B. Tech Program Second Year

Course: Machine Learning

Course Code: DS2231

## Wine Quality Prediction

by

**Agrima Goel (219309075)**

**Divyansh Raiswal (219309166)**

Under the guidance

of

**Name of the mentor – Mr. Venkatesh Gauri Shankar**

**Manipal University Jaipur, India**

Project-based Learning:

Introduction

1. PCA (Principal Component Analysis): PCA is a dimensionality reduction technique that is used to transform high-dimensional data into a lower-dimensional space while preserving most of the variance in the data. It works by identifying the directions in which the data varies the most and projecting the data onto these directions, which are called principal components. PCA is commonly used for feature extraction, data compression, and visualization.

2. Naive Bayes: Naive Bayes is a probabilistic classification algorithm that is based on Bayes' theorem. It assumes that the features of a dataset are independent of each other and calculates the probability of a new data point belonging to a particular class based on the probabilities of the individual features given that class. Naive Bayes is known for its simplicity, scalability, and fast training speed, making it a popular choice for text classification and spam filtering.

3. KNN (K-Nearest Neighbors): KNN is a classification algorithm that is based on the idea that similar data points tend to belong to the same class. Given a new data point, the algorithm searches for its K-nearest neighbors in the training dataset and assigns the majority class of those neighbors as the predicted class for the new point. The K value is a hyperparameter that needs to be tuned for optimal performance.

4. Decision Tree: Decision Tree is a supervised learning algorithm that is used for both classification and regression tasks. It works by partitioning the feature space into a set of rectangular regions based on a set of splitting rules that maximize the information gain or minimize the impurity of the resulting subsets. The tree can be visualized as a set of nested if-else statements that make predictions for new data points.

5. HAC - Hierarchical Agglomerative Clustering: HAC is a type of clustering algorithm used in machine learning and data mining. It is a method for grouping together similar data points or objects based on their similarity or dissimilarity measures. HAC starts with each data point as a separate cluster and recursively merges clusters until a stopping criterion is met, resulting in a hierarchical tree-like structure called a dendrogram.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Methodology**

1. Pandas: Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation.

2. Matplotlib: Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation.

3. scikit-learn, also known as sklearn, is a popular machine learning library for Python. It provides a wide range of tools for data analysis and modeling, including classification, regression, clustering, and dimensionality reduction, among others.

Some of the key features of sklearn include:

1. Consistent API: The library provides a consistent and easy-to-use API for various machine learning algorithms.

2. Extensive documentation: Sklearn has extensive documentation, with examples, tutorials, and explanations of different algorithms and methods.

3. High-performance implementations: Sklearn provides optimized implementations of many machine learning algorithms, making it efficient for large-scale data analysis.

4. Integration with other libraries: Sklearn can be used in conjunction with other libraries like NumPy, Pandas, and Matplotlib, among others, making it easy to incorporate machine learning into your existing data analysis workflows.

5. Active development: Sklearn is an actively developed library, with regular updates and improvements.

Overall, sklearn is a powerful and flexible machine learning library that is widely used in the data science community.

4. Matplotlib is a data visualization library for Python that is widely used for creating a wide variety of plots and graphs. It is a powerful tool that allows you to create static, interactive, and animated visualizations with ease.

Matplotlib provides a wide range of plotting options, including line plots, scatter plots, bar plots, histograms, pie charts, and more. You can also customize your plots with different colours, markers, line styles, fonts, and

other graphical parameters to create beautiful and informative visualizations.

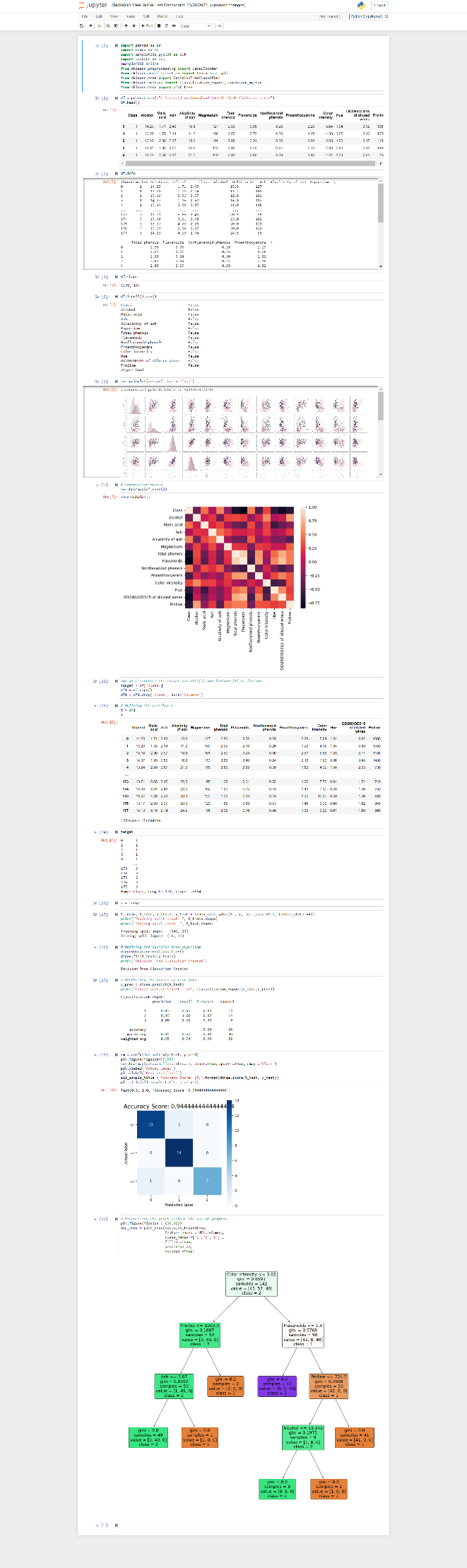
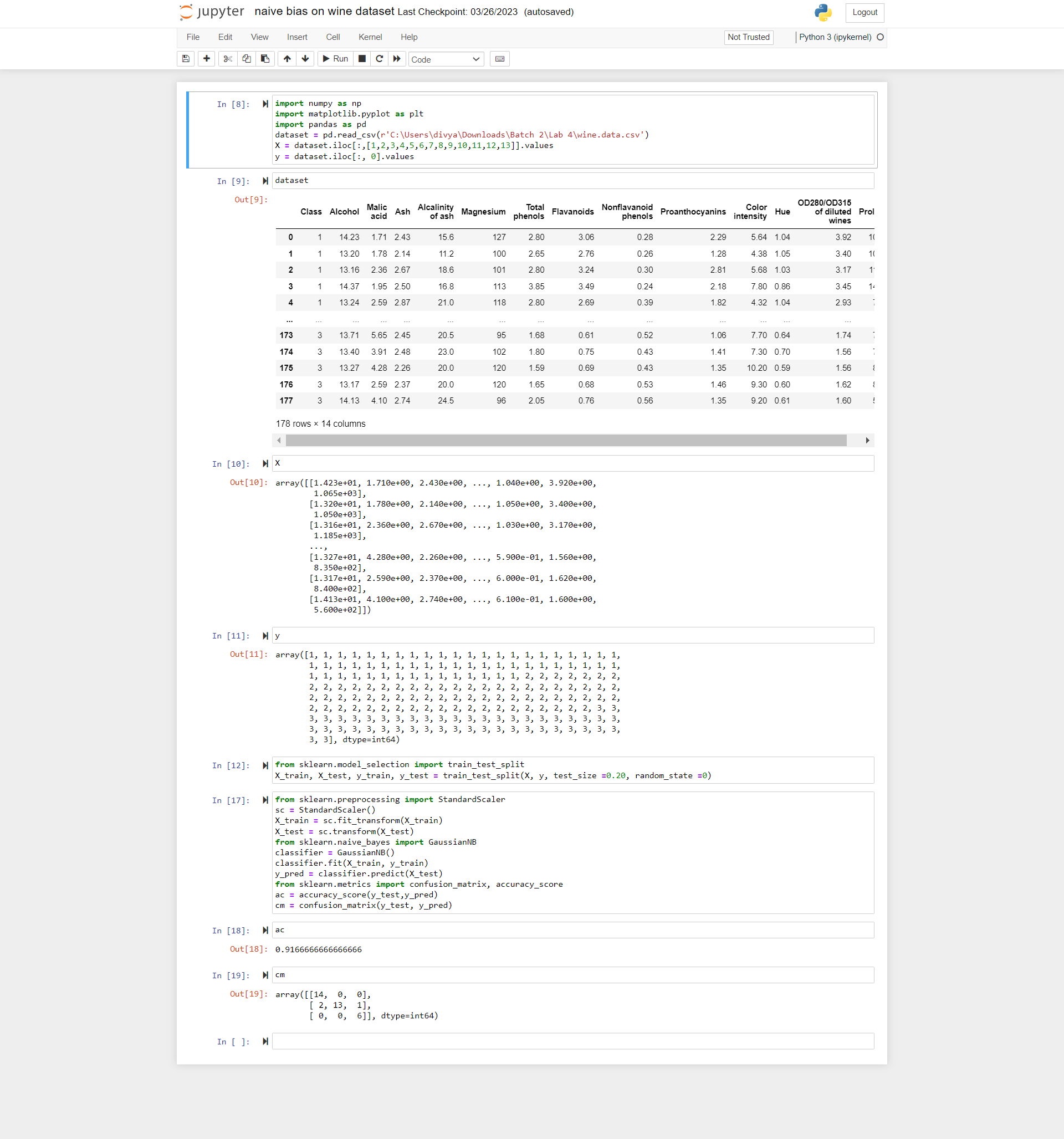
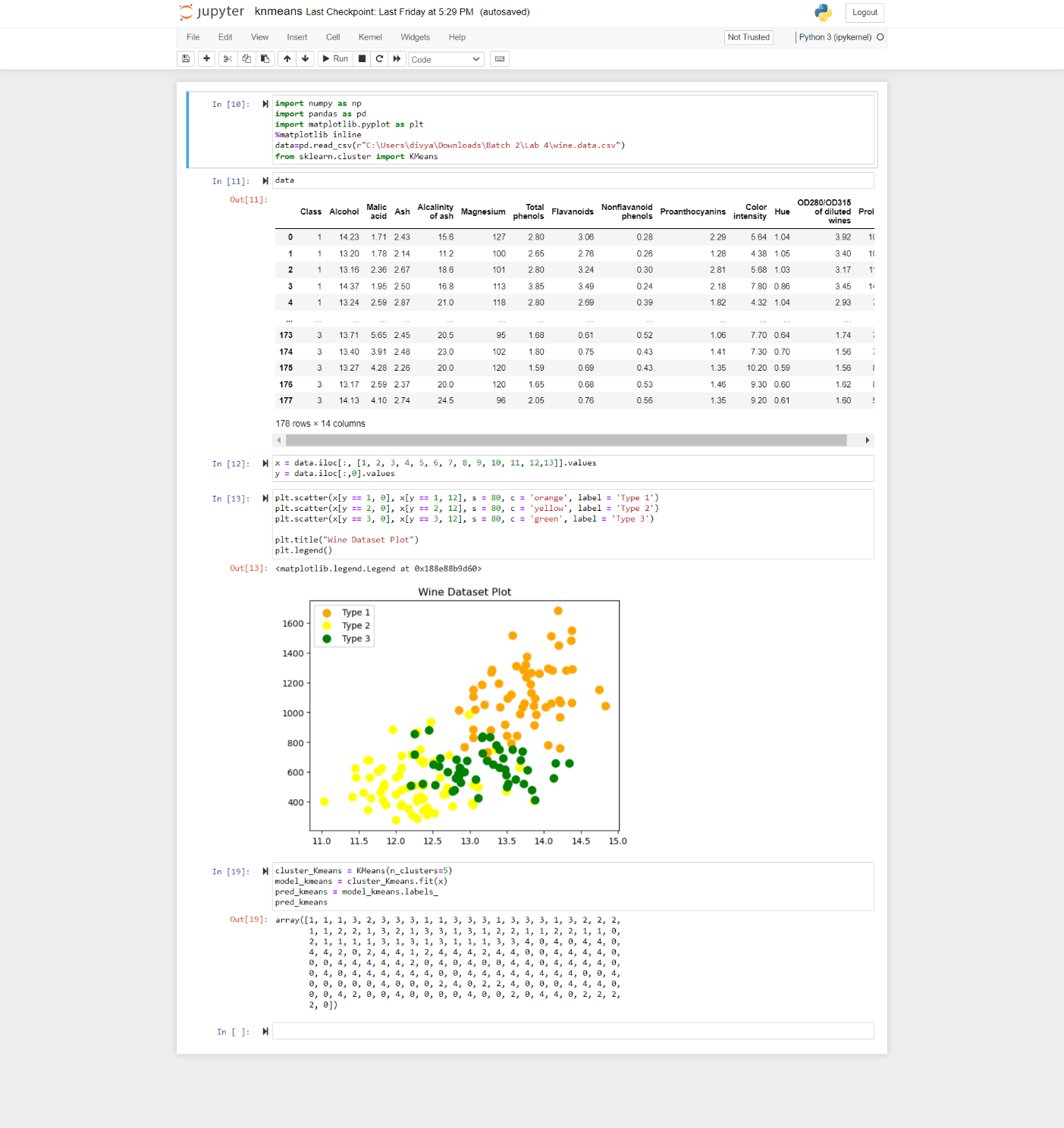
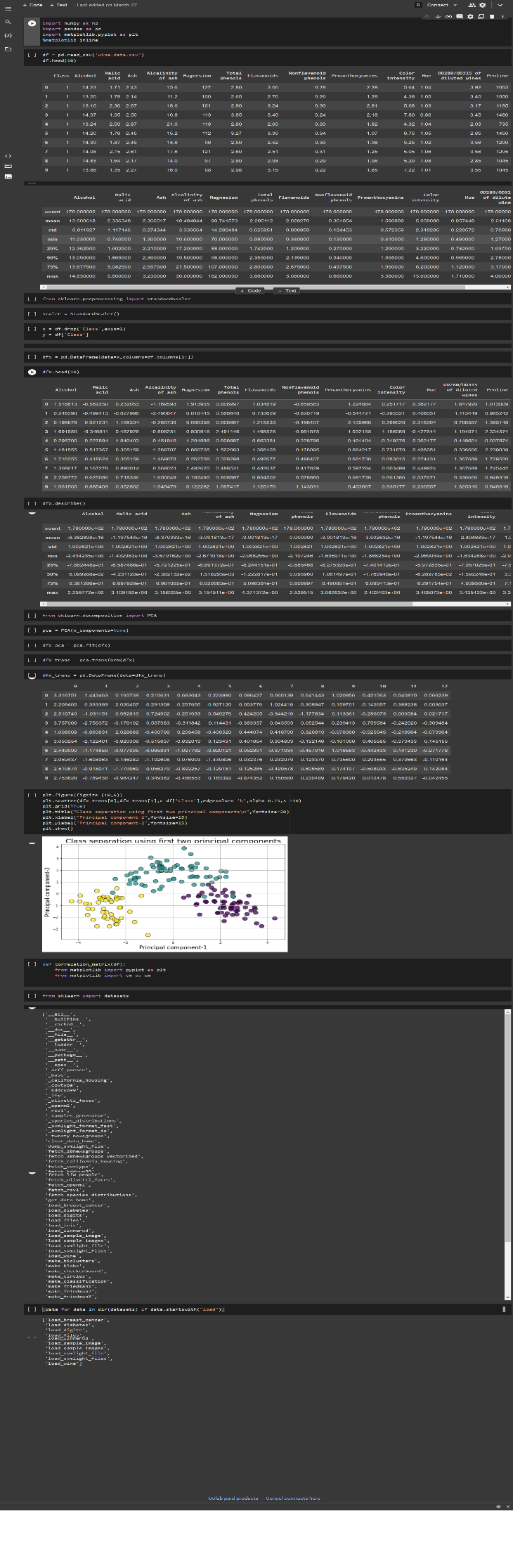
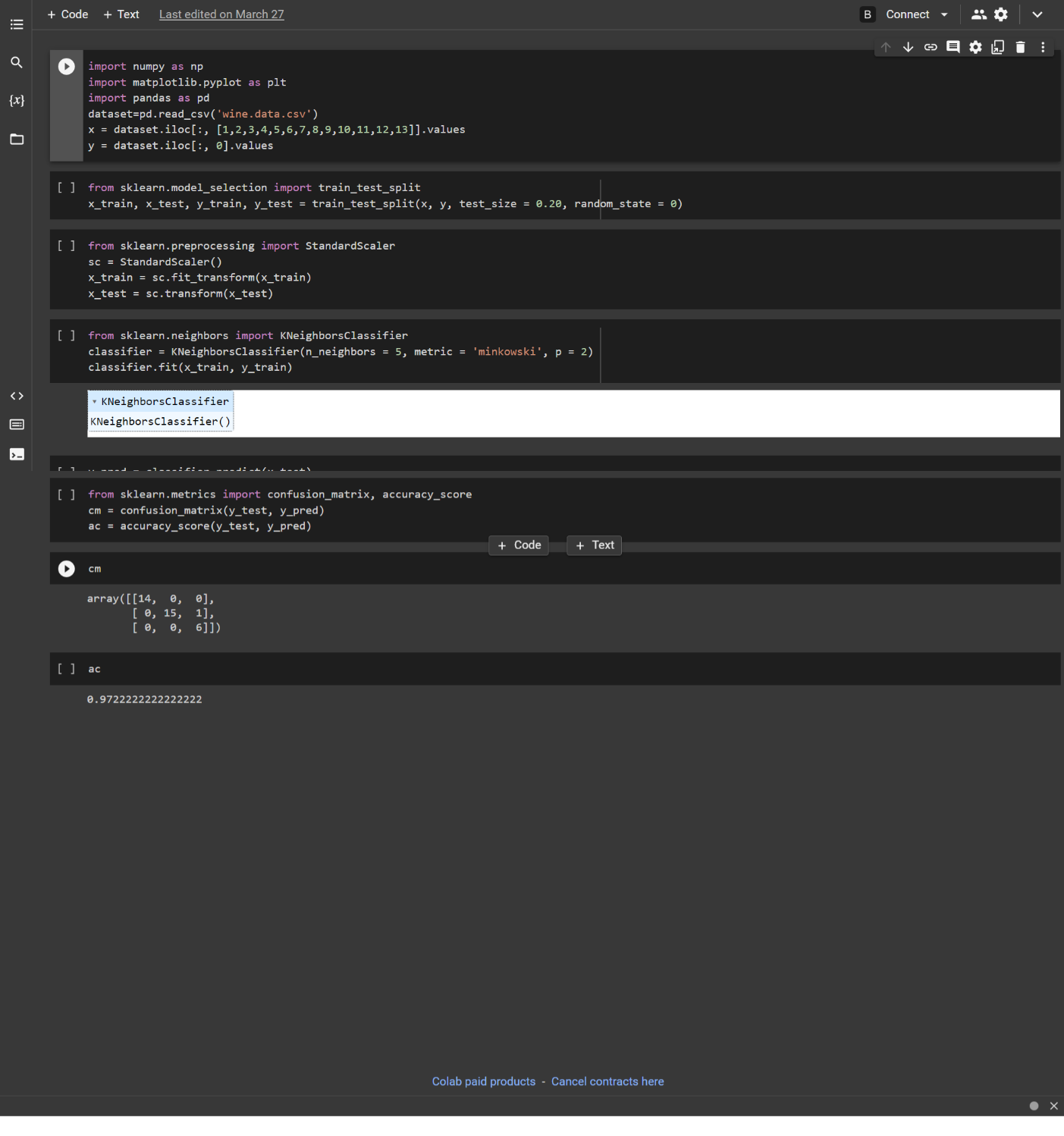
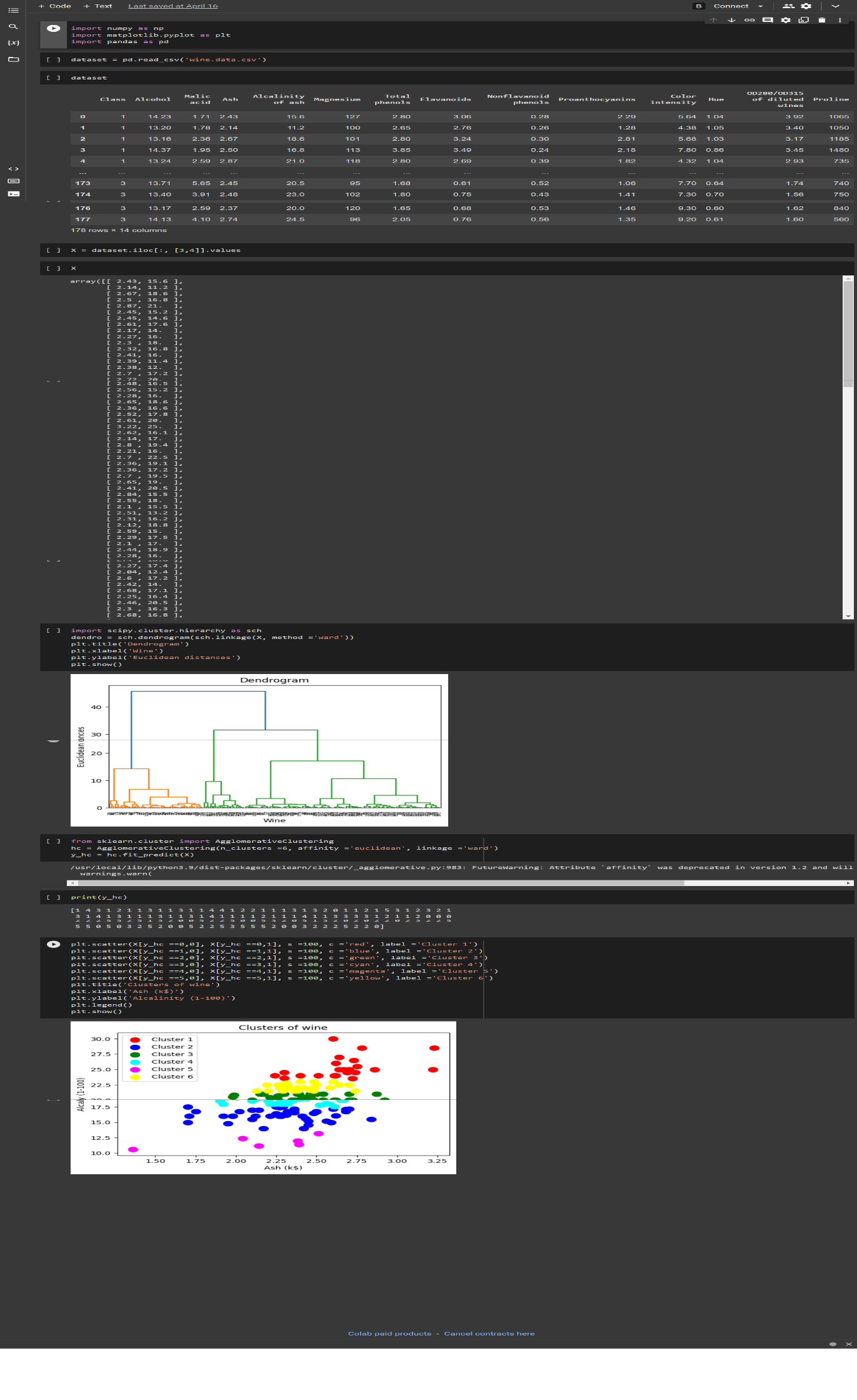
One of the strengths of Matplotlib is its versatility. You can create basic plots with just a few lines of code, or you can create complex visualizations that incorporate multiple data sources and advanced features such as subplots, annotations, and animations.

Another important feature of Matplotlib is its integration with other Python libraries. For example, you can use Matplotlib in conjunction with NumPy, Pandas, or SciPy to create powerful data analysis workflows that combine data manipulation and visualization.

Here are some examples of common plotting tasks that you can achieve with Matplotlib:

* Line plots: use plt.plot() to create line plots of one or more data series .
* Scatter plots: use plt.scatter() to create scatter plots of two or more data series .
* Bar plots: use plt.bar() to create vertical or horizontal bar plots of one or more data series
* Histograms: use plt.hist() to create histograms of one or more data series
* Pie charts: use plt.pie() to create pie charts of one or more data series

Overall, Matplotlib is a powerful and flexible data visualization library that is widely used in scientific computing, data analysis, and machine learning. Its ease of use and customization options make it a great choice for anyone looking to create informative and beautiful visualizations with Python.

****