Women's Institute of Technology & Innovation (WITI)
Intermediate Data Science & Machine Learning
End of Semester II Examination

Course code: CSD 121
Time Allowed: 72 hours
Year of study: 1

Cohort III 2023/2024

General instructions
Attempt ALL Questions

- You have 72 hours to complete this exam. All necessary files for completing the exam are included in the attached zip file. When you submit the examination, I want to see a complete record of your work. Please attach all your python code scripts and any other files you think are necessary to evaluate your work (ensure that the plots and charts/graphs are visible in the environment).
- Remember that you will be graded not only on your answers, but also on your process (such as the efficiency of your code and comments). This exam is designed to assess logical reasoning and problem-solving in addition to coding skills, so if you're unable to answer a question using Python or via coding steps, please explain what steps you would have taken had you known the appropriate Python functions or commands. In general, remember to explain your reasoning and give as complete an answer as you can.
- Follow coding best practices to ensure your code is efficient, easily readable, and the results are reproducible even if the dataset changes over time. Using resources such as Python's internal help files and online resources is encouraged (and often necessary!), but please cite the resources you use, and do not consult with other classmates or people.

1. Using a dataset named "data". Taking advantage of the following libraries and dependencies and any other additional libraries you may deem neccessary: Answer the questions starting from Qn. 2 (Ungraded)

   Standard libraries for data pre-processing and machine learning algorithms:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_ objects as go
from plotly.subplots import make _ subplots
import warnings
warnings.filterwarnings('ignore')
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model _ selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
```

from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from xgboost import XGBClassifier
from catboost import CatBoostClassifier
from sklearn import metrics
from sklearn.metrics import roc_curve
from sklearn.metrics import recall_score, confusion_matrix, precision_score, f1_score, accuracy_score, classification_report
from sklearn.ensemble import VotingClassifier

Standard libraries for data visualization:
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.metrics import f1_score, precision_score, recall_score, fbeta_score
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import ShuffleSplit
from sklearn.model_selection import KFold
from sklearn import feature_selection
from sklearn import model_selection
from sklearn import metrics
from sklearn.metrics import classification_report, precision_recall_curve
from sklearn.metrics import auc, roc_auc_score, roc_curve
from sklearn.metrics import make_scorer, recall_score, log_loss
from sklearn.metrics import average_precision_score

2. Conduct the following data preprocessing on the dataset (10 marks).
   (a) Read the dataset into your environment.
   (b) Visually inspect the missing values in your dataset.
   (c) How big is your matrix?
   (d) How many variables are in the matrix?
   (e) Drop any unneccessary columns in the resulting matrix?
   (f) Using the 'fillna' function in your environment, fill the missing values in the column 'TotalCharges' using the mean values of the same column. (hint: one could first ascertain how many rows are missing data points in the "TotalCharges" column using the following piece of code: "data['TotalCharges'] = pd.to_numeric(data.TotalCharges, errors='coerce'"; and "data.isnull().sum()")
   (g) Transform the labels for the variable "SeniorCitizen" from "0/1" to "No/Yes"
3. Exploratory Data Analysis (EDA): In each of the questions provide a brief write up (not more than two sentences) of your findings (15 marks).
   (a) Using the appropriate variable, explore the distribution of customer churn in the dataset?

(b) How does the customer churn distribution vary across gender in the dataset?

(c) How does the customer churn distribution vary by contract type?.

(d) How does the customer churn distribution vary by payment method?

(e) How does the customer churn distribution vary by internet service and gender?

(f) Visualize and compare the correlation between churn rates and 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'PaperlessBilling', 'MonthlyCharges', 'TotalCharges'

4. Using the underlying objective function for Logistic Regression and K-Nearest Neighbor Cluster algorithm, create the corresponding matrices for inputs ('X') and outputs ('y'). Use the generated inputs and output to classify the churn among customers in the dataset. It is important to note that for columns of :"tenure", 'MonthlyCharges', 'TotalCharges', you may need to first standardize the scales as they may be distributed across wide ranges. (20 marks)

(a) Evaluate your models (logistic and KNN) using the following metrics ["Algorithm", "ROC AUC Mean", "ROC AUC STD", "Accuracy Mean", "Accuracy STD"] (5 marks).

GOOD LUCK!