REVA Academy for Corporate Excellence

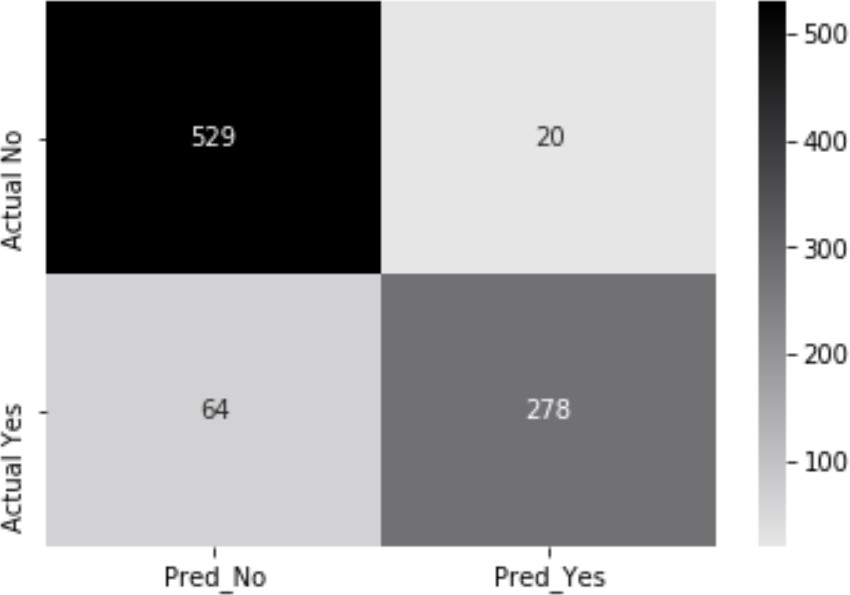
PGDM/MBA in Business Analytics - Batch 06

Trimester I Examination

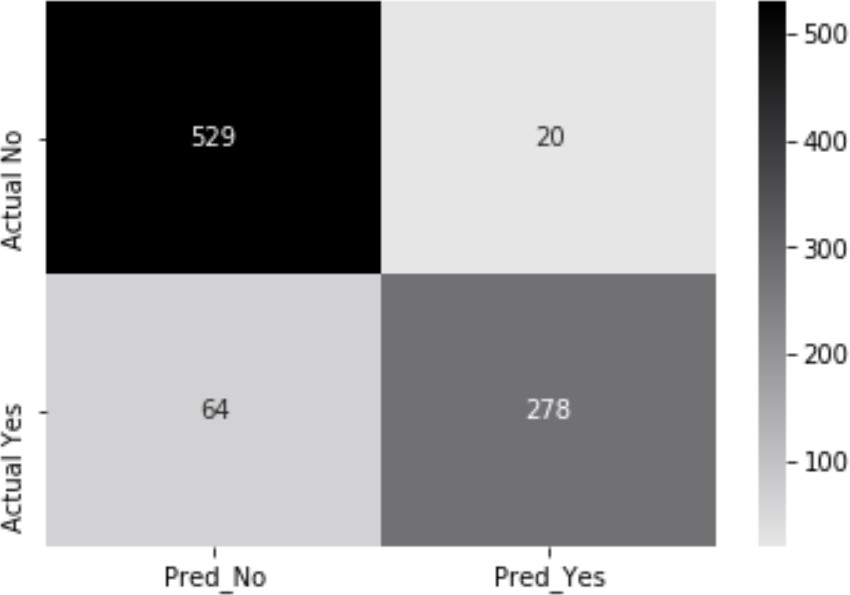
|  |  |  |  |
| --- | --- | --- | --- |
| **Module Name** | Machine Learning | | |
| **Module Code** | PGDBAP19130/ MBA19130 | | |
| **Date** | 22-02-2020 | **Time** | 90 |
| **Exam Type** | Open Book | **Marks** | 50 |
| **Instructions** | Section 1 - Multiple Choice Question.  Section 2 - Case Study. | | |

**Section – A (13 x 2 = 26 Marks)**

1. Which of the following sentence is FALSE regarding regression?
2. It relates inputs to outputs
3. It is used for prediction
4. It may be used for interpretation
5. It discovers causal relationships
6. Which statement about outliers is true?
7. Outliers should be identified and removed from a dataset
8. Outliers should be part of the training dataset but should not be present in the test data
9. Outliers should be part of the test dataset but should not be present in the training data
10. The nature of the problem determines how outliers are used
11. Which of the following is a disadvantage of decision trees?
12. Decision trees are parametric model
13. Decision trees are robust to outliers
14. Decision trees are prone to be over fit
15. None of the above
16. When performing regression or classification, which of the following is the correct way to pre-process the data?
17. Normalize the data → PCA → training
18. PCA → normalize PCA output → training
19. Normalize the data → PCA → normalize PCA output → training
20. None of the above
21. What is the Precision from the below table for class 'Yes'
22. 278/(278+529)
23. 64/(278+64)
24. 278/(278+20)
25. 64/(64+20)



1. What is the Recall from the below table for class 'No'
2. 278/(278+529)
3. 529/(278+64)
4. 278/(278+20)
5. 529/(529+20)

Confusion Matrix

1. In the CRISP-DM framework, what does the DM stand for?
2. Data Mining
3. Data Management
4. Decision Mining
5. Decision Management
6. Association Rule- On a Sunday night in a store, there were a total of 1000 transaction. Out of these 1000 transactions, in 200 transactions Diaper was bought. From the 200 transaction in which Diaper was bought, in 50 transactions Beer was also bought. What is the SUPPORT for buying both Beer and Diaper?
7. 50/200
8. 50/1000
9. 200/1000
10. (50/200)/1000
11. Association Rule- On a Sunday night in a store, there were a total of 1000 transaction. Out of these 1000 transactions, in 200 transactions Diaper was bought. From the 200 transaction in which Diaper was bought, in 50 transactions Beer was also bought. What is the CONFIDENCE for rule “IF buy diapers, THEN buy beer”?
12. 50/200
13. 50/1000
14. 200/1000
15. (50/200)/1000
16. In data pre-processing for a Categorical variable, missing values can be replaced with?
17. Median
18. Mode
19. Mean
20. Either one of the above
21. Which of the following can't be used as a metric for checking spread (variance) of a continuous numerical variable.
22. Range
23. Standard deviation
24. Inter-quartile range
25. None of the above
26. What may not be used as a valid criteria for selecting the right number of components in PCA.
27. The Eigen value should be more than 'one' for a component
28. The Proportion of Variance Explained Criterion
29. The R square should be more than 0.7
30. None of the above
31. Which of the following is an assumption for building a multiple linear regression model.
32. Features are independent from each other
33. Error variance is constant
34. Error is normally distributed
35. All of the above

**Section – B (24 Marks)**

**Using the Breast Cancer data provided in the csv file, you need to build a prediction model for "Propensity to Develop Breast Cancer". In specific, you need to do the following tasks.**

1. What is the modelling technique you chose and why? **[ 2 Marks]**
2. What pre- processing you considered on this data and why? **[ 2 Marks ]**
3. Submit detailed Model output in the HTML format along with all the data cleaning and pre-processing steps. **[ 10 Marks]**
4. What metrics you have computed to evaluate model performance and why? How is the model performance using these metrics? What are the most important features for prediction of Breast Cancer?  **[10 Marks]**

****

---- End of paper ----