ML

\*\*ML Fundamentals\*\*

\*\*Answer the following:\*\*

1. What is Machine Learning? State applications of Machine Learning.

2. Explain the importance of data in the context of Machine Learning.

3. What is Normal Distribution?

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*State the difference between:\*\*

1. Supervised and Unsupervised Learning.

2. Parameterized and Non-parameterized Models.

3. Continuous and Categorical Variables.

4. Prediction and Estimation.

\*\*Answer the following:\*\*

1. Explain Regression and Classification.

2. What do you mean by Training Data and Testing Data?

---

# EDA

\*\*Answer the following:\*\*

1. Why do we need to do scaling?

2. What is the Standard Scaler?

3. Write short notes on skewness and its types and how to detect and reduce it.

4. How will you extract the categorical columns from the set of data frame columns?

---

\*\*Answer the following:\*\*

1. Write a lambda function to convert the string values of a column to numbers.

2. Explain different types of encoding.

---

\*\*Perform the EDA Processing on the dataset\*\* [customers.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/EDA/customers.csv).

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and corresponding output screenshots.

---

### Linear Regression

\*\*Answer the following:\*\*

1. \*\*Define Mean Square Error (MSE)\*\* as a metric for evaluating the performance of regression models.

2. \*\*How does MAE differ from MSE\*\* in terms of measuring the average magnitude of errors between predicted and actual values?

3. \*\*Define the R² Score (Coefficient of Determination)\*\* and its role in evaluating the goodness-of-fit of regression models.

4. \*\*Explain how L1 and L2 regularization methods\*\* help prevent overfitting by adding regularization terms to the cost function.

\*\*Note:\*\* Upload Assignment in PDF format.

---

### Tasks with the Mediclaim Dataset

Refer to the data given at the link with the name \*Mediclaim\* and perform the following tasks (metadata is also provided for your help):

1. \*\*Perform feature engineering on the data.\*\*

2. \*\*Extract X and y values.\*\*

3. \*\*Use OLS to fit the data.\*\*

4. \*\*Analyze the results.\*\*

\*\*Dataset Link:\*\* [Mediclaim Dataset](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/Linear+Regression/mediclaim.csv)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and output screenshots.

---

### Additional Tasks

Refer to the data given at the link with the name \*Mediclaim\* and perform the following tasks:

1. \*\*Perform feature engineering on the data.\*\*

2. \*\*Extract X and y values.\*\*

3. \*\*Use Linear Regression to fit the model.\*\*

4. \*\*Evaluate linear regression metrics.\*\*

5. \*\*Implement Ridge and Lasso Regularization.\*\*

6. \*\*Compare the results with the model created in step 3.\*\*

\*\*Dataset Link:\*\* [Mediclaim Dataset](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/Linear+Regression/mediclaim.csv)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and output screenshots.

---

Sure! Here’s a structured overview of your assignment on Logistic Regression:

---

### Logistic Regression

\*\*Answer the following:\*\*

1. \*\*Define a confusion matrix\*\* and discuss its components in the context of binary classification.

2. \*\*Describe how true positives, true negatives, false positives, and false negatives\*\* are represented in a confusion matrix and their implications for model evaluation.

3. \*\*Define precision and recall\*\* as evaluation metrics for classification models, and explain their relevance in different application scenarios.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. \*\*Describe the ROC curve\*\* and its significance in assessing the performance of binary classification models.

2. \*\*Define overfitting and underfitting\*\* and explain how they occur in logistic regression models.

\*\*Note:\*\* Upload Assignment in PDF format.

---

### Application of F1 Score and ROC Analysis

Given a dataset with the name \*customers.csv\*, explain how you would use the \*\*F1 score\*\* and \*\*ROC analysis\*\* to assess the performance of a logistic regression model. Discuss potential challenges and limitations in model evaluation due to class imbalance.

\*\*Dataset Link:\*\* [customers.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/Logistic+Regression/customers.csv)

---

Sure! Here’s a structured overview for your Decision Tree assignment:

---

### Decision Tree Assignment

\*\*Answer the following:\*\*

1. \*\*Explain the fundamental concept of decision trees\*\* in machine learning and how they are used for classification and regression tasks.

2. \*\*Discuss the advantages and disadvantages of decision trees\*\* compared to other machine learning algorithms, such as logistic regression.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. \*\*Define entropy\*\* as a measure of impurity in decision tree nodes and explain how it is calculated for binary and multi-class classification problems.

2. \*\*Define the Gini index\*\* as an alternative measure of impurity in decision trees and compare it to entropy in terms of computational complexity and sensitivity to class imbalance.

\*\*Note:\*\* Upload the assignment in PDF format.

---

\*\*Answer the following:\*\*

1. \*\*Discuss the concept of pruning\*\* in decision trees and its role in preventing overfitting.

2. \*\*Explain the CART (Classification and Regression Trees) algorithm\*\* as a popular method for constructing decision trees, and outline its key steps.

\*\*Note:\*\* Upload Assignment in PDF format.

---

### Application of Decision Trees

Given a dataset named \*credit.csv\* with imbalanced class distributions and a high-dimensional feature space, discuss the challenges and considerations in using decision trees for classification. Propose strategies for mitigating the impact of class imbalance and feature selection to improve model robustness and generalization performance.

\*\*Dataset Link:\*\* [credit.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/Decision+Tree/credit.csv)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and corresponding output screenshots.

---

Sure! Here’s the text with the dataset link included:

---

Random Forest Assignment

\*\*Answer the following:\*\*

1. Explain the concept of Random Forest in machine learning and how it extends the idea of decision trees.

2. Discuss the advantages of Random Forest over individual decision trees in terms of reducing overfitting and improving predictive performance.

3. Define bootstrap sampling and explain how it is used in constructing individual decision trees within a Random Forest.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. Define feature bagging (or random feature selection) and explain how it is employed in Random Forest to further enhance model diversity.

2. Explain the working principle of AdaBoost and how it adjusts the weights of misclassified data points to focus on difficult-to-classify instances.

\*\*Note:\*\* Upload assignment in PDF format.

---

\*\*Answer the following:\*\*

1. Explain the fundamental differences between bagging and boosting ensemble techniques in machine learning.

2. Define common performance metrics used for evaluating classification models, such as accuracy, precision, recall, and F1-score.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Application of Random Forest\*\*

Given a dataset named \*Customer.csv\* with a mix of numerical and categorical features, discuss the challenges and considerations in using Random Forest for classification or regression tasks.

- Propose strategies for encoding categorical variables and handling missing data to improve model performance and interpretability.

\*\*Dataset Link:\*\* [Customer.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/EDA/customers.csv)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and output screenshots.

---

Sure! Here’s the text with the dataset link included:

---

Random Forest Assignment

\*\*Answer the following:\*\*

1. Explain the concept of Random Forest in machine learning and how it extends the idea of decision trees.

2. Discuss the advantages of Random Forest over individual decision trees in terms of reducing overfitting and improving predictive performance.

3. Define bootstrap sampling and explain how it is used in constructing individual decision trees within a Random Forest.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. Define feature bagging (or random feature selection) and explain how it is employed in Random Forest to further enhance model diversity.

2. Explain the working principle of AdaBoost and how it adjusts the weights of misclassified data points to focus on difficult-to-classify instances.

\*\*Note:\*\* Upload assignment in PDF format.

---

\*\*Answer the following:\*\*

1. Explain the fundamental differences between bagging and boosting ensemble techniques in machine learning.

2. Define common performance metrics used for evaluating classification models, such as accuracy, precision, recall, and F1-score.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Application of Random Forest\*\*

Given a dataset named \*Customer.csv\* with a mix of numerical and categorical features, discuss the challenges and considerations in using Random Forest for classification or regression tasks.

- Propose strategies for encoding categorical variables and handling missing data to improve model performance and interpretability.

\*\*Dataset Link:\*\* [Customer.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/EDA/customers.csv)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and output screenshots.

---

Sure! Here’s the text without HTML tags:

---

Ensemble Techniques Assignment

\*\*Answer the following:\*\*

1. Explain the concept of Gradient Boosting.

2. What is Bootstrapping?

3. Does Gradient Boosting cause Overfitting?

\*\*Note:\*\* Upload assignment in PDF format.

---

\*\*Answer the following:\*\*

1. What is a Voting Classifier?

2. Why are ensemble methods so popular?

3. What is a stump?

\*\*Note:\*\* Upload assignment in PDF format.

---

\*\*Attempt the following:\*\*

Use a Voting classifier to estimate the working of Logistic Regression, Decision Tree, KNN classifier, and also estimate the accuracy.

You can use \*customers.csv\* data or you are free to download data from Kaggle for estimation. For \*customers.csv\*, you are provided with metadata; please read it before doing implementation.

\*\*Dataset:\*\* [customers.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/EDA/customers.csv)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and corresponding output screenshots.

---

Sure! Here’s the text without HTML tags:

---

K-Nearest Neighbors (K-NN) Assignment

\*\*Answer the following:\*\*

1. State the difference between Parametric and Non-Parametric Models.

2. What is the elbow method?

3. How do we determine or decide the value of k?

4. In KNN, how does the choice of K impact the bias-variance trade-off?

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. What is the main drawback of the KNN algorithm?

2. In KNN, how is the class of a new data point determined for classification tasks?

3. What is the primary disadvantage of using a small value of K in KNN?

\*\*Note:\*\* Upload assignment in PDF format.

---

\*\*Attempt the following:\*\*

Take the classified data, such as \*customers.csv\*, and implement the KNN algorithm. You are provided with metadata; please read it before starting the implementation of the algorithm.

\*\*Dataset:\*\* [customers.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/EDA/customers.csv)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshots and corresponding output screenshots.

---

Sure! Here’s the text without HTML tags:

---

Support Vector Machines (SVM) Assignment

\*\*Answer the following:\*\*

1. Explain the SVM in detail.

2. Name some advantages of SVM.

3. What are Support Vectors in SVMs?

4. Compare K-Nearest Neighbours (KNN) and SVM.

\*\*Note:\*\* Upload assignment in PDF format.

---

\*\*Answer the following:\*\*

1. What are Radial Basis Function Kernels?

2. How to select Kernel for SVM?

\*\*Note:\*\* Upload Assignment in PDF format.

---

Here’s the text without HTML tags:

---

Naive Bayes Assignment

\*\*Answer the following:\*\*

1. What is the basic assumption of Naive Bayes Classifier?

2. What are the advantages of choosing Naive Bayes Classifier?

3. What are the disadvantages of choosing Naive Bayes Classifier?

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. What is the impact of outlier in relation to NB classifier?

2. What are the application areas of NB classifier?

3. How can we use categorical values in NB classifier?

\*\*Note:\*\* Upload Assignment in PDF format.

---

Here's the text without HTML tags:

---

PCA Assignment

\*\*Answer the following:\*\*

1. What is curse of dimensionality?

2. Explain the concept of orthogonality.

3. What are eigenvalues and eigenvectors?

\*\*Note:\*\* Please upload assignment in PDF format.

---

\*\*Answer the following:\*\*

1. Define PCA in brief.

2. Why do we need dimensionality reduction? What are its drawbacks?

3. Explain the limitations of PCA.

4. Should one remove highly correlated variables before doing PCA?

\*\*Note:\*\* Upload Assignment.

---

\*\*Attempt the following:\*\*

Apply PCA on \*\*heart\_disease.csv\*\* for implementing binary classification. Please refer to the metadata of heart\_disease data before implementation.

Dataset: [heart\_disease.csv](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/PCA/heart\_disease.csv)

Meta-Data: [Heart\_Disease\_Meta\_Data.txt](https://itv-contentbucket.s3.ap-south-1.amazonaws.com/Exams/ML/PCA/Heart\_Disease\_Meta\_Data.txt)

\*\*Note:\*\* Upload assignment in PDF format with required code screenshot and corresponding output screenshots.

---

Here's the text without HTML tags:

---

K-means Assignment

\*\*Answer the following:\*\*

1. Explain Unsupervised Learning.

2. What are the factors needed to decide the value of k?

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. What are the factors needed to decide the value of k?

2. What is Euclidean distance?

3. What is Manhattan distance?

4. Explain the concept of the Elbow method.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. How can we measure the quality of two clusters?

2. Explain the role of Silhouette Coefficient.

3. Take unsupervised data from Kaggle and implement a clustering algorithm on it.

\*\*Note:\*\* Upload Assignment in PDF format.

---

Here's the text formatted without HTML tags:

---

\*\*Hierarchical Clustering Assignment\*\*

\*\*Answer the following:\*\*

1. What is the concept of clustering in machine learning?

2. Explain the difference between Agglomerative and Partitioning methods of clustering.

\*\*Note:\*\* Upload Assignment in PDF format.

---

\*\*Answer the following:\*\*

1. Elaborate on the difference between algorithms for measuring the distance between two clusters.

2. State some common applications of clustering algorithms.

\*\*Note:\*\* Upload Assignment in PDF format.

---