TECHNO MAIN SALT LAKE

(FORMERLY TECHNO INDIA, SALT LAKE)

Name. Ishaan Singh Mandla

Roll No. 13030822004 Stream. CSE (AI & ML)

Subject Application of Machine Learning in Industries (PCC-AIML601) Semester. 6th

Invigilator's Signature .... Date........................

Part A.

Ans 1) → Two most common supervised tasks:

(i) Classification

(ii) Regression.

Ans 2) → Validation set is used to check the accuracy of the trained model.

Ans 3) →

Ans 4) → AUC value of a perfect classifier is 1.

Ans 5) → Recall is more important for spam email detection, because, it gives the ratio of True Positive over True Positive + False Negative. Meaning Predicted true values over actual true values.

Part B

Ans 6) → train-test-split is a function used to split the dataset into training set and testing set. It is called in python from sklearn.model_selection. It can use 4 parameters, data, target, test_size and random_state.

~~Over~~ ~~when~~ When complex data is used for training set, the model performes well in training set but poorly on test set, this condition is known as overfitting.

When less data ~~of~~ or not accurate amount of data is used for training a model, and ~~a~~ model performs poorly in training set as well as testing set, this is known as underfitting.

Overfitting can be prevented by cleaning the data and handling the missing values before training. Underfitting can be prevented by using a synthetic data along ~~with~~ with the actual data for training the model.

Ans 7)→ when the prediction of model is bent to one side say the positive side, then the model is bias. Because of this bias, the accuracy of the model is reduced.

when there is a huge ~~deflectio~~ deflection in the output for slight change in the input, this is known as variance.

When the bias and variance both are very close to each other, this is known as bias-variance trade-off.

Bias and variance can be reduced by cleaning and preprocessing the ~~d~~ ~~t~~ dataset before training the model. This helps reduce the bias and variance of the model and helps improve accuracy and performance.

Ans 9)→ Confusion Matrix is used to display the accuracy of the model by showing the number of true positives, true negatives, false positives, false negatives values of the predicted outcome. It is an important function of machine learning. as it helps to understand the accuracy ~~of the~~ and performance of the trained model. It displays the above data in a metrics format.

$fn = 5$, $fp = 3$, $tp = 10$, $tn = 82$

$$\text{precision} = \frac{tp}{tp+fp} = \frac{10}{10+3} = \boxed{\frac{10}{13}}$$

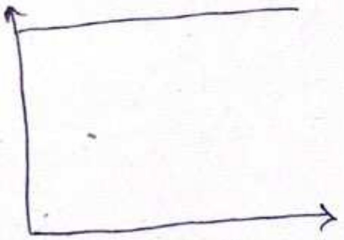
$$\text{Recall} = \frac{tp}{tp+fn} = \frac{10}{10+5} = \frac{10}{15} = \boxed{\frac{2}{3}}$$

$$\text{false negative rate} = \frac{fn}{fn+tn} = \frac{5}{5+82} = \boxed{\frac{5}{87}}$$

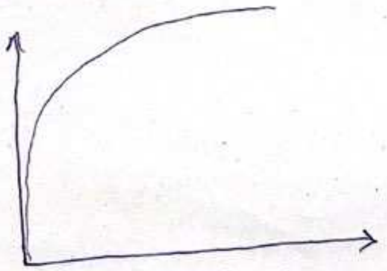
~~false~~ ~~negative rate = fp~~

$$\text{false positive rate} = \frac{fp}{fp+tp} = \frac{3}{3+10} = \boxed{\frac{3}{13}}$$

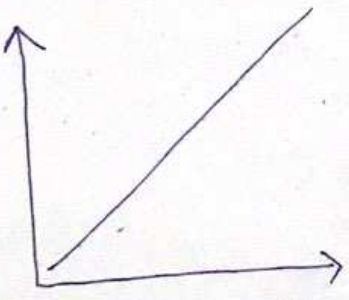
Ans 10) ⇒ AUC stands for Area under the curve. It is ~~calcu~~ calculated to measure the performance of the model. Roc is ~~also~~ also calculated to measure the performance of the model.



Perfect classifier



practical classifier



accurate classifier