**Product Demand Prediction with**

**Machine Learnings-Guidelines**

**NAME : SANTHOSHKUMAR .S E-MAIL: santhoshkumarsasvg@gmail.com**

**REG NO :912721104027**

Abstract

###### Background: Sales forecasting is an important field in the food sector, and it has recently got immense popularity to boost market operations and productivity due to new technologies. The industry has traditionally focused on a conventional statistical model but in the recent years, Machine Learning techniques have received more attention.

###### Objectives: This thesis will help to identify the critical features that influence sales and also an experiment is performed to find the best suitable algorithm for sales forecasting.

###### Methods: Machine Learning Algorithms such as Simple Linear Regression, Gradient Boosting Regression, Support Vector Regression, and Random Forest Regression were considered in this thesis, which they expected to perform well on the issues. An experiment is carried out to determine the efficiency of the algorithms.

###### Results: Algorithms such as Simple Linear Regression, Gradient Boosting Regression, Support Vector Regression, and Random Forest Regression are commonly known for performing better than others, this has been clearly shown that Random Forest Regression is the most appropriate algorithm compared to the others.

###### Conclusions: The Random Forest Regression algorithm performed well after doing all the study when compared with other algorithms. Hence the Random Forest Regression is considered as the best suitable algorithm for forecasting product sales

Aims and Objectives

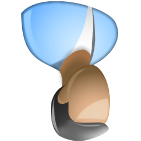
**This thesis aims to develop a Machine Learning model that can predict the sales of products from different outlets. Several objectives were drawn to attain the goal**

Objectives:

* **Converting data into an appropriate form using various preprocessing techniques for the implementation of Machine Learning algorithms**
* **Finding critical features that will most influence sales of the product.**
* **To determine the appropriate Machine Learning algorithm for sales forecasting.**

**Introduction**

**Earlier companies used to produce goods without considering the number of sales and demand. For any manufacturer to determine whether to increase or decrease the production of several units, data regarding the demand for products on the market is required. Companies can face losses if they fail to consider these values while competing on the market. Different companies choose specific criteria to determine their demand and sales**



**C**

**u**

**s**

**t**

**o**

**m**

**e**

**r**



**o**

**r**

**d**

**e**

**r**

**s**

**a**

**p**

**r**

**o**

**d**

**u**

**c**

**t**

**W**

**e**

**b**

**S**

**i**

**t**

**e**



**S**

**e**

**l**

**l**

**e**

**r**

**s**

**S**

**h**

**i**

**p**

**p**

**i**

**n**

**g**

ß

**N**

**t**

**h**

**S**

**e**

**l**

**l**

**e**

**r**

**1**

**2**

**3**

**4**

**N**

**R**

**e**

**v**

**i**

**e**

**w**

**s**

**p**

**r**

**o**

**d**

**u**

**c**

**t**

**Q**

**u**

**a**

**l**

**i**

**t**

**y**

**/**

**s**

**h**

**i**

**p**

**p**

**i**

**n**

**g**

**/**

**S**

**e**

**l**

**l**

**e**

**r**

**R**

**e**

**p**

**o**

**n**

**s**

**e**

**e**

**t**

**c**

**.**

**T**

**r**

**a**

**i**

**n**

**i**

**n**

**g**

**D**

**a**

**t**

**a**

**S**

**T**

**A**

**C**

**K**

**E**

**D**

**G**

**E**

**N**

**E**

**R**

**A**

**L**

**I**

**Z**

**A**

**T**

**I**

**O**

**N**

**S**

**Y**

**S**

**T**

**E**

**M**

**W**

**e**

**b**

**S**

**i**

**t**

**e**

ß

**D**

**e**

**m**

**a**

**n**

**d**

**P**

**r**

**e**

**d**

**i**

**c**

**t**

**i**

**o**

**n**

Figure

1:

General

System

Schema.

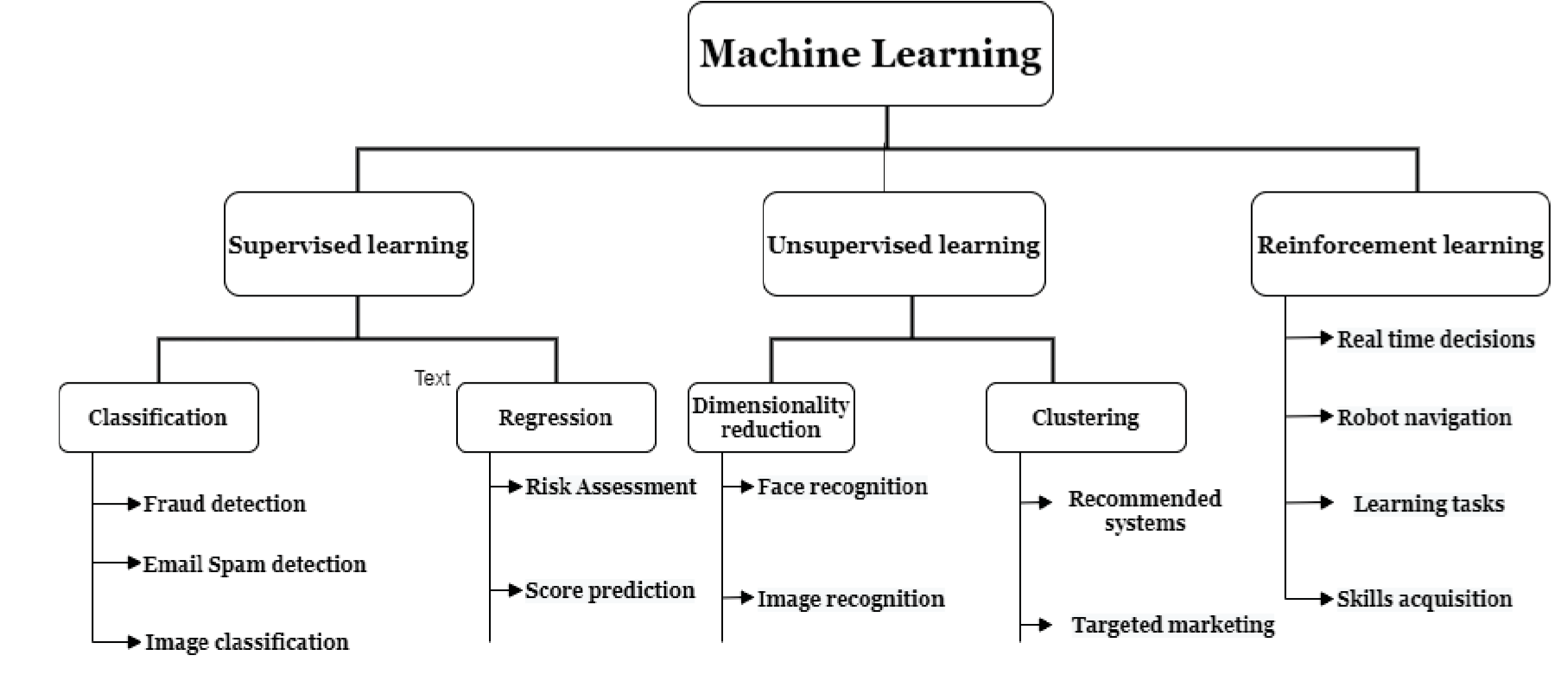
**Background**

**There are several methods for forecasting future demand for the goods and services a business provides. The forecasts are used for planning production and business activities, purchasing materials, inventory management, scheduling work hours, advertising, and often more across most industries. Traditional forecasting approaches were primarily focused on experienced employee opinions or statistical analysis of past data, but in recent years Machine Learning techniques have been implemented with great success in this field.**

**Data Mining**

**Data mining is described as a process for extracting usable data from a larger collection of raw data using statistical, artificial intelligence, Machine Learning and pattern recognition methods[10][11]. Data Mining is increasingly seen as a step in a systematic and iterative process of knowledge discovery, in which automated pattern recognition methods are combined with expert knowledge of the analyst. This process is called the Knowledge Discovery in Databases (KDD) process.**

**Machine Learning**

 **Machine Learning is the area of study which enables machines to learn without being explicitly programmed. Machine Learning is defined as the computer program learns from experience E with respect to some class of tasks T and performance measure P when its performance at tasks in T, as measured by P, strengthens with**

Types of Machine Learning

**Supervised Learning:**

**The most popular model for performing Machine Learning processes is supervised learning. It is commonly used for data where the mapping between input-output data is accurate. Supervised learning is the subset of Machine Learning which concentrates on learning a model of classification or regression, that is, learning from labeled test data.**

**Unsupervised Learning:**

**The data is not explicitly labeled into different classes in the case of unsupervised learning that is there is only unlabeled data. By identifying implicit patterns the model can learn from the data. Unsupervised Learning categorizes the densities, structures, related segments, and other similar properties based on the data.**

**Reinforcement Learning:**

**Reinforcement Learning is a sub-field of Machine Learning. In a given scenario, it is about taking appropriate action to optimize reward. Various algorithms and computers are employed to determine the best possible action or path it will follow in a specific scenario.**

**Data Collection: Gather historical data on product sales, including variables like time, price, marketing efforts, and external factors (e.g., holidays, economic indicators).**

**Data Preprocessing:**

**Clean and preprocess the data, handling missing values, outliers, and encoding categorical variables.**

**Feature Engineering:**

**Create relevant features that can influence product demand, such as seasonality, trends, or lagged variables.**

**Split the Data:**

**Divide the dataset into training and testing sets to evaluate the model's performance.**

**Model Selection:**

**Choose an appropriate machine learning algorithm for demand prediction. Common choices include regression models (e.g., linear regression), time series models (e.g., ARIMA or Prophet), or more advanced techniques like neural networks.**

**Model Training:**

**Train the selected model on the training data, tuning hyperparameters if necessary.**

**Model Evaluation:**

**Assess the model's performance using appropriate metrics (e.g., Mean Absolute Error,Root Mean Squared Error, or others) on the test dataset.**

**Deployment:**

**Deploy the trained model into your application or system to make predictions on new data.**

**Prediction:**

**Use the deployed model to predict future product demand, providing output in the form of predicted demand values or visualizations.**

Program:

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_absolute\_error

X = data[['feature1', 'feature2', ...]]

y = data['demand']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

predictions = model.predict(X\_test)

mae = mean\_absolute\_error(y\_test, predictions)

print(f"Mean Absolute Error: {mae}")

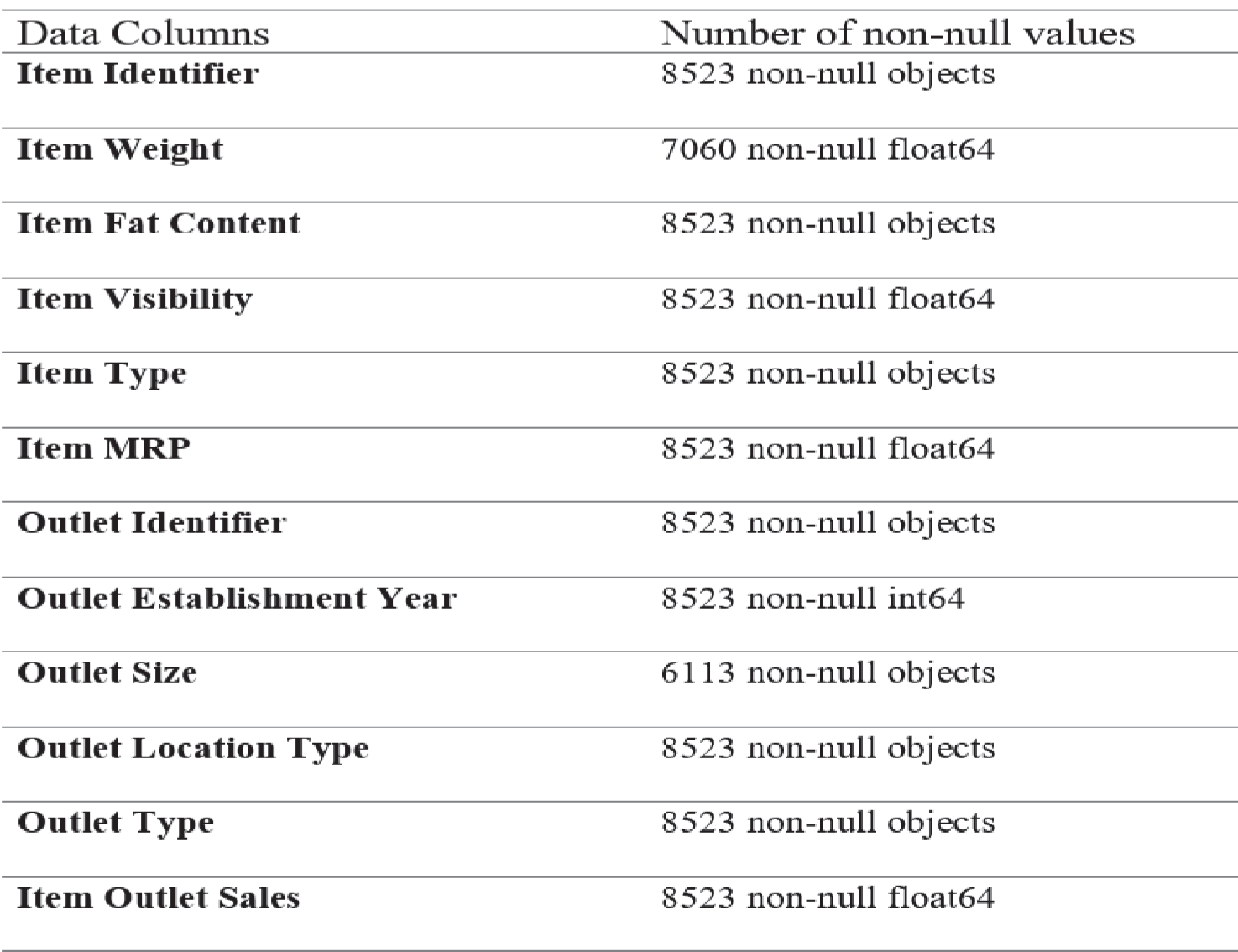
**Feature Selection**

**There are various types of factors that can make the model of Machine Learning more effective on any given task. One of the methods of feature selection is data correlation**



fig:Dataset description

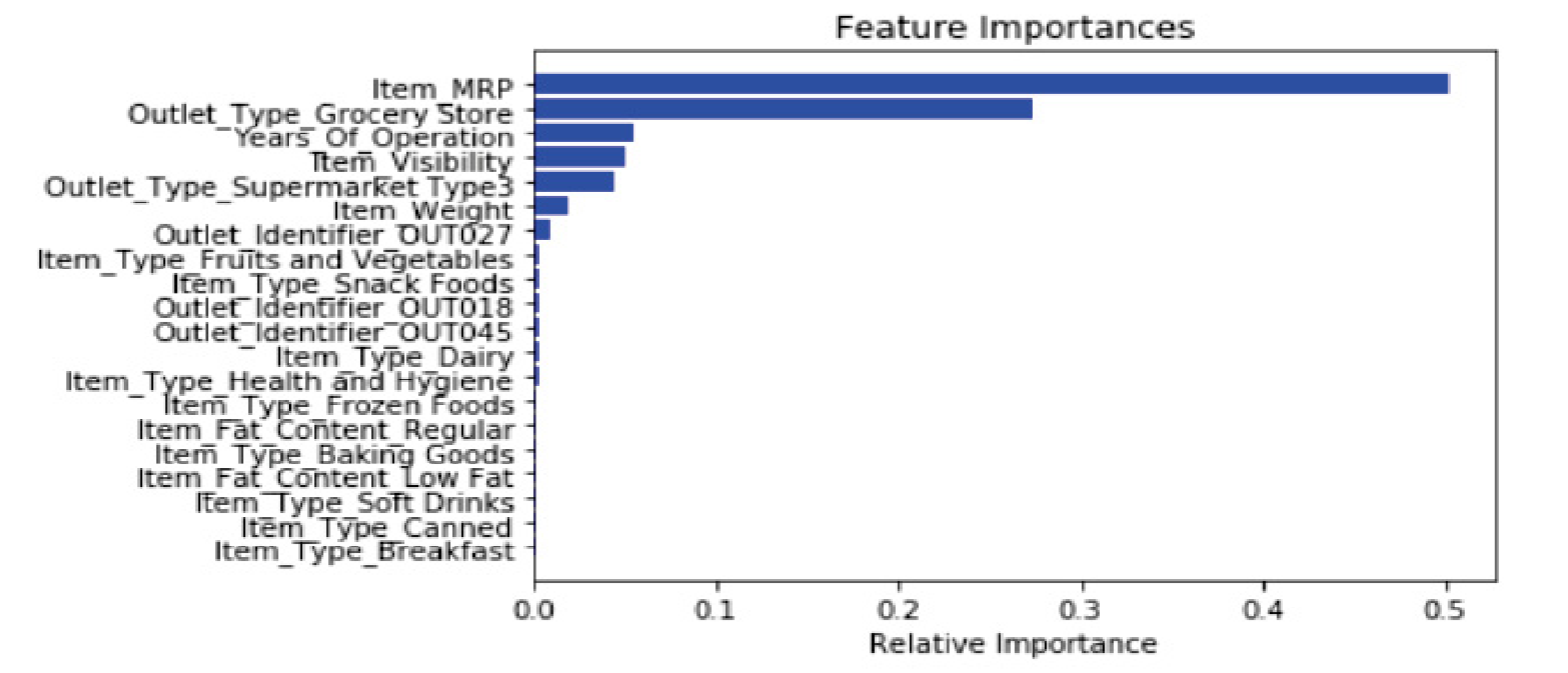
**effective on any given task. One of the methods of feature selection is data correlation**



**Feature Importance**

**Feature Importance refers to a class of approaches for assigning values to input features to a predictive model which determines the relative significance of each factor while forecasting.**

**Feature importance scores provide overview into the model. Most significant scores are determined using a prediction approach that was fitted to the dataset. Inspecting the score of importance gives insight into that particular model and what features are the most essential and least important to the model while making a prediction. This is a type of interpretation of the model that can be carried out for those models that encourage it**



**Data preprocessing**

**Before applying Machine Learning algorithms some of the missing values have been found which can impact the model’s output so this should be handled. The ’item weight’ and ’outlet size’ attributes have 17 percent, and there is 28 percent of missing values**.



G

r

a

d

i

e

n

t

B

o

o

s

t

e

d

T

r

e

e

s



L

i

n

e

a

r

R

e

g

r

e

s

s

i

o

n

**T**

**r**

**a**

**i**

**n**

**i**

**n**

**g**

**D**

**a**

**t**

**a**



R

a

n

d

o

m

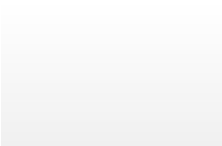
F

o

r

s

t



M

e

t

a

-

R

e

g

r

e

s

s

o

r

1

s

t

P

r

e

d

i

c

t

i

o

n

2

n

d

P

r

e

d

i

c

t

i

o

n

4

t

h

P

r

e

d

i

c

t

i

o

n

L

a

s

t

P

r

e

d

i

c

t

i

o

n



D

e

c

i

s

i

o

n

T

r

e

e

3

r

d

P

r

e

d

i

c

t

i

o

n

Generalization.

Stacked

Figure

2:

EXPERIMENTAL RESULTS

Dataset Definition

**The data used in the experiments was provided from one of the most popular online e-commerce company in {Country Name}. First, standard preprocessing techniques are applied. Some of these techniques include filling in the missing values, removal of missing attributes when a major portion of the attribute values are missing and removal of irrelevant attributes. Each product/good has a timestamp which represents the date it is sold consisting of year, month, week and day information. A product can be sold several times within the same day from both same and different sellers. The demands or sales of a product are aggregated weekly. While the dataset contains 3575 instances and 17 attributes, only 1925 instances remained after the aggregation. Additionally, customers enter the company’s website and choose a product they want.**



**S**

**T**

**O**

**C**

**K**

**<**

**5**

**0**

**Y**

**E**

**S**

**N**

**O**

**3**

**0**

**1**

**0**

**0**

**5**

**0**

**P**

**R**

**I**

**C**

**E**

**<**

**5**

**0**

**0**

**Y**

**E**

**S**

**N**

**O**

**S**

**E**

**L**

**L**

**E**

**R**

**G**

**R**

**A**

**D**

**E**

**>**

**8**

**5**

**Y**

**E**

**S**

**N**

**O**

**P**

**R**

**I**

**C**

**E**

**<**

**7**

**5**

**0**

**Y**

**E**

**S**

**N**

**O**

**7**

**0**

**2**

**0**

**S**

**T**

**O**

**C**

**K**

**O**

**U**

**T**

**Y**

**E**

**S**

**N**

**O**

**1**

**5**

**0**

**6**

**0**

**F**

**A**

**S**

**T**

**S**

**H**

**I**

**P**

**Y**

**E**

**S**

**N**

**O**

**8**

**5**

Result and Discussion

**In this section, we evaluate the proposed model using RMSE evaluation method. After calculating RMSE for single classifiers and SG, we applied analysis of variances (ANOVA) test. It is generalized version of t-test to determine whether there are any statistically significant differences between the means of two or more unrelated groups. We use ANOVA test to show that predictions of the models are statistically different. The training set is divided randomly into 20 different subsets, so that no subset contains the whole training set. Using each of the different subsets and the validation set, the SG model is trained and evaluated on the test set. In the first level of the SG model, various combinations of the four algorithms are used.**

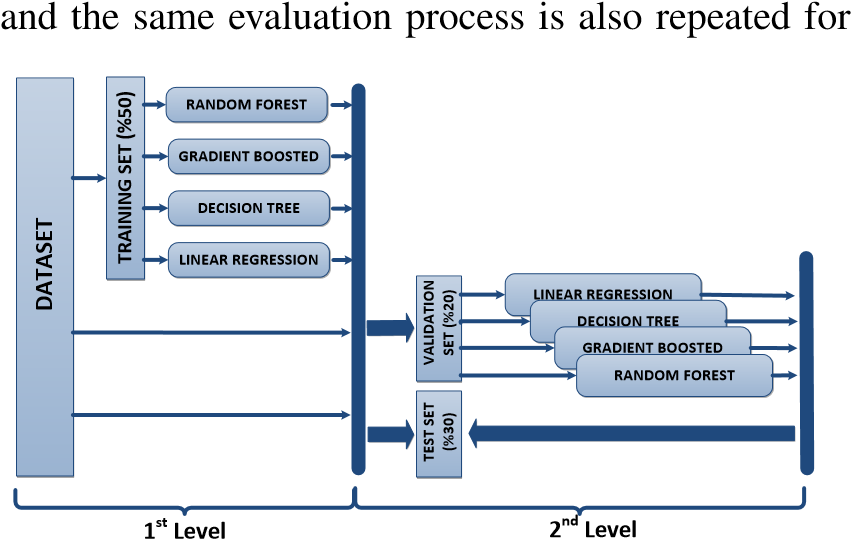


Figure : Stacking Process.

Conclusion :

**Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, sales revenue analysis will help to get the details needed to estimate both the revenue and the income. Different types of Machine Learning techniques such as Support Vector Regression, Gradient Boosting Regression, Simple Linear Regression, and Random Forest Regression have been evaluated on food sales data to find the critical factors that influence sales to provide a solution for forecasting sales. After performing metrics such as accuracy, mean absolute error, and max error, the Random Forest Regression is found to be the appropriate algorithm according to the collected data and thus fulfilling the aim of this thesis**.