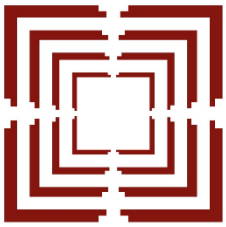

Ahmedabad University

# CSE523 Machine Learning

## Weekly Project Report

## Date: 04-03-2023

**Project title:** Big Mart Sales Prediction

**Group 10**


| Name | Enrolment no. |
|---|---|
| Meet Patel | AU2040010 |
| Dev Patel | AU2040056 |
| Kush Patel | AU2040137 |
| Vatsal Shah | AU2040019 |

1. **Task performed this week**
    - Performed label encoding on the dataset to convert the categorical features to numerical features.
    - Splitting the features and target by removing the “item_outlet_sales” feature from the dataset since “item_outlet_sales” is the target for our model.
    - Splitting the dataset into training and testing set.
2. **Outcomes of task performed**
    - Before applying label encoding:

```
[3] # Showing the first five rows
df_data.head()
```

| | Item_Identifier | Item_Weight | Item_Fat_Content | Item_Visibility | Item_Type | Item_MRP | Outlet_Identifier | Outl |
|---|---|---|---|---|---|---|---|---|
| 0 | FDA15 | 9.30 | Low Fat | 0.016047 | Dairy | 249.8092 | OUT049 | |
| 1 | DRC01 | 5.92 | Regular | 0.019278 | Soft Drinks | 48.2692 | OUT018 | |
| 2 | FDN15 | 17.50 | Low Fat | 0.016760 | Meat | 141.6180 | OUT049 | |
| 3 | FDX07 | 19.20 | Regular | 0.000000 | Fruits and Vegetables | 182.0950 | OUT010 | |
| 4 | NCD19 | 8.93 | Low Fat | 0.000000 | Household | 53.8614 | OUT013 | |

    - After applying label encoding:

```
[30] df_data.head()
```

| | item_identifier | item_weight | item_fat_content | item_visibility | item_type | item_mrp | outlet_identifier | outlet_estab |
|---|---|---|---|---|---|---|---|---|
| 0 | 156 | 9.30 | 1 | 0.016047 | 4 | 249.8092 | 9 | |
| 1 | 8 | 5.92 | 2 | 0.019278 | 14 | 48.2692 | 3 | |
| 2 | 662 | 17.50 | 1 | 0.016760 | 10 | 141.6180 | 9 | |
| 3 | 1121 | 19.20 | 2 | 0.000000 | 6 | 182.0950 | 0 | |
| 4 | 1297 | 8.93 | 1 | 0.000000 | 9 | 53.8614 | 1 | |

    - Splitting the dataset into training and testing sets:

```
[35] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)

print(X.shape, X_train.shape, X_test.shape)

(8523, 11) (6818, 11) (1705, 11)
```

    - We split the dataset into 80% training set and 20% testing set.
    - The dimensions are:
        - Actual dataset: (8523,11)

- Training dataset: (6818, 11)
- Test set: (1705,11)

## 3. Tasks to be performed in the upcoming week

- We will implement the linear regression and XGBoost regressor.