😊 Enjoy Math 506 Exam II 😊 Name: ID: Section:

Questions 1-16: True (A) of False (B)

Q1: Length, weight, and density are all examples of numerical type and ratio measurement scales.

Q2: Examples of measurement levels are categorical and numerical.

Q3: Variables arising from a counting process are called continuous.

Q4: Categorical data, where ordering is not important, are nominal.

Q5: The Second Quartile is always equal to the Median.

Q6: The most common measure of central tendency that is sensitive to extreme values is the Mean.

Q7: When a sample data has a normal distribution, then the measures of central tendency are all equal.

Q8: Measures of variation include Variance and Coefficient of Variation.

Q9: We cannot judge the data distribution from the *normalized* bar graph with response overlay.

Q10: The impact of a categorical predictor on the target is visualized through a *normalized* bar graph with response overlay.

Q11: Normalized histograms with response overlay is useful to bin numerical variables.

Q12: Before model evaluation, a data analyst should make that the test data set be balanced.

Q13: Data dredging is an essential step in the EDA phase.

Q14: Resampling is used for balancing the training set.

Q15: When the model is highly nonlinear, a data analyst may need 90% of the data for training the model.

Q16: Data science methodology does not follow the statistical inference approach.

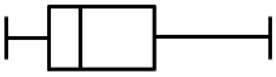
Q17: The measure that shows the variation relative to the mean is the

a) Standard Deviation b) Variance c) Mode d) Coefficient of Variation

Q18: If Median=200, Mode=200, and Mean=120, then the shape of distribution is

a) Symmetric b) Left-Skewed c) Right-Skewed d) Ordinal

Q19: The following Box Plot is



a) Left-Skewed b) Right-Skewed c) Symmetric d) Nominal

Q20: A normally distributed sample covers 99.7% within how many standard deviations from the mean?

a) 1.5 b) 2 c) 2.5 d) 3

Use the given table to answer Q21 to Q23.

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	**335**	**65**	**400**

Q21: Of invoices with errors, the portion of the Small Amount is

a) **30.77** b) **5.00** c) **10.53** d) **170**

Q22: The proportion of invoices with no errors is

a) **16.25** b) **50.75** c) **83.75** d) **335**

Q23: Of Medium Amount invoices, the proportion that has errors is

a) **71.43** b) **61.54** c) **10.00** d) **28.57**

Q24: Which piece of code do you include to normalize a Bar-Graph with Response Overlay?

a) **crosstab1=pd.crosstab(df['Age'], df['Respose'])**
b) **crosstab1=crosstab.div(crosstab.sum(1),axis=0)**
c) **crosstab1=pd.crosstab(df['Age'], df['Respose'],'normalize'=True)**
d) **crosstab1=crosstab.div(crosstab.sum(1),axis=1)**

Q25: The output of: plt.hist([Age_yes, Age_No], bins=10, stacked=True)

a) **a normalized histogram with response overlay**
b) **a non-normalized histogram with repose overlay**
c) **two histograms with response overlay**
d) **two normalized histograms with age overlay**

Q26: A training set has a binary target with 250-Yes records, and 700-No records. How many Yes-records do you resample in order to have the Yes-records 35% of the rebalanced training set.

a) **120** b) **124** c) **127** d) **136**

Q27: The Gini Index used in the CART method is

a) **strictly binary** b) **non-binary** c) **binary and non-binary** d) **Enropy**

Q28: $H(X) = -\Sigma_j \, p_j log_2(p_j)$ **is called the**

a) **Enropy of** X b) **Gini Index** c) **Information Gain** d) **Heaviside function**

Q29: The minimum value of $H(X) = -\Sigma_j\, p_j log_2(p_j)$ is achieved when values of p_j equal to

a) 1 or 2 b) 0 or 1 c) 0.5 or 2 d) 1 or infinity

Q30: Random Forests determines the final classification of a record by considering

a) Multiple trees b) More nodes c) More leaf nodes d) An optimal root node

Q31: A modeling technique that takes several models' output into account to arrive at a single answer is called

a) C5.0 b) Ensemble c) Random Forests d) CART

Q32: Which object sets the number of decision trees in Random Forests to be 20?

a) n_estimators=20 b) num_trees=20 c) trees_forests=20 d) criterion=20

Q33: Given: TP=30, FP=20, TN=40, and FN=10, then the accuracy of All_Positive_Model is

a) 30% b) 40% c) 50% d) 60%

Q34: Given: FP=10, FN=20, TAP=60, and TAN=40, then Specificity×Sensitivity =

a) 75% b) 66.7% c) 50% d) 24.6%

Q35: Use the data-driven cost matrix to calculate the profit per record.

a) 3$ b) 2.8$ c) 1.1$ d) 0.9$

TN=40; cost=0$	FP=30; cost=3$
FN=20; cost=0$	TP=10; profit=20$

Wish you a good luck 😊