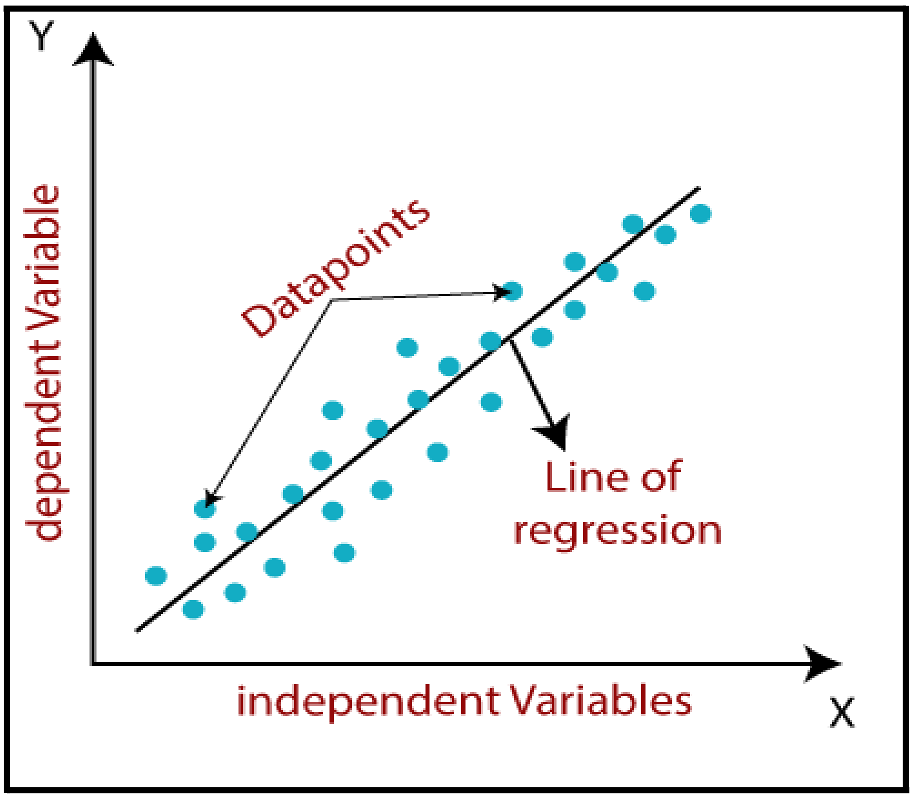
Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as**sales, salary, age, product price,**etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or moreindependent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the valueof the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



**Random Forest Regression Models:**

Random Forest is a popular machine learning algorithm that belongs to the supervisedlearning technique. It can be used for both Classification and Regression problems in ML. Itis based on the concept of**ensemble learning,**which is a process ofcombining multiple classifiers to solve a complexproblem and to improve the performance of the model.

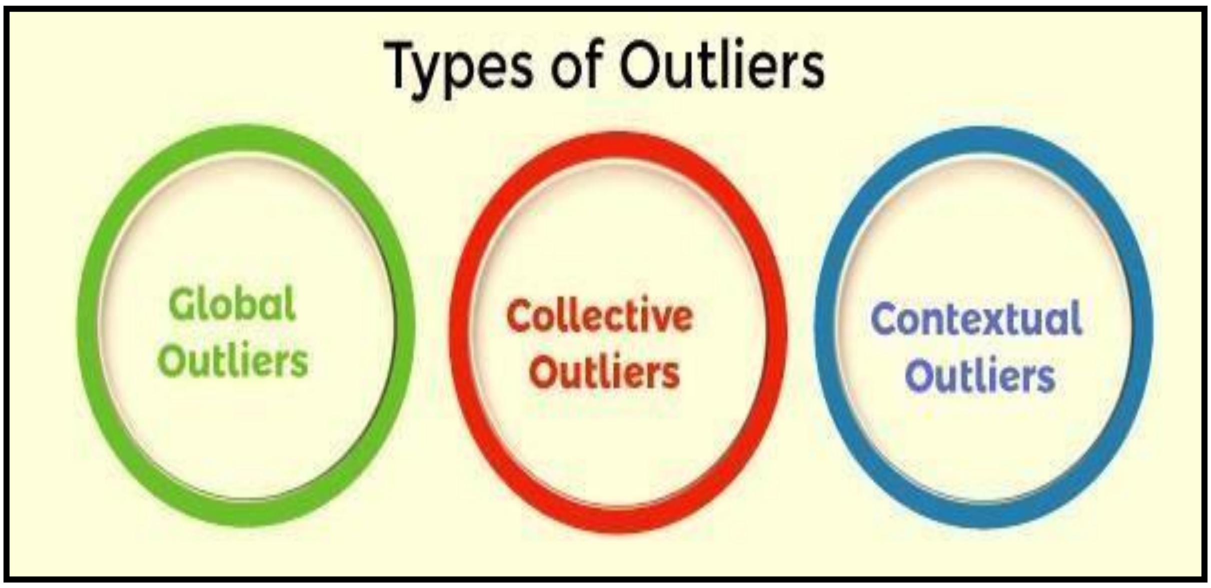
As the name suggests,**"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictiveaccuracy of that dataset."**Instead of relying on one decision tree, the random forest takesthe

prediction from each tree and based on the majority votes of predictions, and it predicts thefinal output.

**The greater number of trees in the forest leads to higher accuracy and prevents theproblem ofoverfitting.**

**Outlier:**

Themajor thing about the outliers is what you do with them. If you are going toanalyze any task to analyzedata sets, you will always have some assumptions based onhow this data is generated. If you ndsome data points that are likely to contain some form of error, then these are definitely outliers, and depending on the context, you want toovercome those errors. The data mining process involves the analysis and prediction of datathat the data holds. In 1969, Grubbs introduced the first definition of outliers.

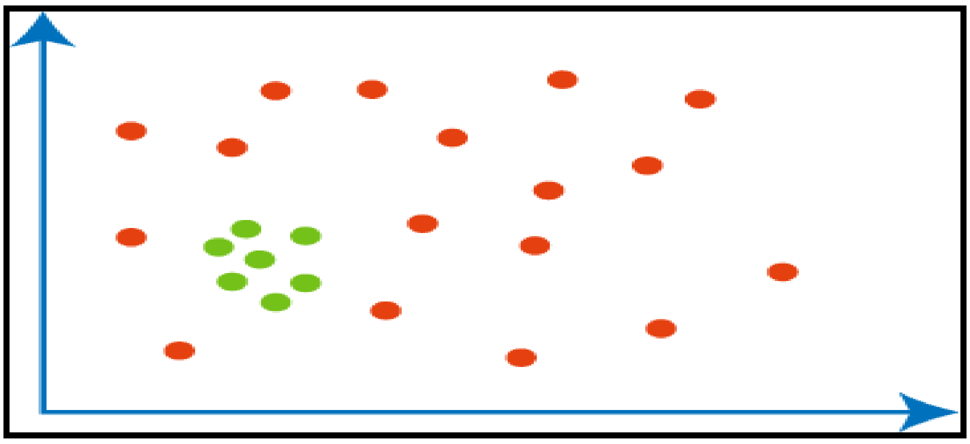


**Global Outliers**

Global outliers are also called point outliers. Global outliers are taken as the simplest form of outliers. Whendata points deviate from all the rest of the data points in a given data set, it is known as the global outlier. In most cases, all the outlier detection procedures are targeted to determine the global outliers. The green data point is the global outlier.

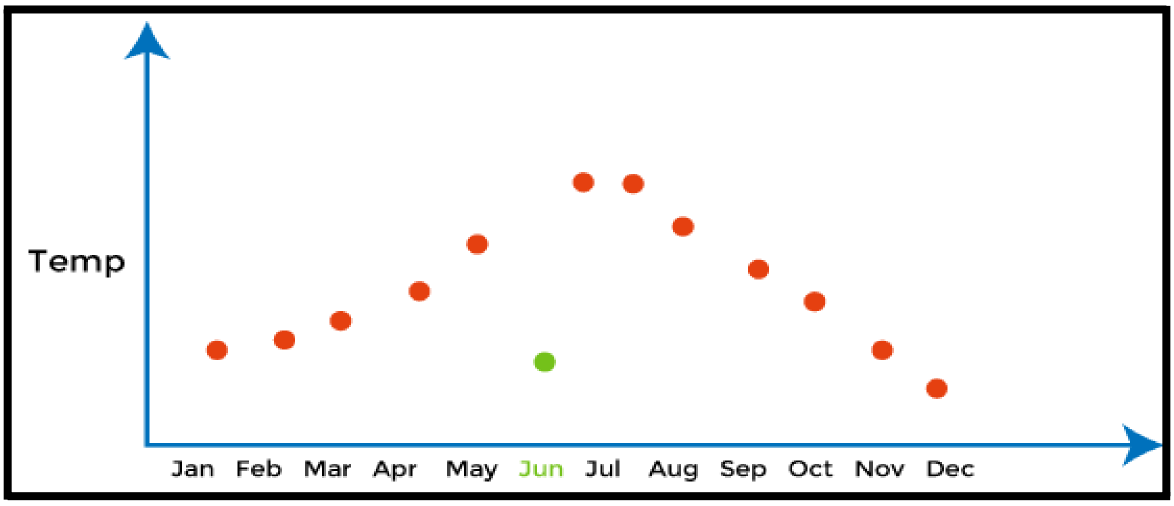
**Collective Outliers**

In a given set of data, when a group of data points deviates from the rest of the data set is called collectiveoutliers. Here, the particular set of data objects may not be outliers, but when you consider the data objectsas a whole, they may behave as outliers. To identify thetypes of different outliers, you need to go through background information about the relationship between the behavior of outliers shown by different data objects. For example, in an Intrusion Detection System, the DOS package from one system to another is taken as normal behavior. Therefore, if this happens with the various computer simultaneously, it is considered abnormal behavior, and as a whole, they are called collective outliers. The greendata points asa whole represent the collective outlier



### Contextual Outliers

As the name suggests, "Contextual" means this outlier introduced within a context. For example, in the speech recognition technique, the single background noise. Contextual outliers are also known as Conditional outliers. These types of outliers happen if a data object deviates from the other data points because of any specific condition in a given data set. As we know, there are two types of attributes of objectsof data: contextual attributes and behavioral attributes. Contextual outlier analysis enables the users to examine outliers in different contexts and conditions, which can be useful in various applications. For example, A temperature reading of 45 degrees Celsius may behave as an outlier in a rainy season. Still, it will behave like a normal data point in the context of a summer season. In thegiven diagram, a green dot representing the low-temperature value in June is a contextual outlier since the same value in December isnot an outlier.



### Haversine:

The Haversine formula calculates the shortest distance between two points on a sphere using their latitudes and longitudes measured along the surface. It is important for use in navigation.

### Matplotlib:

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

### Mean Squared Error;

The**Mean Squared Error (MSE)**or**Mean Squared Deviation (MSD)**of an estimator measures the average of error squares i.e. the average squared difference between theestimated values and true value. It is a risk function, corresponding to the expected value ofthe squared error loss. It is always non − negative and values close to zero are better.

**Conclusion:**

In this way we have explored Concept correlation and implement linear regression andrandomforest regression models.