Questions for "Machine Learning with Python" set by Drs. Abbas and Shihab

Question 1: [12 marks]

Consider the dataset entitled: "googleplaystore_new.csv" *available on Canvas*, and answer the following in ONE "Jupyter Notebook" File

A. Check for any missing values. If any, display all rows that contain null values, then delete them. Then, check for duplicates. If any, delete the redundant rows. **[1 mark]**
B. Show the average rating per each 'Category' using the appropriate plot. **[1 mark]**
C. Create a new column called 'Size in bytes' (numeric) and convert the entries from 'Size' column (M means megabyte and k means kilobytes). **[1 mark]**
D. Create a new column called 'Numeric_installs' (numeric) and convert the entries from 'Installs' column (remove '+' and ','). **[1 mark]**
E. Delete the following columns: "Type", "Price", "Genres", "Android Ver", "Size", "Installs" and "App" and save the new CSV file as "Updated_googleplaystore.csv". **[1 mark]**
F. Show the correlation table amongst all numeric features and plot them on a heatmap. For each pair with a high correlation (>0.70 or <-0.7), draw a scatter plot and provide a reasonable explanation of the plot/correlation. **[1 mark]**
G. Using all remaining features find the best regression model to predict "rating" (use the standard training/test partition without cross-validation). **[4 marks]**
H. Drop "Installs_Num" and "Size in Bytes" columns and repeat the previous task, what do you conclude? **[2 marks]**

Question 2: [18 marks]

Consider the dataset* entitled: "20221013_ted_talks.csv" *available on Canvas*, and answer the following in ONE "Jupyter Notebook" File

A. Check for any missing values. If any, display all rows that contain null values, then delete them. Then, check for duplicates. If any, delete the redundant rows. **[1 mark]**
B. How many unique values are in 'title', 'speaker', 'recorded_date', 'published_date' and 'event'? **[1 mark]**
C. Generate a new column labelled 'year' by extracting the year from the 'published_date' column, and subsequently remove all the columns specified in the preceding question. **[1 mark]**
D. We want to predict the number of "views" based on the "likes", "duration", and "year" ONLY (consider the 'year' feature as categorical not continuous value). Find the regression model that best fits the model using the standard train/test split along with a cross-validation approach. Do you think the model is reliable? Did you have any 'lucky' and/or 'unlucky' partitions? **[5 marks]**

E. We want to predict the number of "views" based on the "likes" ONLY using "Matplotlib's RegPlot". What is the best regression function degree that fits the model *visually.* **[2 marks]**

F. Based on the results from the previous part, what do you conclude? Could you prove your conclusion by adjusting and repeating the experiment of part D? **[4 marks]**

G. Although predicting the 'year' is a classification problem, let's try to solve it through a regression problem. Using "likes", "duration", & "reviews" and the Train/Validation/test partition along with the Cross-validation approach, build the best model to predict the 'year'. How would you interpret the predicted value (e.g. 2013.98)? **[4 marks]**

*Dataset information:
Source: https://www.kaggle.com/datasets/miguelcorraljr/ted-talks-2022

Attribute	Description	Data Type
talk_id	Talk identification number provided by TED	int
Title	Title of the talk	string
Speaker	Speakers in the talk	dictionary
recorded_date	Date the talk was recorded	string
published_date	Date the talk was published to TED.com	string
Event	Event or medium in which the talk was given	string
Duration	Duration in seconds	int
Views	Count of views	int
Likes	Count of likes	int

Question 3 [20 marks]

Consider the dataset provided in this link:
https://www.kaggle.com/datasets/harbhajansingh21/persistent-vs-nonpersistent

The dataset uses patient information, provider attributes and clinical and treatment factors to predict the likelihood of patients continuing to use a particular medication over an extended period as prescribed by their healthcare provider. Persistency is an important metric for pharmaceutical companies to track because it indicates how successful they are in retaining patients on their medications.

Study the dataset carefully and familiarise yourself with each column. Consider the type of data in each column.

Now examine the solution provided in this link:
https://www.kaggle.com/code/harbhajansingh21/logistic-regression-vs-svm-hypothesis-testing

This solution carries out some exploratory data analysis, splits the data into training and test parts and trains a Logistic Regression model and an SVM model. Study this notebook carefully.

Answer the following questions:

Q3.1. Are the independent variables suitable to the problem? Can you spot any potential biases in the data collection? **[2 marks]**

Q3.2. What are the key findings of the logistic regression model? How well does it predict continued medication use? **[4 marks]**

Q3.3. How has multicollinearity been addressed in the model? Are there any highly correlated independent variables that could affect the model's accuracy? **[4 marks]**

Q3.4. What are the limitations of the data and model? **[2 marks]**

Q3.5. Are there any external factors not included in the model that could influence medication adherence? **[2 marks]**

Q3.6. Train a KNN model and compare its performance to that of the Logistic Regression model. **[6 marks]**

Submission format: Word document for your answers to Q3.1–Q3.5 and a Jupyter notebook, with Markdown copious comments, for Q3.6.