Date _/_/_

Assignment

Answer:1 Machine learning is different from traditional programming in that traditional programming relies on explicitly programmed instructions to perform a task, while machine learning involves training a model on data to make decisions or predictions without explicit instructions. Some applications of machine learning include image and speech recognition, natural language processing and predictive modeling.

With respect to the term task, experience and performance:

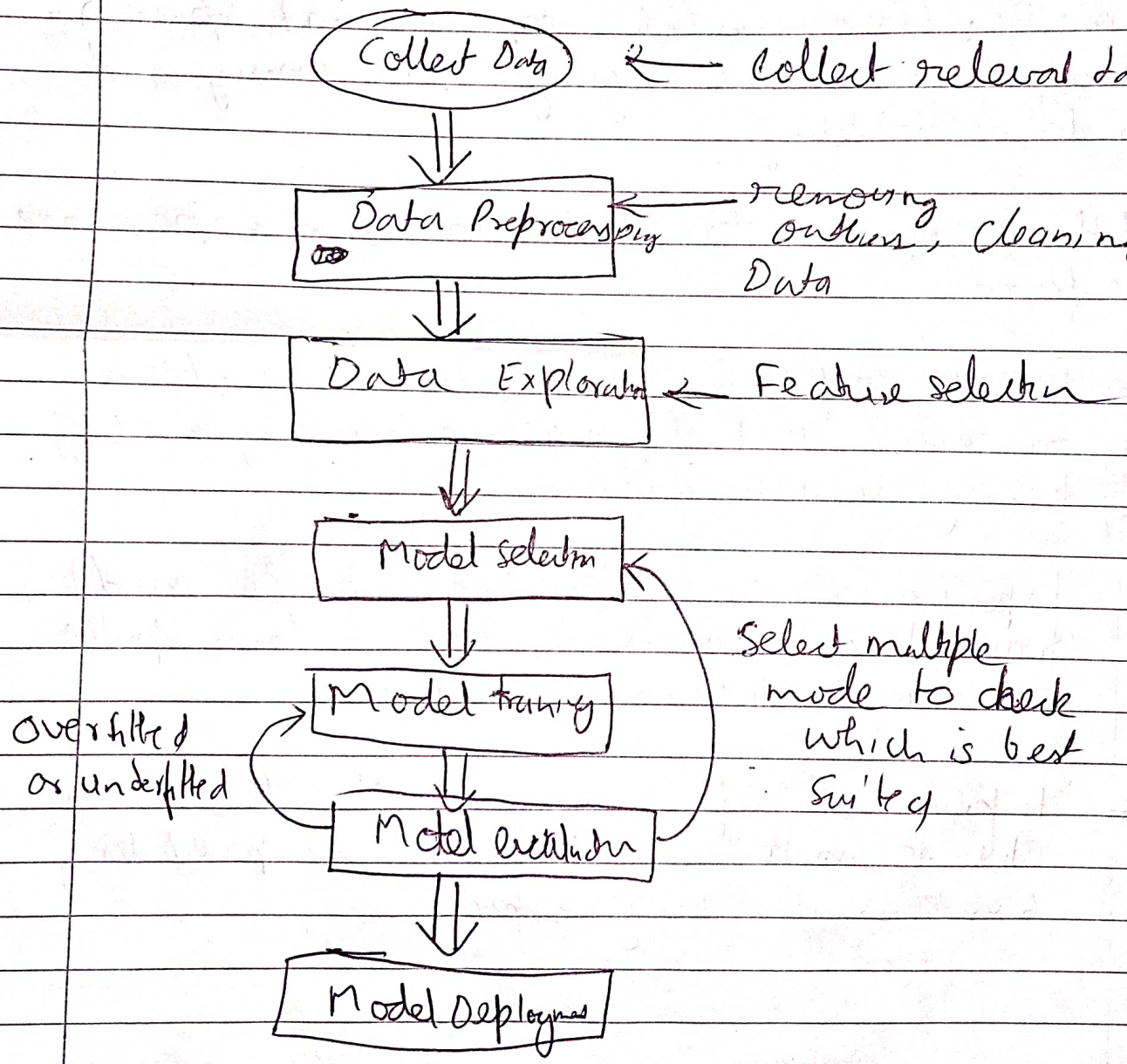
Task refers to the specific problem or decision that the machine learning model is being used to address.

Experience refers to the data that the model is trained on, which is used to learn patterns and make decisions or predictions.

Performance refers to how well the model is able to make accurate decisions or predictions based on its training experience.

Date

Answer:2 (I) Data Collection –
(II) Data preprocessing
(III) Data Exploration
(IV) Model Selection
(V) Model training
(VI) Model Evaluation
(VII) Model Deployment



Date _/_/_

Answer:3 There are mainly three types of machine learning approaches:

1) Supervised learning: In this type of learning, the model is trained on labeled data, where the correct output is already known. The model makes predictions based on this labeled data.
e.g — regression, decision tree, and random forest

2) Unsupervised Learning: In this type of learning, the model is not given any labeled data and instead has to find patterns and structure in data by itself.
Examples of unsupervised learning clustering, dimensionality reduction and anomaly detection

3) Reinforcement learning:
In this type of learning, the model learns to make decision through trial-and-error by interacting with an environment. The model receives rewards or penalties based on its actions and learns to optimize its decision-making over time.
example - Q-Learning and SARSA.

Types of supervised learning:
(i) Regression — Continuous target variable
(ii) Classification — Categorical target variable

Date _ / _ /

Answer: In machine learning, it is important to evaluate the performance of a model on unseen data to ensure that it generalizes well to new cases. This is typically done by splitting the available data into three sets: training, testing and validation data.

1. Training Data — This is the dat used to train the model. The model's parameters are adjusted during the training process to minimize the error on this data. The model learns to make predictions based on the patterns present in the training data.

2. Testing Data: This is set of data that is used to evaluate the performance of the model. The model is tested on this data after it has been trained. The test data is typically used to estimate the generalization error of the model. Which is an indicator of how well the model will perform on new, unseen data.

3. Validation Data: This is set of Data that is used to tune the model's hyperparameters. Hyperparameters are the parameters of the model that are not learned during the training process. Such as the number of hidden layers in a neural network, the learning rate, etc. Validation data is used to evaluate the performance of the model with different hyperparameter settings and the best performing setting are chosen before the model is tested on the test data

Date _/_/_

Answer:5 For classification problem :

① Accuracy : This is the proportion of correctly classified instance out of all instances it is a good metrices for balanced datasets, but it can be misleading for imbalanced datasets

② Precision : Precision is the proportion of true positive predictions out of all positive predictions. It is a good metric for problems where false positive are more costly than false positive.

③ Recall : Recall is the proportion of true positive predictions out of all positive instances. It is a good metric for problems where false negatives are more costly than false positive

④ F1-Score — f_1-score is a harmonic mean of precision and recall. It is used when you want balance precision and recall.

⑤ AUC-ROC curve — Area Under the Receiver Operating Characteristic curve is a graphical representation of the performance of a binary classifier. It plots the true positive rate against false positive rate. Area under curve is used to measure overall performance of the classifier

Date _ / _

for regression —

1) Mean Absolute Error (MAE): This is the average between the predicted and actual values.

2) Mean Squared Error: This is the average of the squared difference between the predicted and actual values.

3) R-Squared: R-Squared is a statistical measure that represents the proportion of the variance for a that's explained by an independent variable(s) in a regression model

4) Root mean squared error: This is the square root of the mean squared error. It has same units as the original data and is commonly used to compare different models or method.

Answer: 6 K-fold technique: is the procedure in which we take specific value of K and making K division of that data. Cross validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data.

Date _/_/_

Answer:7 A confusion matrix is a table used to evaluate the performance of a classification model. It displays the number of correct and incorrect predictions made by the model. The matrix provide four outcom. true positive (TP), False positive (FP), true Negative (TN) and false negative (FN), which is used to calculate various metrics such as accuracy, precision, recall and F1 score.

Predicted	Actual: has	Actual: NO
has	TP	FP
Not	FN	TN

=> Performance of the model total 40 balls
=> 30 balls are span

Predicted	Actual: span	Actual: not span
span	25	6
not span	5	4

i) => Classification (Supervised learning)

ii) Answer

iii) Accuracy $= \frac{25+9}{40} = \frac{29}{40} = \frac{TP+TN}{TP+TN+FP+FN}$

Precision $= \frac{TP}{\text{Total positive Predicted}} = \frac{TP}{TP+FP} = \frac{25}{25+6} = \frac{25}{31}$

Recall $= \frac{TP}{\text{Actual Positive}} = \frac{25}{TP+FN} = \frac{25}{30}$

Date _/_/__

$$F_1 \text{ Score} = \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$= \frac{\text{Precision} * \text{Recall}}{2}$$

$$= \frac{\text{Precision} * \text{Recall}}{2 \text{Precision} \times \text{Recall}}$$

$$= \frac{\frac{25}{31} + \frac{25}{30}}{2 \times \frac{25}{31} \times \frac{25}{30}}$$

$$F_1 - \text{Score} = \frac{25 \times 30 + 31 \times 25}{2 \times 25 \times 25}$$

(v) Because In that case to detecting spam is nice but detecting not spam is worst so. Might be data is imbalanced so that we cannot always rely on accuracy.

Answer: Overfitting occurs when when a machine learning model is too complex and learns the training data too well, resulting is poor generalization performance on unseen data.

Underfitting occurs when a machine learning model is too simple and does not capture the underlying patterns in the data, resulting in poor performance on both training and testing.