Python資料分析與機器學習應用
Data Analysis and Machine Learning with Python
Midterm Test

The total score of the test is 150 points, with a maximum score of 120 points.

Please answer your questions in a new document (e.g., doc, pdf, ppt) in the order of the questions. You can answer in either English or Chinese. Once completed, please upload it to NTU Cool before the end of the test (before the system closes).

Note 1: Avoid uploading your answers at the last minute to prevent network congestion or any unforeseen issues that may cause failure to upload them. For example, discrepancies between the NTU Cool system time and the school clock time. DO NOT accept late submissions or make-up submissions.

Note 2: You can open books and use AI tools for assistance, but you CANNOT directly copy the answers they provide. In other words, please try to describe the answers in your own words after understanding them.

I. (10%) **True or False Questions (Please answer T for True or F for False):**

1. In data analysis, using the Python Pandas package makes it convenient to handle structured data.

2. In data analysis, data cleaning only refers to removing missing values from raw data.

3. In machine learning, we typically divide data into training and testing data sets to evaluate the performance of the model.

4. Supervised learning is a machine learning approach where the model is trained based on labeled training data.

5. Decision tree is a supervised learning method used for both classification and regression tasks.

II. (10%) **Multiple Choice Questions (Please answer a, b, c, d):**

1. Which method can be used for statistical analysis in Python?

a) describe()

b) read_csv()

c) plot()

d) fit()

2. Which is the common algorithm used for classification tasks in machine learning?

a) Linear Regression

b) K Nearest Neighbors

c) K-means

d) PCA

3. Which function is used for handling missing values in data analysis?

a) dropna()

b) fillna()

c) isnull()

d) all of the above

4. Which method is used for grouping data in data analysis?

a) Groupby()

b) Split()

c) Filter()

d) Map()

5. Which chart can be used to visualize the distribution of data?

a) histogram

b) boxplot

c) scatter plot

d) all of the above

III. (80%) **Programming Questions:**

1. (20%) Please examine the code to answer the questions below:
   a. (10%) What is the purpose of the following code? (Please provide a detailed description beyond simply stating "data analysis")
   b. (10%) Is there a simpler way to achieve the same result? (Please provide the code in either an ipynb or py file)

```
import pandas as pd

df = pd.read_csv("question.csv")

data_dict = {}

for i in range(df.shape[0]):
    category = df.loc[i, "Category"]
    if (df.loc[i,'Group'] not in data_dict):
        data_dict[df.loc[i,'Group']] = {}
    data_dict[df.loc[i,'Group']][category] = df.loc[i,
'Value']

group_list = []
categoryA_list = []
categoryB_list = []
categoryC_list = []
categoryD_list = []
categoryE_list = []

for k, v in data_dict.items():
    group_list.append(k)
```

```
    if "A" in v:
        categoryA_list.append(1)
    else:
        categoryA_list.append(0)

    if "B" in v:
        categoryB_list.append(1)
    else:
        categoryB_list.append(0)

    if "C" in v:
        categoryC_list.append(1)
    else:
        categoryC_list.append(0)

    if "D" in v:
        categoryD_list.append(1)
    else:
        categoryD_list.append(0)

    if "E" in v:
        categoryE_list.append(1)
    else:
        categoryE_list.append(0)

data_new = {"Group": group_list, "A": categoryA_list,
"B": categoryB_list, "C": categoryC_list, "D":
categoryD_list, "E": categoryE_list}
df_new = pd.DataFrame.from_dict(data_new)

df_new["A_Count"] = df_new["A"]
df_new["B_Count"] = df_new["B"]
```

```
df_new["C_Count"] = df_new["C"]
df_new["D_Count"] = df_new["D"]
df_new["E_Count"] = df_new["E"]

result = df_new[df_new["D_Count"] == 0].shape[0]
print(result)
```

2. (35%) In class, we utilized three features from the Titanic dataset: "Pclass," "Sex," and "Age," as features / predictors (X), with the "Survived" column serving as the target label / variable (y). With data from nearly 900 passengers, we trained eight classification supervised learning models (e.g., Logistic Regression, K-NN, SVC, Gaussian Naive Bayes, Multinomial Naive Bayes, Decision Tree, Random Forest, and XGBoost) to predict the survival of Jack and Rose from the movie Titanic.

Now, we have obtained an updated passenger list (comprising over 1,300 passengers) and additional column meanings from titanic.csv and the definitions of the following data dictionary and variable notes.

Data Dictionary

Variable	Definition	Key
Survived	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children traveled only with a nanny, therefore parch=0 for them.

If we add two new features, SibSp and Parch, as new features / predictors (X), resulting in a total of five features, and redefine the data for Jack and Rose as follows:
- Jack, a 20-year-old male, a poor child who won a third-class ticket at a casino and boarded the ship with a friend;
- Rose, a 17-year-old female, a daughter of nobility, boarded the ship with her mother and fiancé.

Please answer the following questions:
a. (24%) What are the survival results for Jack and Rose using the eight classification models? (Please provide the code in either an ipynb or py file)

b. (3%) Are there any models with significantly different results?
c. (8%) Given the above, what do you think could be potential reasons for the differences in results?

3. (25%) The motor products manufactured by a certain company have a certain life cycle. Please analyze the provided data on motor life cycles to answer the following questions: (Please provide the code in either an ipynb or py file)

a. (5%) What is the average lifespan of the batch of motors? Please calculate the mean and explain its significance.
b. (5%) Please plot a histogram of motor lifespans to show the distribution across different lifespan intervals.
c. (5%) How many products in the batch have reached or exceeded the expected lifespan? The expected lifespan is 1000 hours.
d. (5%) Please calculate the standard deviation of the motor lifespans and explain its significance.
e. (5%) Does the lifespan of the batch of motors follow a normal distribution? Please briefly explain and provide the results of relevant statistical tests. For example, you can use the Shapiro–Wilk test in Scipy (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html).

Data for 10 motors:

Motor ID	Lifespan (hours)
1	980
2	1050
3	990

4	1100
5	1000
6	980
7	1020
8	990
9	950
10	1020

IV. (50%) **Essay Questions:**

1. (10%) Please explain the importance of data cleaning in the process of data analysis, and list three common data cleaning techniques, providing examples of their application scenarios.

2. (10%) Please explain the significance of overfitting in machine learning and provide methods to avoid overfitting.

3. (10%) Please explain the working principles of the decision tree and the random forest models, compare their advantages and disadvantages, and provide examples of suitable application scenarios for each.

4. (10%) Please discuss common data visualization tools in data analysis, such as Matplotlib and Seaborn, explaining their pros and cons and suitable usage methods.

5. (10%) Is it appropriate to use random forest regression and decision tree regression to detect anomalies? Please elaborate carefully.