Q.1 **[Total: 25 marks]**

(a) What is the relationship between Artificial Intelligence (AI), Machine Learning (ML) and Artificial Neural Network (ANN)?

Outline the main characteristics of each discipline and how they are related.

Indicate what are the major "seasons" in the development of AI history and point out the reasons that led to reduced interest during the AI winters.

[10 marks]

(b) What are the main branches of Machine Learning? For each branch, describe the distinctive aspects and provide relevant examples of practical applications.

List the different sample uses of machine learning, define what are Regression and Classification problem by highlighting the differences and provide example of their application in maintenance applications.

Indicate some techniques that can be used to solve Classification problems.

[5 marks]

(c) Explain what are underfitting and overfitting in the context of Machine Learning and sketch a diagram that illustrates the issue, highlighting the bias-variance trade-off.

Describe some techniques that can be used to mitigate overfitting.

Define what are Training, Test and Validation set and discuss possible ways to distribute data within the three sets.

[10 marks]

Q.2 **[Total: 25 marks]**

You are in charge of analysing data relative to delays in a metro system. The management of the company that operates the trains asks you to provide a report that may help in understanding the performance of the railway network.

The data are stored in a pandas DataFrame whose first five rows are illustrated in Fig. Q.2 (a).

The info() method for the DataFrame prints the following information:

```
Data columns (total 11 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   Date       119999 non-null  object
 1   Time       119999 non-null  object
 2   Day        119999 non-null  object
 3   Station    119999 non-null  object
 4   Code       119998 non-null  object
 5   Min Delay  119999 non-null  float64
 6   Min Gap    119999 non-null  float64
 7   Bound      93843 non-null   object
 8   Line       119537 non-null  object
 9   Vehicle    119999 non-null  float64
 10  Cause      119996 non-null  object
```

In particular, the columns “Date” and “Time” indicate the date (dd/mm/yyyy format) and the time (hh:mm format) (note: these are stored as string objects in the given DataFrame), when an event has occurred. Each event is associated with a “Code” and it may correspond to a delay, whose value in minutes is specified by “Min Delay”. The event is also associated with a certain “Cause”. The affected “Bound”, “Line” and “Vehicle” are also available in the dataset.

Using functions from the pandas and plotly libraries, write python code to:

(a) perform some cleaning and pre-processing of the dataset:

(i) print for each column in the DataFrame the respective number of missing values;

(ii) drop all the entries containing missing values;

(iii) remove all the rows for which the column ‘Line’ has a value other than: {“BD”,”YU”,”SHP”,”SRT”};

(iv) select only the rows with “Min Delay” greater than zero;

(v) add a column “Delay Type” to the DataFrame whose value is ‘short’ for delays between 1 and 5 minutes, ‘medium’ for delays between 5 and 15 minutes and ‘long’ for delays greater than 15 minutes, respectively.

[10 marks]

(b) perform some exploratory data analysis to:

(i) visualize the delay based on the day of the week;

(ii) determine how many delays of type ‘short’, ‘medium’ and ‘long’ are present, aggregated based on the day of the week;

(iii) create an interactive histogram and visualize the number of delay events per “Delay Type” based on the day of the week (hint: in plotly.express.histogram, use the barmode option as *barmode="group"*);

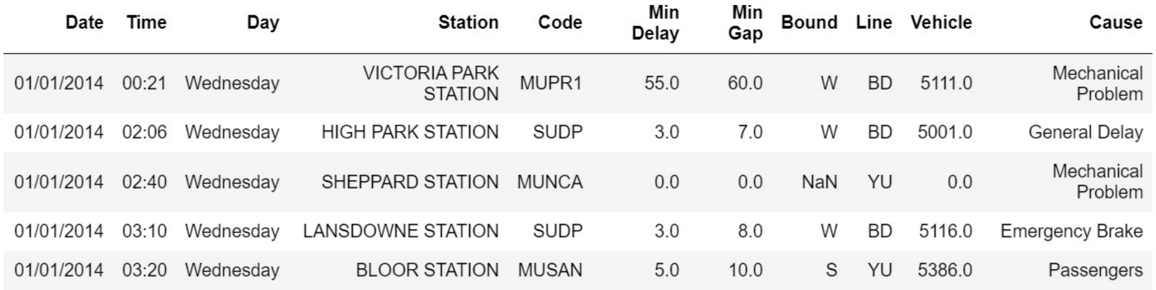
(iv) for the line ‘YU’, find the cumulated number of minutes of delay associated with each cause;

(v) create an interactive histogram to illustrate the total hours of delays for each cause;

(vi) create an interactive histogram to visualize the total hours of delays in each month;

(vii) create an interactive histogram to visualize, for the direction ‘S’, the cumulated number of events per hour of the day.

[15 marks]



Date	Time	Day	Station	Code	Min Delay	Min Gap	Bound	Line	Vehicle	Cause
01/01/2014	00:21	Wednesday	VICTORIA PARK STATION	MUPR1	55.0	60.0	W	BD	5111.0	Mechanical Problem
01/01/2014	02:06	Wednesday	HIGH PARK STATION	SUDP	3.0	7.0	W	BD	5001.0	General Delay
01/01/2014	02:40	Wednesday	SHEPPARD STATION	MUNCA	0.0	0.0	NaN	YU	0.0	Mechanical Problem
01/01/2014	03:10	Wednesday	LANSDOWNE STATION	SUDP	3.0	8.0	W	BD	5116.0	Emergency Brake
01/01/2014	03:20	Wednesday	BLOOR STATION	MUSAN	5.0	10.0	S	YU	5386.0	Passengers

Fig. Q.2 (a)

Q.3 **[Total: 25 marks]**

You are the chief data scientist in a company that is in charge of performing the maintenance of a railway fleet. At the end of the service, the maintenance team reports the state of the train as 'good' or 'degraded'. If the train is 'good', it is made available for service for the next day, otherwise it is sent to the workshop for reparation. Suppose that the state of health of the train is associated with two real variables x$_1$, x$_2$ ϵ [0, 10], measured from the train. You are given a set of 8 samples *i*=1,…8 reported below, where the train health status is represented by the binary variable *y* ϵ {"good", "degraded"} (assume the class degraded as positive, i.e. "*degraded*"=1).

i	$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
1	1.3	5.5	*degraded*
2	8.2	6.0	*degraded*
3	6.2	4.5	*good*
4	0.5	3.4	*good*
5	7.8	1.3	*degraded*
6	6.5	9.2	*degraded*
7	1.5	7.1	*good*
8	0.9	8.9	*good*

(a) consider the logistic model $\boldsymbol{p}_\theta\left(y^{(i)} | \, \boldsymbol{x}^{(i)}; \boldsymbol{\vartheta}\right) \sim \sigma\left(\boldsymbol{\vartheta}^T \boldsymbol{x}^{(i)}\right) = \sigma(\vartheta_1 x_1^{(i)} + \vartheta_2 x_2^{(i)})$, where $\sigma(\cdot)$ = sigmoid function associated with the standard logistic curve and $\boldsymbol{\vartheta}^T = [\vartheta_1, \vartheta_2]$ are the coefficients of the logistic regression. Using a decision boundary (probability threshold) of 0.5, calculate the accuracy, precision, sensitivity and F1-score of the model with coefficients (i) $\boldsymbol{\vartheta}^T = [2, -1]$ and (ii) $\boldsymbol{\vartheta}^T = [1, -1]$, respectively.

[20 marks]

(b) On the x$_1$-x$_2$ plane, draw (i) each sample of the given dataset (using a different marker to represent the associated class) and (ii) the two straight lines corresponding to $\boldsymbol{\vartheta}^T \boldsymbol{x} = 0$ for the coefficient values given in **(a)**, respectively, to graphically represent the classification of the points.

[5 marks]

Q.4 **[Total: 25 marks]**

(a) Given the following data points:

$$p_1 = \begin{bmatrix} 7.5 \\ 4 \end{bmatrix},\ p_2 = \begin{bmatrix} 1.7 \\ 9.9 \end{bmatrix},\ p_3 = \begin{bmatrix} 3.1 \\ 3.1 \end{bmatrix},\ p_4 = \begin{bmatrix} 8 \\ 4.5 \end{bmatrix},\ p_5 = \begin{bmatrix} 2.2 \\ 0.1 \end{bmatrix}, p_6 = \begin{bmatrix} 7.8 \\ 0.5 \end{bmatrix}$$

compute the first iteration of the k-means algorithm assuming that the data shall be partitioned into two clusters (k=2) and that the initial centroids are associated with the points:

$$c_1 = \begin{bmatrix} 0.1 \\ 5.1 \end{bmatrix}, \qquad c_2 = \begin{bmatrix} 3.8 \\ 2.8 \end{bmatrix}$$

Indicate to what clusters each point is assigned and compute the new coordinates of the centroids after the first iteration. Sketch a diagram illustrating the data points, the clusters, and the centroids.

[8 marks]

(b) A MultiLayer Perceptron (MLP) Artificial Neural Network with three layers contains:

- 5 neurons in the input layer
- 10 neurons in the hidden layer
- 4 neurons in the output layer

For a given input, calculate how many sum and multiplication operations are needed in one forward propagation step. Provide the number of operations per layer as well as the total number.

[2 marks]

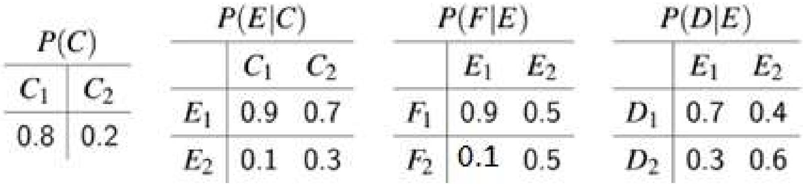
(c) Explain the basic network architecture of the AutoEncoders (AE) and how it can be beneficial for anomaly detection.

[15 marks]

Q.5 **[Total: 25 marks]**

(a) Consider a graphical network with four binary variables C, D, E and F.

The probability tables are given as follows:



$P(C)$

C_1	C_2
0.8	0.2

$P(E|C)$

	C_1	C_2
E_1	0.9	0.7
E_2	0.1	0.3

$P(F|E)$

	E_1	E_2
F_1	0.9	0.5
F_2	0.1	0.5

$P(D|E)$

	E_1	E_2
D_1	0.7	0.4
D_2	0.3	0.6

(i) Construct the network model.

(ii) Evidence shows that F = F1. Using a bottom-up propagation of Pearl's tree algorithm, update the inference for C.

[15 marks]

(b) Construct a Bayesian network based on the qualitative statistical dependencies described as follows:

1. Uneven road (UR) increases the chance of suspension damage (SD).
2. Inadequate tyre pressure (TP) can lead to both passenger discomfort (PD) and fuel inefficiencies (FI).
3. The absence of adequate type pressure (TP) can be manifested in either passenger discomfort (PD) of fuel inefficiencies (FI).

[4 marks]

Q5 continued on next page …

(c) Consider the 3 node Markov chain graphical model in Fig. Q.5 (c).

Each node represents a binary variable. Calculate the marginal probability distribution for node 'a' by using variable elimination. Comment on the computational complexity.

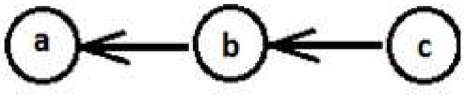


Fig. Q.5 (c) **[6 marks]**