TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**INTRODUCTION TO MACHINE LEARNING**

MID-TERM REPORT INTRODUCTION TO MACHINE LEARNING

STUDENT 1: 522H0120 – Nguyễn Đình Việt Hoàng

STUDENT 2: 521H0473 – Nguyễn Minh Quân

LECTURER: TS. Lê Anh Cường

**THÀNH PHỐ HỒ CHÍ MINH**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NHẬP MÔN HỌC MÁY**

BÀI BÁO CÁO GIỮA KỲ MÔN NHẬP MÔN HỌC MÁY

SINH VIÊN 1: 522H0120 – Nguyễn Đình Việt Hoàng

SINH VIÊN 2: 521H0473 – Nguyễn Minh Quân

GIẢNG VIÊN HƯỚNG DẪN: TS. Lê Anh Cường

**THÀNH PHỐ HỒ CHÍ MINH**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn thầy Lê Anh Cường vì đã giảng dạy tụi em nhiệt tình môn Nhập môn học máy. Chúng em muốn bày tỏ sự cảm kích sâu sắc với sự tận tâm và kiến thức chuyên môn mà thầy đã chia sẻ với chúng em. Qua những buổi học của thầy chúng em đã có cơ hội hiểu biết thêm về các khía cạnh cơ bản của Machine Learning qua việc giải thích chi tiết và áp dụng thực tế. Thầy đã giúp chúng em nắm vững kiến thức và áp dụng chúng vào thực tế. Cuối cùng, chúng em xin cảm ơn đến thầy Lê Anh Cường vì sự tận tâm và sự hỗ trợ quý báu trong suốt quá trình học môn Nhập môn học máy. Những kiến thức và kỹ năng mà chúng em đã học được sẽ luôn có giá trị và ảnh hưởng đến sự phát triển của chúng em trong tương lai. Chúng em xin chân thành cảm ơn thầy Lê Anh Cường và chúc thầy sức khỏe, thành công và hạnh phúc.

*TP. Hồ Chí Minh, ngày 30 tháng 3 năm 2024*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Nguyễn Minh Quân*

*Nguyễn Đình Việt Hoàng*

**PHIẾU ĐÁNH GIÁ CỦA GIẢNG VIÊN HƯỚNG DẪN**

Tên giảng viên hướng dẫn:

Ý kiến nhận xét:

Điểm tổng theo phiếu đánh giá rubrik:

*TP. Hồ Chí Minh, ngày 29 tháng 3 năm 2024*

*Giảng viên hướng dẫn*

*(Ký tên và ghi rõ họ tên)*

**BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Nhóm chúng em xin cam đoan đây là bài báo cáo của riêng chúng em và được sự hướng dẫn của thầy Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong bài báo cáo này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong bài báo cáo còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào nhóm chúng em xin hoàn toàn chịu trách nhiệm về nội dung bài báo cáo giữa kỳ học kỳ 2/2023-2024 của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng em gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 30 tháng 3 năm 2024*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Nguyễn Minh Quân*

*Nguyễn Đình Việt Hoàng*

MỤC LỤC

**CÂU 1. 1**

1.1 Present the steps in solving a classification problem or regression problem using machine learning methods. 1

1.2 Perform the steps in question 1 on a classification or regression problem using different machine learning methods. Evaluate and compare the results of applying these methods; 1

1.3 Find solutions to improve the accuracy of the models mentioned in question 1

**CÂU 2. 2**

2.1 Let's study yourself and present what is the overfitting problem of machine learning models? Describe overfitting phenomena………………………………………………2

2.2 Please present solutions to avoid overfitting, from general solutions to solutions for each specific machine learning method (including learned machine learning methods and you can be expanded to other learning methods). ………………………………………2

2.3 Present the problems mentioned in questions 1 and 2 through one or more real data sets (for classification and/and regression problems). Note that you must choose data sets and machine learning method where overfitting occurs. ………………………………..2

**DANH MỤC TÀI LIỆU THAM KHẢO 11**

**PHỤ LỤC**

**QUESTION 1:**

(1) Present the steps in solving a classification problem or regression problem using machine learning methods.

Outline the process of tackling a classification problem using machine learning techniques.

Step 1 - Data Gathering:

Begin by collecting relevant data pertaining to the issue at hand. This data can be sourced from various outlets such as databases, CSV files, or .data files.

Step 2 - Data Preprocessing:

Data often requires refinement before being inputted into a machine learning model. Preprocessing tasks may involve eliminating noisy data, handling missing values, normalizing data, encoding categorical variables, and dividing the data into training and testing sets.

Step 3 - Model Selection:

Depending on the problem's nature (classification or regression) and the data's specific requirements, choose an appropriate machine learning model. Options include decision trees, linear regression, support vector machines (SVM), among others.

Step 4 - Model Training:

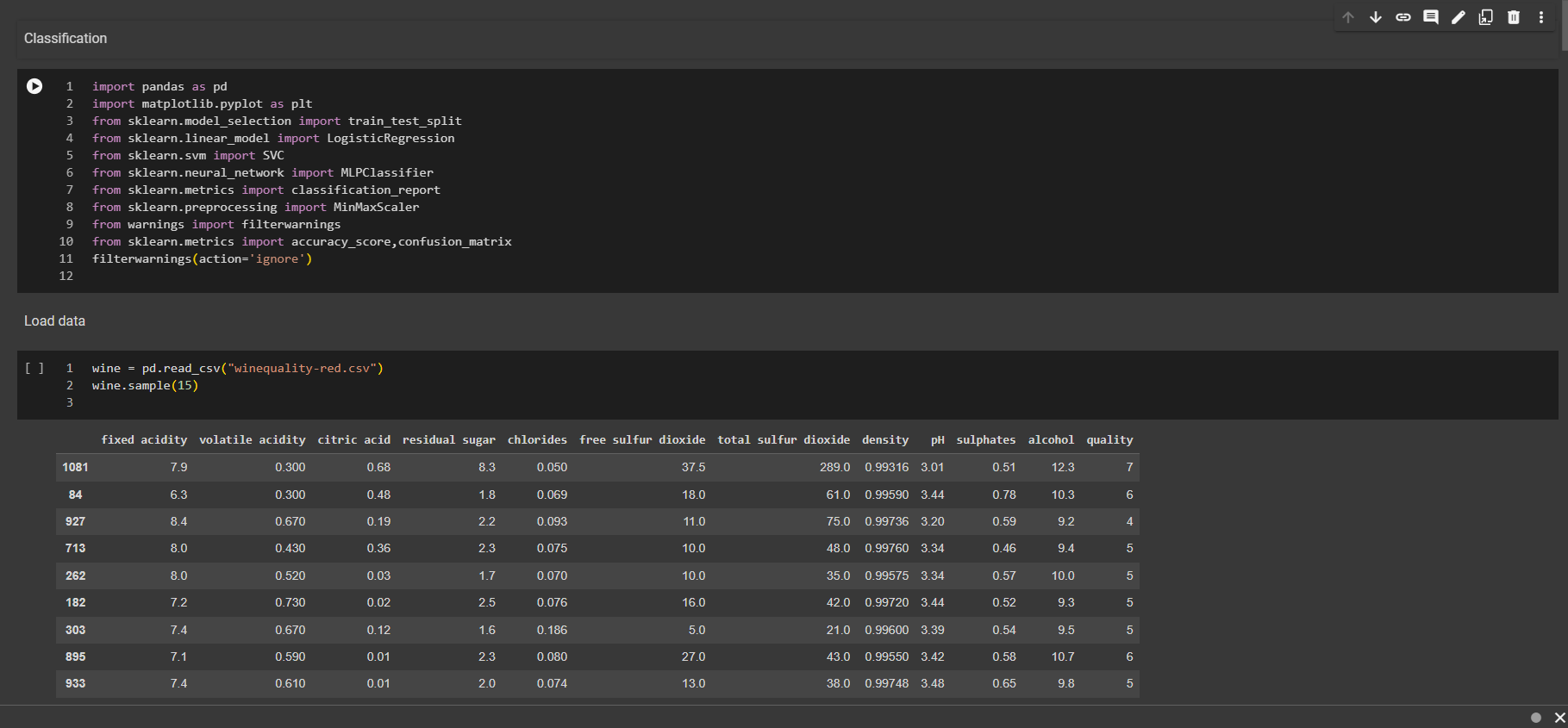
Utilize the training dataset to teach the model how to associate input features with labels (for classification) or target values (for regression). This typically involves optimizing the model's parameters to minimize a loss function.

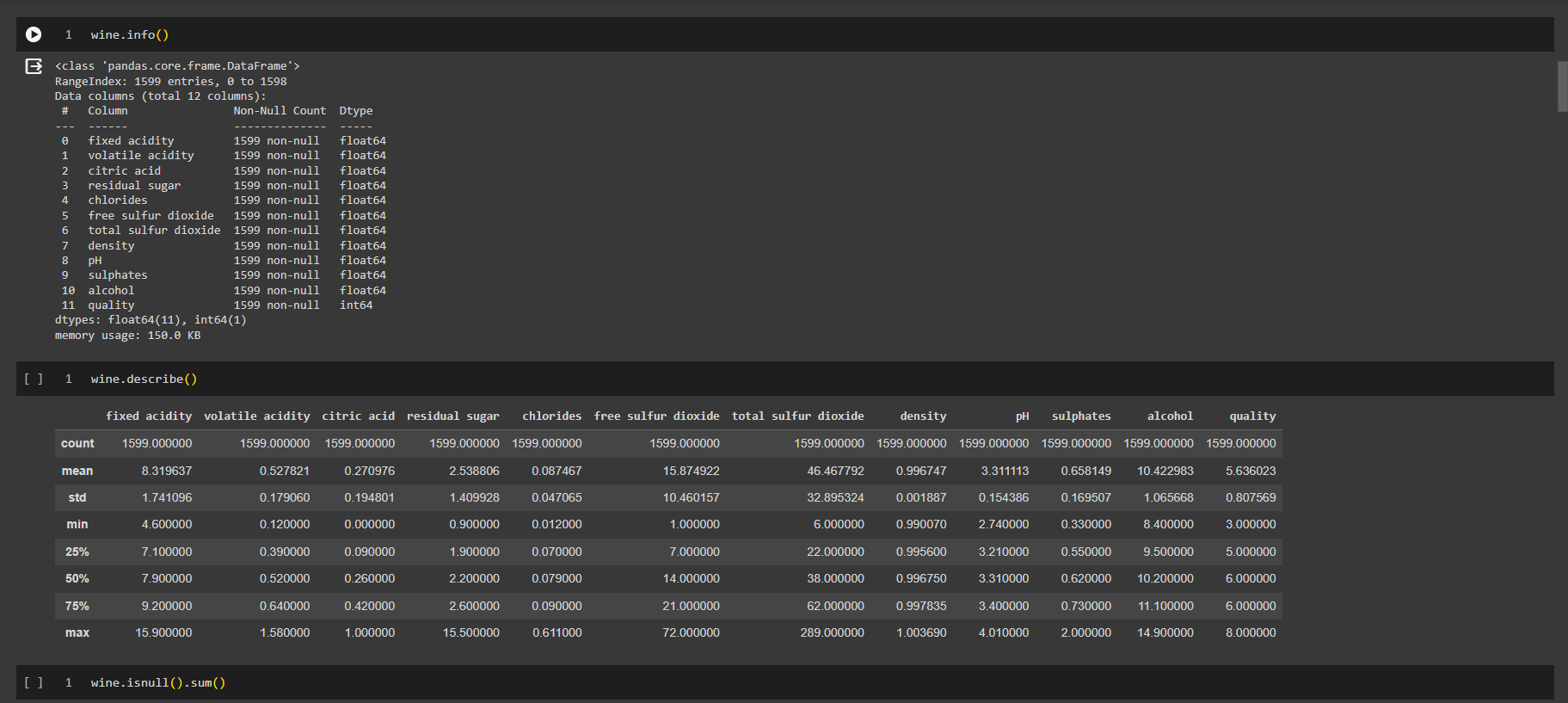
Step 5 - Model Evaluation:

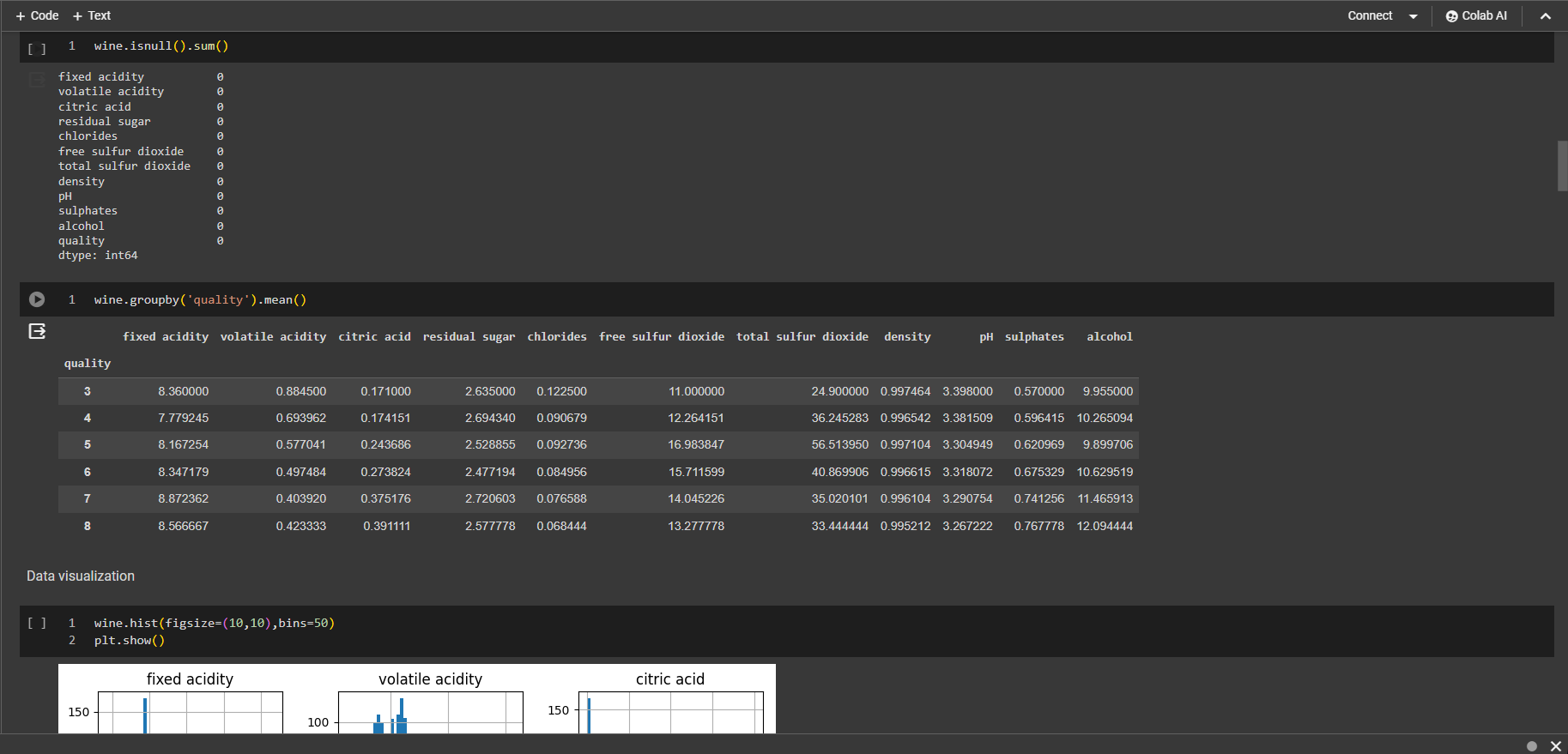
After training the model, assess its performance on a separate testing dataset that it hasn't encountered before. This evaluation aids in gauging the model's ability to generalize to new data. Common metrics for classification problems include accuracy, balanced accuracy (F1-score), and confusion matrix, while regression problems often employ metrics like mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R-squared).

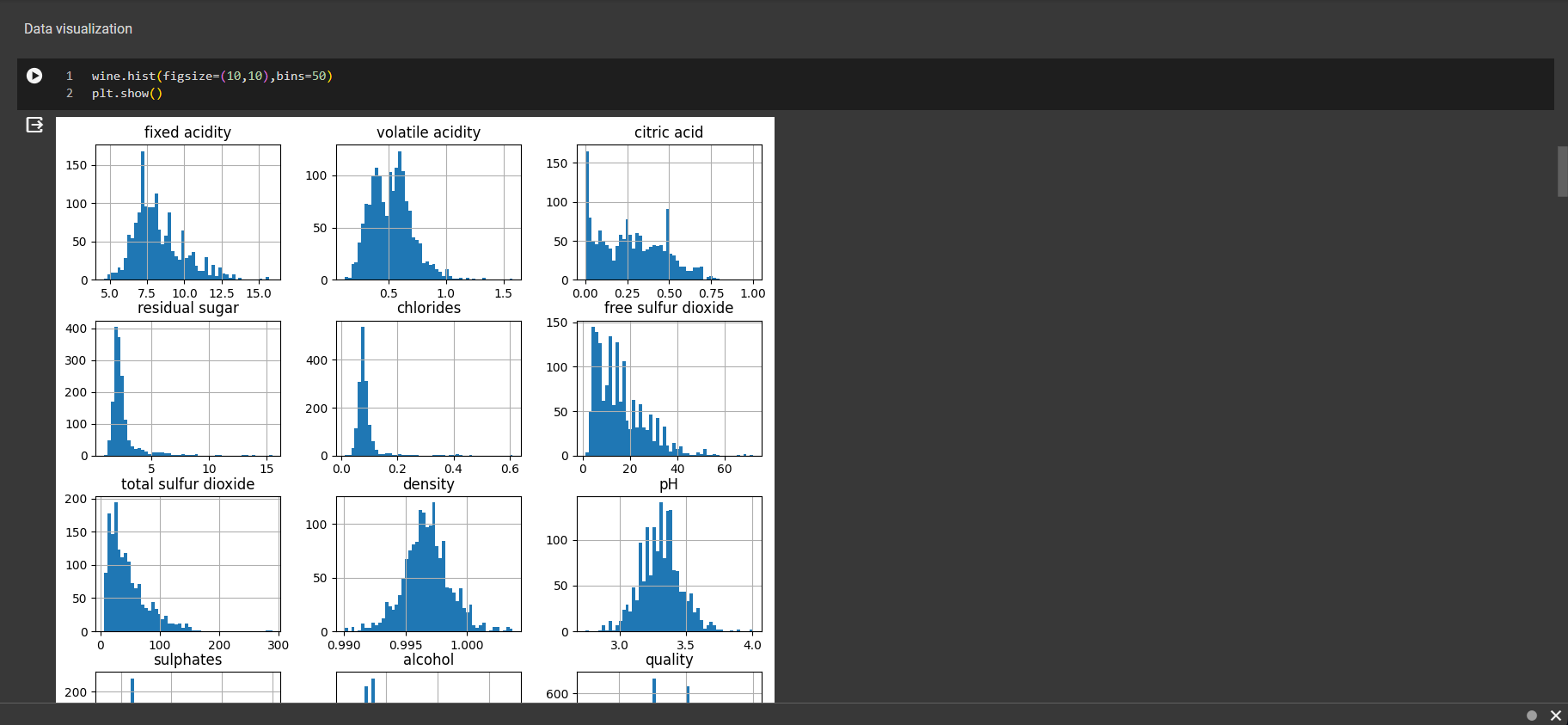
(2) Perform the steps in question 1 on a classification or regression problem using different machine learning methods. Evaluate and compare the results of applying these methods;

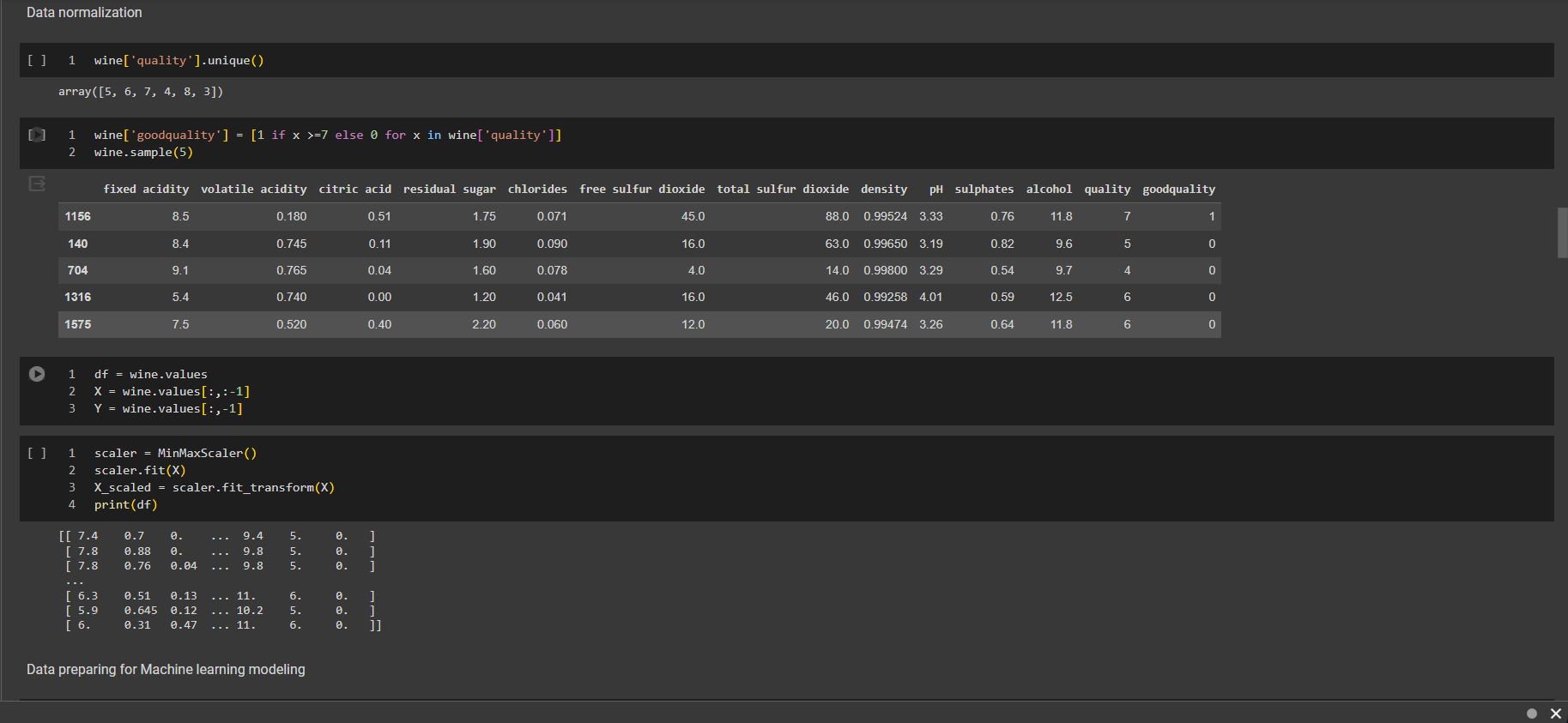
(3) Find solutions to improve the accuracy of the models mentioned in question (2).

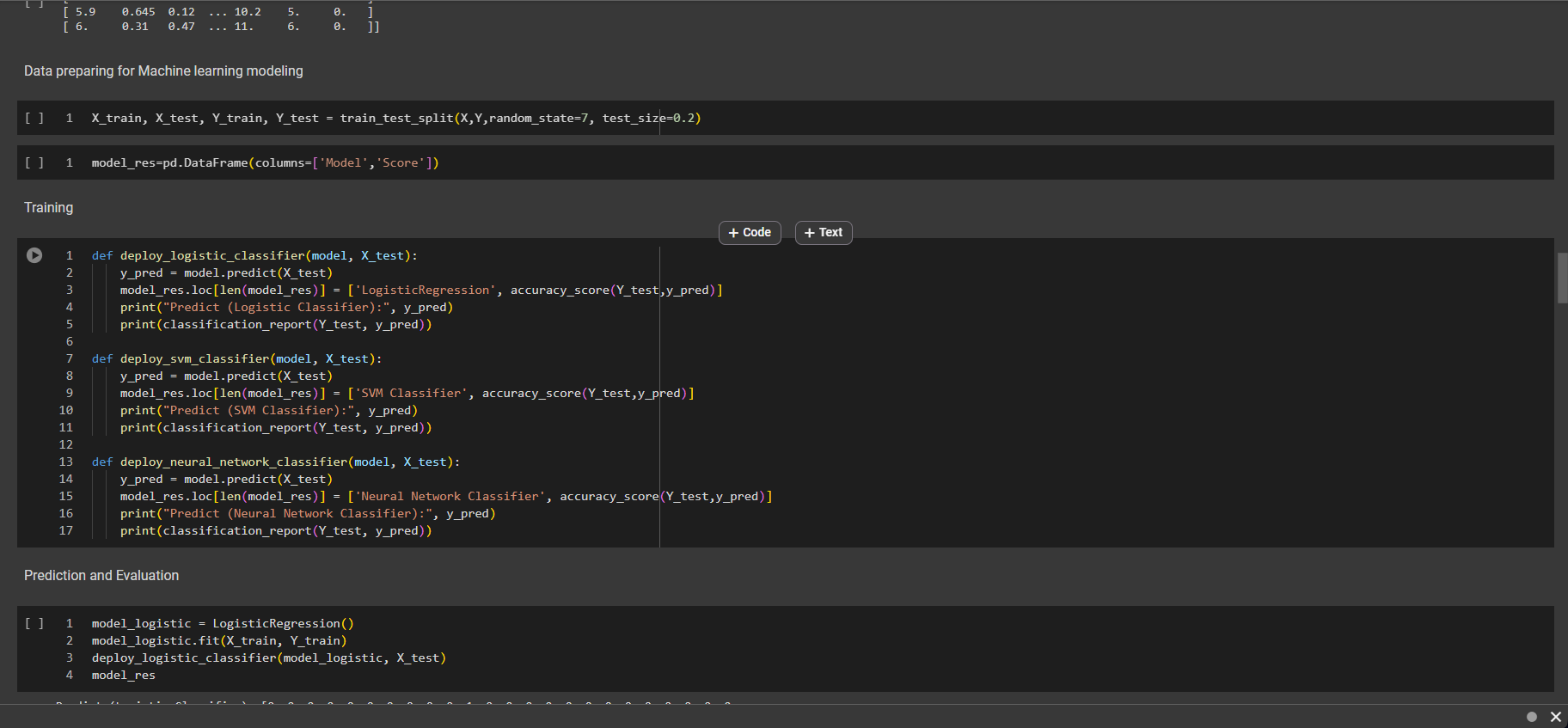


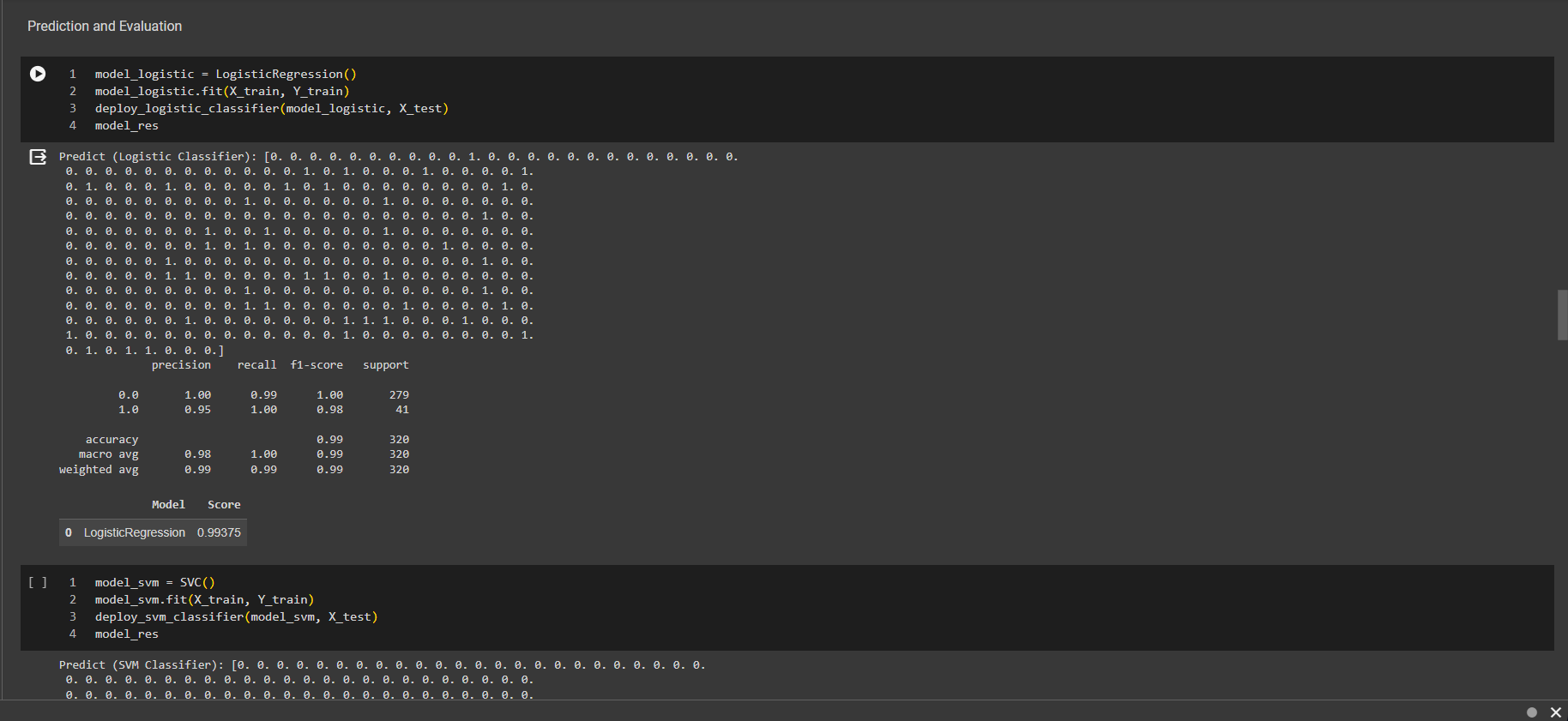


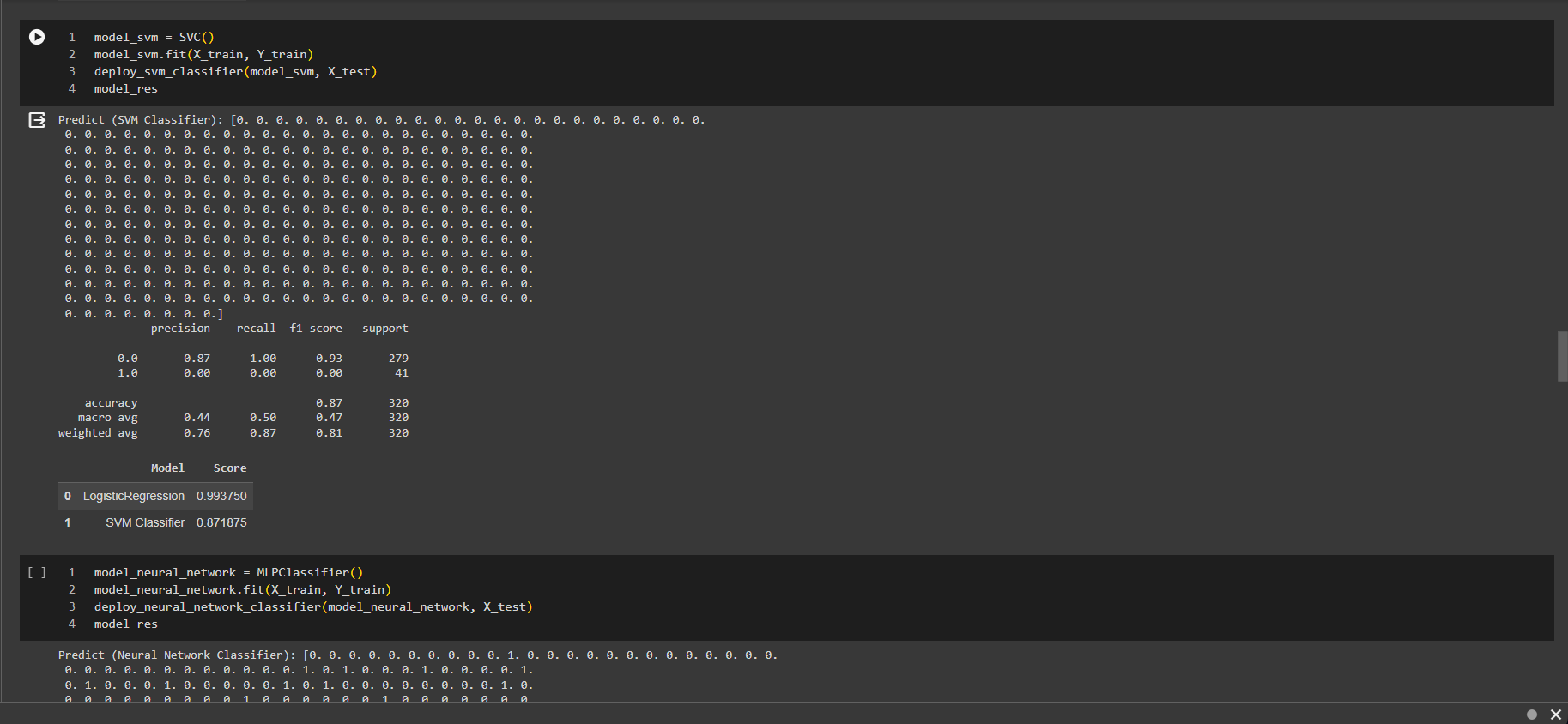


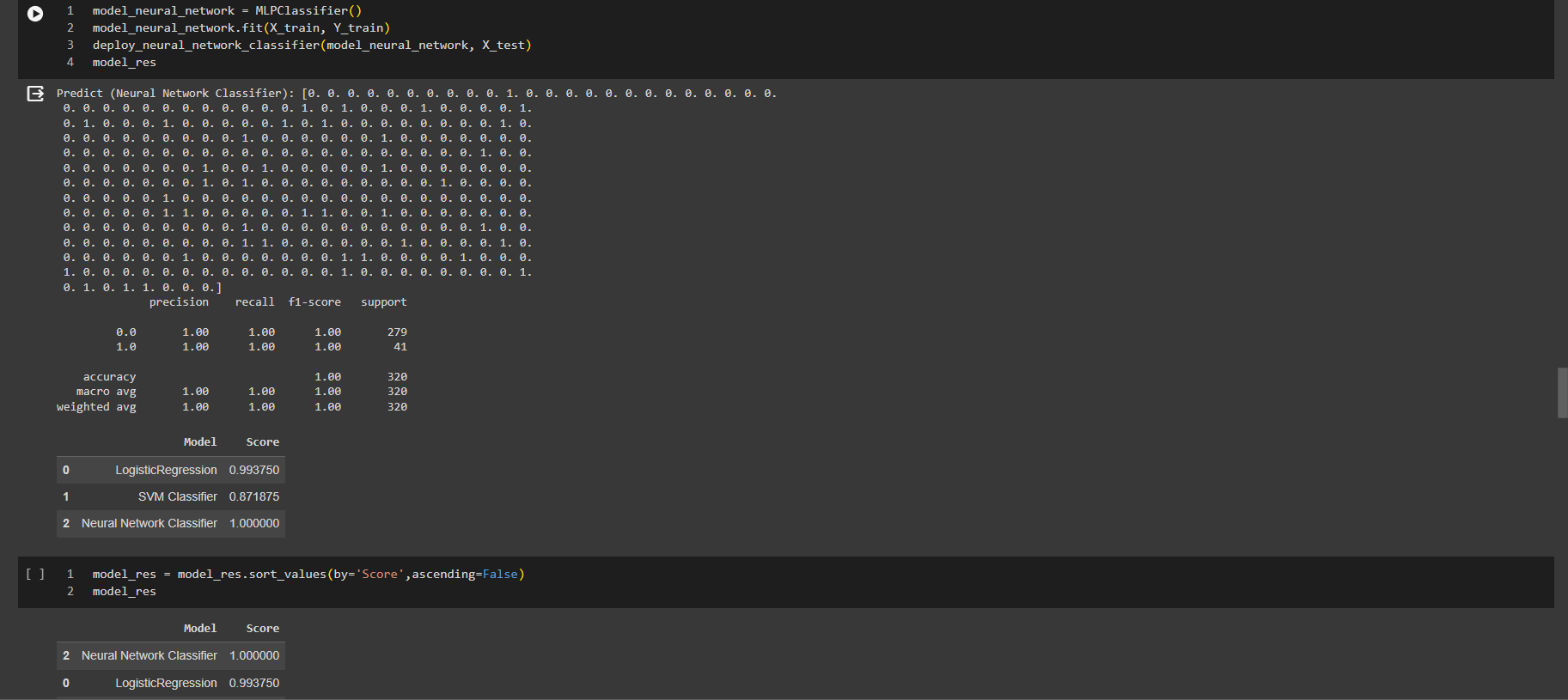


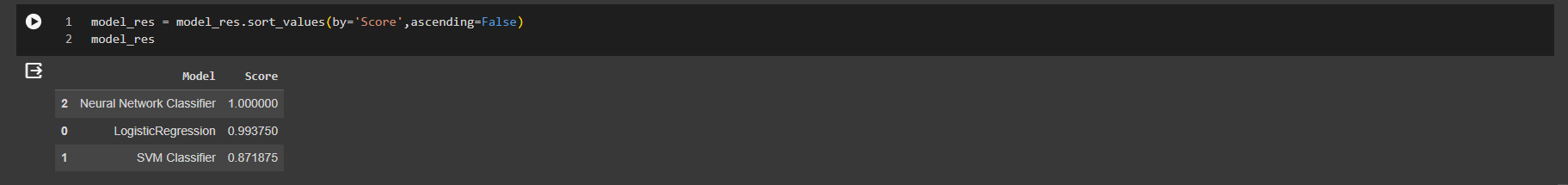












**QUESTION 2:**

(1) Let's study yourself and present what is the overfitting problem of machine learning models? Describe overfitting phenomena.

Explain the overfitting issue in machine learning models and describe the phenomenon.

a. Overfitting in Machine Learning Models:

Overfitting arises when a machine learning model captures intricate details and noise from the training data to such an extent that it performs poorly on unseen data. Essentially, the model becomes overly complex, identifying spurious patterns that do not extend beyond the training dataset. This leads to a high variance in predictions, resulting in superior performance on the training set compared to the test set or any new data. (Overfitting xảy ra khi một mô hình học máy nắm bắt các chi tiết tinh vi và nhiễu từ dữ liệu huấn luyện đến mức nó hoạt động kém trên dữ liệu chưa thấy. Cơ bản, mô hình trở nên quá phức tạp, nhận biết các mô hình giả mạo không mở rộng ra khỏi tập dữ liệu huấn luyện. Điều này dẫn đến sự thay đổi lớn trong các dự đoán, dẫn đến hiệu suất vượt trội trên tập huấn luyện so với tập kiểm tra hoặc bất kỳ dữ liệu mới nào.)

b. Causes of Overfitting:

• Excessive features: An abundance of features relative to observations can cause the model to tightly fit to the training data's specific details. (• Quá nhiều đặc trưng: Sự dư thừa của các đặc trưng so với các quan sát có thể khiến mô hình khớp chặt chẽ với các chi tiết cụ thể của dữ liệu huấn luyện.)

• Model complexity: A highly intricate model with numerous parameters is more prone to overfitting. (• Độ phức tạp của mô hình: Một mô hình phức tạp với nhiều tham số có xu hướng overfitting hơn.)

• Inadequate training data: A limited dataset amplifies the risk of overfitting as the model may not encounter enough data variance to accurately learn underlying patterns. (• Dữ liệu huấn luyện không đủ: Một tập dữ liệu hạn chế làm tăng nguy cơ overfitting vì mô hình có thể không gặp đủ sự thay đổi dữ liệu để học một cách chính xác các mô hình cơ bản.)

(2) Please present solutions to avoid overfitting, from general solutions to solutions for each specific machine learning method (including learned machine learning methods and you can be expanded to other learning methods).

Propose methods to mitigate overfitting, encompassing general strategies and techniques tailored to specific machine learning methods. (Đề xuất các phương pháp giảm thiểu overfitting, bao gồm các chiến lược chung và các kỹ thuật được điều chỉnh cho các phương pháp học máy cụ thể.)

a. General Strategies: (a. Các chiến lược chung:)

• Simplify the model: Opt for a simpler model with fewer parameters, such as linear models, if they suit the problem adequately. (• Đơn giản hóa mô hình: Chọn một mô hình đơn giản hơn với ít tham số, như các mô hình tuyến tính, nếu chúng phù hợp với vấn đề.)

• Cross-validation: Employ cross-validation methodologies like k-fold cross-validation to ensure stable model performance across different data subsets. (• Cross-validation: Sử dụng các phương pháp cross-validation như k-fold cross-validation để đảm bảo hiệu suất mô hình ổn định trên các tập con dữ liệu khác nhau.)

• Data augmentation: Expand the size and diversity of the training set by introducing slightly modified copies of existing data or newly generated synthetic data to alleviate overfitting. (Tăng cường dữ liệu: Mở rộng kích thước và đa dạng của tập huấn luyện bằng cách giới thiệu các bản sao được chỉnh sửa nhẹ của dữ liệu hiện có hoặc dữ liệu tổng hợp mới được tạo ra để giảm thiểu overfitting.)

• Feature selection/reduction: Decrease input feature count using methods like Principal Component Analysis (PCA) or manually selecting pertinent features to reduce complexity. (• Lựa chọn/giảm đặc trưng: Giảm số lượng đặc trưng đầu vào bằng cách sử dụng các phương pháp như Principal Component Analysis (PCA) hoặc lựa chọn thủ công các đặc trưng liên quan để giảm độ phức tạp.)

b. Specific Techniques for Machine Learning: (b. Các kỹ thuật cụ thể cho học máy:)

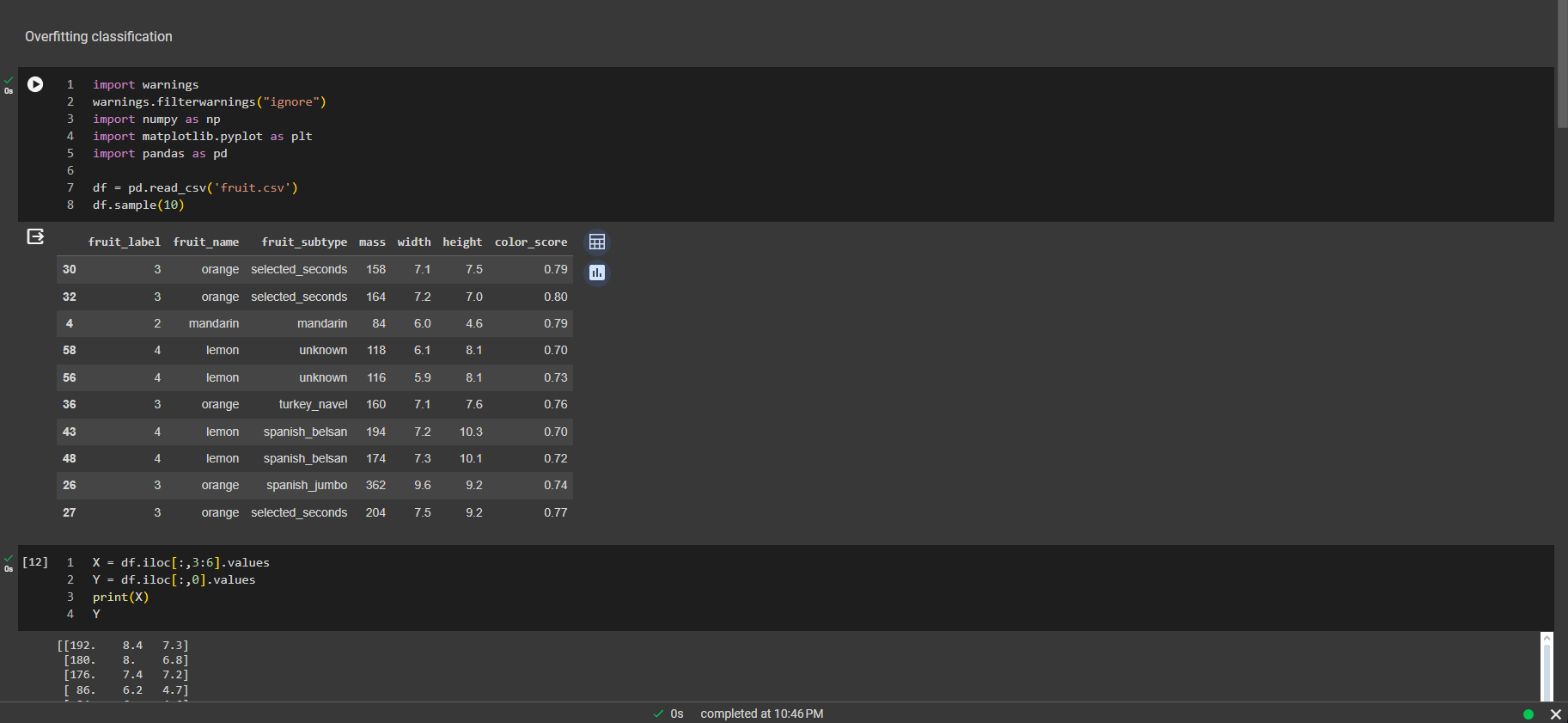
• Decision Trees: Implement pruning techniques, specify a maximum tree depth, or set minimum samples per leaf to regulate growth. (• Cây quyết định: Thực hiện các kỹ thuật cắt tỉa, xác định độ sâu tối đa của cây, hoặc đặt số mẫu tối thiểu cho mỗi lá để điều chỉnh sự phát triển.)

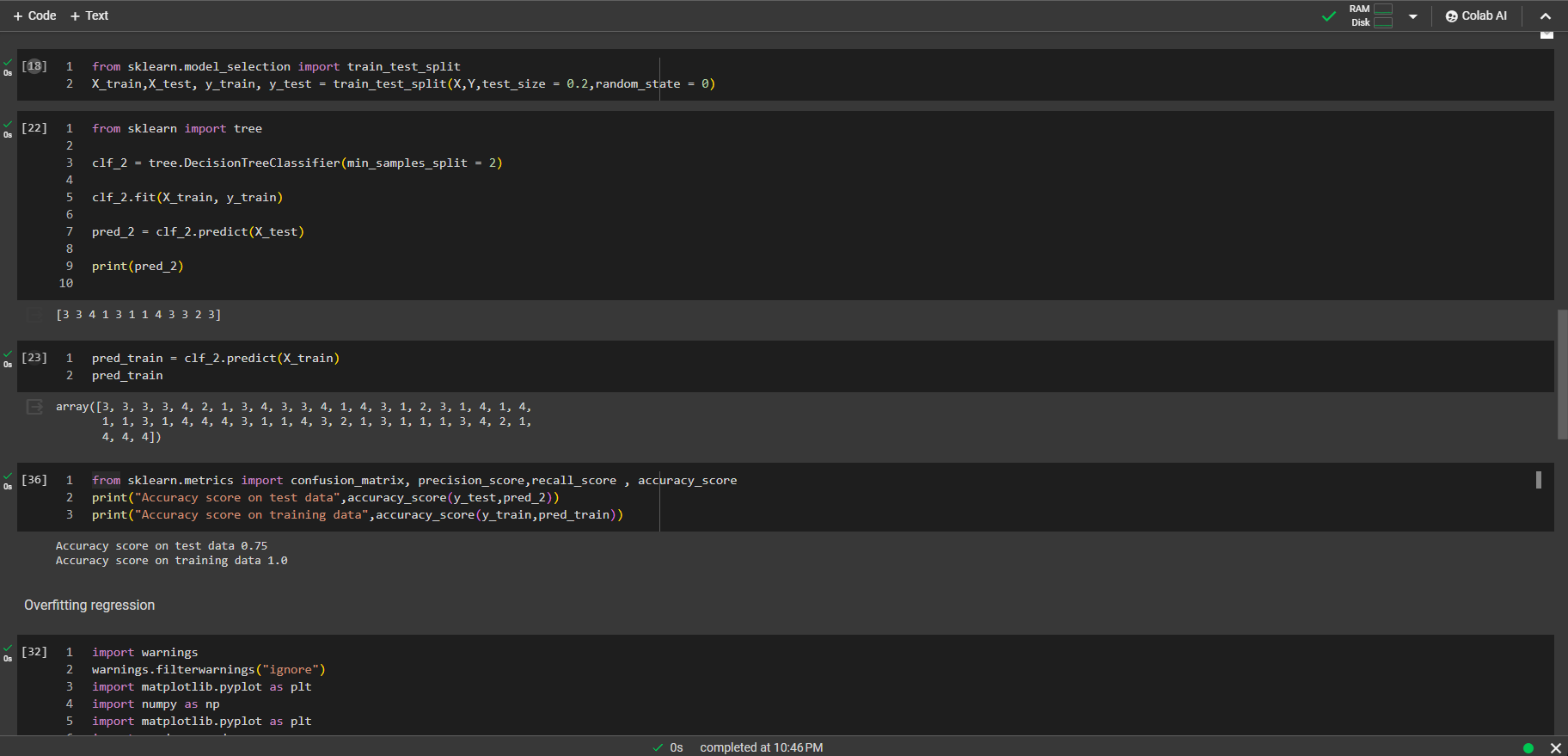
• Support Vector Machines (SVMs): Opt for an appropriate kernel and regularization parameter (C) to prevent overfitting. Utilizing simpler kernels, like linear, can be beneficial for certain datasets. (• Máy vector hỗ trợ (SVMs): Chọn một kernel phù hợp và tham số điều chuẩn © để ngăn chặn overfitting. Sử dụng các kernel đơn giản hơn, như tuyến tính, có thể có lợi cho một số tập dữ liệu.)

• K-nearest neighbor: Determine the number of neighbors (k) via cross-validation to ensure the model does not overemphasize noise in the data prematurely. (• K-nearest neighbor: Xác định số lượng hàng xóm (k) thông qua cross-validation để đảm bảo rằng mô hình không nhấn mạnh quá sớm nhiễu trong dữ liệu.)

(3) Present the problems mentioned in questions 1 and 2 through one or more real data sets (for classification and/and regression problems). Note that you must choose data sets and machine learning method where overfitting occurs.

Overfitting Classification





Overfitting Regression

