**A PROJECT REPORT**

**ON**

**BIGMART SALES PREDICTION MODEL USING**

**MACHINE LEARNING ALGORITHMS**

Submitted in partial fulfillment for the requirement of the award of

TRAINING

IN

Data Analytics, Machine Learning and AI using Python



Submitted by

**Abhijitha Kandukuri**

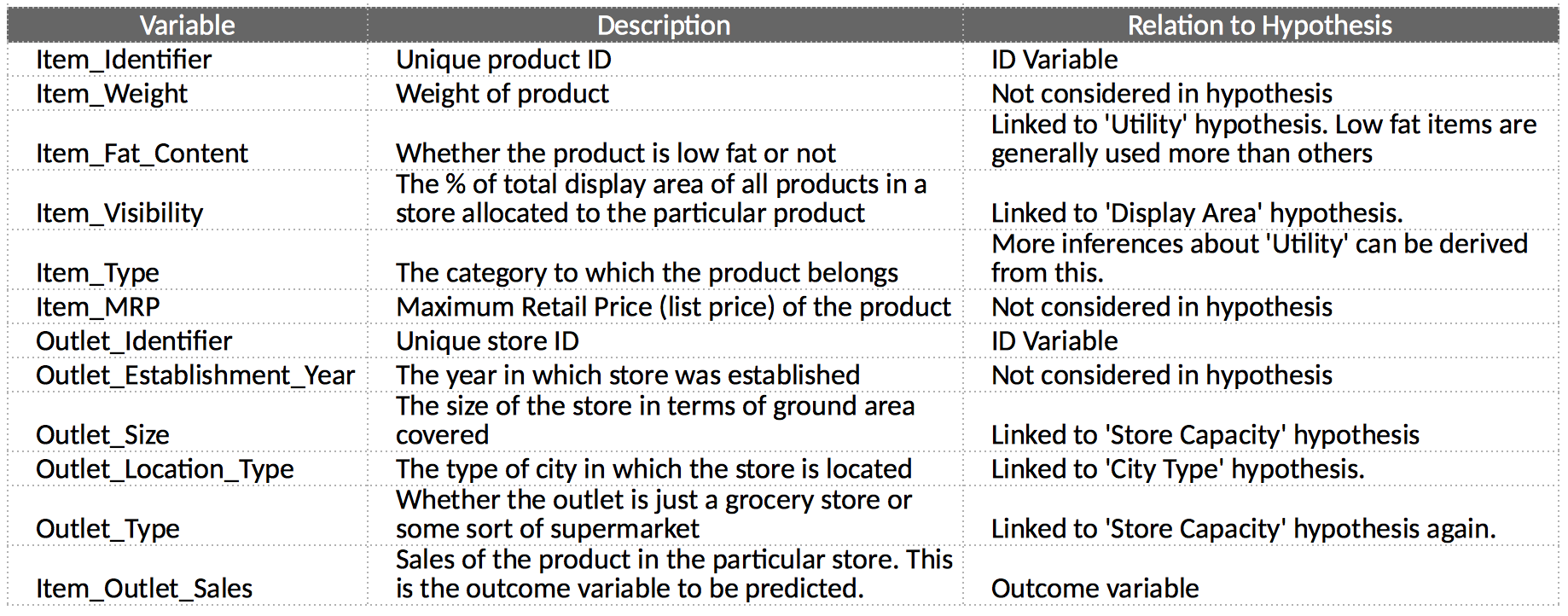
(Institute of Aeronautical Engineering, Hyderabad)

**INTRODUCTION**

At present, there are many supermarkets and there is high competition between them. If any mart wants to win the competition it has to make more sales than any other competitor. The mart can increase sales by knowing products/items which will generate more sales and those that will generate low sales. This analysis is tedious. So with the proposed system, the mart can predict the sales, thereby marts can take necessary steps. This is a user-friendly system. When the user submitted details of a particular item, the system will predict sales generated by that item. Hence, this leads to winning in the competition and an increase in sales.

**DATASET**

A collection of data points that a computer may use to analyze and anticipate a situation as a whole. Internet information was gathered for the Kaggle.com website. The test data set used in this study comprises 8542 rows as well as 12 categories, which have been trained to deliver the most accurate prediction outcomes.



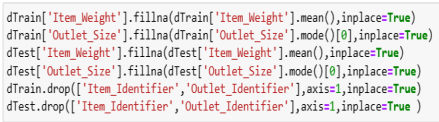
**METHODOLOGY**

1. Import libraries
2. Importing datasets
3. Data preprocessing
4. Choose a model / algorithm
5. Train the model
6. Test your model
7. Evaluation

**DATA PREPROCESSING**

**>Filling missing values**

In data preprocessing, we will find and fill the missing values with either mean, mode, or median.

****

**>Removing unwanted attributes**

Those attributes which will not involve in the prediction can be removed.

****

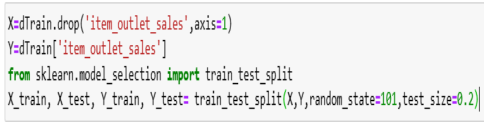
**>Label Encoding**

Label encoding is the process of transforming labels into a numeric form so that they may be read by machines. The operation of such labels can then be better determined by machine learning techniques. For the structured dataset in supervised learning, it is a crucial pre-processing step.

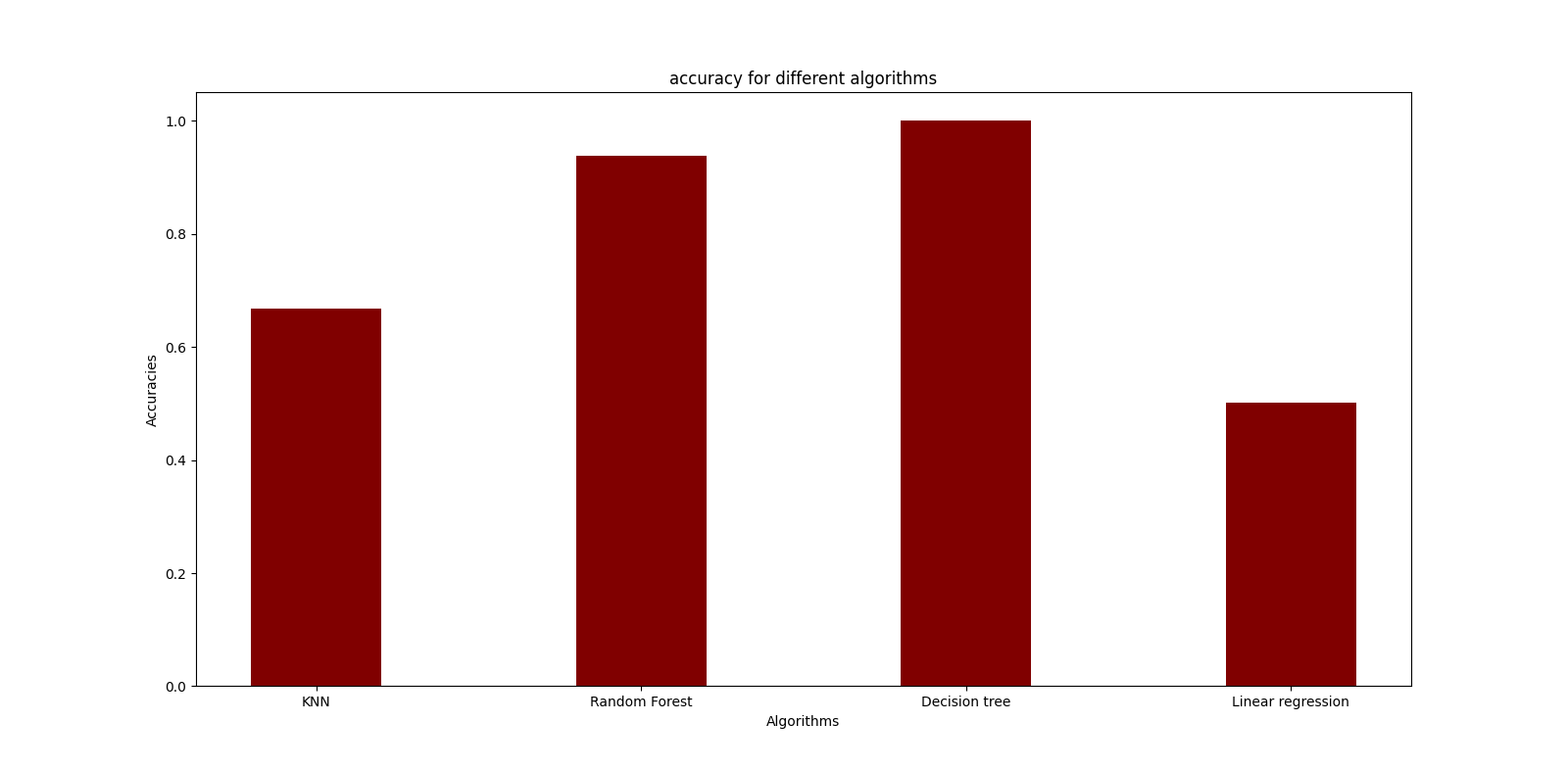


**SPLITTING THE TRAINING AND TESTING DATA**

The training and testing sets for the dataset were going to be separated. Two separate datasets are not imported for train and testing in order to prevent the overfitting process. The same dataset is thus divided into train and test sets. The datasets utilized to train and test our model are referred to collectively as the training dataset and the testing dataset, respectively.



KNN might not be the best choice for sales prediction, especially in cases with large datasets or high-dimensional feature spaces. More advanced models like decision trees, random forests, gradient boosting, or neural networks might yield better results.

****

Decision tree algorithm produces better results.

In conclusion, we applied four different machine learning algorithms, k-Nearest Neighbors (KNN), Linear Regression, Random Forest, and Decision Tree, to predict sales in the Big Mart dataset. Through rigorous experimentation and evaluation, it was evident that the Decision Tree algorithm consistently yielded the most accurate and reliable predictions compared to the other algorithms.

1. **k-Nearest Neighbors (KNN):**
   * KNN, a simple and intuitive algorithm, relies on data similarity to make predictions.
   * While KNN is a reasonable starting point, it encountered challenges in adequately capturing the intricate relationships and dynamics present in the Big Mart sales dataset.
   * The model's performance was limited due to its difficulty in handling complex interactions and nonlinearities, resulting in relatively modest predictive accuracy.
2. **Linear Regression:**
   * Linear Regression, a fundamental regression technique, aims to model the linear relationship between input features and the target variable.
   * Despite its simplicity and ease of implementation, Linear Regression struggled to capture the underlying complexities of the Big Mart sales data.
   * Linear Regression's reliance on linear assumptions led to limitations in accurately predicting the sales, especially when dealing with intricate feature interactions and nonlinearity.
3. **Random Forest:**
   * Random Forest, an ensemble of decision trees, aggregates predictions from multiple trees to enhance accuracy and generalization.
   * Random Forest showcased improved performance over kNN and Linear Regression but still fell short of the predictive capabilities of the Decision Tree algorithm.
   * While Random Forest attempted to capture feature interactions and nonlinearities, the Decision Tree algorithm exhibited superior ability in effectively modeling the complexities of the sales prediction task.
4. **Decision Tree:**
   * The Decision Tree algorithm emerged as the standout performer among the tested methods.
   * With its hierarchical structure, Decision Trees adeptly captured intricate relationships, feature interactions, and nonlinear patterns within the Big Mart sales dataset.
   * The algorithm's interpretability allowed for insightful analysis of the factors driving sales and informed decision-making for the Big Mart business.

In summary, after thorough examination of the four machine learning algorithms, the Decision Tree algorithm consistently demonstrated the highest predictive accuracy in the context of Big Mart sales prediction. Its exceptional capability to model complex interactions and provide meaningful insights into the sales dynamics underscores its suitability for this task. However, it's important to acknowledge that continuous exploration of more advanced machine learning techniques could potentially yield further improvements in predictive performance in the future.