**IMPLEMENTATION**

**MODULES:**

* Data Collection
* Dataset
* Data Preparation
* Model Selection
* Analyze and Prediction
* Accuracy on test set
* Saving the Trained Model

**MODULES DESCSRIPTION:**

**Data Collection:**

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms

Data set Link: https://www.kaggle.com/shivan118/big-mart-sales-prediction-datasets

**Dataset:**

The dataset consists of 8523 individual data. There are 12 columns in the dataset, which are described below.

**1.Item**Identifier ---- Unique product ID  
**2.ItemWeight** ---- Weight of product  
**3.Item**Fat**Content** ---- Whether the product is low fat or not  
**4.Item**Visibility ---- The % of the total display area of all products in a store allocated to the particular product  
**5.ItemType** ---- The category to which the product belongs  
**6.Item**MRP ---- Maximum Retail Price (list price) of the product  
**7.OutletIdentifier** ---- Unique store ID  
**8.Outlet**Establishment**Year** ---- The year in which the store was established  
**9.Outlet**Size ---- The size of the store in terms of ground area covered  
**10.OutletLocation**Type ---- The type of city in which the store is located  
**11.\*OutletType** ---- Whether the outlet is just a grocery store or some sort of supermarket  
**12.Item**Outlet**Sales** ---- sales of the product in t particular store. This is the outcome variable to be predicted.

**Data Preparation:**

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)

Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data

Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis

Split into training and evaluation sets

**Model Selection:**

We used decision tree regression machine learning algorithm , We got a accuracy of 95.7% on test set so we implemented this algorithm.

**Decision tree regression**

**Decision Tree** is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. The branches/edges represent the result of the node and the nodes have either:

Conditions [Decision Nodes]

Result [End Nodes]

The branches/edges represent the truth/falsity of the statement and take makes a decision based on that in the example below which shows a decision tree that evaluates the smallest of three numbers:

**Decision Tree Regression:** Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

**Analyze and Prediction:**

In the actual dataset, we chose only 9 features:

**1.ItemWeight** ---- Weight of product  
**2.Item**Fat**Content** ---- Whether the product is low fat or not  
**3.Item**Visibility ---- The % of the total display area of all products in a store allocated to the particular product  
**4.ItemType** ---- The category to which the product belongs  
**5.Item**MRP ---- Maximum Retail Price (list price) of the product  
**6.Outlet**Establishment**Year** ---- The year in which the store was established  
**7.Outlet**Size ---- The size of the store in terms of ground area covered  
**8.OutletLocation**Type ---- The type of city in which the store is located  
**9.\*OutletType** ---- Whether the outlet is just a grocery store or some sort of supermarket

**Accuracy on test set:**

We got an accuracy of 95.80% on test set.

**Saving the Trained Model:**

Once you’re confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 (or) .pkl file using a library like pickle .

Make sure you have pickle installed in your environment.

Next, let’s import the module and dump the model into .pkl file