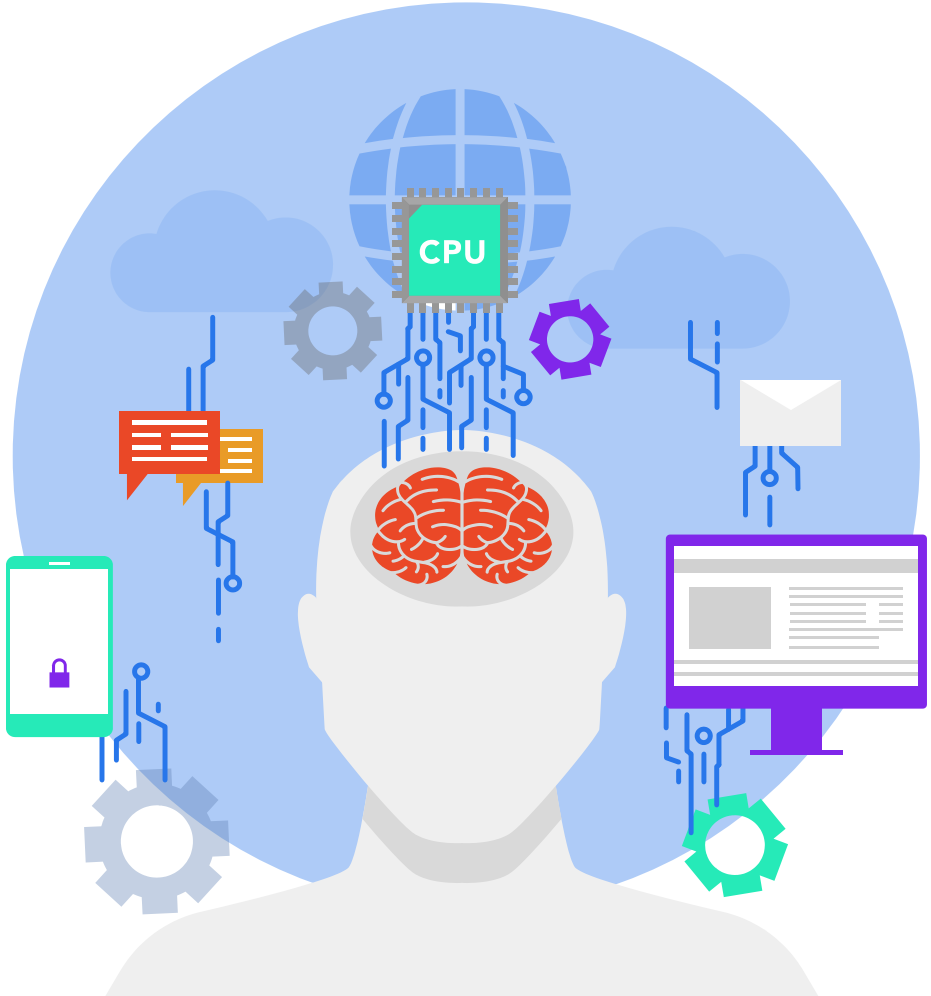


Start up Profit Prediction using Machine Learning

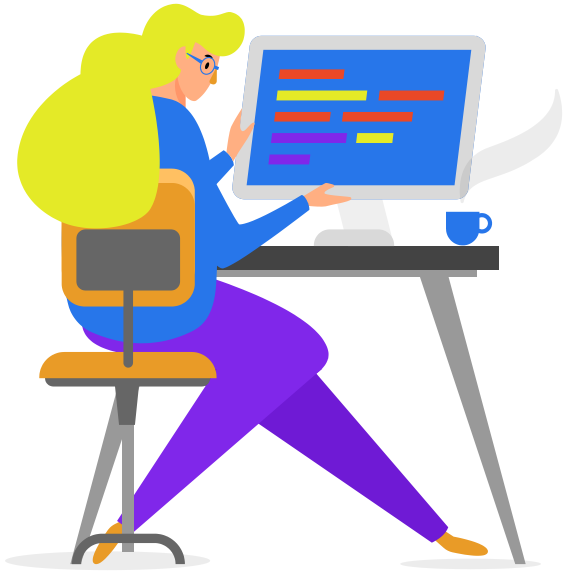
Abstract

What makes a startup successful? How to define success?

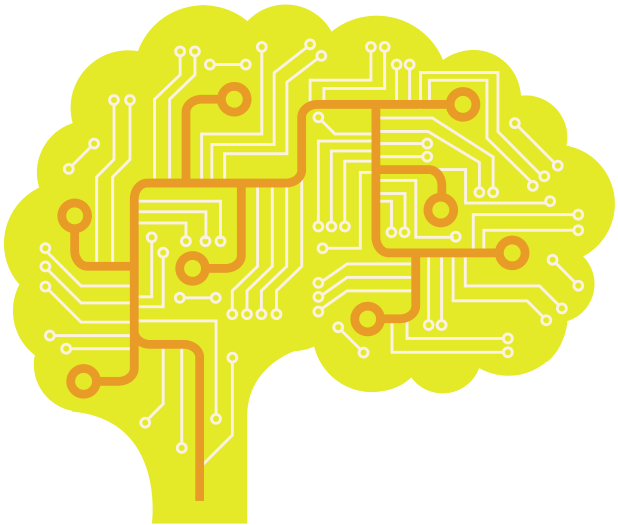
Small and medium sized businesses (SMEs) have been considered to be the driving force of technological innovation, economic flexibility and growth while creating new job. Hence, the success of these companies is in the interest and favor of society. This project aims at constructing an appropriate quantitative model to predict the profit of a company.

Entrepreneurs and management teams of the firms that operate in disruptive areas like blockchain applications and cryptocurrencies face unique business risks and uncertainties compared to those of traditional established companies. This project approaches the start-up's success prediction through profit prediction. The given data is used for making profit prediction.

Table of contents



Introduction



The profit of a company depends on a lot of parameters and values. We have used Regression algorithms of machine learning to find the profit of a company. For the given dataset, we have implemented 4 different algorithms to find the algorithm best-suited.

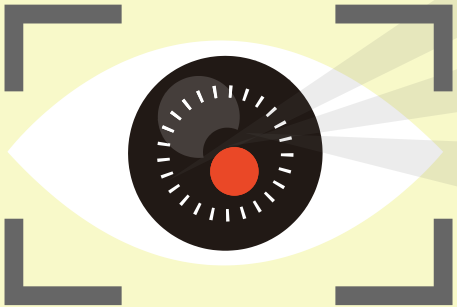
The algorithms used are

01 Multiple Linear Regression
02 Random forest
03 SVM
04 Decision tree

After finding the different efficiency scores, we found that multiple linear regression is the best performing algorithm.

Existing method

The existing system to solve this problem of start-up success prediction used the following algorithms.



Decision Tree

This tree-structured classifier gave a validation accuracy score of 95.5%.

Random Forest

This gathering learning technique gave a validation accuracy score of 93.4%.

Support Vector

The popular supervised learning algorithm gave a validation accuracy score of 85.3%.

Proposed method

Multiple linear regression is employed because

Linear relationship
between dependent and each independent variable

01

Multicollinearity
is not exhibited by the data values

02

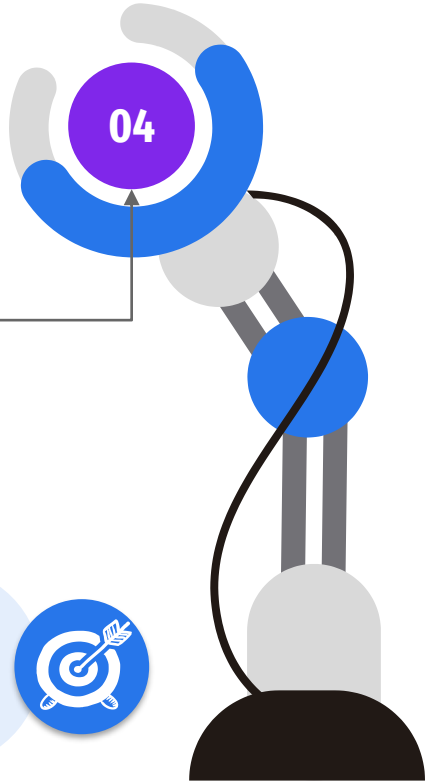
Homoscedasticity
the residual variance is constant

03

Multivariate normality
residuals are normally distributed

04

And this gave a whooping validation accuracy score of 96.1%



System architecture

Collect the data

The dataset given is analyzed

Prepare the data

The attributes - administration which played no role in prediction were removed

Train a model

Training the model using the training set and testing the model using the testing set.

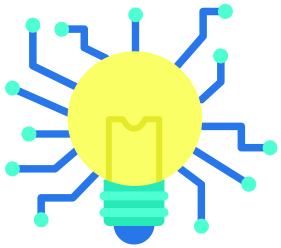
Evaluate the model

Predicting the profit

Improve

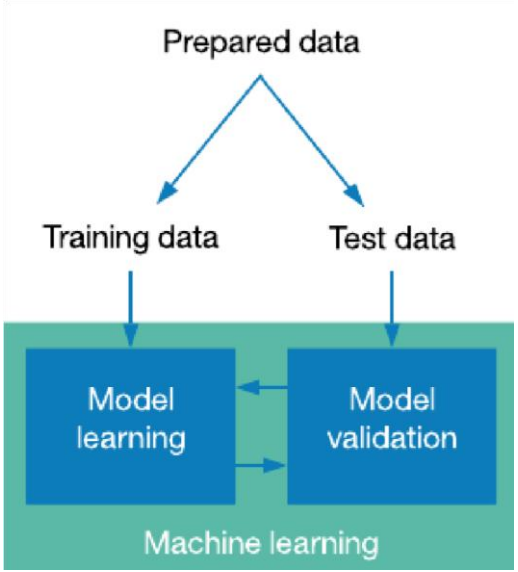
Residual anaysis is carried out

Methodology



Collection of Datasets

Initially, we collect a dataset for our startup profit prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model.



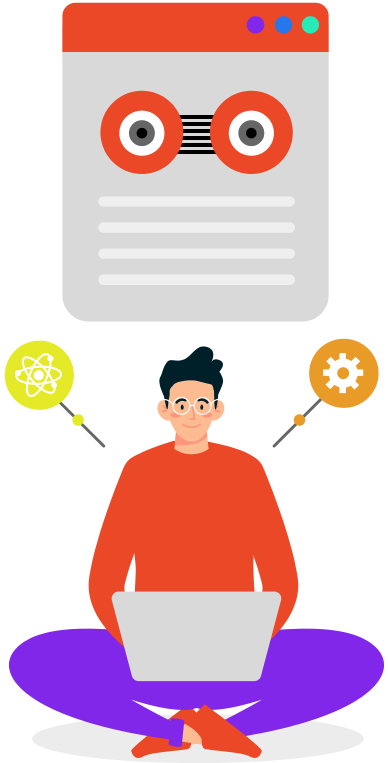
Methodology

Data Preparation

Source : Startup Profit Prediction dataset

https://drive.google.com/file/d/1Z7RKmScBO7n9vcDIG3Xeo853Ics4QFaF/view

This dataset was built by augmenting datasets of R & D spend, administration and marketing spend available for a company. The classification goal is to predict the profit value of a company if the value of its R & D spend, administration and marketing spend are given.



Methodology

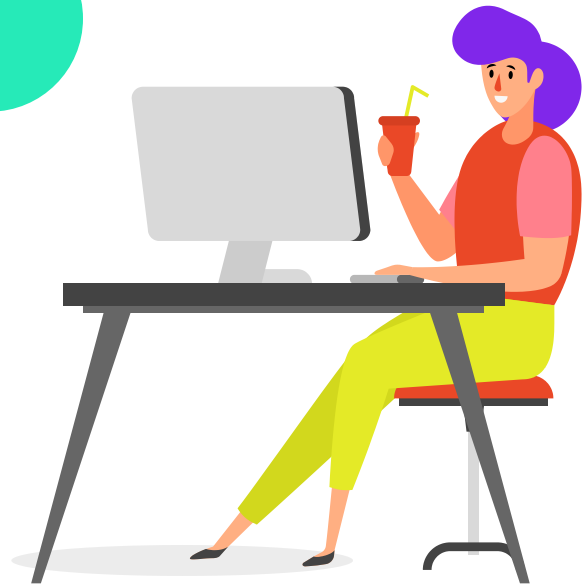
Selection of Attributes

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system.

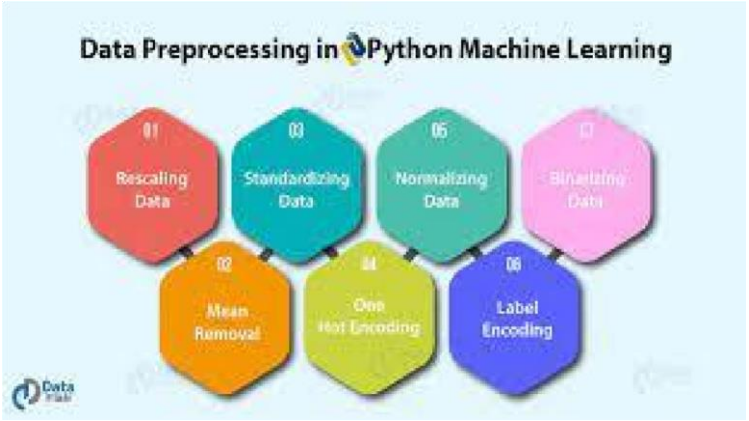
01 R & D Spend

02 Administration

03 Marketing Spend



Methodology

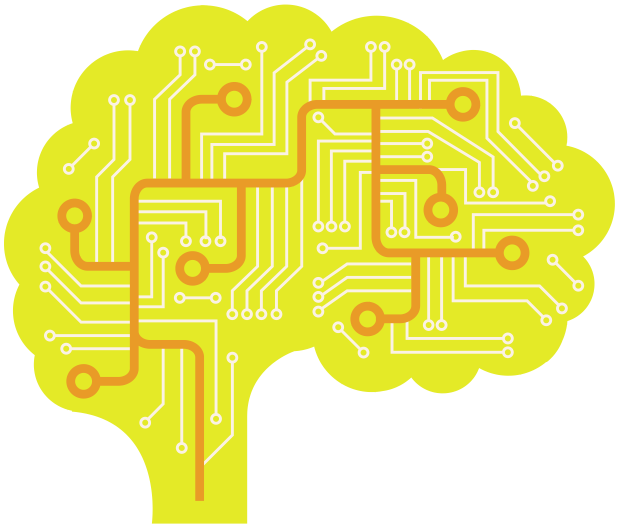


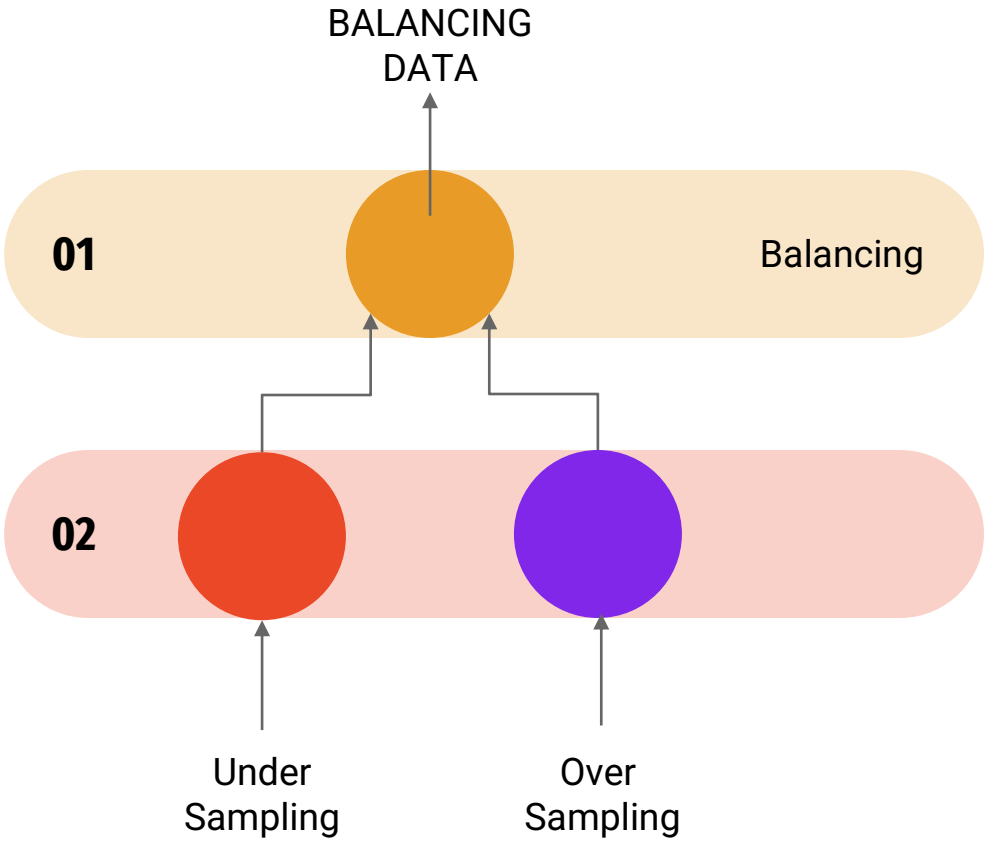
Pre-processing of data

Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.

Methodology

Balancing of data





Methodology

01

Multiple Linear Regression

Prediction of Output

02

Random Forest

04

Support Vector

05

Decision Tree

ML Classsification Algorithms

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

ML Classification Algorithm

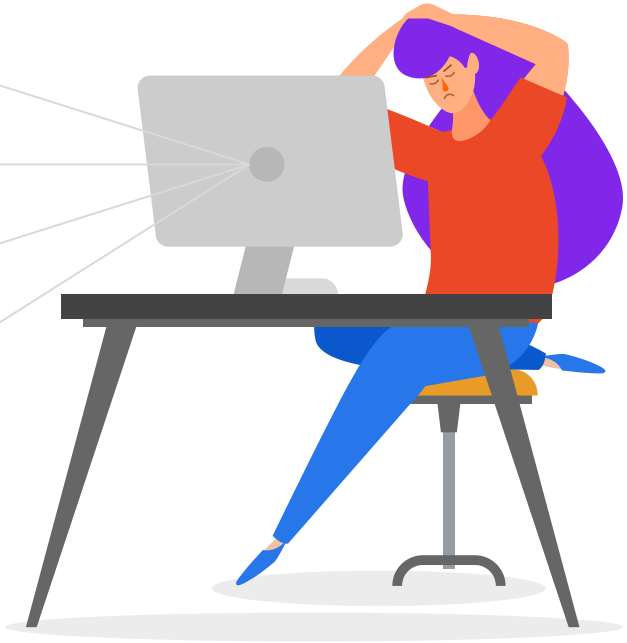
Working of Random Forest

01 Select random K data points from the training set.

02 Build the decision trees associated with the selected data points (Subsets).

03 Choose the number N for decision trees that you want to build. Repeat Step 1 & 2.

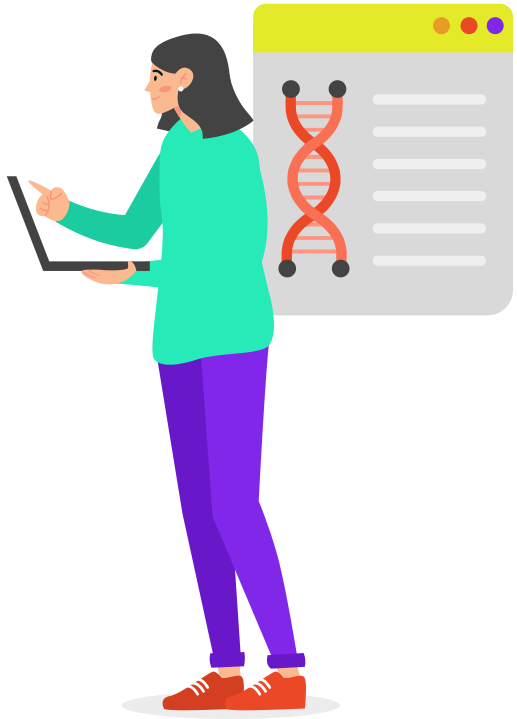
04 For new data points, find the predictions of each decision tree, and assign data points to the category that wins the majority n the new votes.



ML Classification Algorithm

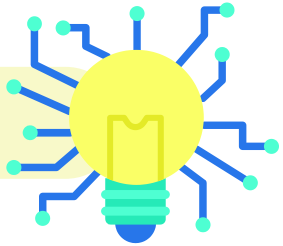
SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.



ML Classification Algorithm

Support Vector advantages vs disadvantages



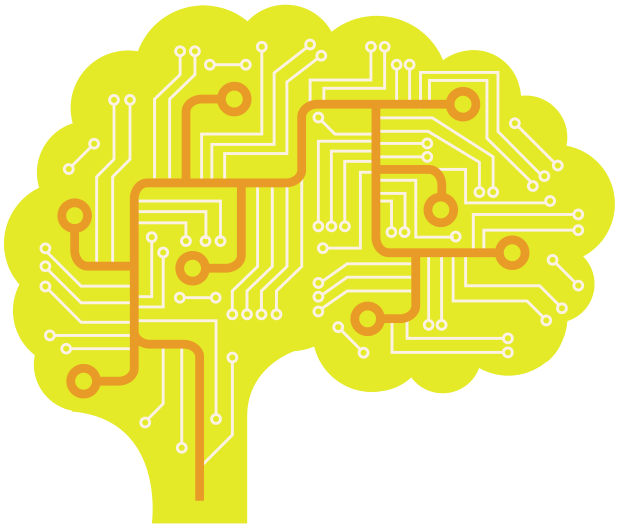
Advantages

- Continuous improvement
- Lots of applications
- Trend identification
- Pattern identification

Disadvantages

- Data acquisition
- Time and space
- Time-consuming
- High error possibilities
- Algorithm selection

ML Classification Algorithm



Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision Tree, there are two nodes, which are the Decision Node and Leaf Node.

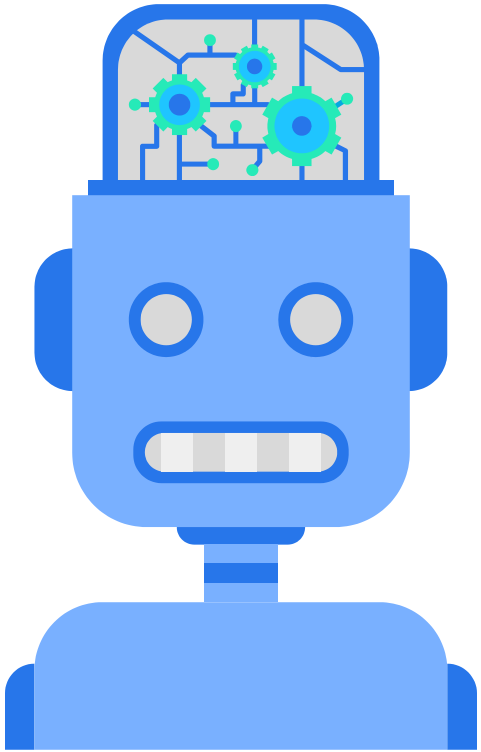
ML Classification Algorithm

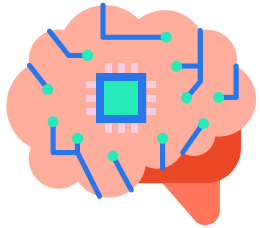
Reasons for choosing decision tree

01 Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand

02 The logic behind the decision tree can be easily understood because it shows a tree-like structure

ML Classification Algorithm





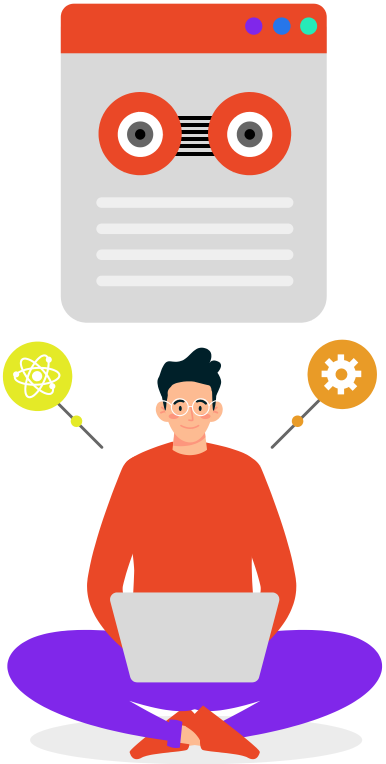
Multiple Linear Regression

Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

ML Classification Algorithm

Applications of Multiple Linear Regression

Effectiveness of Independent variable on prediction:

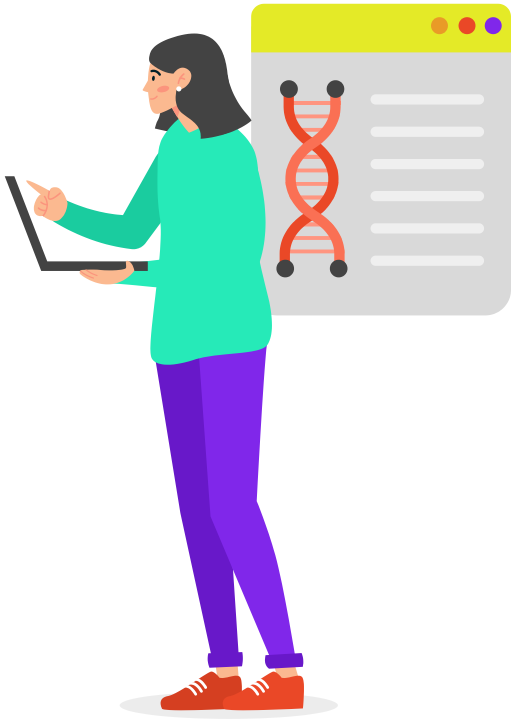


Predicting the impact of changes

Implementation

RESULT AND ANALYSIS

We have implemented 4 different types of algorithm on this data set and analyzed the percentage of score for different types of algorithms. Firstly we have split the dataset into two part. One is train data other is test data. From the data set we have taken 25% data as test data and other 75% data as train data. Then we have implemented algorithms using sklearn. In this dataset we have used Decision tree, Random forest, SVM and Multiple Linear Regression in this dataset. After using these 4 algorithms we get different types of success rate score for each algorithm. The best algorithm according to the success rate is Multiple Linear Regression and Decision tree as we have got 96.07% for MLR and 95.50% for decision tree success rate by using this algorithm.



Model Evaluation and Selection

Predicted success rate 96.07%

Multiple Linear Regression

```
lr = LinearRegression()
lr.fit(X_train, y_train)
r2_score = lr.score(X_test, y_test)
print(f"Training Accuracy Score: {lr.score(X_t
rain, y_train) * 100:.1f}%")
print(f"Validation Accuracy Score: {lr.score(X
_test, y_test) * 100:.1f}%")
```

Predicted success rate 95.50%

Decision Tree Regression

```
dt = DecisionTreeRegressor()
dt.fit(X_train, y_train)
print(f"Training Accuracy Score: {dt.s
core(X_train, y_train) * 100:.1f}%")
print(f"Validation Accuracy Score: {dt.
score(X_test, y_test) * 100:.1f}%")
```

04 Random forest regression

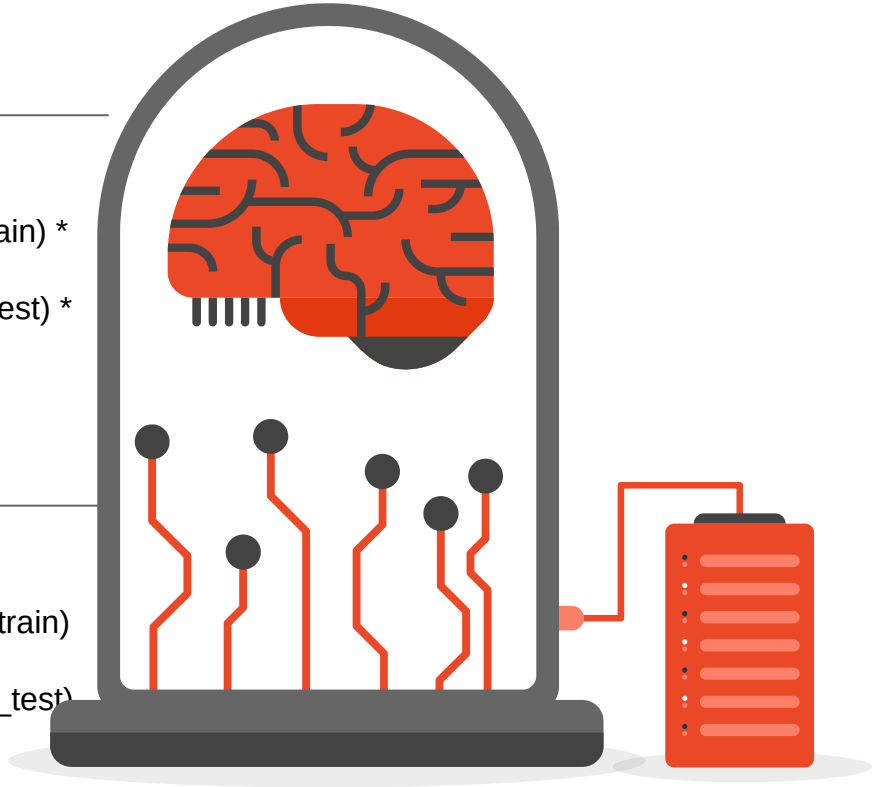
93.40%

```
rf = RandomForestRegressor()
rf.fit(X_train, y_train)
print(f"Training Accuracy Score: {rf.score(X_train, y_train) *
100:.1f}%")
print(f"Validation Accuracy Score: {rf.score(X_test, y_test) *
100:.1f}%")
```

05 Support Vector Machine regression

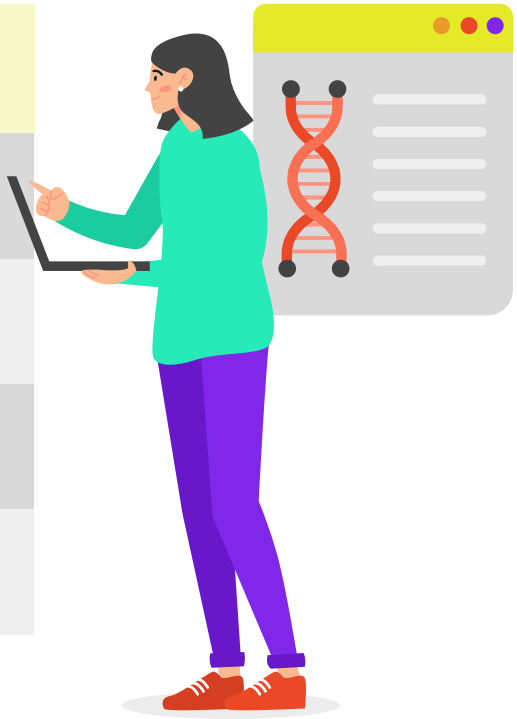
85.30%

```
svr = SVR()
svr.fit(X_train, y_train)
print(f"Training Accuracy Score: {svr.score(X_train, y_train)
* 100:.1f}%")
print(f"Validation Accuracy Score: {svr.score(X_test, y_test)
* 100:.1f}%")
```



Comparison between models

Models and their accuracy scores	
Multiple Linear Regression	96.07%
Decision tree	95.50%
Random forest Regression	93.40%
Support-vector machine	85.30%



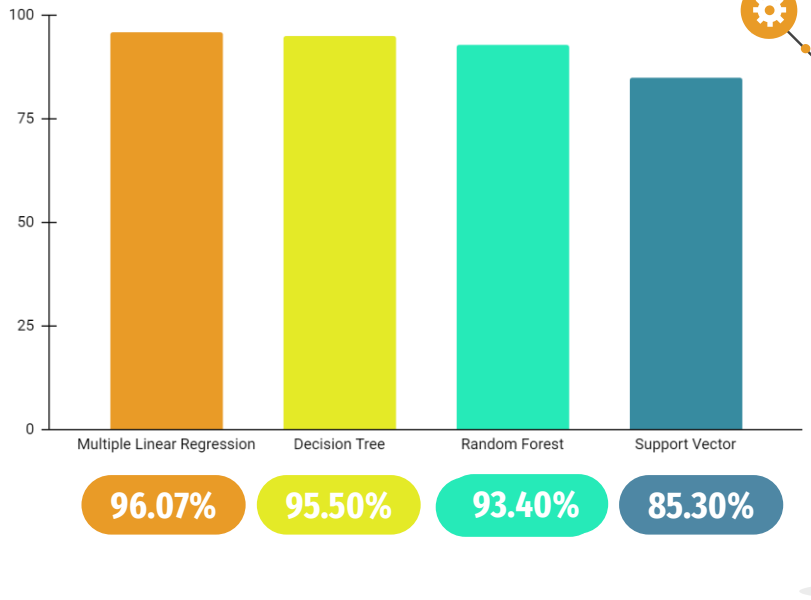
Machine Learning Infographics

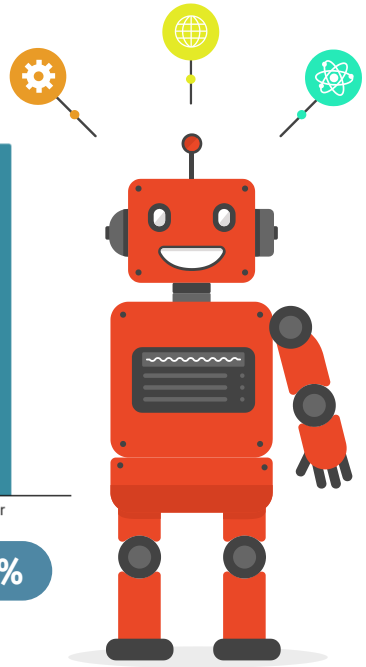
01 Multiple Linear
most efficient for the given data

02 Decision Tree
gives almost the same result as that of MLR however less efficient than that

03 Random Forest
less satisfactory

04 Support vector
least efficient





Conclusion

What makes us to conclude MLR as a best fit?

While working on this dataset we have come across different types of problems and challenges and we have overcome them by learning the solution. We have worked on this type of problem keeping in mind on the usage of different Machine learning algorithm and give benefit to others. This dataset is a real life dataset and we have come to an assumption that if anyone want best result from it, he should take the Multiple Linear Regression in consideration as it has the highest accuracy rate for this types of dataset. If anyone wish to start a startup with his unique idea, following the way generated from the **MLR** will be the best for him. We hope this findings of ours will help others and add some value in machine learning industry.