# Supervised Learning

Project - 60 Marks

---

**General Instructions:**

1. Submission of all the parts is expected in 1 notebook only
2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only
3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.
4. If output for any code cell is missing, 50% marks will be deducted.
5. If any kind of plagiarism from any source is found, it will not be evaluated and zero (0) Marks will be given.

---

## ---------- Part 1 ----------30 Marks

**Context:**

Medical research university X is undergoing a deep research on patients with certain conditions. University has an internal AI team. Due to confidentiality the patient's details and the conditions are masked by the client by providing different datasets to the AI team for developing a AIML model which can predict the condition of the patient depending on the received test results.

**Data Description:**

The data consists of biomechanics features of the patients according to their current conditions. Each patient is represented in the data set by six biomechanics attributes derived from the shape and orientation of the condition to their body part.

1. P_incidence
2. P_tilt
3. L_angle
4. S_slope
5. P_radius
6. S_degree
7. Class

**Project Objective:**

Demonstrate the ability to fetch, process and leverage data to generate useful predictions by training Supervised Learning algorithms.

- **Steps and Tasks:**
  1. **Data Understanding: 5**
     a. Read all the 3 CSV files as DataFrame and store them into 3 separate variables. [1 Mark]
     b. Print Shape and columns of all the 3 DataFrames. [1 Mark]
     c. Compare Column names of all the 3 DataFrames and clearly write observations. [1 Mark]
     d. Print DataTypes of all the 3 DataFrames. [1 Mark]

e. Observe and share variation in ‘Class’ feature of all the 3 DaraFrames. [1 Mark]

**2. Data Preparation and Exploration: 5**

a. Unify all the variations in ‘Class’ feature for all the 3 DataFrames. [1 Marks]
For Example: ‘tp_s’, ‘Type_S’, ‘type_s’ should be converted to ‘type_s’

b. Combine all the 3 DataFrames to form a single DataFrame [1 Marks]

**Checkpoint: Expected Output shape = (310,7)**

c. Print 5 random samples of this DataFrame [1 Marks]

d. Print Feature-wise percentage of Null values. [1 Mark]

e. Check 5-point summary of the new DataFrame. [1 Mark]

**3. Data Analysis: 10**

a. Visualize a heatmap to understand correlation between all features [2 Marks]

b. Share insights on correlation. [2 Marks]

i. Features having stronger correlation with correlation value.

ii. Features having weaker correlation with correlation value.

c. Visualize a pairplot with 3 classes distinguished by colors and share insights. [2 Marks]

d. Visualize a jointplot for ‘P_incidence’ and ‘S_slope’ and share insights. [2 Marks]

e. Visualize a boxplot to check distribution of the features and share insights. [2 Marks]

**4. Model Building: 6**

a. Split data into X and Y. [1 Marks]

b. Split data into train and test with 80:20 proportion. [1 Marks]

c. Train a Supervised Learning Classification base model using KNN classifier. [2 Marks]

d. Print all the possible classification metrics for both train and test data. [2 Marks]

**5. Performance Improvement: 4**

a. Tune the parameters/hyperparameters to improve the performance of the base model. [2 Marks]

b. Clearly showcase improvement in performance achieved. [1 Marks]
For Example:

i. Accuracy: +15% improvement

ii. Precision: +10% improvement.

c. Clearly state which parameters contributed most to improve model performance.
What could be the probable reason? [1 Marks]

## ---------- Part 2 ----------30 Marks

**Context:**

A bank X is on a massive digital transformation for all its departments. Bank has a growing customer base whee majority of them are liability customers (depositors) vs borrowers (asset customers). The bank is interested in expanding the borrowers base rapidly to bring in more business via loan interests. A campaign that the bank ran in last quarter showed an average single digit conversion rate. Digital transformation being the core strength of the business strategy, marketing department wants to devise effective campaigns with better target marketing to increase the conversion ratio to double digit with same budget as per last campaign.

**Data Description:**

The data consists of the following attributes:

1. ID: Customer ID
2. Age Customer's approximate age.
3. CustomerSince: Customer of the bank since. [unit is masked]
4. HighestSpend: Customer's highest spend so far in one transaction. [unit is masked]
5. ZipCode: Customer's zip code.
6. HiddenScore: A score associated to the customer which is masked by the bank as an IP.
7. MonthlyAverageSpend: Customer's monthly average spend so far. [unit is masked]
8. Level: A level associated to the customer which is masked by the bank as an IP.
9. Mortgage: Customer's mortgage. [unit is masked]
10. Security: Customer's security asset with the bank. [unit is masked]
11. FixedDepositAccount: Customer's fixed deposit account with the bank. [unit is masked]
12. InternetBanking: if the customer uses internet banking.
13. CreditCard: if the customer uses bank's credit card.
14. LoanOnCard: if the customer has a loan on credit card.

**Project Objective:**

Build a Machine Learning model to perform focused marketing by predicting the potential customers who will convert using the historical dataset.

- **Steps and Tasks:**
  1. **Data Understanding and Preparation: 5**
     - **a.** Read both the Datasets 'Data1' and 'Data 2' as DataFrame and store them into two separate variables. [1 Mark]
     - **b.** Print shape and Column Names and DataTypes of both the Dataframes. [1 Mark]
     - **c.** Merge both the Dataframes on 'ID' feature to form a single DataFrame [2 Marks]
     - **d.** Change Datatype of below features to 'Object' [1 Mark]
       'CreditCard', 'InternetBanking', 'FixedDepositAccount', 'Security', 'Level', 'HiddenScore'.
       *[Reason behind performing this operation:- Values in these features are binary i.e. 1/0. But DataType is 'int'/'float' which is not expected.]*

2. **Data Exploration and Analysis: 5**
    a. Visualize distribution of Target variable 'LoanOnCard' and clearly share insights. [2 Marks]
    b. Check the percentage of missing values and impute if required. [1 Mark]
    c. Check for unexpected values in each categorical variable and impute with best suitable value. *[Unexpected values means if all values in a feature are 0/1 then '?', 'a', 1.5 are unexpected values which needs treatment ]* [2 Marks]

3. **Data Preparation and model building: 10**
    a. Split data into X and Y. [1 Mark]
    *[Recommended to drop ID & ZipCode. LoanOnCard is target Variable]*
    b. Split data into train and test. Keep 25% data reserved for testing. [1 Mark]
    c. Train a Supervised Learning Classification base model - Logistic Regression. [2 Mark]
    d. Print evaluation metrics for the model and clearly share insights. [1 Mark]
    e. Balance the data using the right balancing technique. [2 Mark]
    f. Again train the same previous model on balanced data. [1 Mark]
    g. Print evaluation metrics and clearly share differences observed. [2 Mark]

4. **Performance Improvement: 10**
    a. Train a base model each for SVM, KNN. [4 Marks]
    b. Tune parameters/hyperparameters for each of the models wherever required and finalize a model. [3 Mark]
    c. Print evaluation metrics for final model. [1 Mark]
    d. Share improvement achieved from base model to final model. [2 Mark]

---

**Submission Format:**

1. .ipynb *(Jupyter Notebook)* **and**
2. .html *(Jupyter Notebook > File > Download as > HTML)*

**5 Marks will be deducted if submission in any of the formats is missing.**

---