Birla Institute of Technology and Science, Pilani

Comprehensive examination May 2023 - Question paper

Course name: Predictive Analytics
Time: 3 PM – 6 PM (3 hours)

Course code: MPBA G513
Total marks: 35

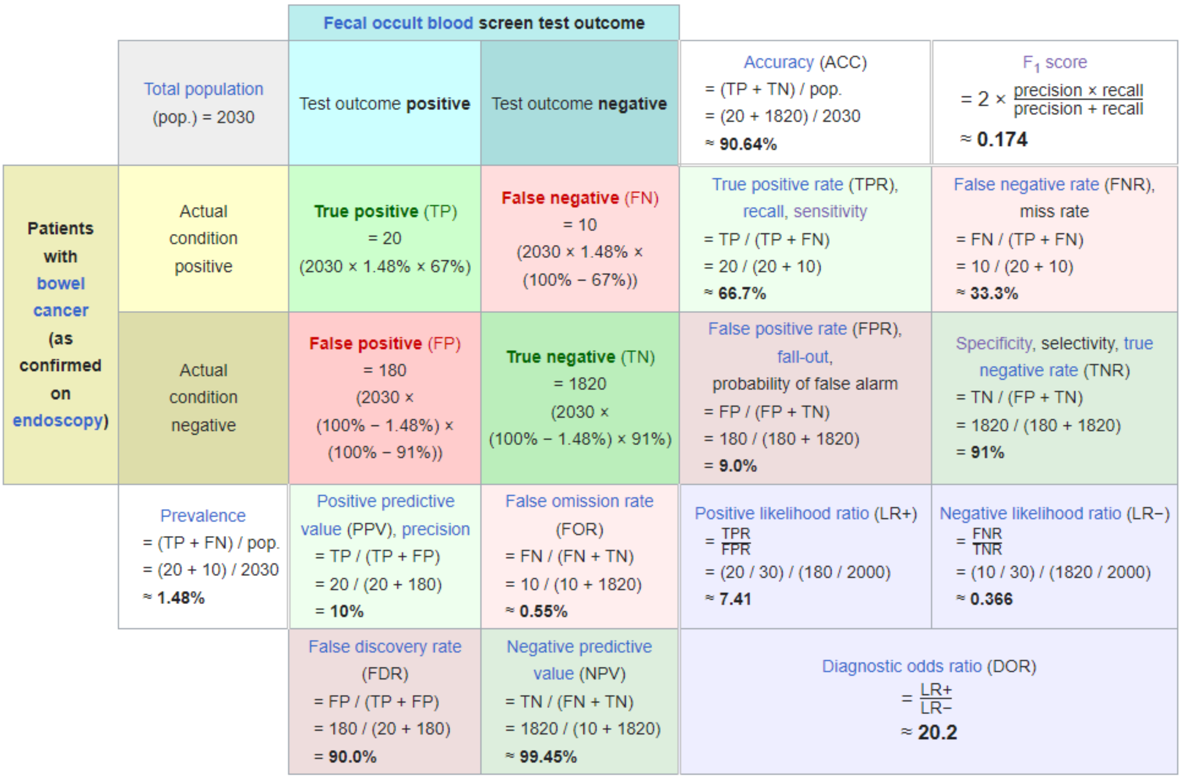
Note: Attempt all the questions

1. True. ROC is plotted between TPR and FPR. TPR is known as 'Recall; and FPR is calculated as 1-Specificity. 1

2. F1-score is harmonic mean of 'Precision (PPV)' & 'Recall or Sensitivity (TPR)'. 1

$$F_1\,\text{score} = \frac{2\,\text{PPV}\times\text{TPR}}{\text{PPV}+\text{TPR}}$$

3. 4



		Fecal occult blood screen test outcome			
	Total population (pop.) = 2030	Test outcome **positive**	Test outcome **negative**	Accuracy (ACC) = (TP + TN) / pop. = (20 + 1820) / 2030 ≈ **90.64%**	F_1 score $= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ ≈ **0.174**
Patients with bowel cancer (as confirmed on endoscopy)	Actual condition positive	**True positive** (TP) = 20 (2030 × 1.48% × 67%)	**False negative** (FN) = 10 (2030 × 1.48% × (100% − 67%))	True positive rate (TPR), recall, sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **66.7%**	False negative rate (FNR), miss rate = FN / (TP + FN) = 10 / (20 + 10) ≈ **33.3%**
	Actual condition negative	**False positive** (FP) = 180 (2030 × (100% − 1.48%) × (100% − 91%))	**True negative** (TN) = 1820 (2030 × (100% − 1.48%) × 91%)	False positive rate (FPR), fall-out, probability of false alarm = FP / (FP + TN) = 180 / (180 + 1820) = **9.0%**	Specificity, selectivity, true negative rate (TNR) = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**
	Prevalence = (TP + FN) / pop. = (20 + 10) / 2030 ≈ **1.48%**	Positive predictive value (PPV), precision = TP / (TP + FP) = 20 / (20 + 180) = **10%**	False omission rate (FOR) = FN / (FN + TN) = 10 / (10 + 1820) ≈ **0.55%**	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$ = (20 / 30) / (180 / 2000) ≈ **7.41**	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$ = (10 / 30) / (1820 / 2000) ≈ **0.366**
		False discovery rate (FDR) = FP / (TP + FP) = 180 / (20 + 180) = **90.0%**	Negative predictive value (NPV) = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.45%**	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ ≈ **20.2**	

DOR = 20.2
MCC = 0.23

4. The K-Means clustering terminates with identification of clusters in a dataset, where each cluster creates a decision boundary having the cluster's centroid. This is often represented as Voronoi diagram with Voronoi cells, each Voronoi cell represents the subregion of the predictor space, the nucleus of the Voronoi cell represents the centroid. 2
Read more: https://en.wikipedia.org/wiki/K-means_clustering

5. Dendrogram 1

6. True. Learn more: 1

https://stats.stackexchange.com/questions/269428/reverse-the-sign-of-pca

7. False. Hierarchical clustering using complete linkage measures the maximal intercluster dissimilarity, whereas single linkage measures the minimal intercluster dissimilarity. 1

8. False. There is no such difference in terms of number of splitting of the dendrogram's internal nodes. 1

9. False. Maximal Margin Classifier technique is not used commonly as it is extremely sensitive to individual data point and is a predecessor of Support Vector Classifier. 1

10. A predictor space is the total area which is subdivided into several sub-regions as per the splitting rules defined by decision trees. https://medium.com/all-about-ml/decision-trees-796395a2d37b 1

11. The sign of f(x) predicts the class label (A/B), whereas the magnitude provides the purpose of confidence with which a data point can be assigned to a class. 1

12. False. Maximal Margin Classifier (MMC) uses hard margin hyperplane as compared to Support Vector Classifier (SVC - uses soft margin). This is the reason why MMC is extremely sensitive to newly added data points as it impacts the hyperplane and therefore affects its prediction. 2

    MMC's biggest drawback is its hard margin hyperplane, as it does not allow a single data point to be on the wrong/opposite side of the hyperplane and therefore requires the dataset to be perfectly linearly separable. A soft margin is a feature of Support Vector Classifier (SVC) which allows a little bit of tolerance of a few data points to be on the other side of the hyperplane and therefore makes the SVC more robust as compared to MMC

13. C (Classification margin ~ tolerance) and gamma are two hyperparameters. C controls the softness of margin and gamma controls the non-linearity of the kernels for support vector machines. 1

14. LDA considers all the data points for prediction whereas SVM only considers support vectors (data points falling on the separating hyperplane with margin) 1

15. False. A SVM with a polynomial kernel with degree d = 1 reduces itself to become a Support Vector Classifier which only works with linearly separable datasets. 1

16. LabelEncoder() function is available in Python's sklearn module and is used to convert string variables into numeric values. It uses alphabetical sorting for conversion, with that logic 'Churner' becomes 0 and ;Non-churners' become 1 as their numeric representation. 2

17. False. Decision trees generally outperform when decision boundary is non-linear. https://medium.com/all-about-ml/decision-trees-796395a2d37b 1

18. Bootstrap aggregation or Bagging is a method to take repeated sampling 2

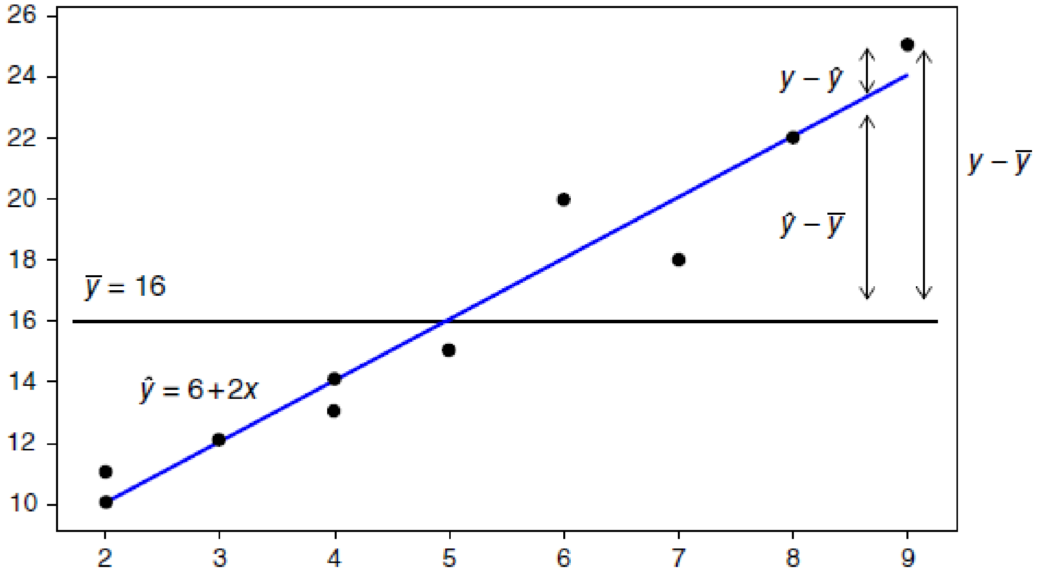
from a dataset to create several sample from the limited dataset while allowing replacement (same data point can be taken several times). While Bagging is performed with all the predictors, Random Forests only consider k predictors out of total n predictors to avoid growing correlated trees; therefore RF prefer small but more number of trees.

19. False. Association rule mining can be used even when target label is available for training or not, hence it can be used for supervised/unsupervised learning. 1

20. WCSS is Within Cluster Sum of Squares score, a metric to measure the sum of the distance between each point with its cluster center. It is used along with elbow method to obtain the optimum number of clusters in a given dataset. 1

21. y = actual value 2
ŷ = predicted value
ȳ = average value

SSE = Sum of Squared Error
SSR = Sum of Squares Regression
MSE = Mean Squared Error
MSR = Mean Squared Regression

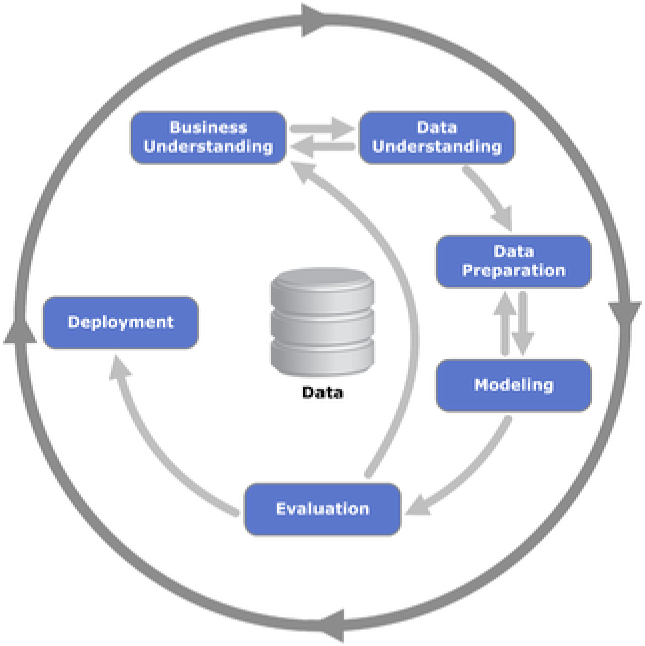


SST = SSR + SSE

$$\text{SST} = \sum_{i=1}^{n} (y - \bar{y})^2 \qquad \text{SSR} = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 \qquad \text{SSE} = \sum (y - \hat{y})^2$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

22. CRISP-DM = Cross-Industry Standard Process for Data Mining. 1



23. Bayes Theorem 2

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Naive Bayes is called naive because it assumes that each input variable is independent having no correlation with other variables.

24. random_state=31 will perform shuffling of data points but with the seed value to be 31 (same every time), providing the same indices for the data points for modelling and testing purposes. 1

25. A Q-Q plot is known as the Quantile-Quantile plot and is used to test for the normality of the distribution. A straight line in a Q-Q plot proves that the distribution is normal 1

26. Z-scaling is performed using Python’s StandardScaler() function available in scikit-learn module. 1