***Supervised machine learning classification: Customer Churn prediction***

***Quiz:***

1. **What are the dimensions of the dataset?**

9000 rows and 17 columns

9000 rows and 14 columns

10000 rows and 14 columns

900 columns and 10 rows

2. **How many null values are present in the dataset?**

10% of the dataset presents null values

1% of the dataset presents null values

5% of the dataset presents null values

All other answers are incorrect

3. **What kind of data is NumOfProducts?**

Int64

Int32

Float

String

Factor

4. **Are NumOfProducts and Age the same kind of data?**

Yes

No

5. **Which of the following statement is true?**

Credit score and Balance are continuous variables.

Credit score and Age are the only continuous variables in the dataset.

Gender is a discrete variable.

6. **Which countries are included in this dataset?**

Select the correct answer.

Germany, the USA, and Canada

Spain and Germany

France

France, Spain, and Germany

All other answers are incorrect

7. **Determine the quantile 25 and 75 of the variable age for men and women using groupby.**

Female 31 and 44; Male 33 and 42

Female 33 and 46; Male 30 and 44

Female 32 and 45; Male 32 and 43

Female 34 and 43; Male 31 and 41

8. **How do you calculate the mean, min, and max values as well as some percentiles (25th, 50th, or median and 75th) for a given dataset?**

Select the correct answer.

df.describe()

df.mean()

df.sum()

9. **Compute the number of people per country that has a credit card.**

*There could be more than just one correct answer.*

df[['Geography', 'HasCrCard']].groupby('HasCrCard').**sum**()

Copy

df[['HasCrCard']].groupby('Geography').**sum**()

Copy

df[['Geography','HasCrCard']].groupby('Geography').**sum**()

Copy

df[['Geography', 'HasCrCard']].groupby('Geography').count()

10. **What is the mean salary for men over 40 years old?**

100678

100451

149756

100183

100449

11. **Correlation**

Based on the correlation analysis of the dataset, which variable has the highest correlation with the target column?

HasCrCard

Age

Balance

12. **What is the correlation between Balance and the target?**

0.11

0.28

-0.05

13. **Is the credit score positively associated with the target?**

Yes

No

14. **What type of data is a scatter plot typically used to represent or analyze in data visualization and statistical analysis?**

A scatter plot uses dots to represent values for two different continuous variables.

A scatter plot uses lines to represent values for two different numeric variables.

A scatter plot uses bars to represent values for two different variables.

15. **Visualization**

What type of plot would you use to compare the credit score distribution for customers who have churned the bank versus those who have not? Please select the figure that shows the correct representation.

sns.boxplot(data=df,x='Exited',y='CreditScore')

sns.displot(data=df,x='Exited',y='CreditScore')

sns.scatterplot(data=df,x='Exited',y='CreditScore')

16. **What type of information can you gain from a box plot in statistical analysis and data visualization?**

*There could be more than just one correct answer.*

Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.

Box plots show the quartiles 5, 25, 75 and 95

They are built to provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers.

17. **What is the appropriate figure or chart to represent the number of classes for the 'Exited' variable?**

*There could be more than just one correct answer.*

sns.countplot(x="Exited", data=df)

sns.catplot(x="Exited", data=df, kind="count")

sns.catplot(x="Exited", data=df)

sns.barplot(x="Exited", data=df)

18. **Density plot**

Compared to overlapping histograms, overlapping density plots generally do not present the same issues, as the continuous density lines assist the viewer in distinguishing between the different distributions. This is because the smooth lines of the density plot allow for a more intuitive understanding of the shape of the data, even when multiple distributions are being presented simultaneously

Select the code that shows the balance distribution by country.

sns.displot(data=df, x="Balance", y="Geography", kind="kde")

sns.histplot(data=df, x="Balance", hue="Geography", kind="kde")

sns.displot(data=df, x="Balance", hue="Geography")

sns.displot(data=df, x="Balance", hue="Geography", kind="kde")

19. **Which of the following statements are true about normalization?**

Data normalization is used to restrict a certain range of values.

An important step before building the model is to normalize the features regardless of the model selected.

You would not be normalizing the whole datasets together.

20. **Encode categorical variables**

A machine learning algorithm needs to be able to understand the data it receives. There are plenty of methods to encode categorical variables into numeric and each method comes with its advantages and disadvantages. Which is the correct way to encode the variables gender and geography?:

Count Encoding

Label encoder

One hot encoder

20. **Classification or Regression**

Based on this, you should select wheater this scenario is a classification or a regression problem.

Classification

All other answers are incorrect

Regression

21. **Split train and test**

Select the correct way to split the dataset in 30% test and 70% train.

*There could be more than just one correct answer.*

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.70, random\_state=0)

X\_train, X\_val, Y\_train, Y\_val = train\_test\_split(X, Y, test\_size=0.30, random\_state=0)

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.20, random\_state=0)

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.30, random\_state=0)

22. **Confusion Matrix**

We ask you to build a predictive model that answers the question: “what sorts of people were more likely to commit churn?”.

Which of the following statements of the confusion matrix are true?

*There could be more than just one correct answer.*

False negative is the number of people that the model predicts will churn and do not.

True positive is the number of people that the model predicts will churn and do.

False positive is the number of people that the model predicts will not churn and do.

True negative is the number of people that the model predicts will not churn and do not.

23. **XGBoost**

Which of the following statements of the model are true about XGBoost? There could be more than just one correct answer.

XGBoost creates decision trees on randomly selected data samples, gets a prediction from each tree, and selects the best solution by voting

Normalized the features when using XGBoost is unnecessary

One of the key advantages of XGBoost is its ability to handle missing data and large datasets efficiently

Standardizing the dataset is always relevant

24. **XGBClassifier**

Train an XGBoost with the following parameters: objective='"binary:logistic" and random\_state=42 and calculated the accuracy for the training set.

The accuracy is 0.87

The accuracy is 0.05

The accuracy is 0.90