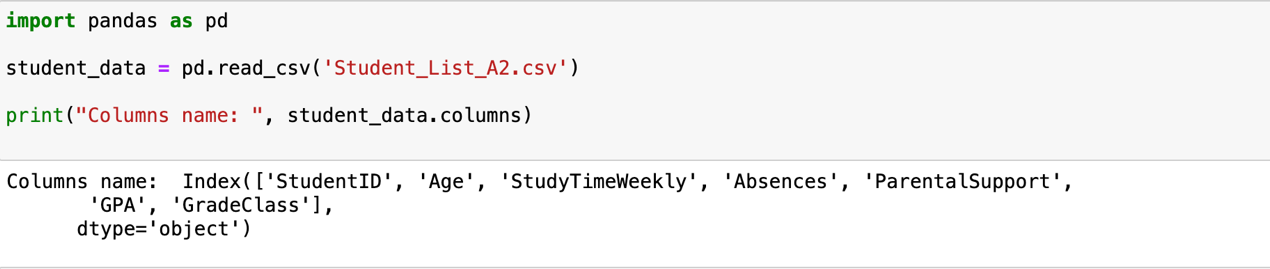
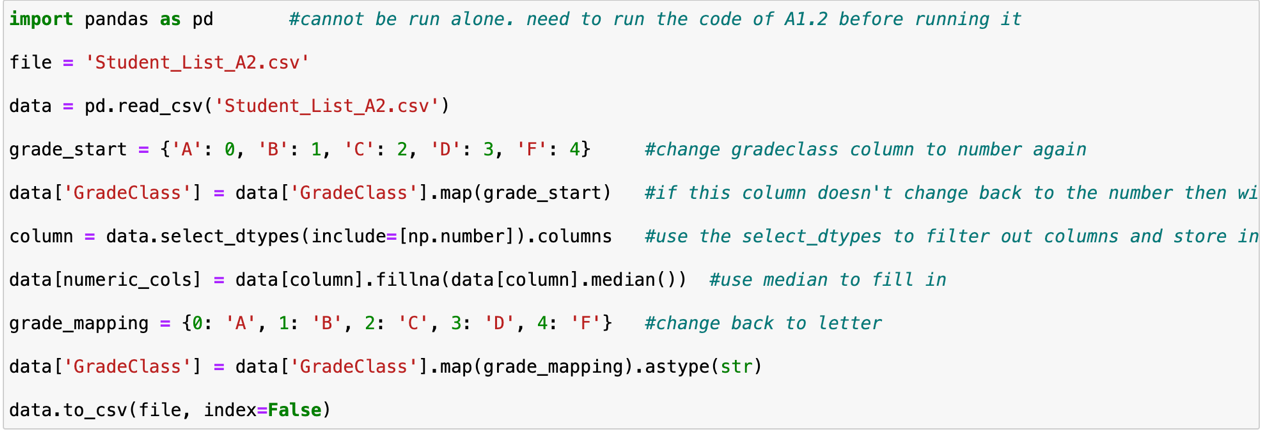
# FIT1043 Assignment 2-34550720

## A1. Data Wrangling

A1.1:

A1.2:文本

描述已自动生成

A1.3:

(PS: This code does not run alone. need to finish running A1.2. Need run one by one.)

Filling with a median:

Robustness: The median is not sensitive to extreme values and better reflects the central trend of the data.

Applicability: Applies to numeric data, especially when the data has outliers.

A1.4:图片包含 文本

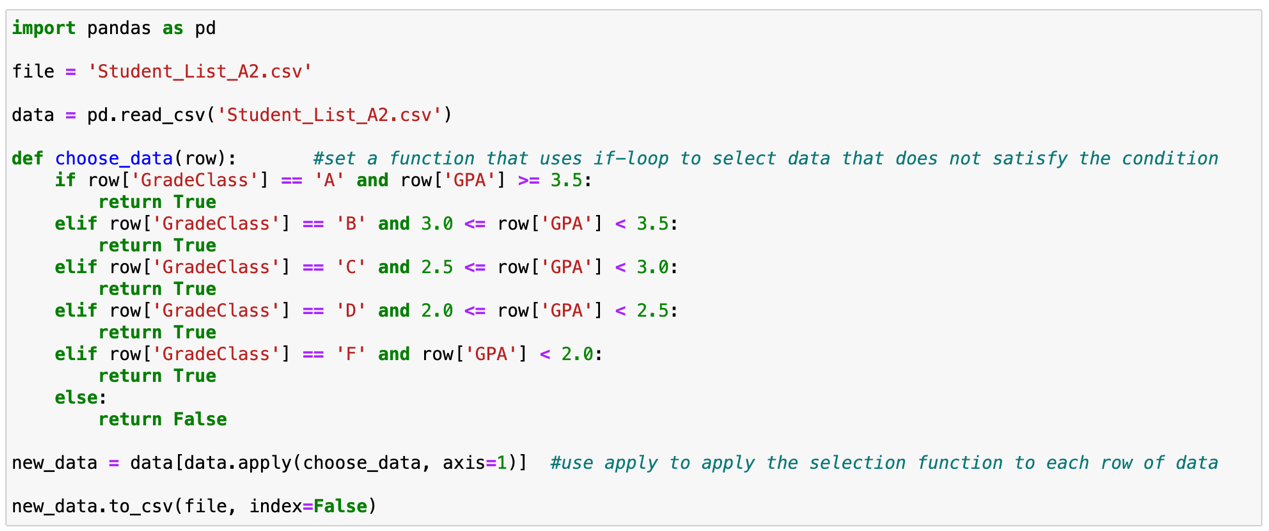
描述已自动生成

Remove outliers:

Improve model performance: Outliers can interfere with machine learning models, resulting in overfitting or increased bias.

Data consistency: Removing outliers can make data more consistent and reduce analysis biases caused by abnormal data.

Enhanced data quality: Outliers are usually errors or anomalies in the data collection process and removing them can improve the reliability of the data.

A1.5:

Filter out all rows that do not meet the criteria. This means that rows whose GPA does not match GradeClass will be deleted. For example, if A row has a GPA of 3.8 but GradeClass is 'B', the row is deleted because it does not meet the condition that GPA >= 3.5 should correspond to 'A'.

## A2. Supervised Learning

A2.1:

**Supervised machine learning**

Supervised learning concept Supervised learning is a basic type of machine learning where the goal is to learn models from labeled data to make predictions about new data. Supervised learning has the following key elements:

***Features:*** Input data used for prediction. It is usually expressed in matrix form, with each row representing a sample and each column representing a feature.

***Labels:*** Indicates the class or target value to which the sample belongs. Supervised learning models make predictions by learning the relationship between input features and labels.

***Training Set:*** A data set of known features and labels for the training of the model.

***Test Set:*** A data set that is used to verify the performance of a model and contains known labels but is not used in training.

**Partitioning of data sets**

In supervised learning, dividing a data set into a training set and a test set is a very important step. A common division is to use 80% of the data for training and 20% for testing.

***Training Set:*** Used to train a model that learns the mapping between features and labels in the training set.

***Test Set:*** Used to evaluate the generalization ability of the model, the test set is not involved in model training, and is used to verify the performance of the model on unknown data.

**Common algorithms for supervised learning**

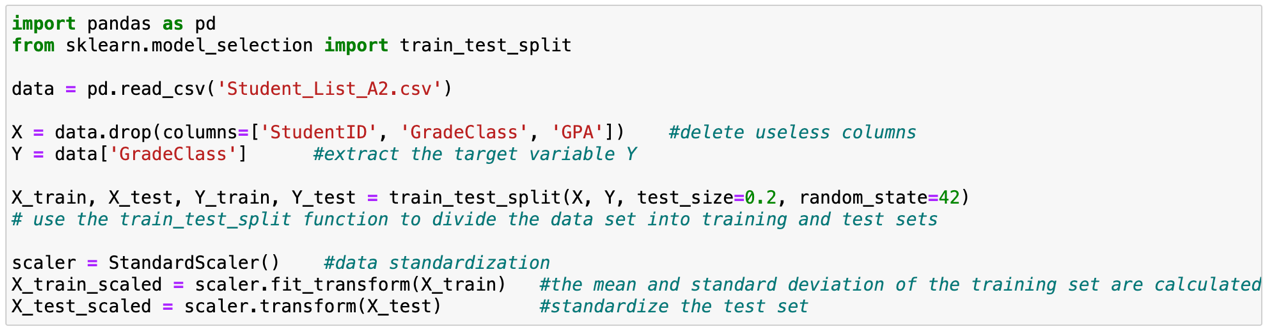
***Classification:*** used to predict discrete class labels, such as support vector machines (SVMS), decision trees, random forests, etc.

***Regression:*** Used to predict continuous numerical labels, such as linear regression, decision tree regression, etc.

**Feature selection and data preparation**

In supervised learning, it is very important to select the right features. Feature selection can affect the accuracy and training speed of the model. In the data preparation stage, it is necessary to remove irrelevant features, process missing values, and encode features.

A2.2 & A2.3:



***X:*** X is the feature of the data set. With the drop method, we delete the columns ['StudentID', 'GradeClass', 'GPA'], leaving the remaining columns as features.

* A StudentID is a unique identifier for a student that has no practical meaning for the classification task and is usually removed.
* GradeClass is the target variable, which is what we're predicting, so it won't be a feature either.
* GPA represents the Grade Point Average of students. According to the previous task description, this column has been used to generate GradeClass, so the two columns are repeated (GradeClass is generated according to GPA) and do not need to be used as a feature.

***Y:*** Y is the target variable, also called the label, which is what we want the model to predict. The GradeClass column is selected, which represents the student's grade classification (such as A, B, C, etc.).

***test size=0.2:*** indicates that 20% of the samples in the data set are divided into test sets, and the remaining 80% are divided into training sets.

## A3. Classification (training)

A3.1.a:

Data Normalization and Standardization:

Scaling Data Standardization refers to the Data, make it have the same scale, to better deal with machine learning model. Normalization (also called feature scaling) is crucial when working with machine learning algorithms that are sensitive to the scale of the input data. The common methods of Standardization are:

Z-score standardization: The mean of a feature is reduced to zero, and the standard deviation is normalized. It is suitable for most machine learning algorithms, such as SVM and linear regression.

A3.1.b:

In sklearn.preprocessing, the most common function used for normalization is StandardScaler(). This scaler standardizes features by removing the mean and scaling to unit variance. This ensures that all features are treated equally by the model.

A3.2.a:

SVM is a supervised learning model for classification and regression, often used for classification tasks. The core idea is to find a hyperplane (or decision boundary) that maximizes the Margin between two types of data points. It looks for the optimal hyperplane in a high-dimensional space to classify data points as clearly as possible. SVM include:

Hyperplanes: decision boundaries used to separate different classes of data.

Support Vectors: The sample points closest to the hyperplane that determine the direction and position of the hyperplane.

Kernel Function: When the data cannot be segmented linearly, SVM maps the data to the high-dimensional space through the kernel function, so that the data can be segmented linearly in the high-dimensional space. Common kernel functions are linear kernel, polynomial kernel, and RBF kernel.

A3.2.b:

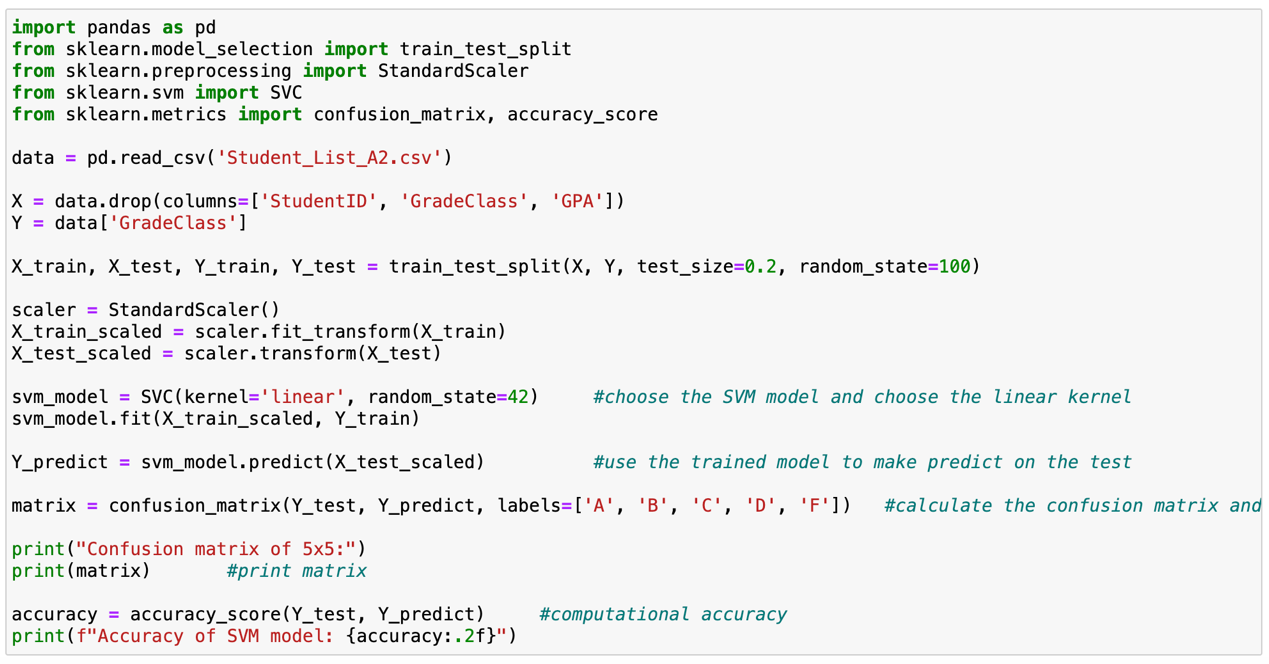
Kernel Function: When the data cannot be segmented linearly, SVM maps the data to the high-dimensional space through the kernel function, so that the data can be segmented linearly in the high-dimensional space. Common kernel functions are linear kernel, polynomial kernel, and RBF kernel.

Linear Kernel: Used when the data is linearly separable.

Polynomial Kernel: Useful when the data is polynomially separable.

Radial Basis Function (RBF) Kernel: A popular kernel for non-linear classification tasks. It maps the data into infinite-dimensional space.

A3.2.c



First, the students' data is read from the CSV file and then stored.

X is the feature and Y is the target variable.

The data set is then divided into a training set and a test set, with the test set being 20% as required.

After many tests and changes, I found that when random\_state=100, the prediction accuracy is the highest. So, I use random\_state=100.

The training set is standardized using StandardScaler, the mean and standard deviation are calculated and applied to the training set.

The same normalization is performed on the test set, but the mean and standard deviation are not recalculated.

Create a linear kernel support vector machine classifier. The SVM model was trained using the training set data.

A3.3

I used the decision tree in task A3.3.

A decision tree is a flowchart-like structure where internal nodes represent features, branches represent decision rules, and leaf nodes represent results. The model learns to divide the data according to the most important features, at the nodes where the information gain is maximized.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

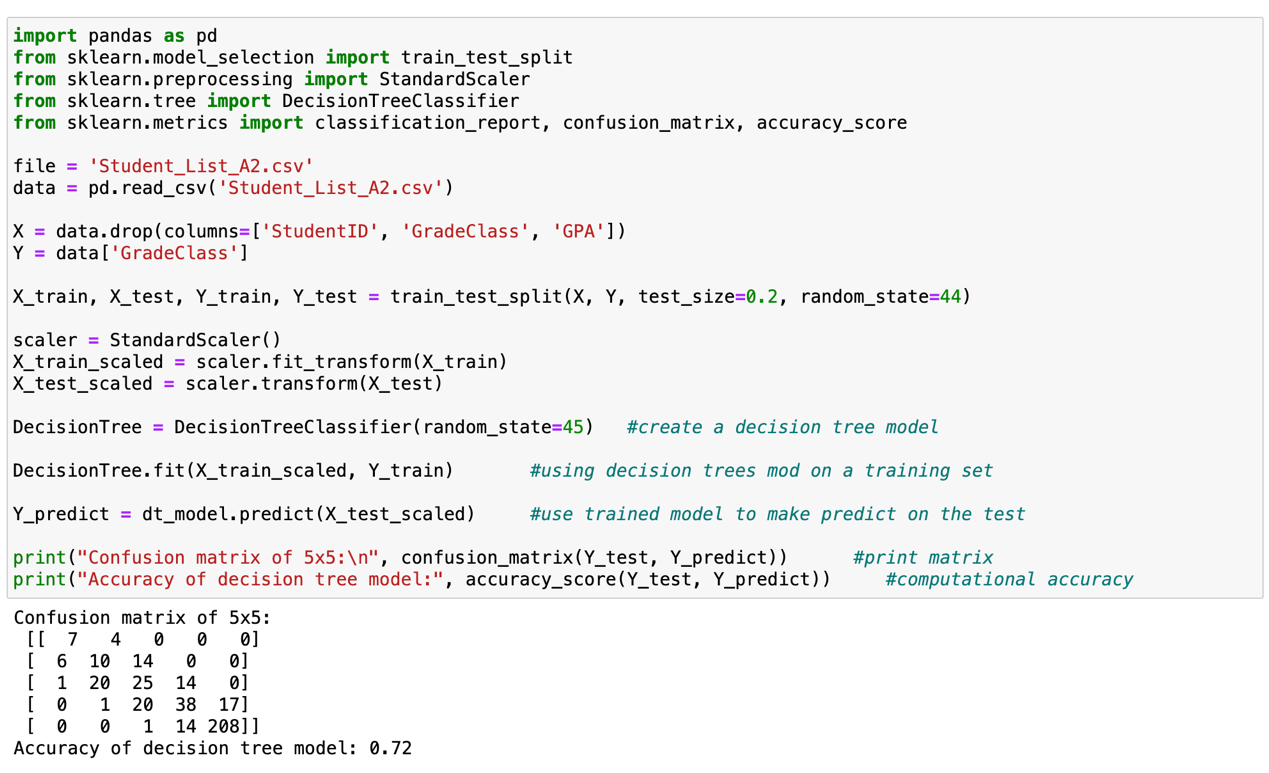
The overall flow of the code is similar to SVM. However, I need to change the model used, replacing the SVM model with a decision tree model.

After my test, in the decision tree model, the prediction accuracy is relatively high when random\_state=45, so I changed 100 to 45.

## A4. Classification (prediction)

A4.1 & A4.2

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

The first image shows the 5x5 matrix of the SVM model with an accuracy of 78%

The second image shows a 5x5 matrix of the decision tree with an accuracy of 72%.

A4.3

**Analyze decision tree model:**

Advantages:

Easy to understand and explain: The structure of the decision tree is similar to the human decision-making process and is easy to visualize and understand.

Multiple types of data can be processed: both numerical and categorical data can be processed.

There are fewer data preprocessing requirements: there is no need to standardize or normalize the data, and missing values can be processed.

Ability to handle multiple output problems: decision trees can predict multiple target variables simultaneously.

Disadvantages:

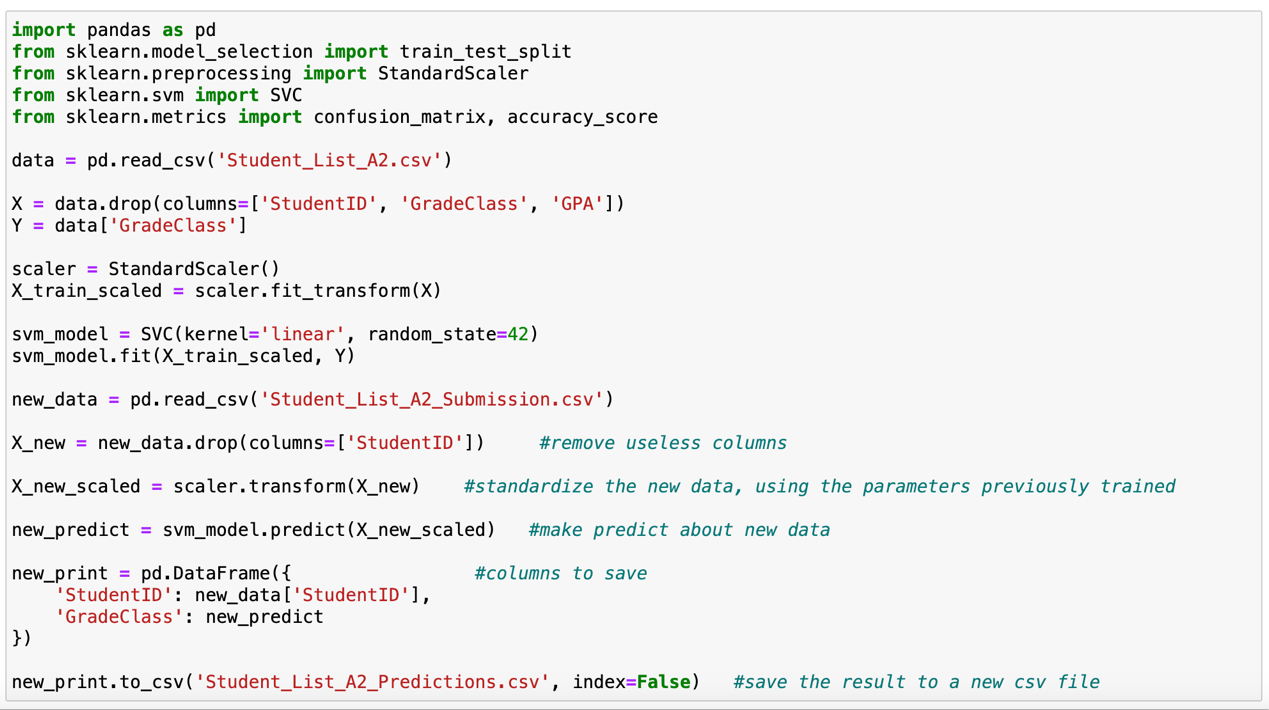
Easy to overfit: Decision trees tend to overfit training data, especially when the depth of the tree is large.

Sensitive to small changes in the data: Small changes in the data can result in a completely different tree structure, which can affect the stability of the model.

Multi-category features: When there are more features and more categories, the decision tree is more inclined to choose features with more values for splitting.

**Choose the SVM model and explain why:**

Compared with SVM model, decision tree model is less complex, because the basic structure of SVM model is hyperplane and support vector. Although the training and prediction speed of SVM model are lower than that of decision tree model, the prediction accuracy is higher. Because there are four data to be processed, the multi-dimensional SVM model is dominant. Compared with decision tree model, SVM model has lower overfitting risk and stronger anti-noise ability. So, the final choice is the SVM model.

A5. Independent evaluation表格

描述已自动生成

There are 11 'A' and none of them are predict correct

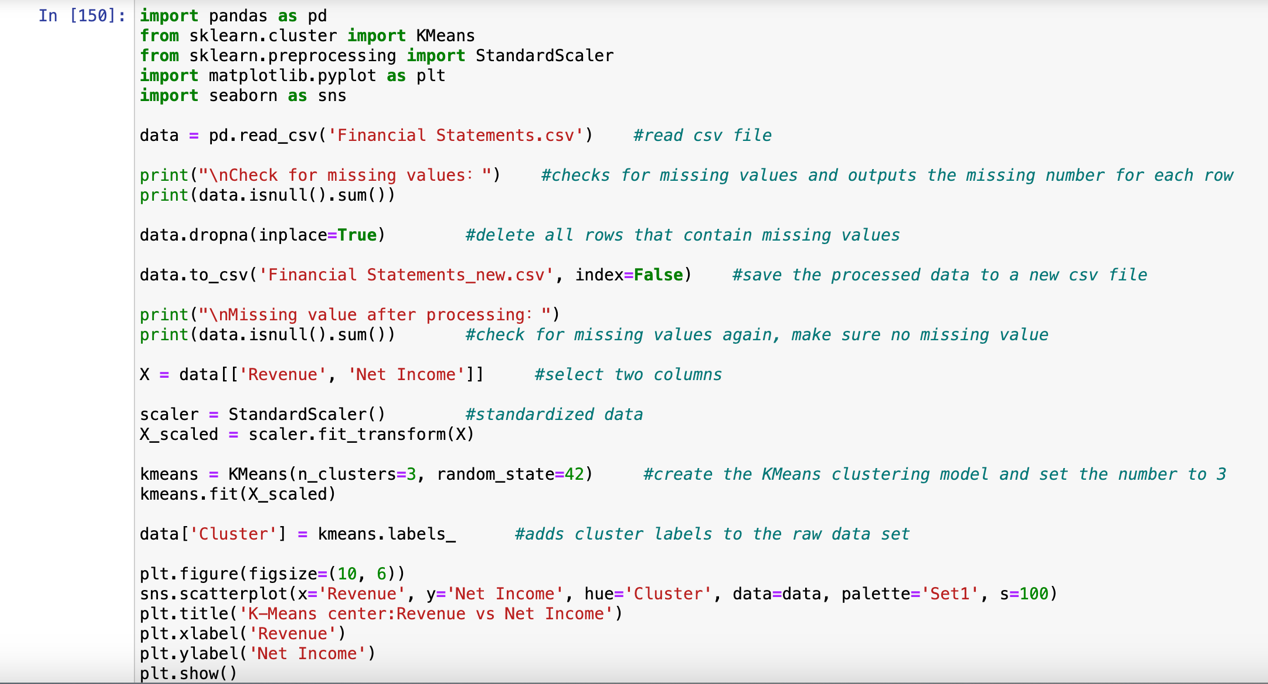
表格

中度可信度描述已自动生成表格

描述已自动生成

I used SVM model for prediction, but I found that there was no 'A' in the predicted student grades. Through the 5x5 confusion matrix before the SVM model, it is obvious that during training and testing, the prediction accuracy of 'A' option is 0. So there is no 'A' option for this prediction.

# Task B: Selection of Dataset, Clustering and Video Preparation

形状

描述已自动生成形状

描述已自动生成图表, 散点图

描述已自动生成