PyJaipur Summer of Algorithm - ML Test 1 (EDA)

I. Which one is correct for missing data handling process?
   A. Dropping incomplete rows
   B. Dropping variables
   C. Value imputation
   D. All of the above

II. Which one is not related to data validation?
   A. Data type validation
   B. Range and Constraint validation
   C. data tagging
   D. Option 1 and 2

III. How will you display statistics of dataframe(df)?
   A. df.display_stat()
   B. df.describe()
   C. df.stats()
   D. None of the above

IV. Exploratory Data Analysis (EDA) is:
   A. The stage at which the data are described by the traditional measures of central tendency, spread and distribution shape
   B. Especially appropriate for nominal data
   C. A set of statistical methods specially designed for exploring a small, unruly data set and identifying any abnormalities in distribution or highly unusual scores
   D. Of limited value because no formal statistical tests are made

V. There is a rise in the average global temperature and the average global life expectancy. Can we comment about the relationship between the two?

A. They have absolutely no statistical relationship

B. Both the features are negatively correlated and one causes the other

C. Both the features are positively correlated but neither causes the other

D. None of the features are correlated but one causes the other

VI. For extracting columns "Salary" and "Age" from a data-set which of these command is correct?

A. x=df.iloc["Salary","Age"]

B. x=df.loc["Salary","Age]

C. x=df["Salary"]+df["Age"]

D. x = df["Salary"+"Age"]

Also, x=df[[“Salary, “Age”]], works fine.

VII. Which among mean and median is affected by outliers and why?

A. Median is affected by outliers because one big outlier can significantly change the position of the median

B. Mean is affected by outliers because it is the average of all values so it will be highly affected by the presence of outliers

C. Since outliers affect median, mean is also affected

D. None of the them are affected

VIII. How to deal with missing data?

A. Remove the data with missing values

B. Change the missing value to some default value

C. Replace it by mean of values

D. Leave it as it is to maintain the integrity of the dataset

Although, the answer provided by the mentor who made the question is c but don't take it as a rule. You will come across a lot of datasets where choosing other options will help the model learn better.

IX. Box plot finds the following in the data set.

A. Mode, Maximum, Minimum, Median, First quartile, Third Quartile
B. Mode, Maximum, Minimum
C. Maximum, Minimum, Median, First quartile, Third Quartile
D. Maximum, Minimum, Median, First quartile, Fourth Quartile, Second Quartile

X. Which of the following statements is true?

A. The standard deviation can take a negative value
B. The variance and standard deviation are always appropriate descriptive measures for any set of continuous or scale data
C. The variance and standard deviation are measures of spread or dispersion
D. If each of a set of scores is multiplied by a constant (say 2), the value of the standard deviation increases fourfold

XI. If a distribution is positively skewed, which of the following statements is true?

A. A square root transformation will make the distribution more skewed
B. A square transformation will make the distribution less positively skewed
C. A logarithmic transformation will tend to symmetrise the distribution
D. An antilogarithmic transformation such as exponentiation will reduce the positive skewness

XII. A competition in Kaggle is going on currently on blindness detection. Suppose that you participate in the contest, build a classification model and achieve a test accuracy of 97%. Should you be satisfied? Why/Why not?

A. Yes. A 97% accuracy is state-of-the-art in regard to blindness detection

B. No. Since these datasets are imbalanced, the model might learn to only predict the majority class (not blind) instead of minority class (blind)

C. Yes. If a model is able to learn to predict with 97% accuracy, then there is low chance for failure.

D. No. the model has overfit.

XIII. John has a test score of 64, which is at the 90th percentile. This means that:

A. 5% of people score lower than 64

B. 90% of people score lower than 64

C. 10% of people score lower than 64

D. 90% of people score higher than 64

XIV. What are the useful constraints to visualize three dimensions data?

A. Color

B. Size

C. Shape

D. All of the above

XV. How will you convert gender feature of dataframe(df) into numerical values?(where male:0 and female:1)

A. a = {'male':0,'female':1}
df['gender']=df['gender'].map(a)

B. a = {'male':0,'female':1}
df['gender'] = df['gender'].classify(a)

C. a = {'male':0,'female':1}

```
    df['gender'] = df['gender'].classify(a)
```

D.
```
a = {'male':0,'female':1}
df['gender'] = df['gender'].classify(a)
```