

**Project Report:** Artificial Intelligence Lab

**Course Code:** CSE316

**Team Members:** Abdun Nafi Annafi  
 212-15-14755  
 Tasnuva Muhtasim  
 212-15-14761  
  
**Submitted To:** Ms. Nasima Islam Bithi  
 Lecturer  
 Dept. of CSE  
 Daffodil International University

**Introduction:**The main goal of this project is to preprocess the given dataset by handling null values, performing encoding, outlier handling and scaling.We have used various machine learning algorithms on the preprocessed data and their accuracies are compared to find out the best model for the dataset

**Dataset:**The dataset we worked on contains our dependent variable which is Sales and the rest are independent variables.

**Preprocessing:**    
In our project we have used preprocessing techniques to make it suitable to train for a model.  
Null value handling: We have handled the null value we found in the postal code column using dropna function. This was necessary because most of the classification models don't work because of the null value in the dataset.

Column drop: Our model will show the sales prescription, we have dropped all the unnecessary columns. The columns we categorized as unnecessary

● Columns related to customers, we can predict the sales with the product ID.

● Product name as we have product id

● The country as the country column has only one value. This action was necessary to reduce the column sizes and the memory.

Outlier handling: We removed the outliers from the sales column. this was important because outliers may distract the model from the real solution.

**Feature Engineering:**In order to improve model performance by the creation, transformation, or selection of features, feature engineering is essential to machine learning. It manages categorical data, decreases dimensionality, captures temporal connections, mitigates data imbalances, handles non-linearity, and deals with noisy data.  
1) Encoding: In this project we have used onehot encoding because there are multiple categorical data. That's why to convert them to numerical values we have to use onehot encoding.  
2) Outlier Handling: In this project we have handled the outliers using trimming method to make the dataset more suitable for training the model.  
3) Scaling: Here we have used standard scaling to enhance algorithm performance.

**Models:**We use different types of machine learning algorithms to get a clear and distinct result which helps us to get the meaningful insight we want from our dataset. The algorithms we used in this project are:

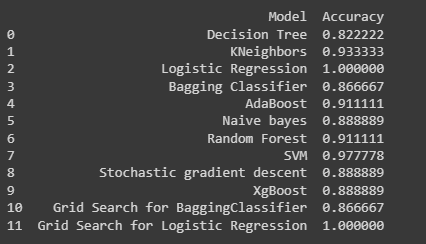
1. Decision Tree(0.84)
2. KNeighbours Classifier(0.93)
3. Logisitic Regression(1)
4. Begging Classifier(0.93)
5. AdaBoost Classifier(0.91)
6. Bernoulli Naive Bayes(0.88)
7. Random Forest Classifier(0.88)
8. SVM Classifier(0.97)
9. Stochastic gradient descent(0.1)
10. XgBoost Classifier(0.88)
11. Grid Search for BaggingClassifier(0.91)
12. Grid Search for LogisitcRegression(0.1)

**Evaluation on Models:**

After applying the classification models, best accuracy was found on the model logistic regression.

Regression models are kown to be best for finding out numerical values. Here, our dependent variable is sales. Therefore, Logistic Regression model giving the outmost accuracy makes sense

**Comparison Table:**



**Conclusion:**

According to comparison matrices Logistic regression is best.