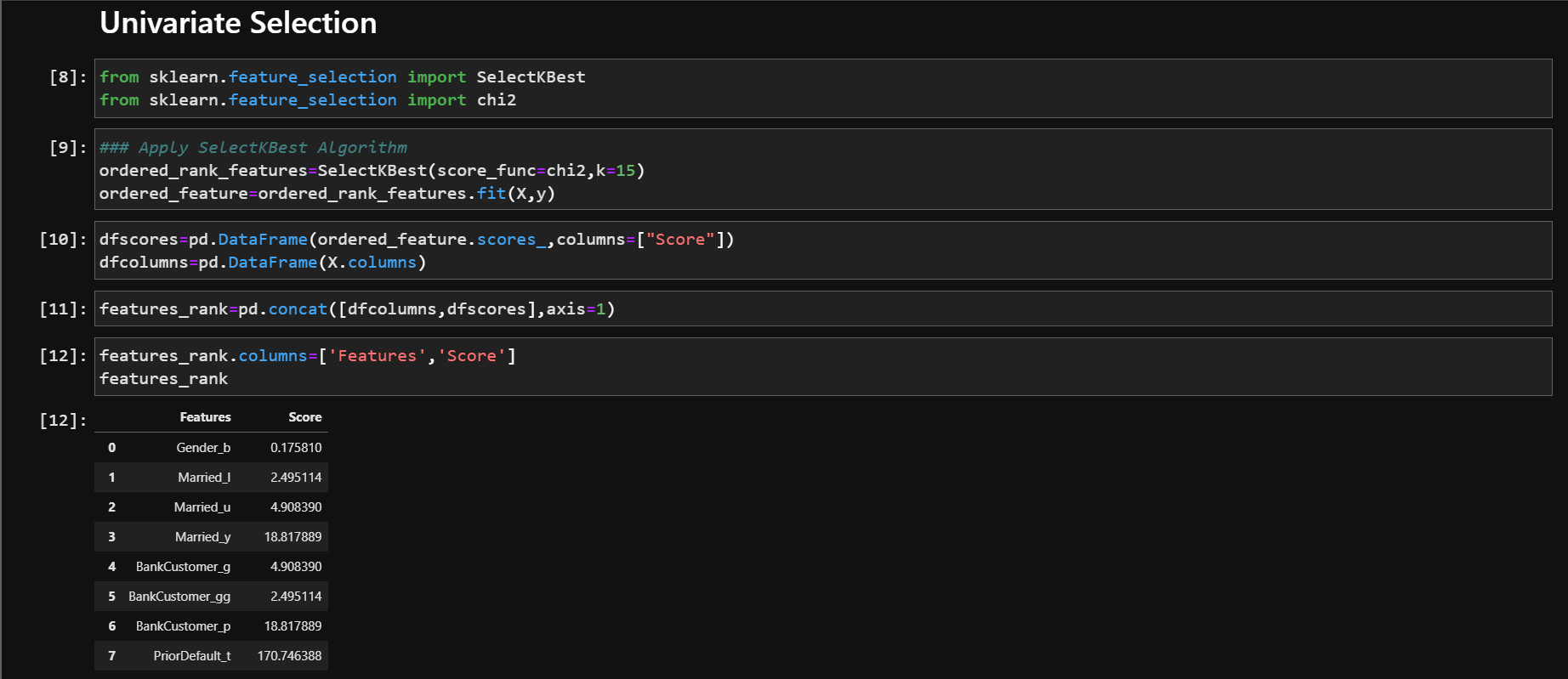


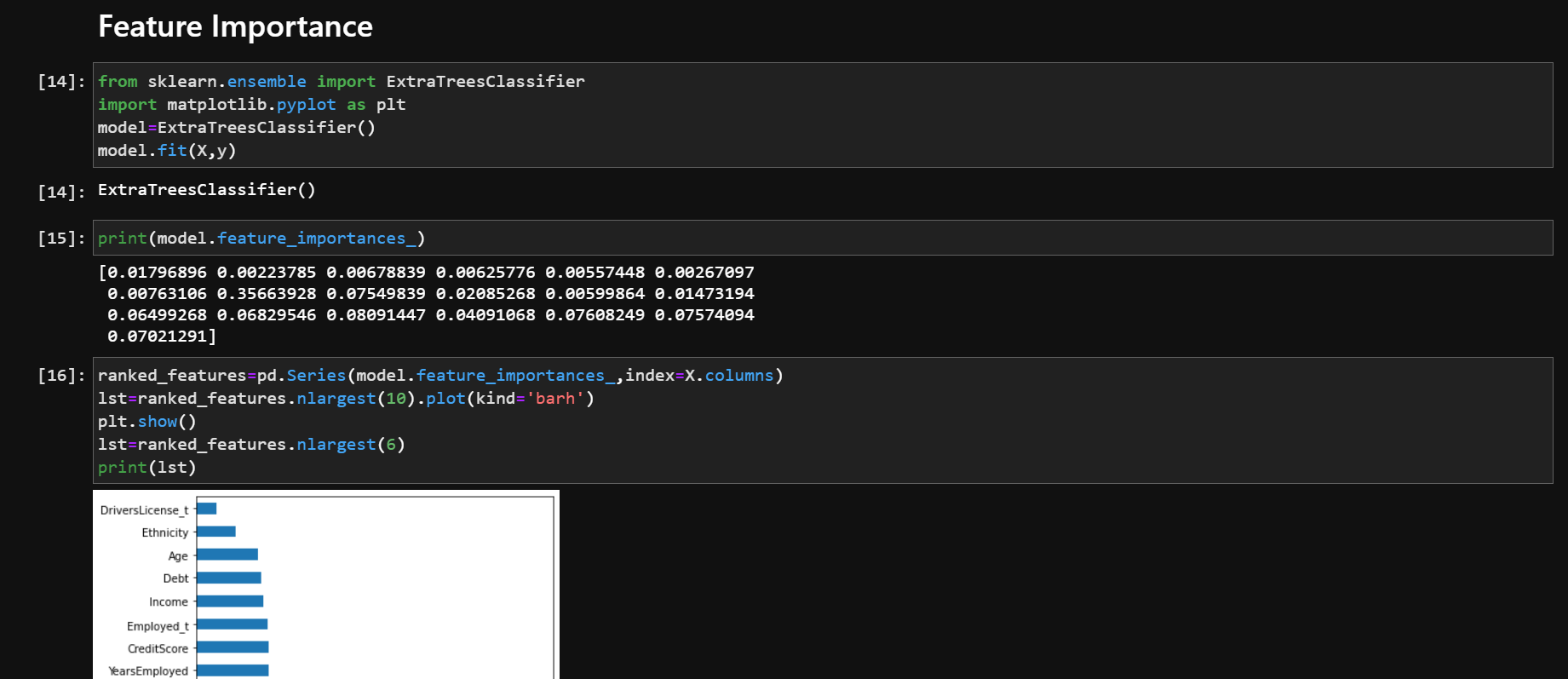




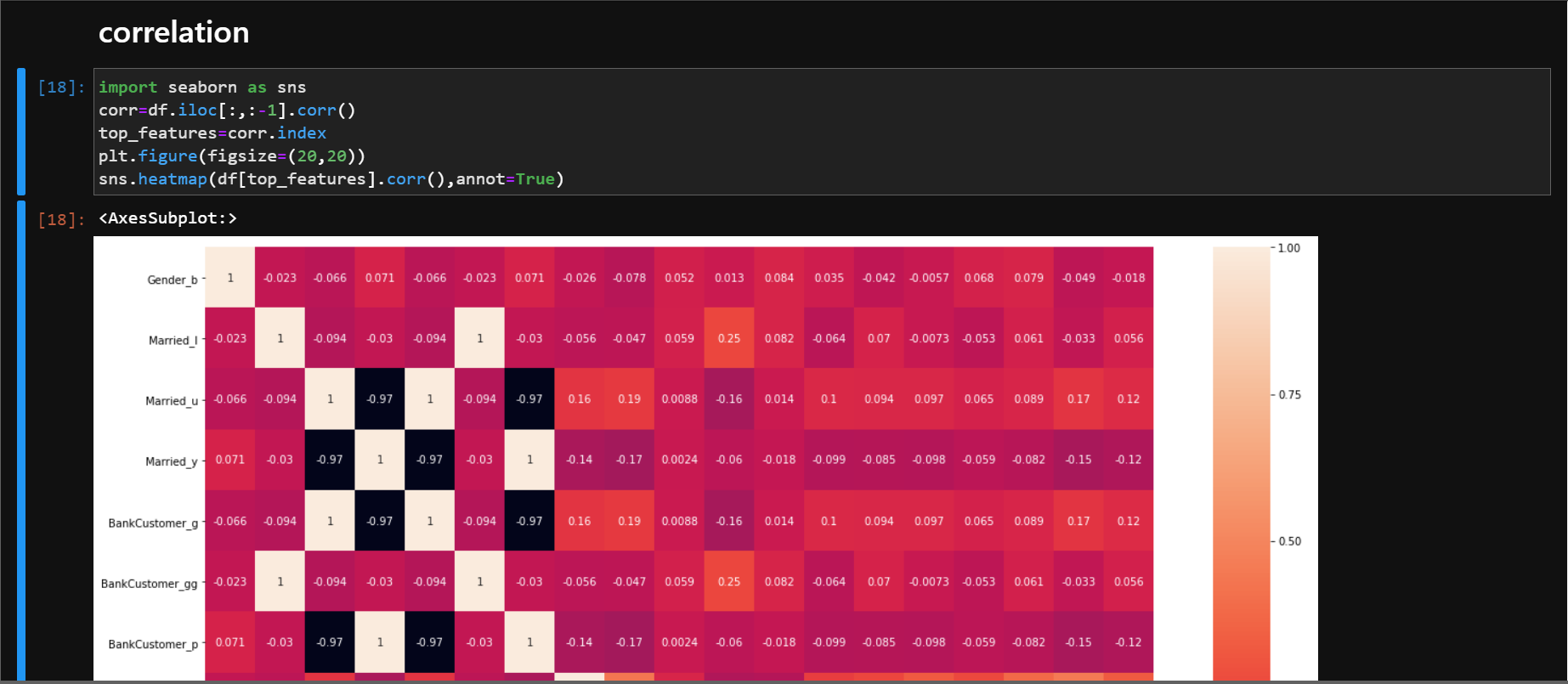


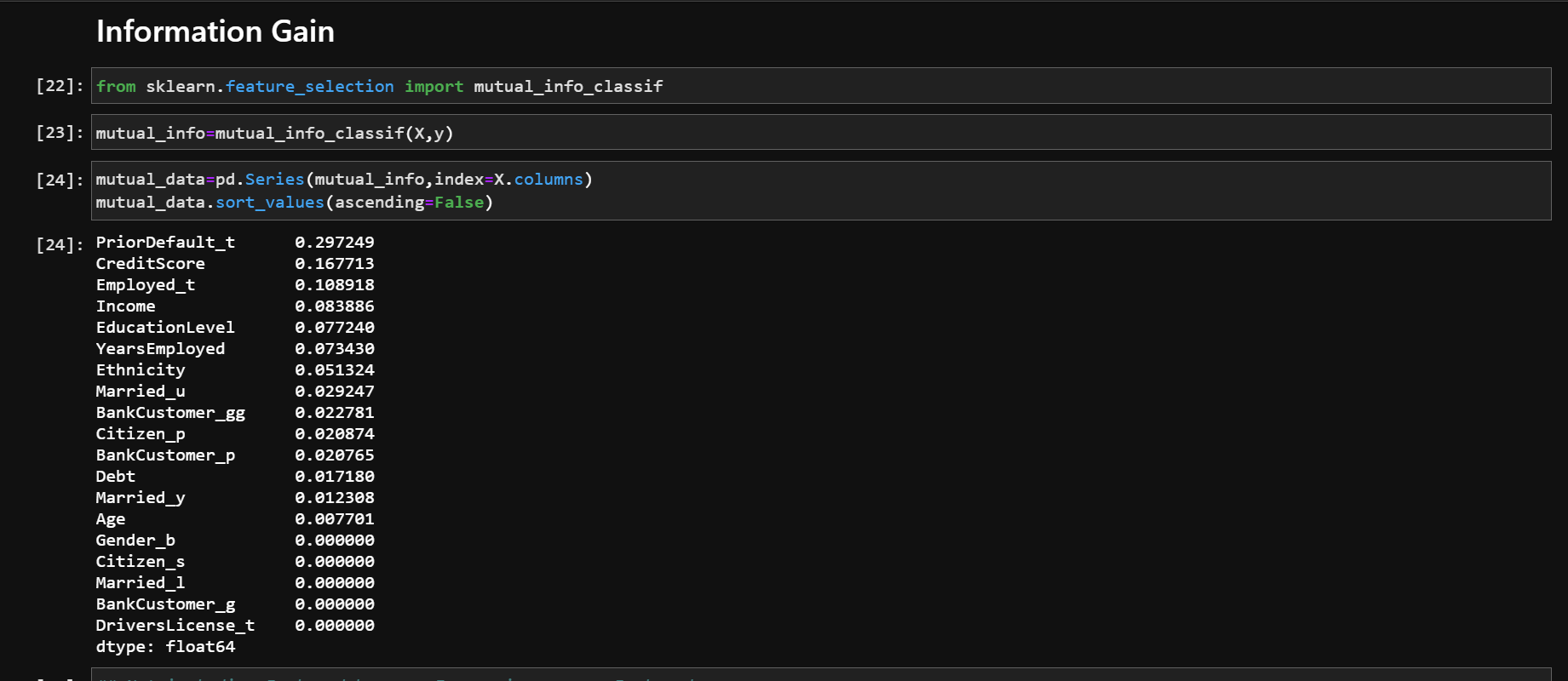
**Feature Selection**

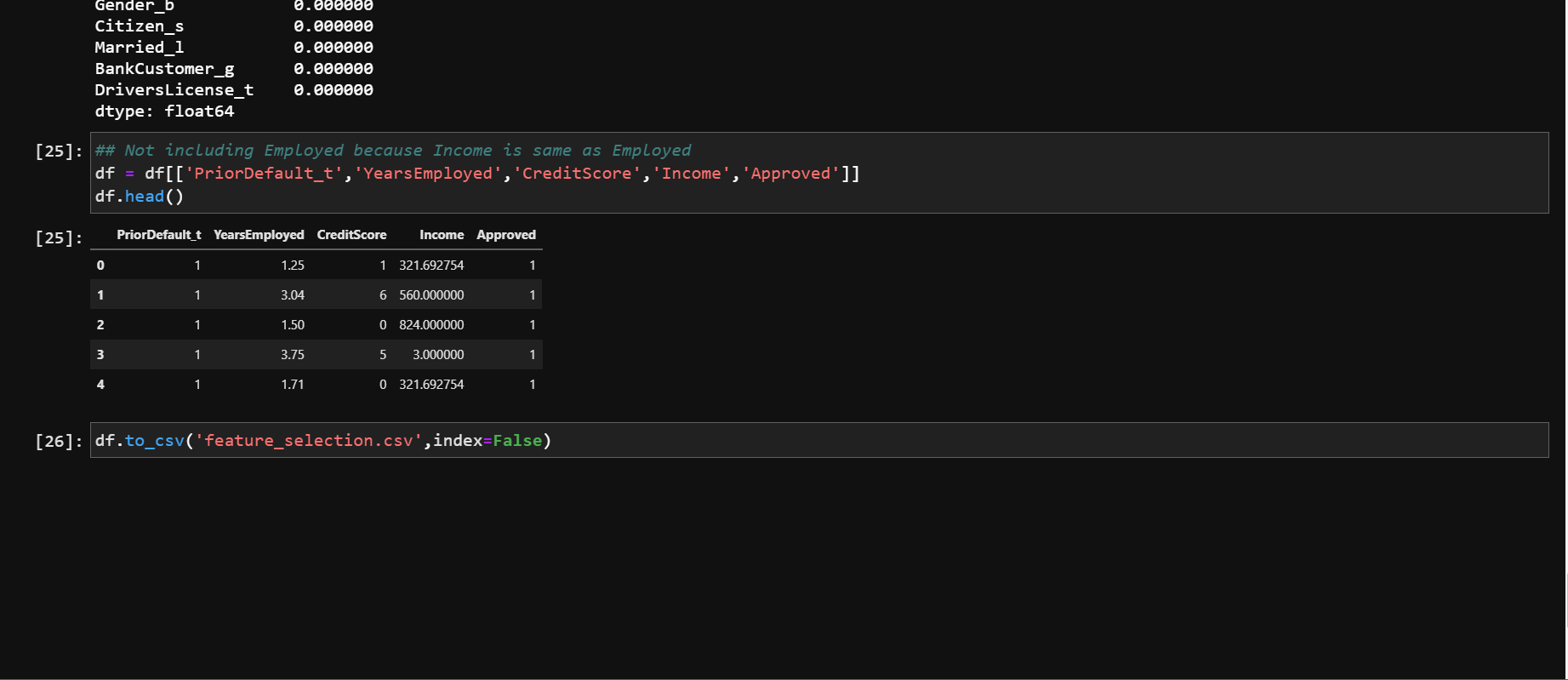




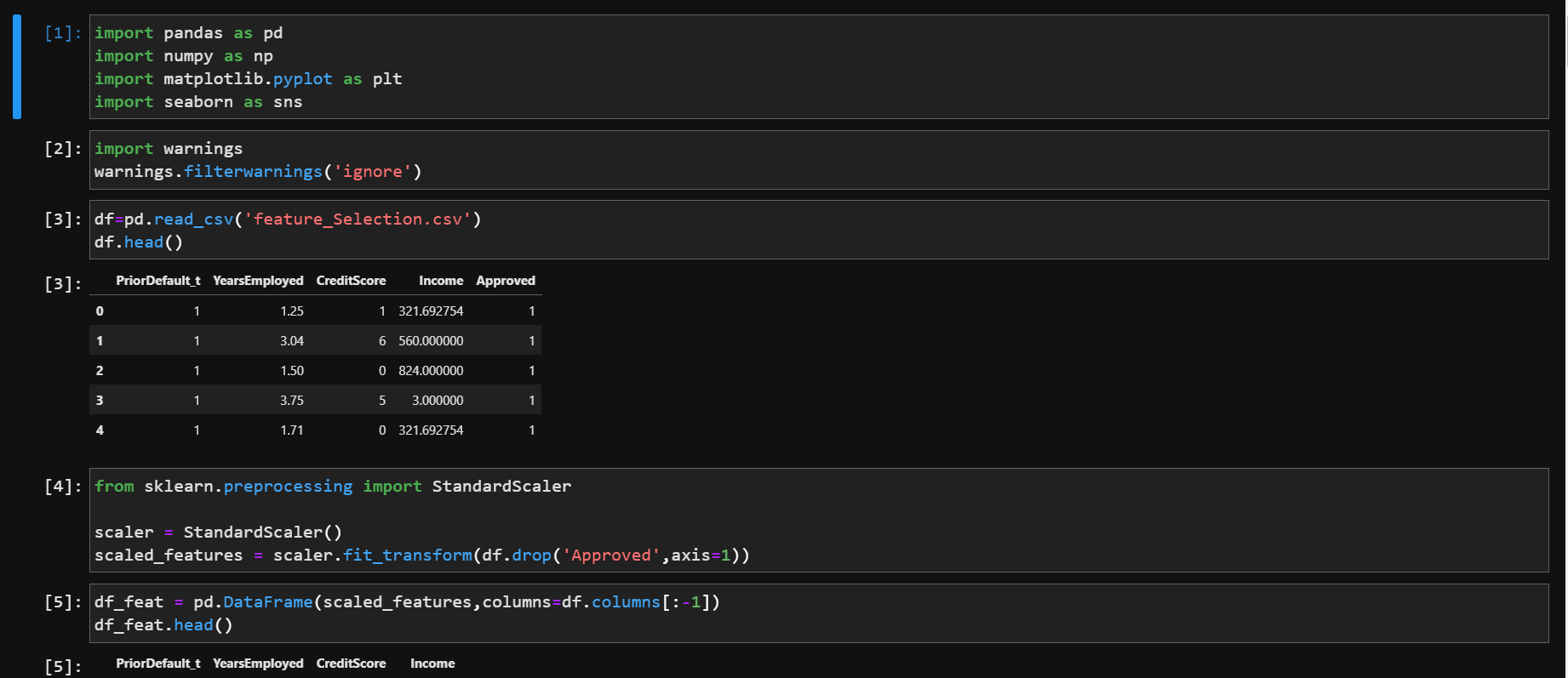






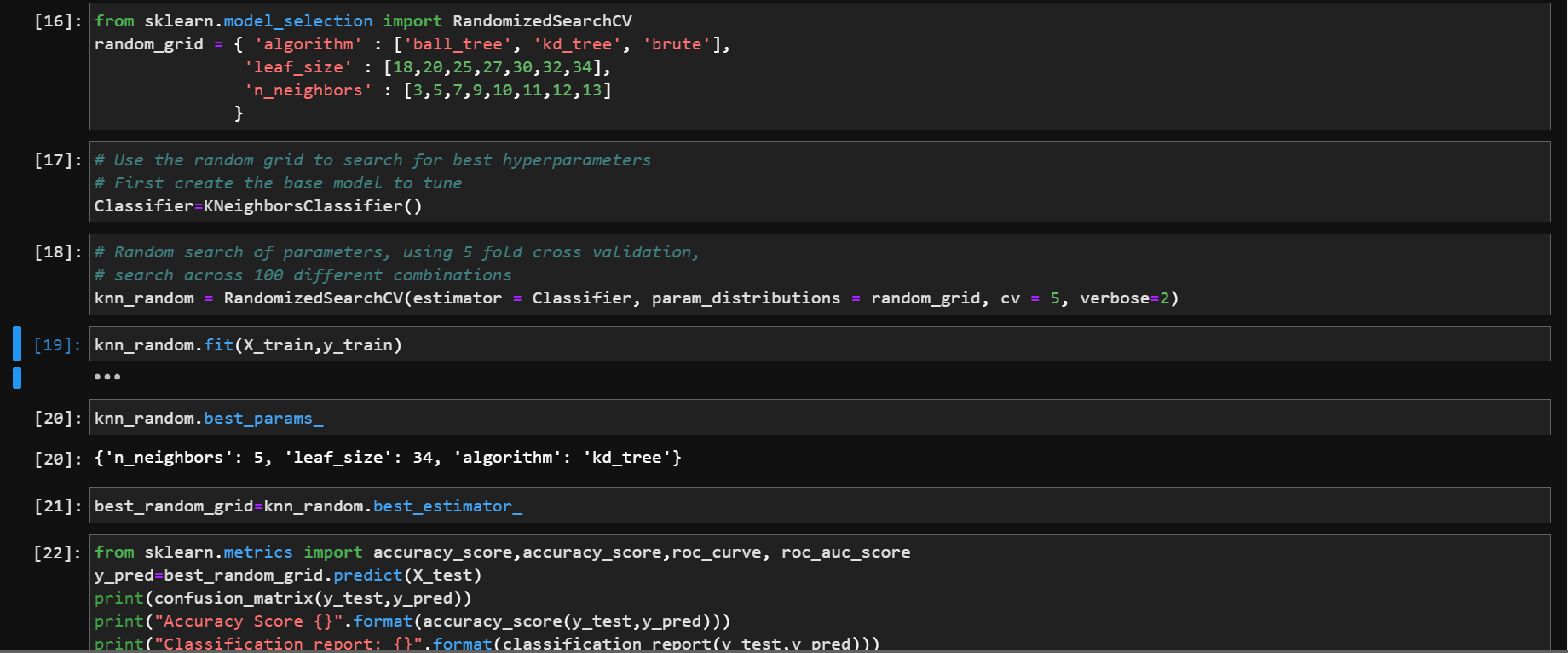


**Training with KNN**

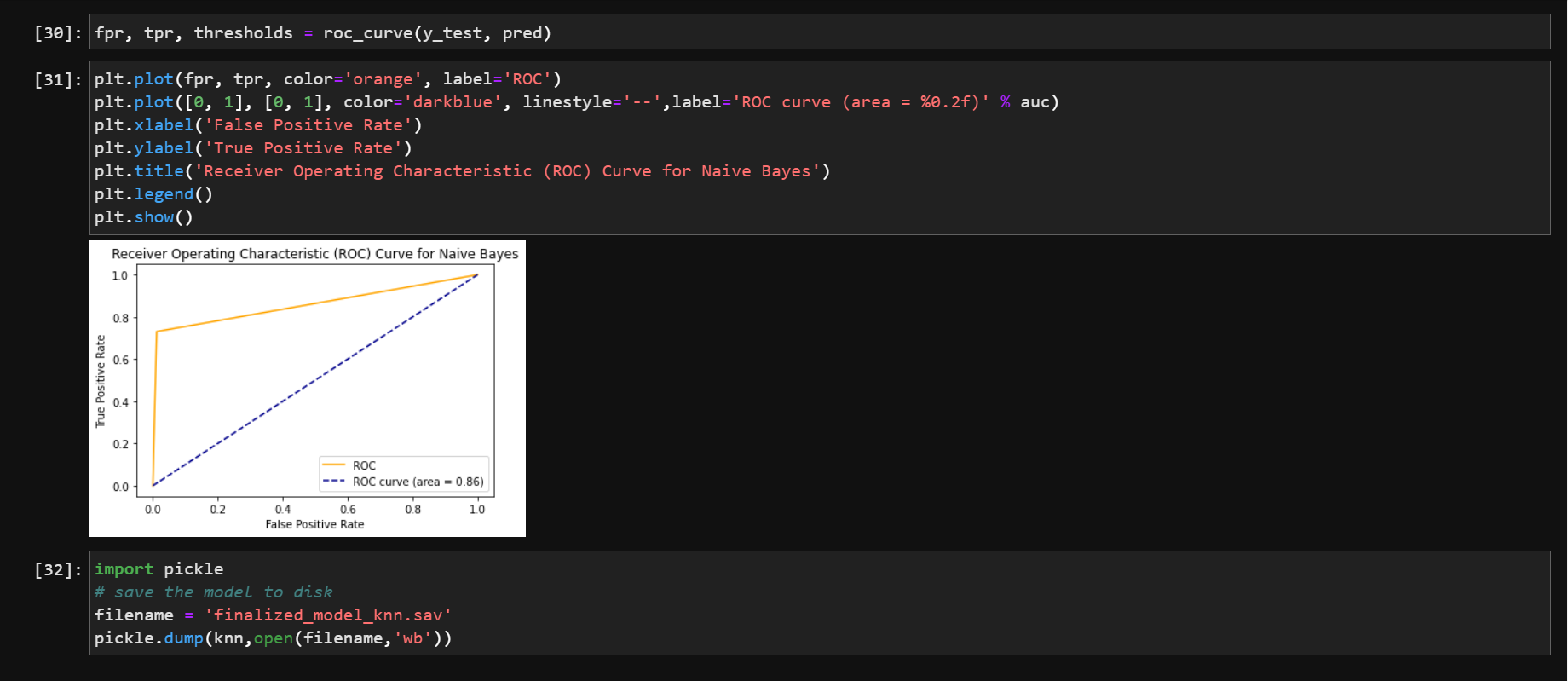






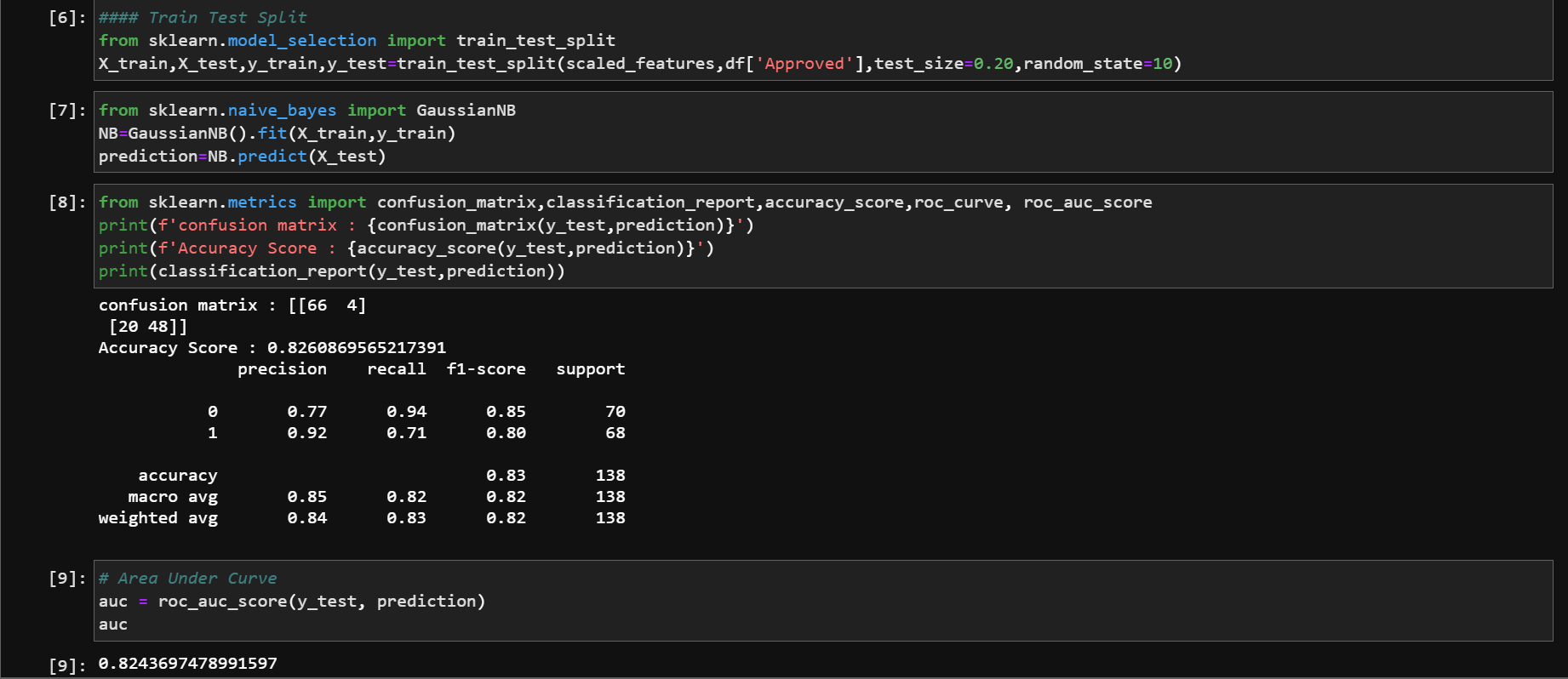


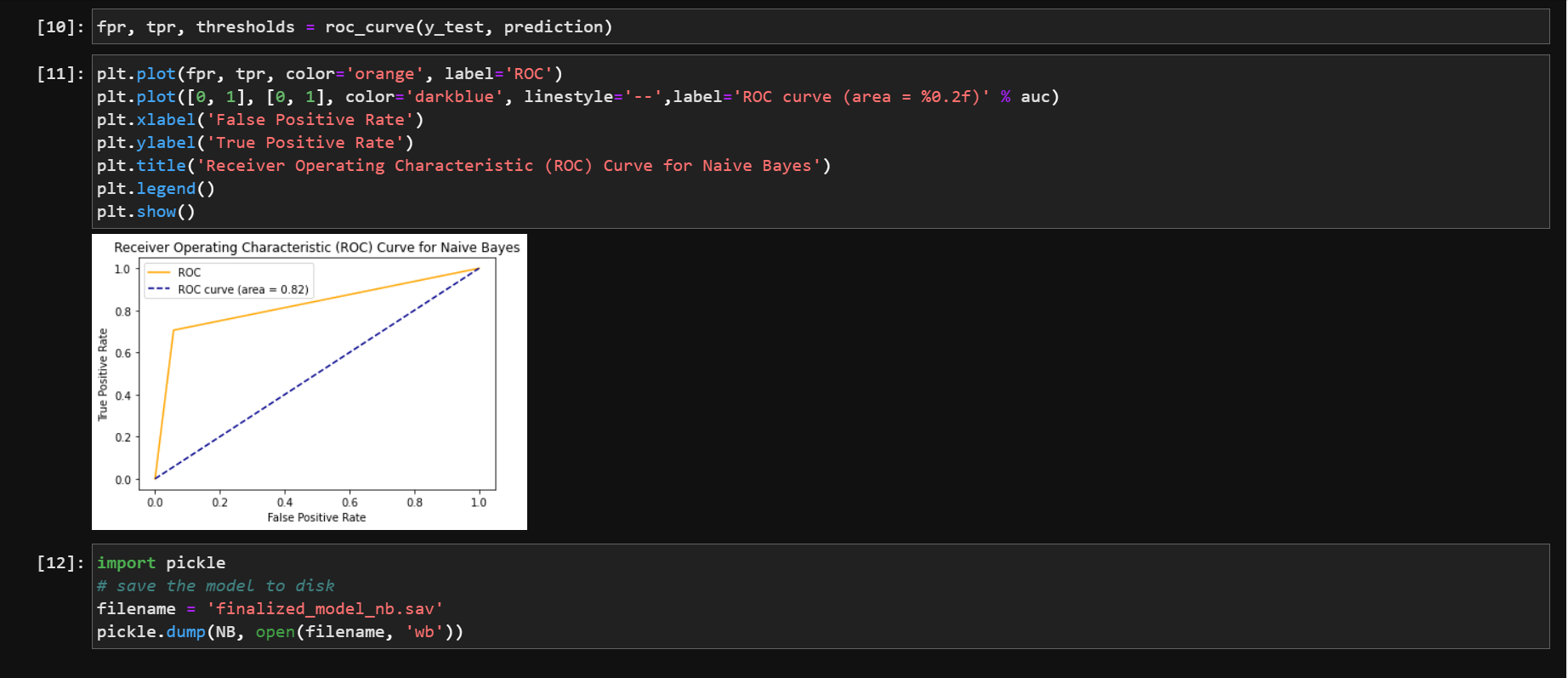




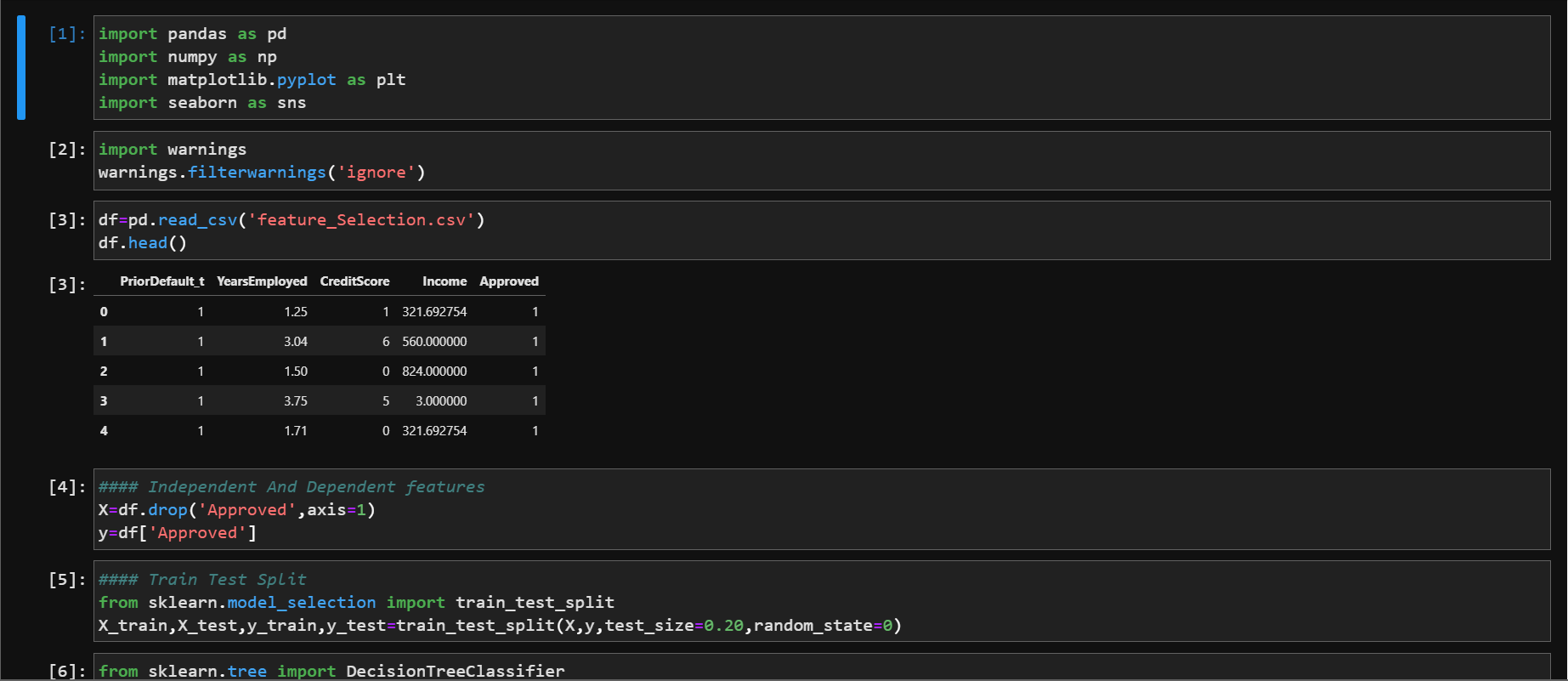
**Training with Naive Bayes**

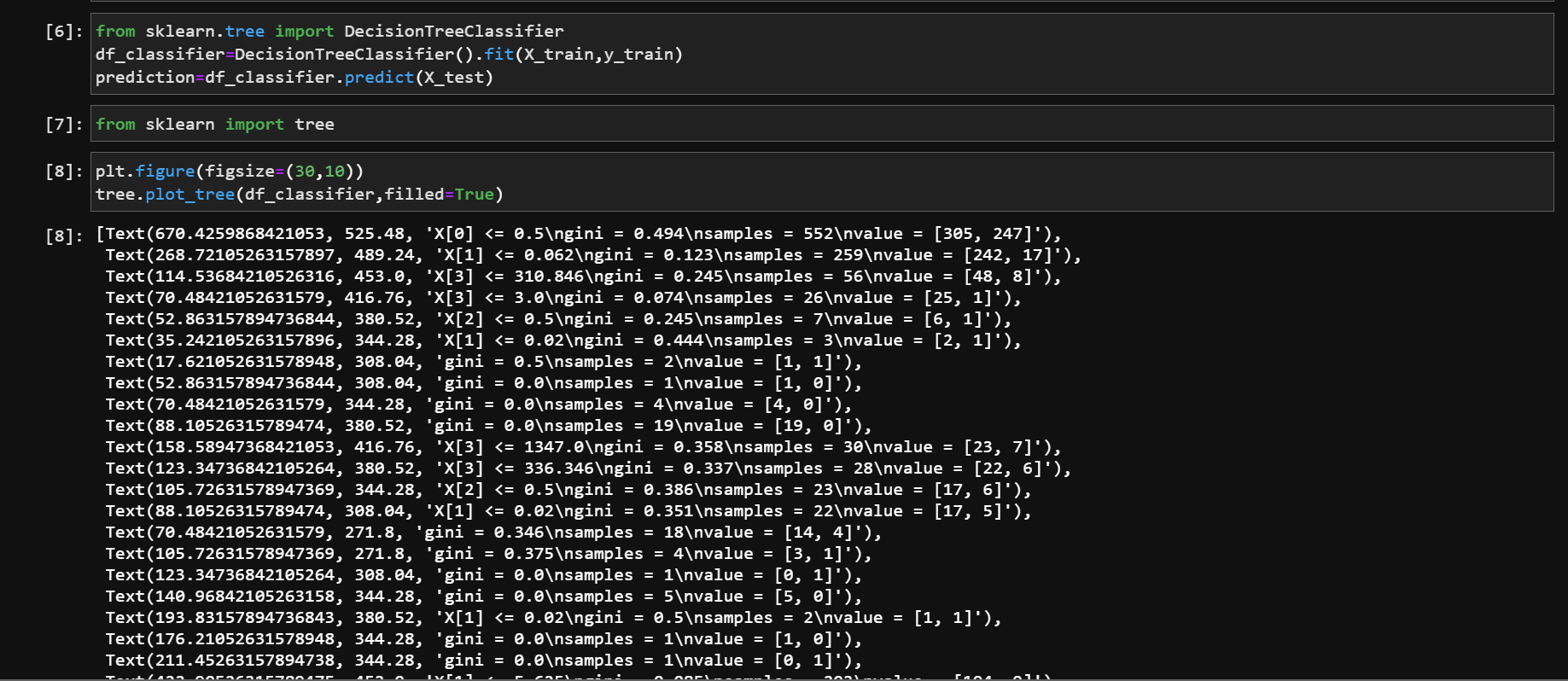


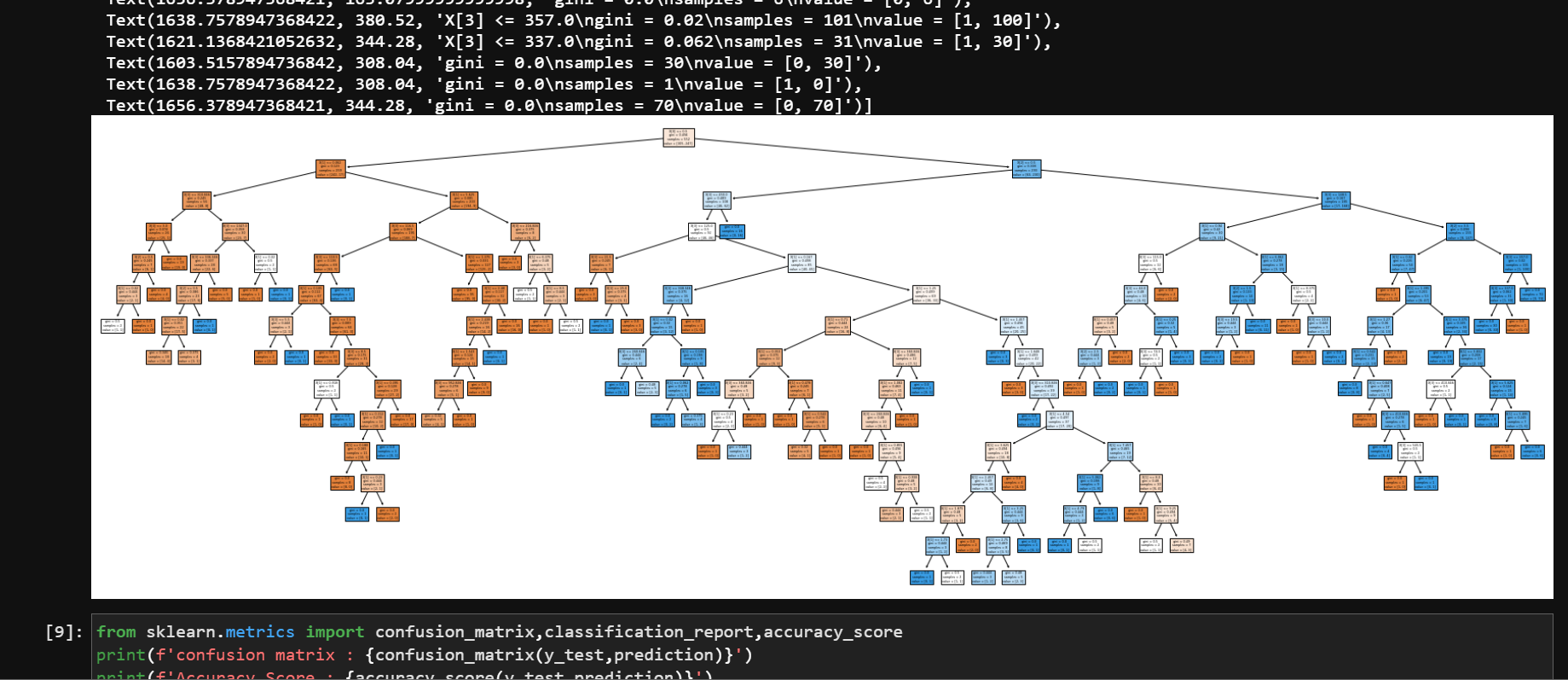


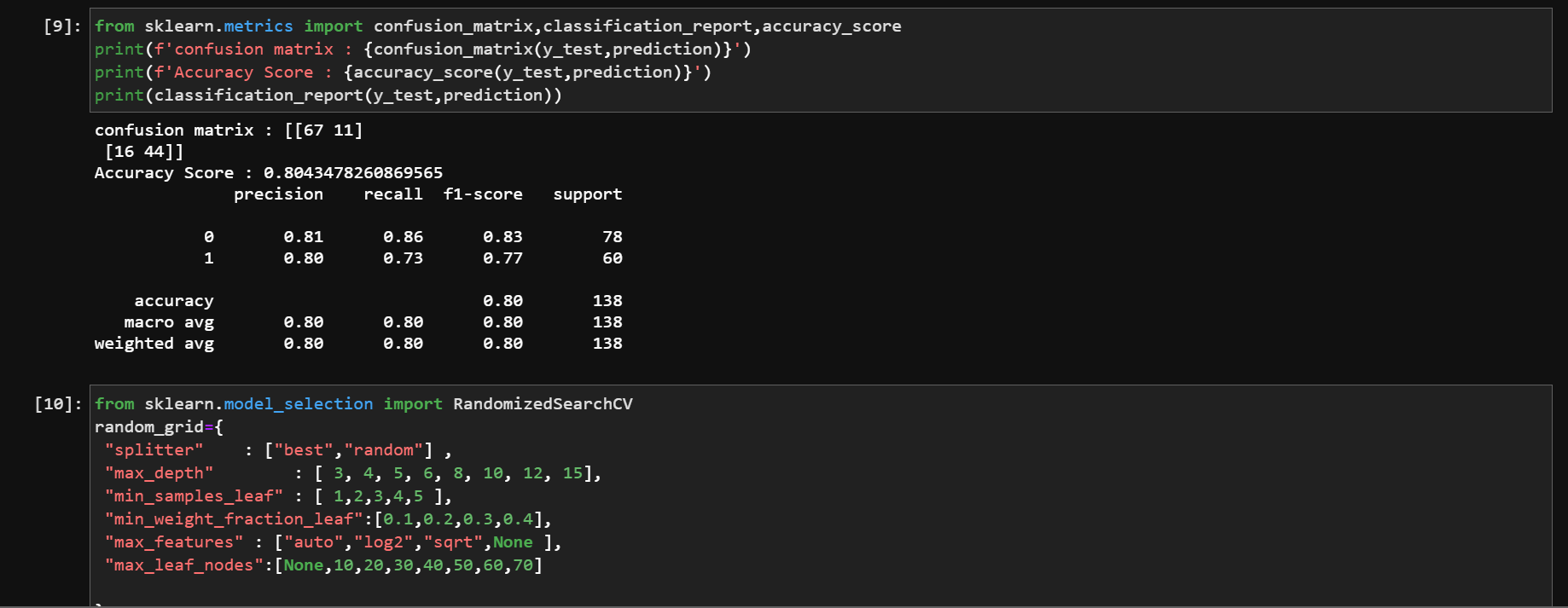


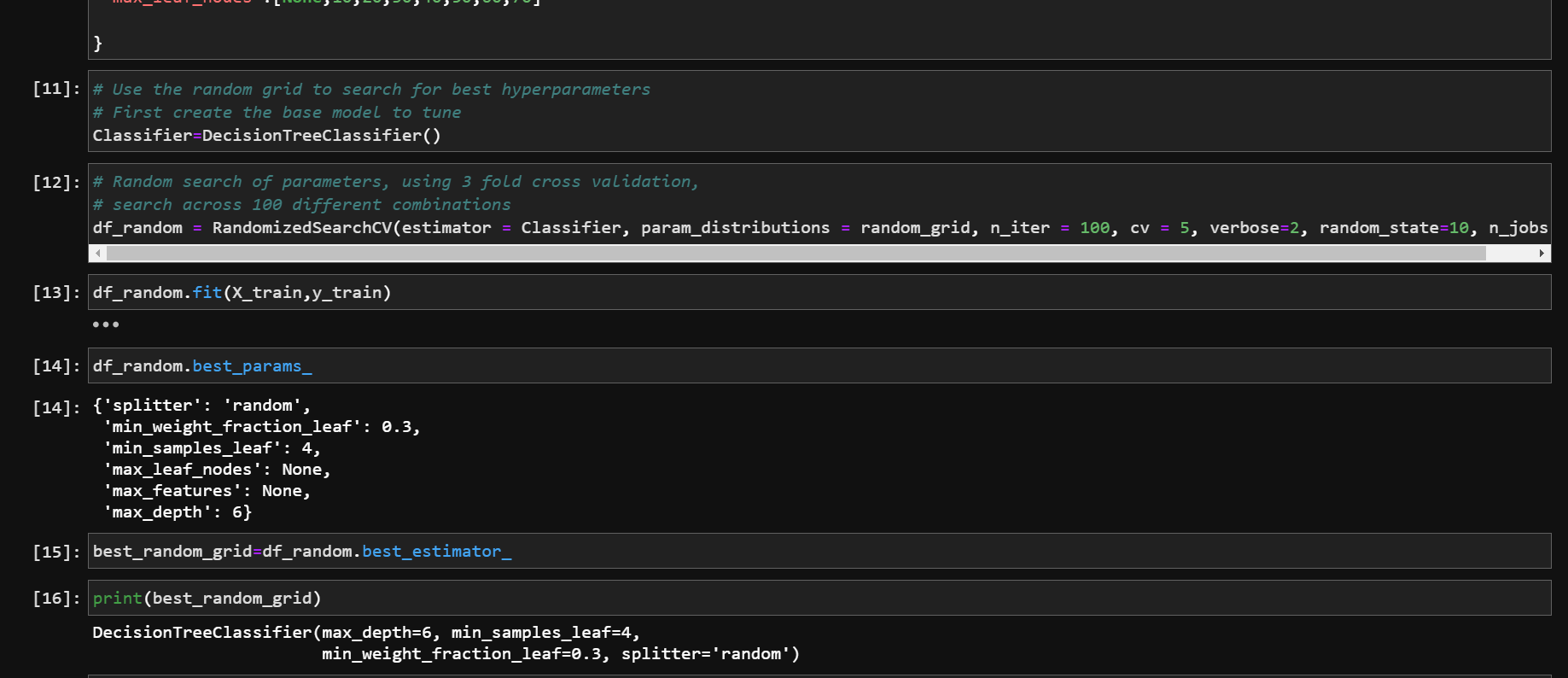
**Training with Decision Tree**

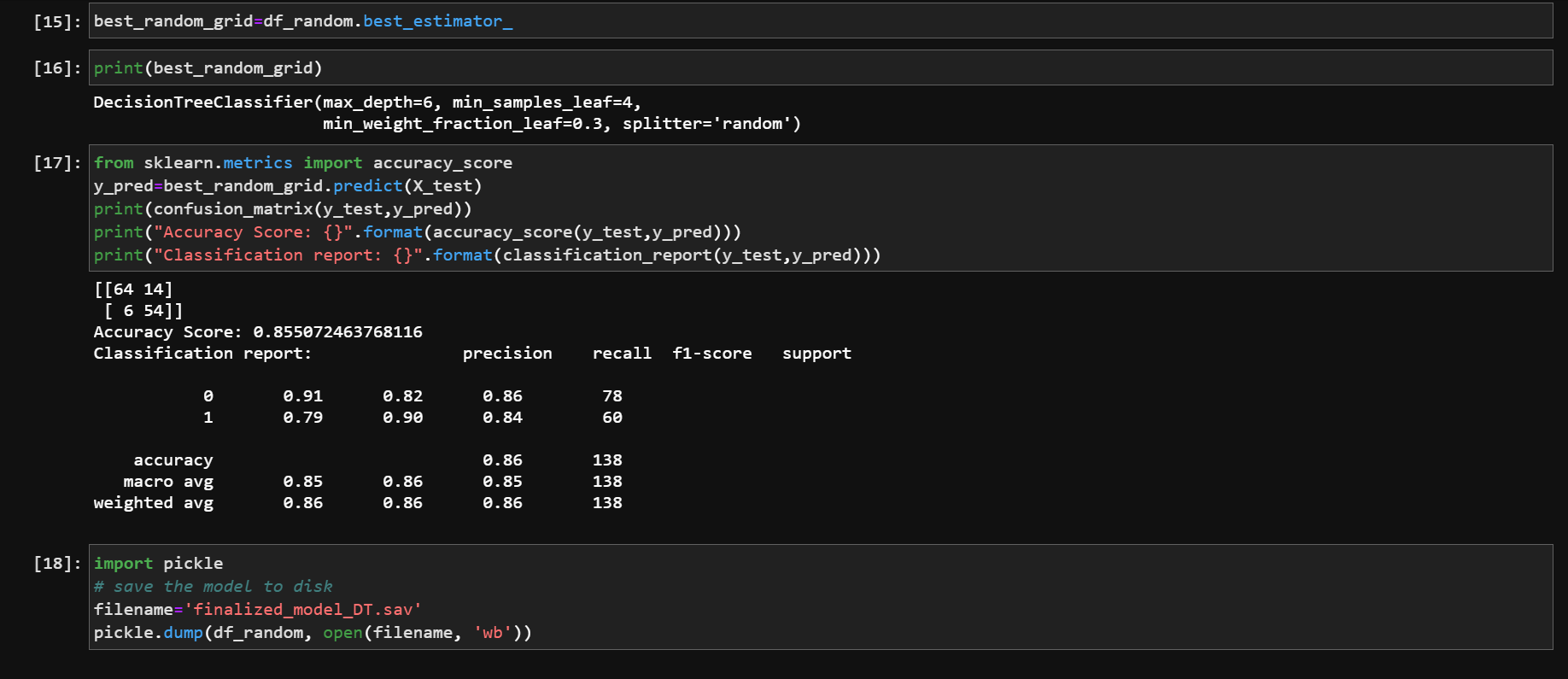




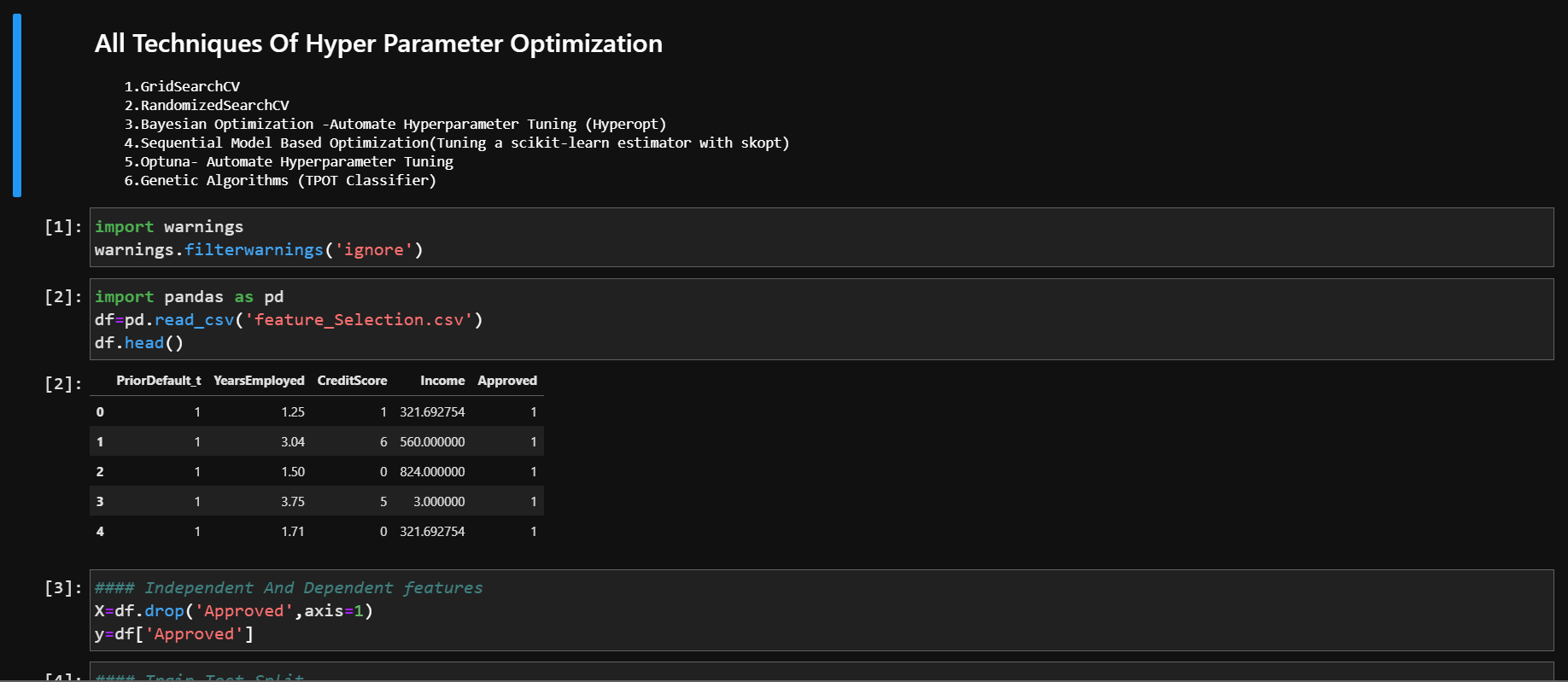


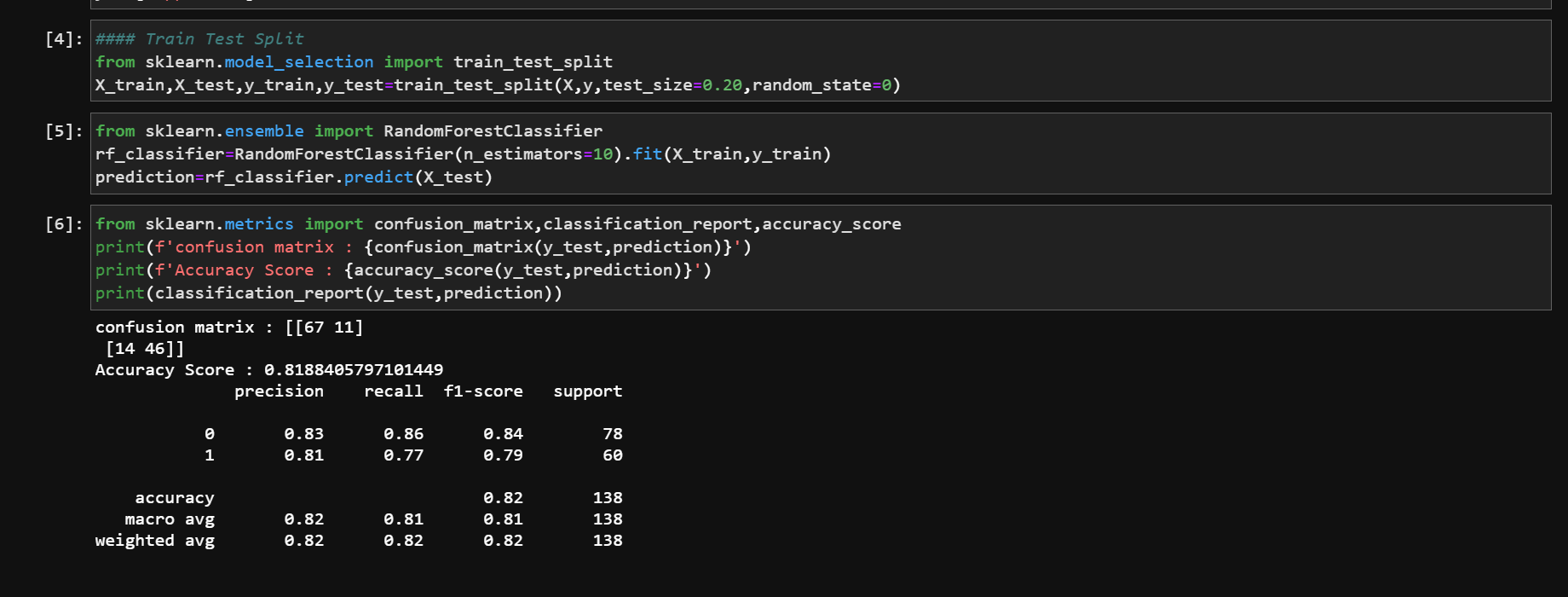




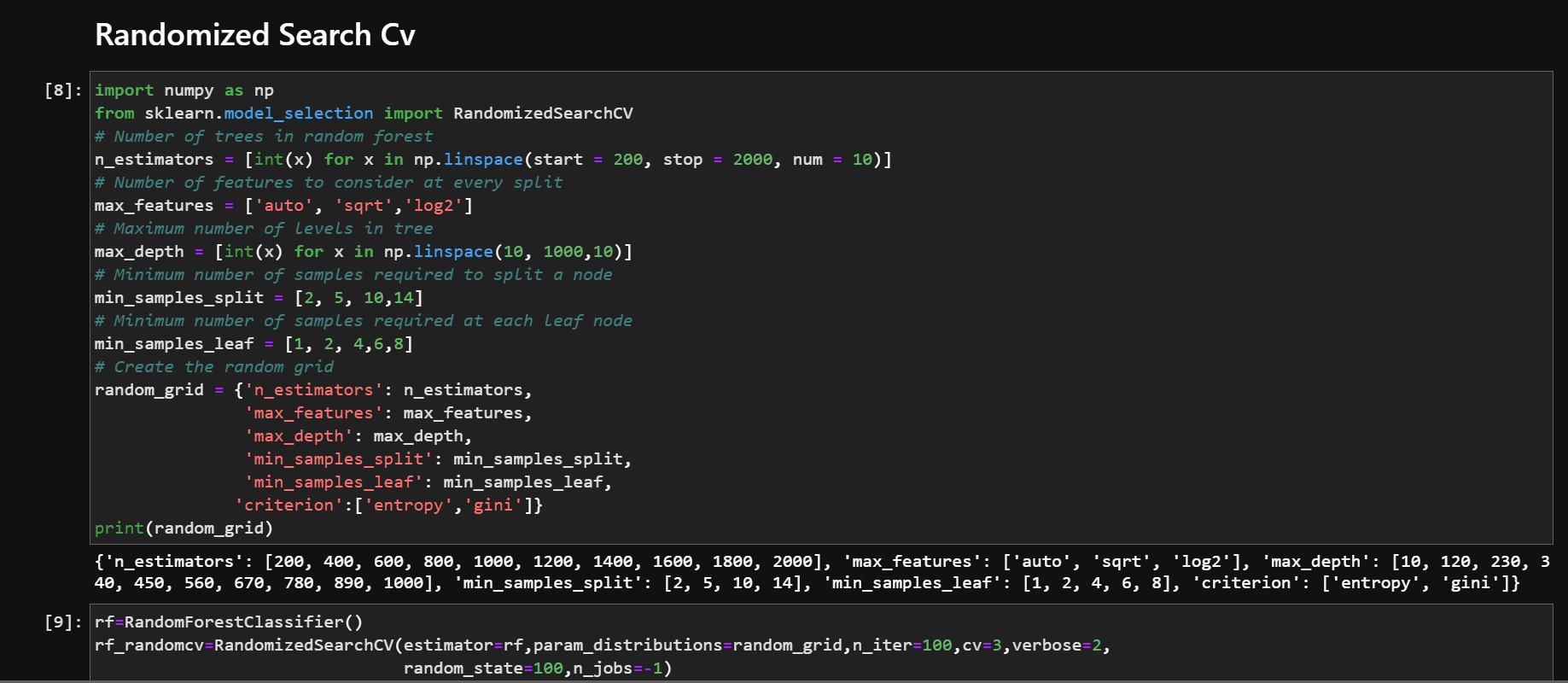


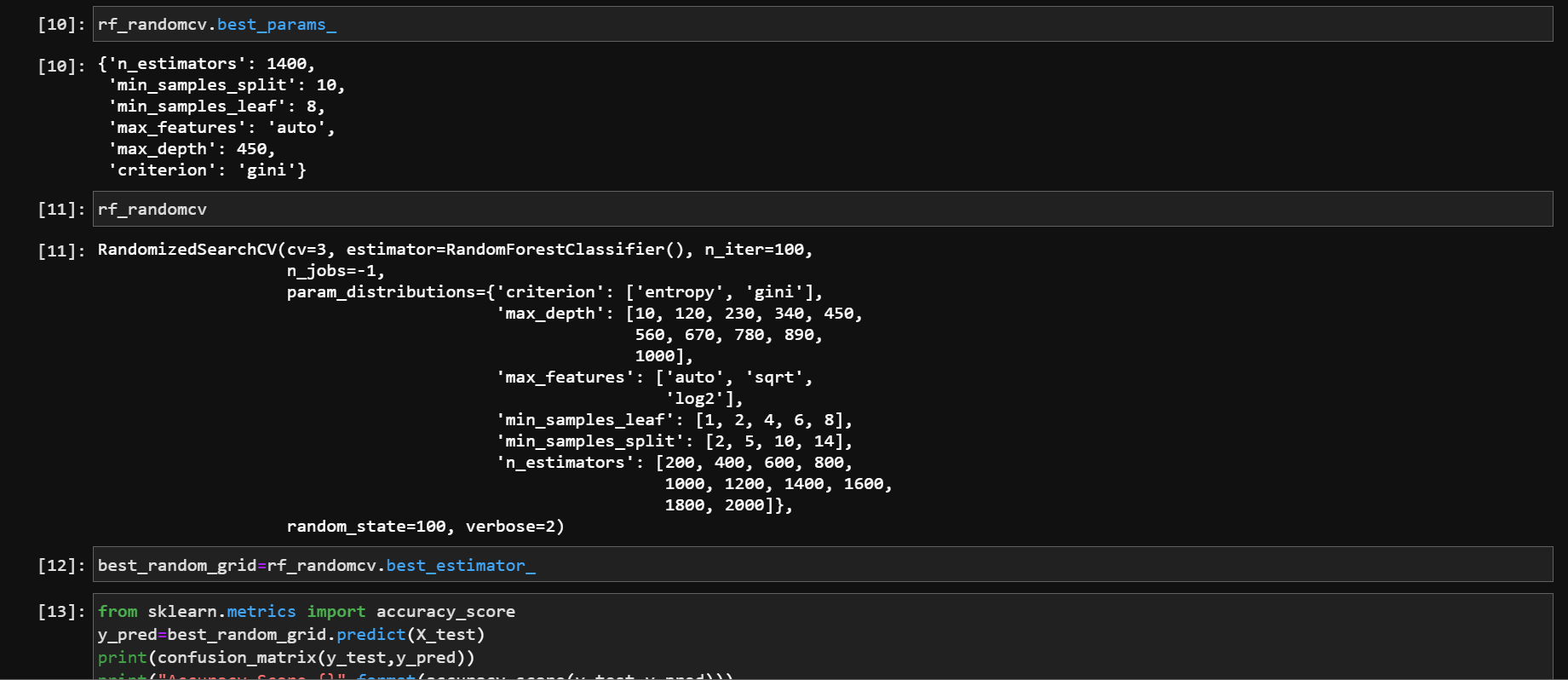
**Training with Random Forest and Hyper parameter tuning**

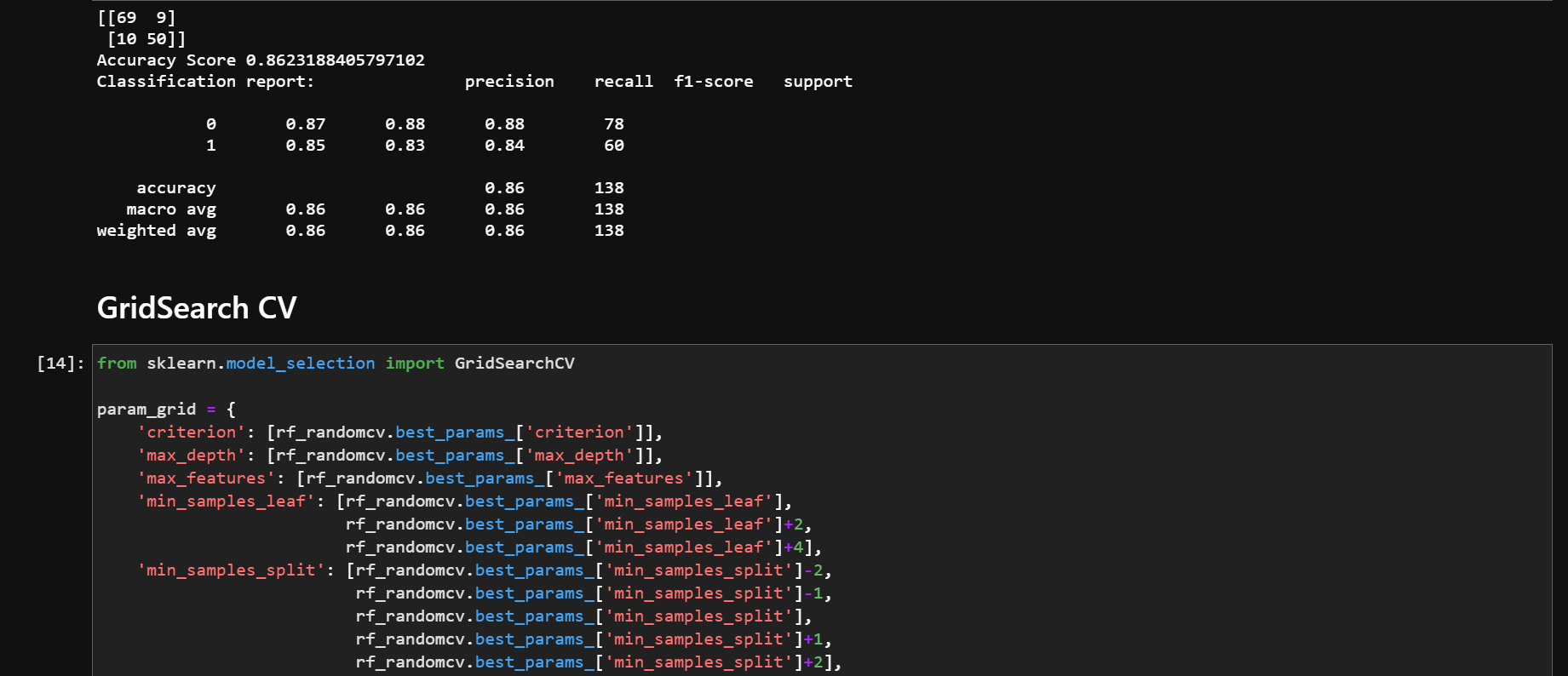


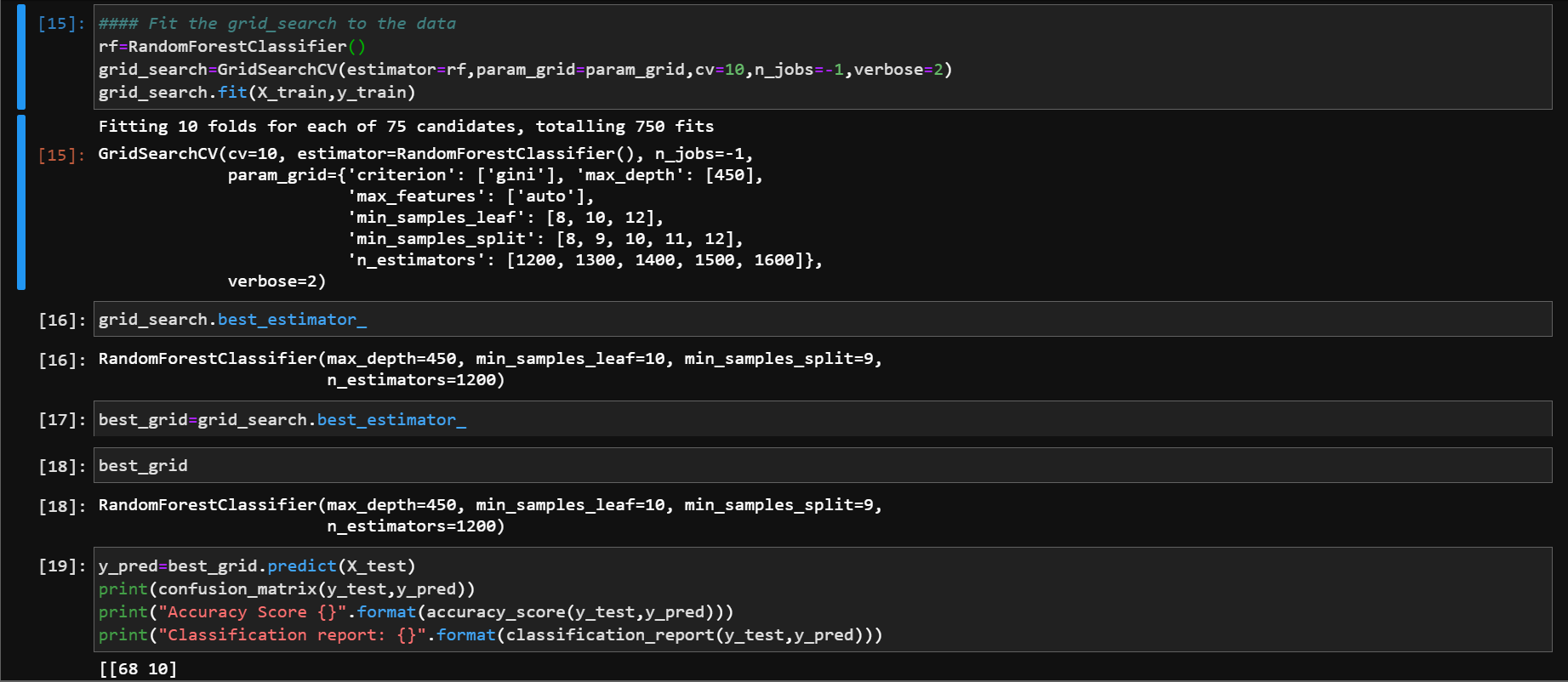


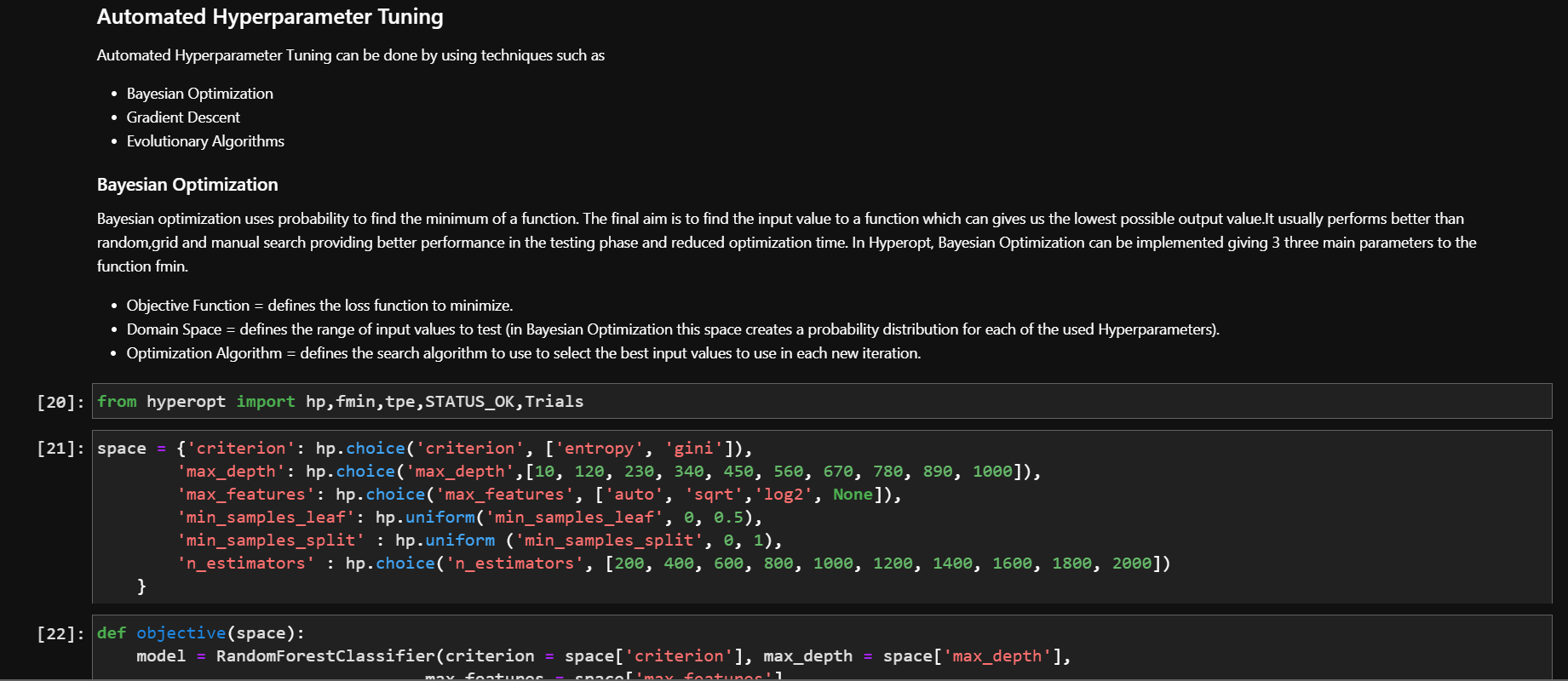


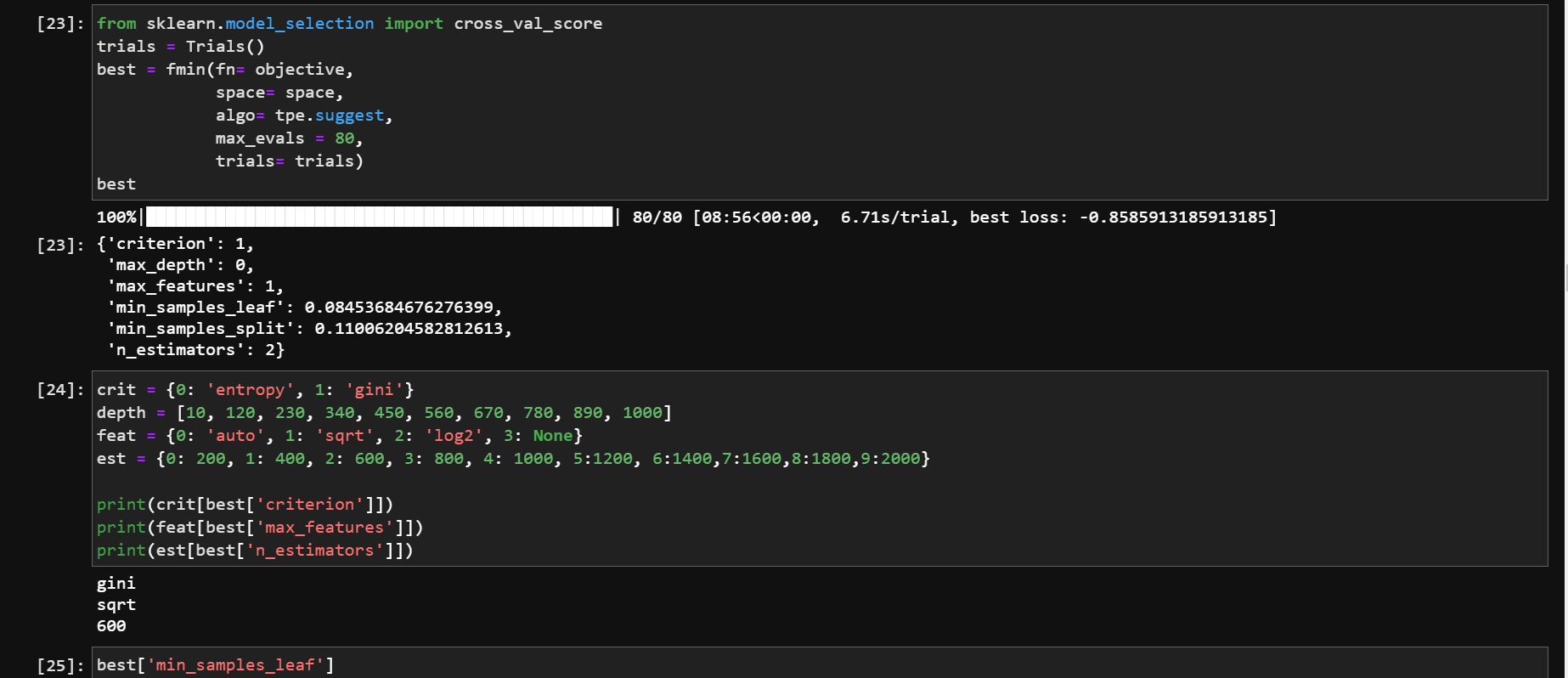


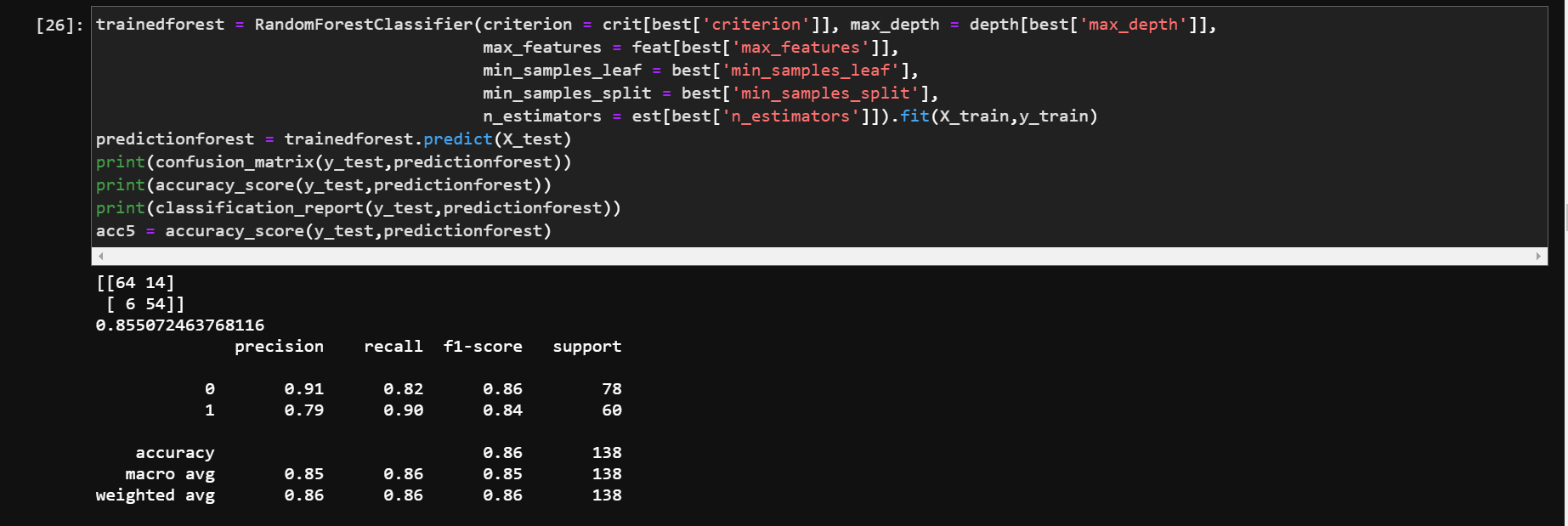


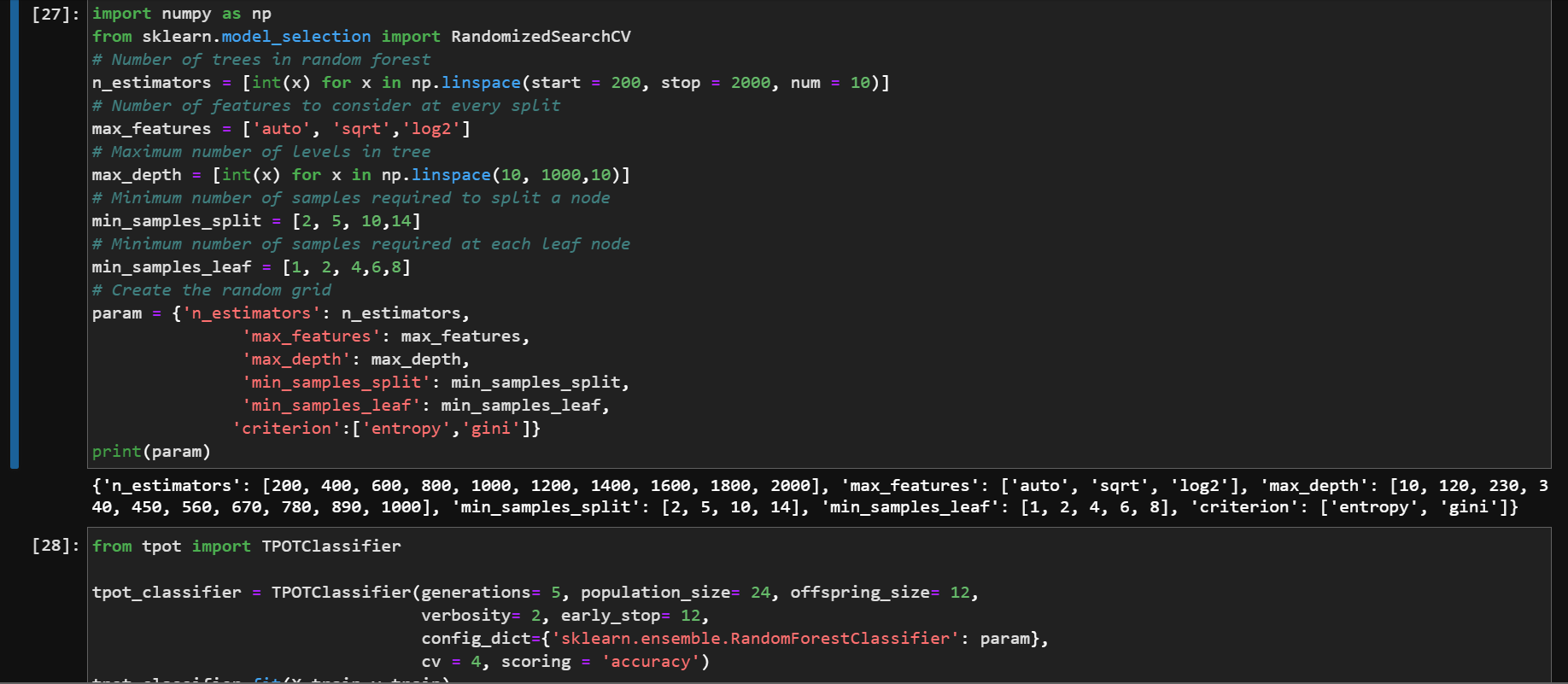


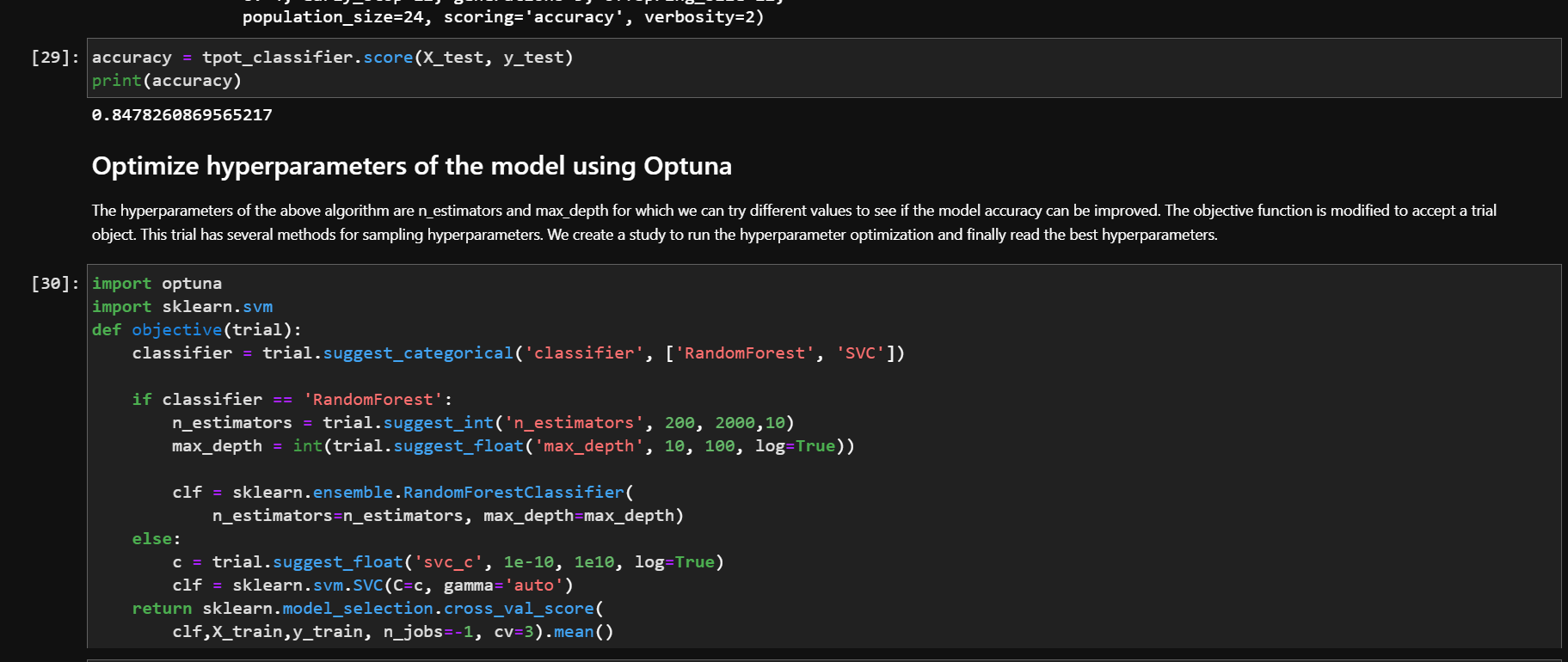


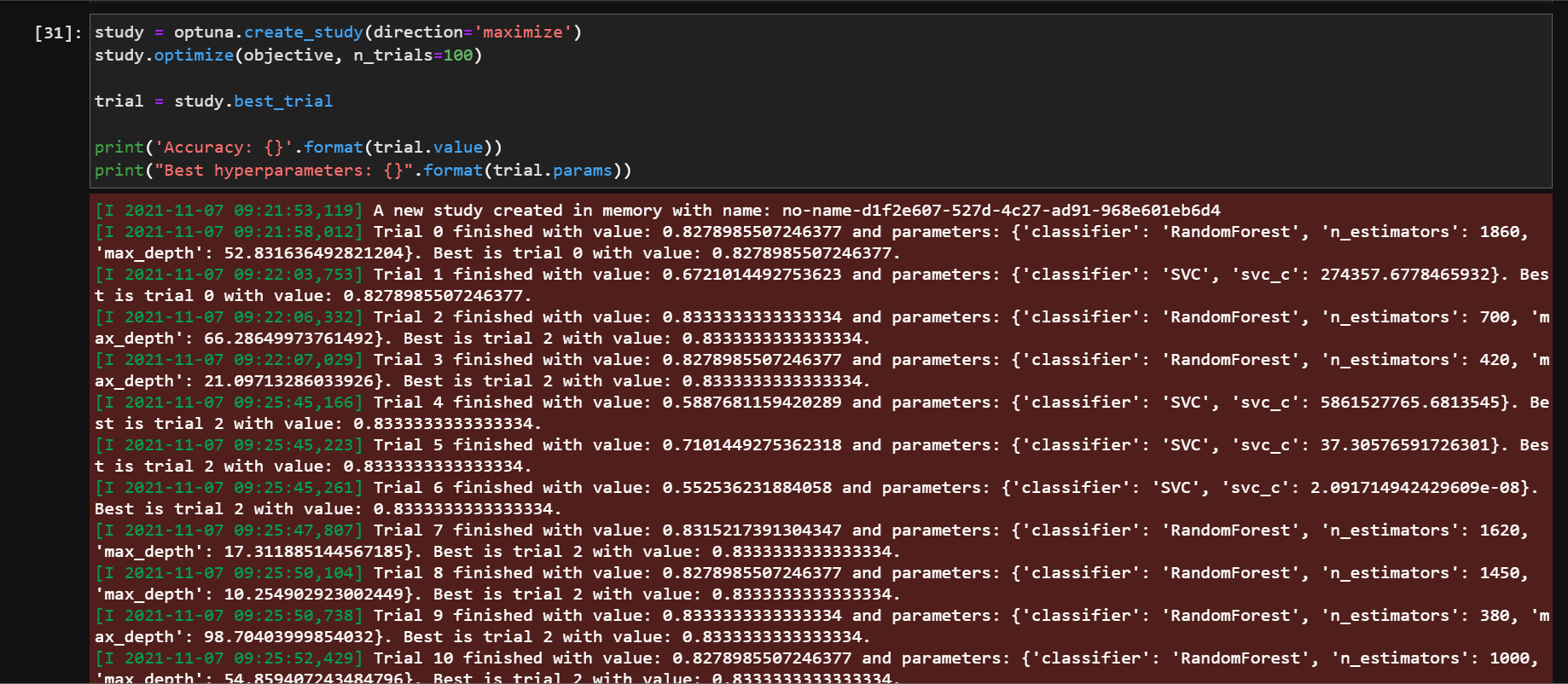


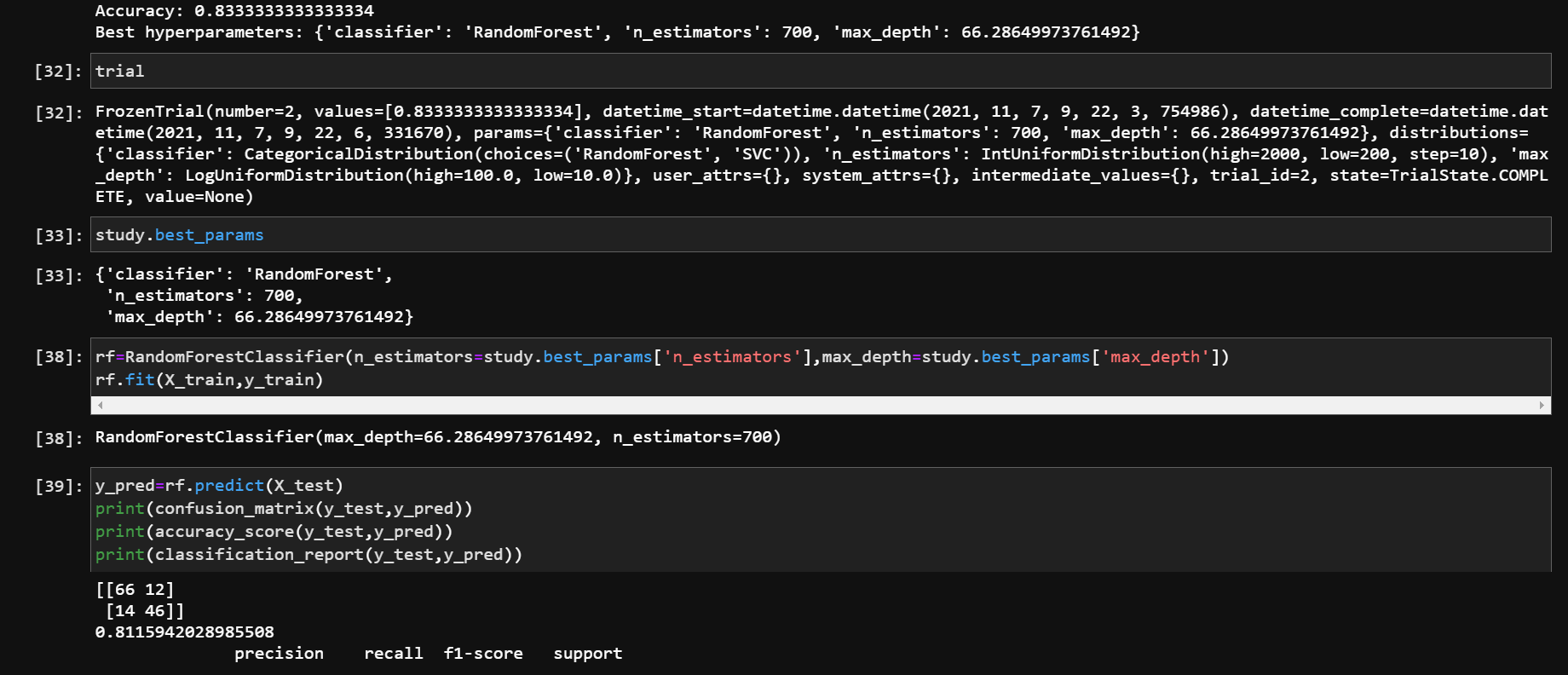


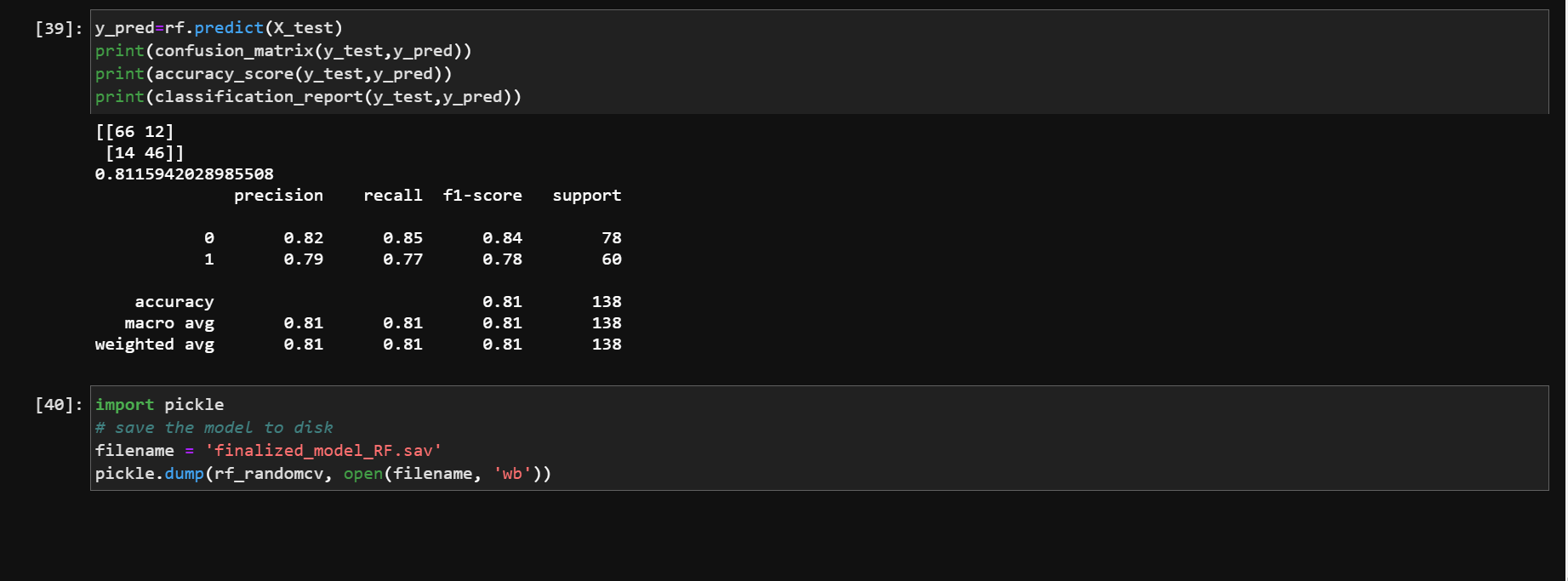




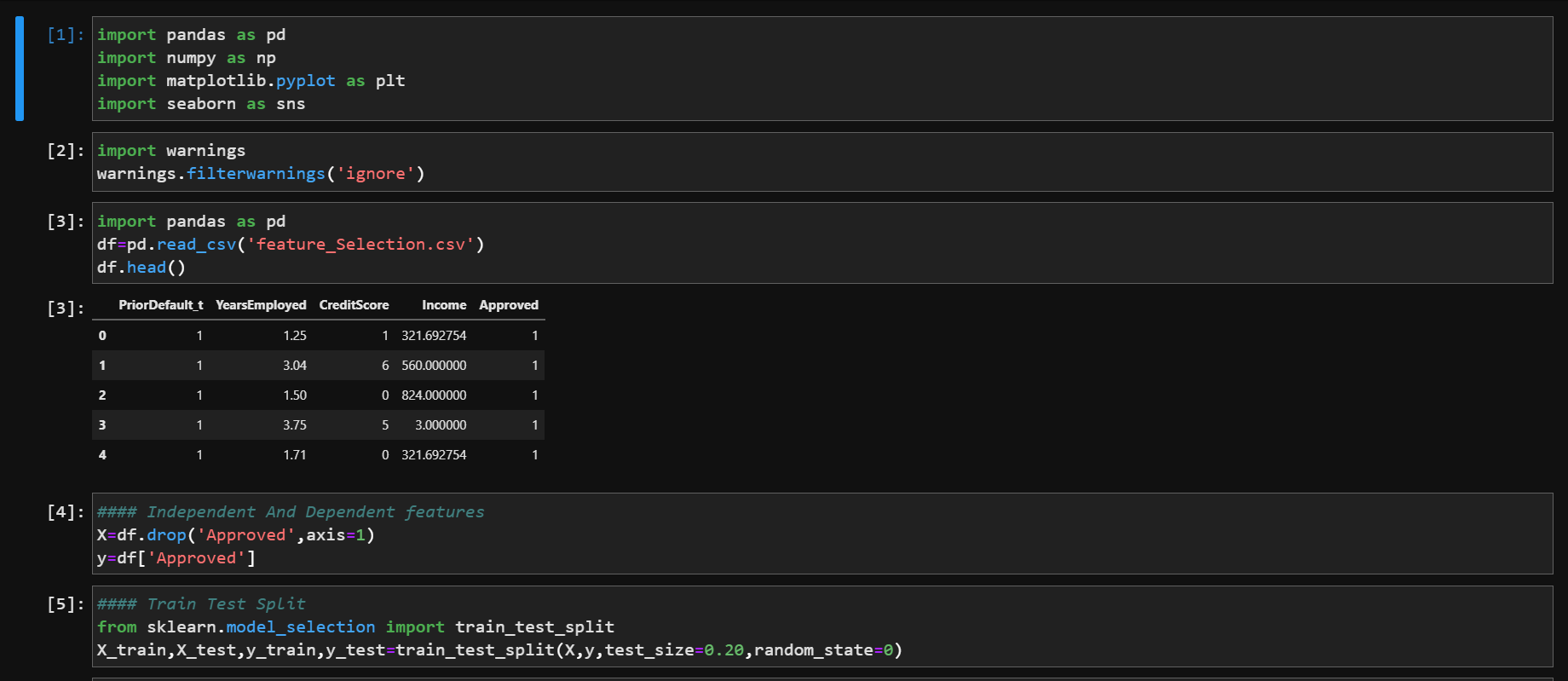


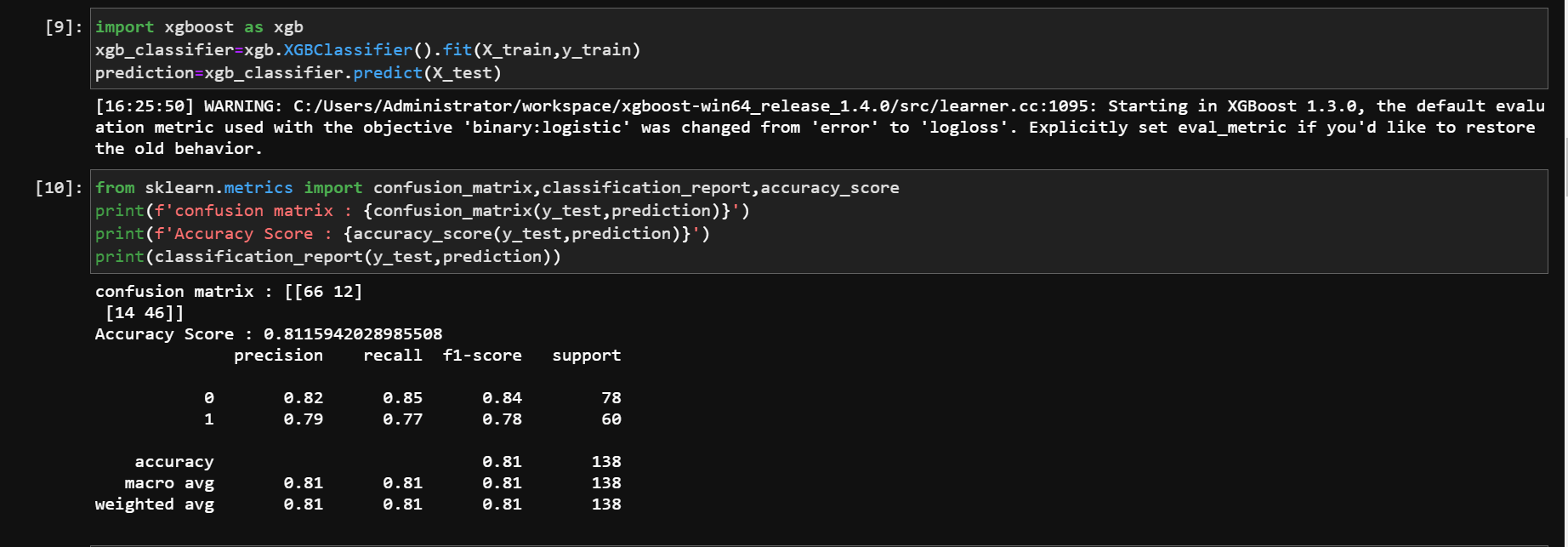


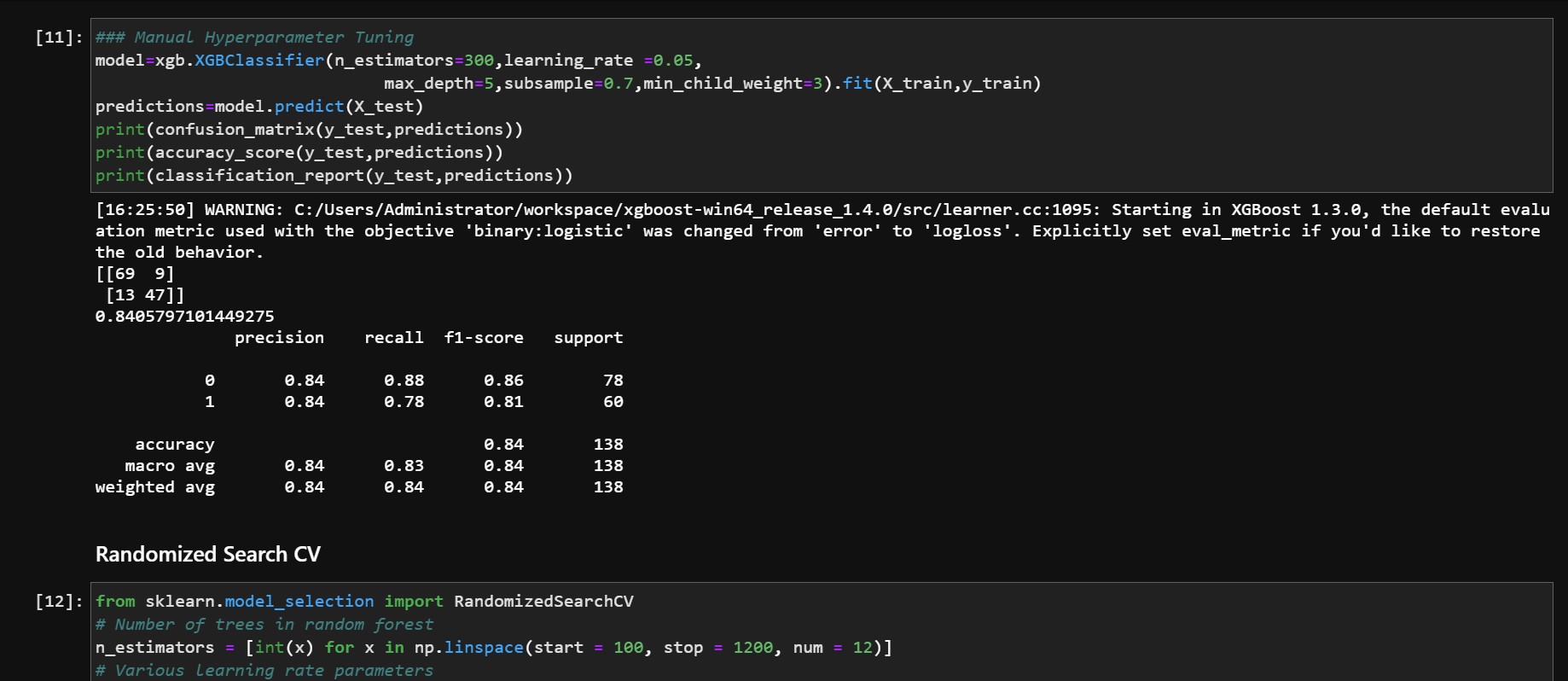


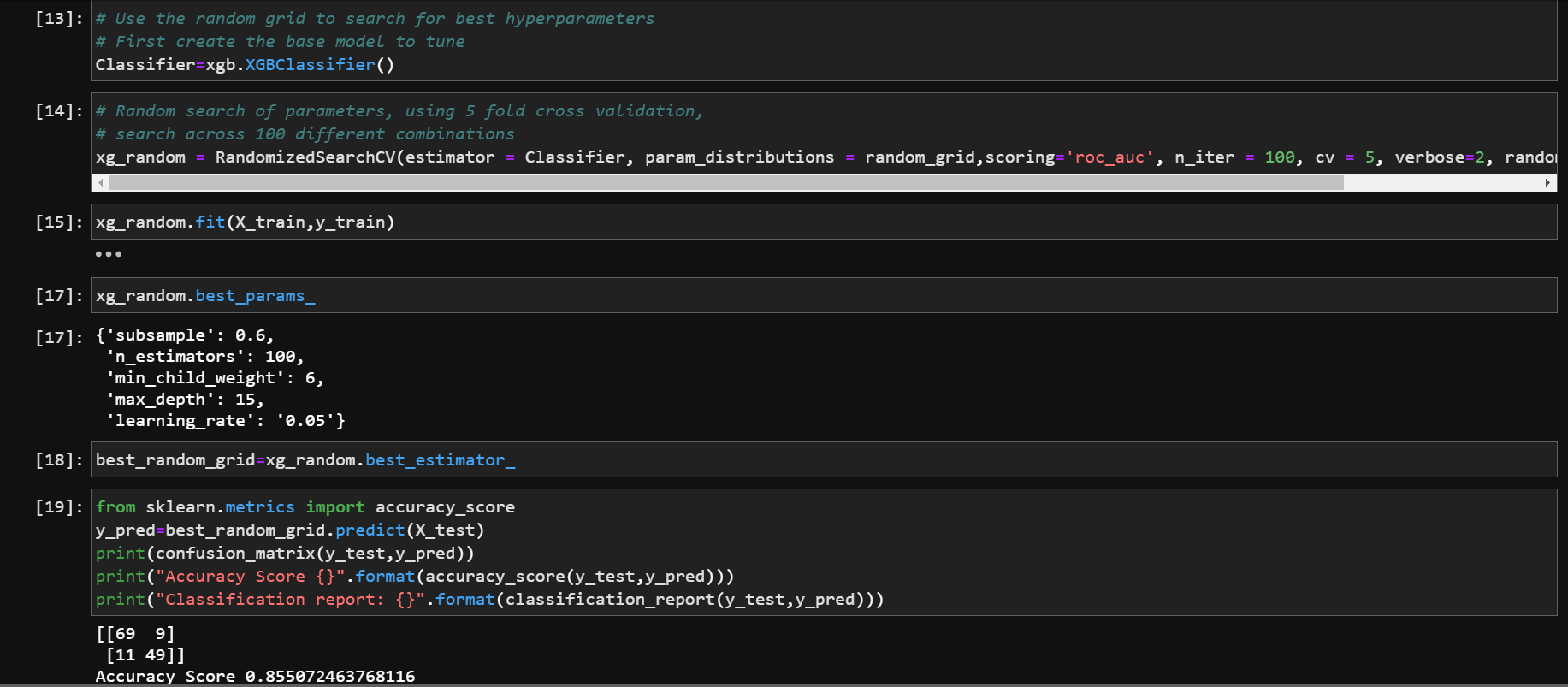


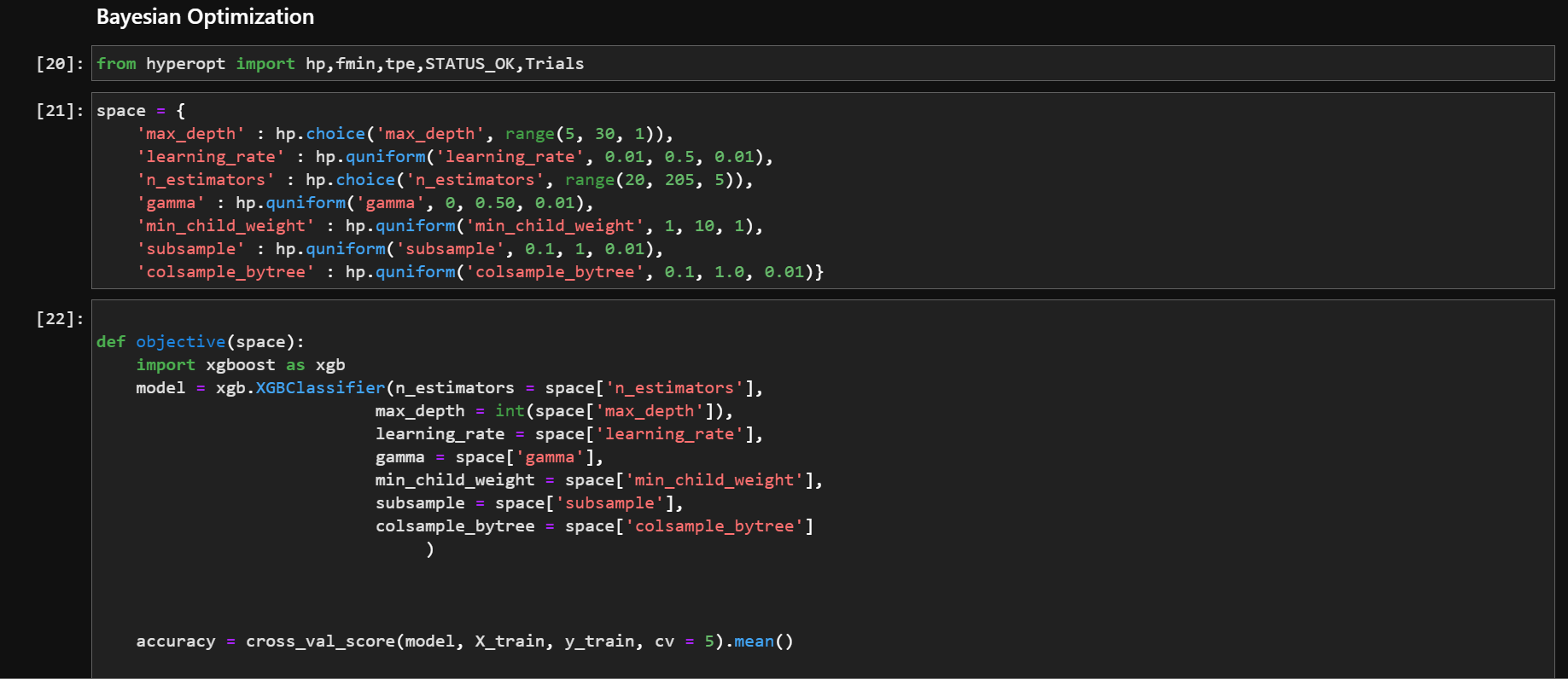
**Training with Xgboost**

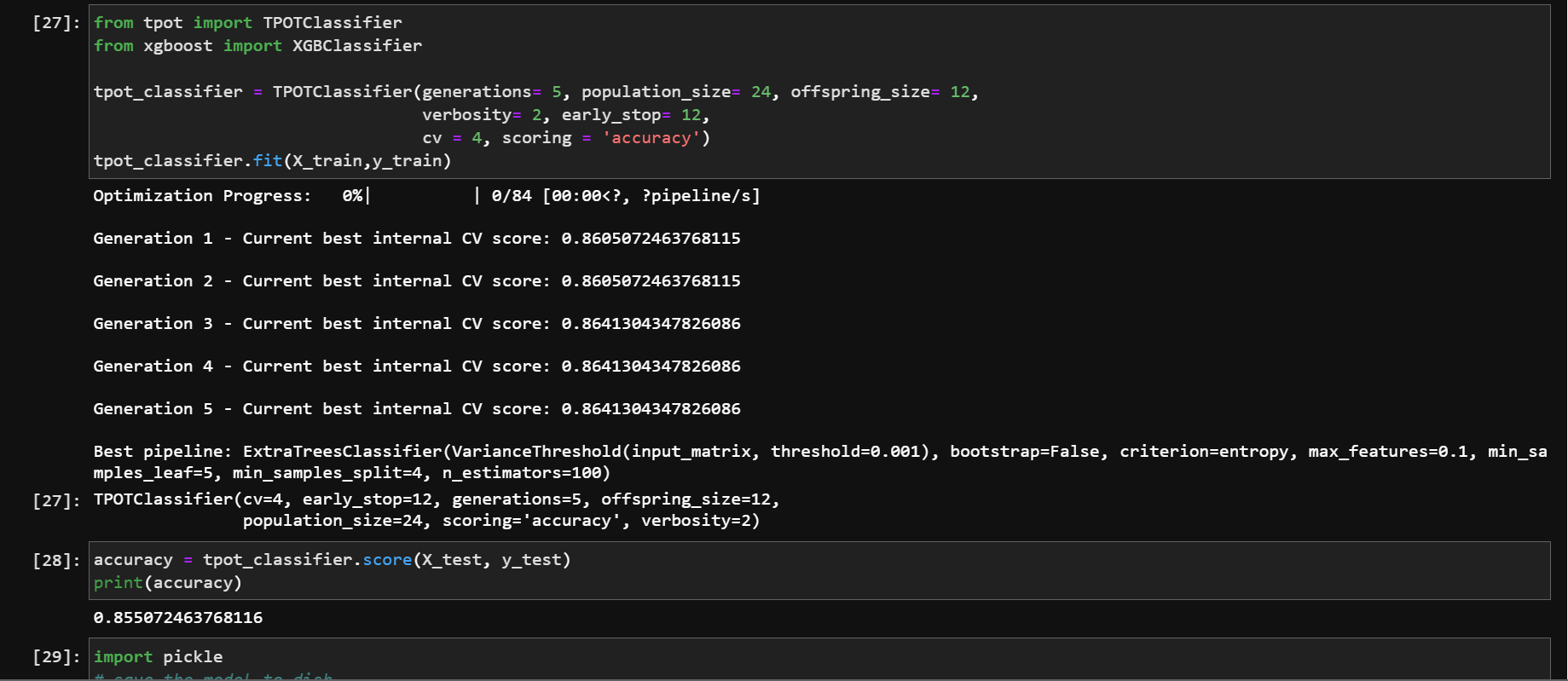












**7.2 Testing**