SCHOOL OF COMPUTING, TECHNOLOGY AND APPLIED SCIENCES

CAS4192 MACHINE LEARNING TECHNIQUES

FINAL SEMESTER EXAMINATION

TUESDAY, 26 NOVEMBER 2024

09:00 - 12:00 HOURS

TIME ALLOWED : WRITING – THREE HOURS : READING – 5 MINUTES

INSTRUCTIONS:

1. Section A: this question is **compulsory** and must be attempted.
2. Sections B: Answer **THREE (3)** questions from this section.
3. This examination paper carries a total of **100 marks**.
4. Candidates must **not turn this page** until the invigilator tells them to do so.

SECTION A: Question 1 is compulsory and must be attempted

NOTE: Your answers should NOT be brief. Give detailed essay style reasoned answers with structured paragraphs and headings.

Question 1

Using the model code and its corresponding output given below, critically study both and answer the following questions:

a) Compare the performance metrics (confusion matrix, accuracy score, and classification report) generated by the SVM model. Which metric is the most informative for assessing the model's effectiveness, and why? **(10 Marks)**

b) Critically study the results produced by the box plot graphs and interpret the results in the context of the given model. **(10 Marks)**

c) Explain how the encoding of categorical variables impact the SVM model's ability to classify large_purchase accurately and identify any potential limitations or biases introduced during this step. **(10 Marks)**

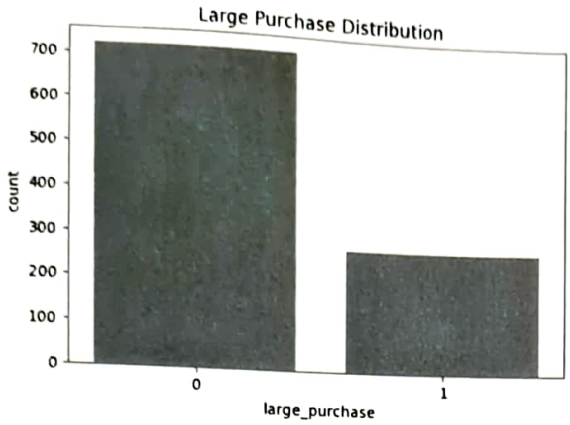
d) Based on the analysis of the prediction results, propose a strategy to improve the model's performance and suggest additional features or preprocessing steps that could enhance its predictive accuracy. **(10 Marks)**

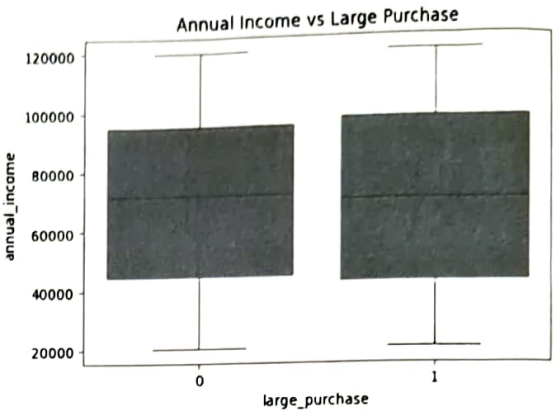
Python Model Code

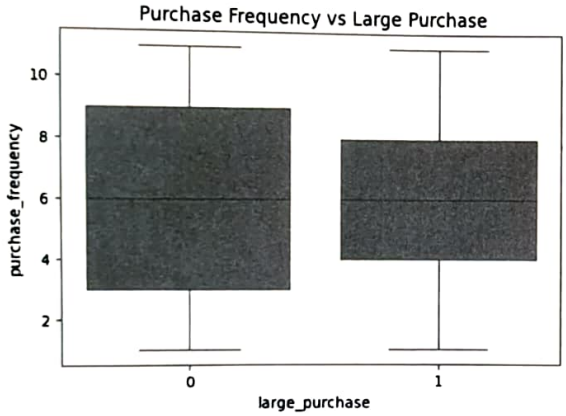
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# Simulating data for retail customer behavior
np.random.seed(42)
n_samples = 1000
data = {
    'age': np.random.randint(18, 65, n_samples),
    'annual_income': np.round(np.random.uniform(20000, 120000, n_samples), 2),
    'purchase_frequency': np.random.randint(1, 12, n_samples),
    'membership_type': np.random.choice(['Basic', 'Premium', 'VIP'], n_samples),
    'online_shopping': np.random.choice(['Yes', 'No'], n_samples),
    'loyalty_program': np.random.choice(['Yes', 'No'], n_samples),
    'large_purchase': np.random.choice([0, 1], n_samples, p=[0.7, 0.3])
}

# Creating a DataFrame
```

```
df = pd.DataFrame(data)

# Encoding categorical variables
df_encoded = pd.get_dummies(df, drop_first=True)

# Splitting data into features (X) and target (y)
X = df_encoded.drop('large_purchase', axis=1)
y = df_encoded['large_purchase']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Scaling features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Training the SVM model
svm_model = SVC(kernel='linear', C=1.0, random_state=42)
svm_model.fit(X_train, y_train)

# Making predictions
y_pred = svm_model.predict(X_test)

# Evaluating the model
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

print("\nAccuracy Score:")
print(accuracy_score(y_test, y_pred))

# Visualization
plt.figure(figsize=(6, 4))
sns.countplot(data=df, x='large_purchase')
plt.title('Large Purchase Distribution')
plt.show()

plt.figure(figsize=(6, 4))
sns.boxplot(data=df, x='large_purchase', y='annual_income')
plt.title('Annual Income vs Large Purchase')
plt.show()

plt.figure(figsize=(6, 4))
sns.boxplot(data=df, x='large_purchase', y='purchase_frequency')
plt.title('Purchase Frequency vs Large Purchase')
plt.show()
```

Output








```
Classification Report:
              precision    recall  f1-score   support

           0       0.74      1.00      0.85       149
           1       0.00      0.00      0.00        51

    accuracy                           0.74       200
   macro avg       0.37      0.50      0.43       200
weighted avg       0.56      0.74      0.64       200


Accuracy Score:
0.745
```

Total: [40 Marks]

SECTION B: Attempt any THREE questions in this section

Question 2

a) Convolutional Neural Networks (CNNs) are a type of artificial neural network specifically designed for processing and analyzing grid-like data. Describe any three applications of CNNs. **(6 Marks)**

b) Evaluation metrics in machine learning are measures used to assess the performance of models. Compare and contrast accuracy and precision metrics **(6 Marks)**

c) With the aid of an example, evaluate how increasing the size of the training dataset impacts a machine learning model that is overfitting. **(8 Marks)**

Question 3

[Total: 20 Marks]

Machine learning models are algorithms or mathematical frameworks that learn patterns and relationships from data to make predictions, classifications or decisions without explicit programming. These models process input data, adapt through training, and generalize to unseen scenarios. They are the core tools in machine learning systems for solving various tasks.

a) Support Vector Machines (SVMs) are supervised machine learning algorithms used for **classification**, **regression**, and **outlier detection**. They are particularly effective for high-dimensional datasets and when the relationship between features and labels is non-linear. With the a aid of a diagram, explain how SVMs work. **(6 Marks)**

b) Model selection refers to the process of choosing the most appropriate machine learning model for a given problem. Justify why model selection is important. **(6 Marks)**

c) A decision tree is a supervised machine learning algorithm used for classification and regression tasks. Write short notes on the following aspect of decision trees. **(8 Marks)**

Information Gain
Gini Index
Pruning
Entropy

[Total: 20 Marks]

Question 4

Machine learning can be categorized into three main types. Supervised learning involves training models with labeled data to predict outcomes. Unsupervised learning analyzes unlabeled data to uncover hidden patterns, such as clusters or associations. Reinforcement learning focuses on optimizing actions by learning from feedback through rewards or penalties.

a) Machine learning (ML) can be broadly categorized into different types based on the learning approach and the nature of the task. Identify and describe the three main types of machine learning types. **(6 Marks)**

b) K-Means is an unsupervised machine learning algorithm used for clustering, where data points are grouped into K clusters based on their similarity. List the seven steps of the K-means algorithms. **(8 Marks)**

c) Naive Bayes is a supervised machine learning algorithm based on Bayes' Theorem, widely used for classification tasks. Identify and just three areas where you would use this algorithm. **(6 Marks)**

[Total: 20 Marks]

Question 5

a) Exploratory Data Analysis (EDA) is a critical step in the data analysis and machine learning process. Justify why it is important to conduct EDA before Machine Learning models are trained and deployed. **(6 Marks)**

b) Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment. Describe any four key features of RL, **(8 Marks)**

c) Notebooks are interactive computational environments that combine code, text visualizations and data exploration in a single interface. Explain why it is important to use notebooks in Machine Learning Projects. **(6 Marks)**

[Total: 20 Marks]

END OF EXAMINATION