WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

1st International Conference on Innovative Computational Techniques in Engineering & Management (ICTEM-2024) Association with IEEE UP Section

Supervised Machine Learning: A Review on Regression Technique

Shivangi Verma[1], Chhavi Chaudhary[2]

[1] Graphic Era Deemed University, Dehradun, Uttarakhand, mca.shivangi@gmail.com
[2]Teerthankar Mahaveer University,Moradabad , cvishnoi.9@gmail.com
DOI: https://doi.org/10.55248/gengpi.6.sp525.1911

ABSTRACT:

We know that today Machine Learning is advanced success in several sectors such as intelligent controls, decision making, speech reorganization, natural language, and computer graphics and despite the requirement to evaluate and read data. Machine learning is an fourth developed revolution, scientific field. Today Machine learning is an hot topics in the framework of computing and is widely useful in various application areas. We know that the work of regression model is provided the relation between dependent and independent variables. This research begin the most important concepts of Regression analysis as a numerical process consisting of a set of machine learning methods including data processing and regularization, first start by introducing the supervised Machine learning then we expand inspiration for algorithms used in data splitting. In which we focus on the regression techniques and its methods and in last we conclude the research with an real life applications.

KEYWORDS: MACHINE LEARNING (ML), DEEP LEARNING (DL), SUPPORT VECTOR MACHINE(SVM).

1. INTRODUCTION

Machine learning is the revise of computer algorithms that systems provided the facility to automatically learn and develop from experience. It is generally seen as a sub-field of artificial intelligence that occupy the improvement of algorithms and statistical models that allow computers to improve their performance in tasks during knowledge. The algorithms of Machine learning are used to make decisions automatically without any external parties. At that time, machine learning is used in many applications like speech reorganization, face detection, frauds detection, images processing, Natural language processing, Medical field, Google maps etc. it is used to make assessment based on previous data. By algorithm on the bases of previous data, algorithm to find the data and produces the relationship between variables. There are three types of machine learning, including supervised learning, unsupervised learning and reinforcement learning. Supervised learning are based on continuous data while unsupervised learning based on classification of data. Reinforcement learning provides the best possible pathway that should take in specific circumstances.

ML used the different instances of dataset i.e. is represented by same set of features. We known that feature of ML may be continous, categorical or binary form. if instances are given in same label of data that's learning is known as Supervised learning algorithm(Table 1). in unsupervised learning algorithm instances are given in unlabeled of data[1]. Reinforcement learning [2] provides the best possible pathway that should take in specific circumstances. The training information provides the knowledge system by the exterior parties is in the form of scalar reinforcement learning. Continuous ML applications that can be set up as supervised learning. in this paper we talk about the Regression techniques and it gives output in the form of ordered values and continuous values.

Table 1. Instances values of ordered data

	Dataset values			
Case	Data value 1	Data value 2	Data value n	Classes
1	###	##	###	Bad
2	###	##	##	Good
3	###	#	#	Good
……	……	……	……	……..
……	……	……..	……	……..

Case	Data value 1	Data value 2	Data value n	Classes
1	######	######	####	Bad
2	######	#####	#####	Good
3	#####	###	###	Good
……	…….	……	……	……..
…….	…….	……...	……	……..

We have limited our journal, published books, references and conferences. According to author Dutton [3] provide the review of machine learning and De Montanans [4] also presented the historical view of ML. we know that this single article cannot presented the review of all the regression technique. We all know about ML is a subfield of AI that involves the improvement of algorithm and statistical representation of data that enables computer to improves their performance. According to Tom Michel (1999) "A computer program is said to learn from experience E with respect to some class of tasks T and performance measures P, its performance at tasks T as measured by P improved with Experience E".

Supervised learning is applied while the dataset is in the structure of input variables and output values. The algorithm finds out the mapping function between the input and output. the machine is learn with the input values and equivalent to its output. For example, user takes the input values i.e. image of any animals. Firstly, the machine understands the pictures, together with the color, eyes, shape, and size. Machine takes the input from the dataset and recognized the object and gives desired output related to input which is given by user. These types of method are known as object recognition in supervised machine learning. **The main aim of the supervised learning technique to detect the Risk evaluation, Fraud Detection, Spam filter,** etc. the supervised learning is divided into two types:

1.1. Classification

The output in the form of categorical i.e. yes or no, positive or negative

1.2. Regression

The output in the form of continuous value i.e. price.

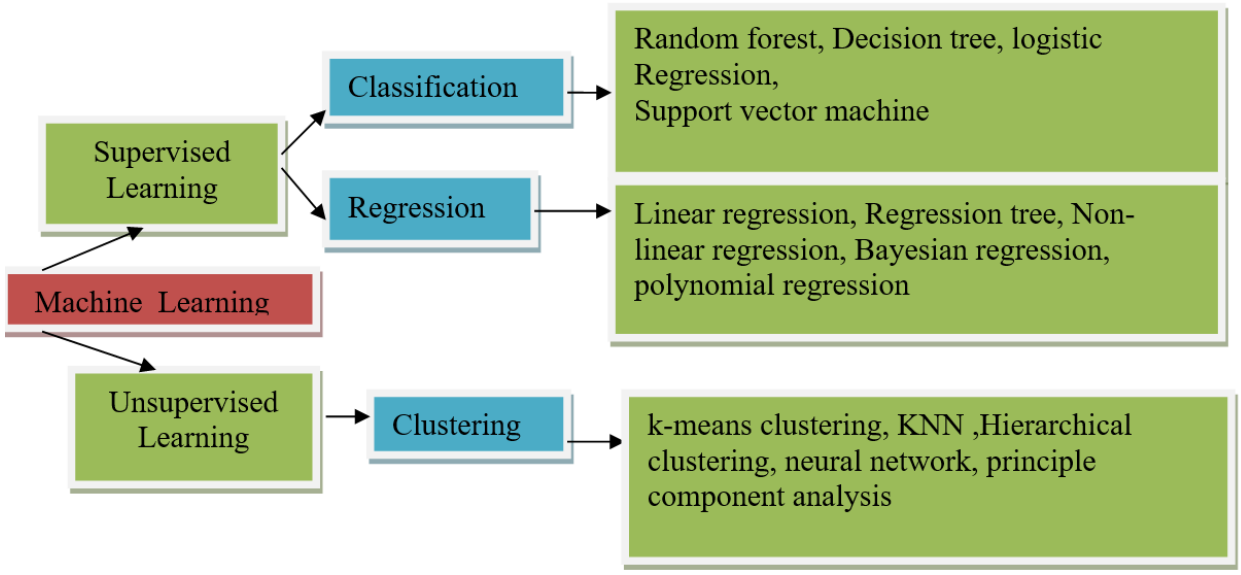

Figure 1. overview of Machine Learning

We know that different types of algorithm are used in machine learning to solve the difficulty of datasets. User will check according to our dataset, which type of algorithm will be useful for our dataset? So in this paper we talk about algorithm which comes under the supervised learning algorithm. In our next section cover issues of supervised learning such as data pre-processing and extract data. In section 3 discuss about regression techniques and statically techniques covers in section 4and section 5 describe the polynomial regression and t last concludes our work.

2. GENERAL ISSUES OF SUPERVISED LEARNING

Machine learning is trained the labeled data or continuous data and on the basic of that data, machine give the output. the main aim of supervised algorithm is to find the mapping function between input variable (x) and output variable(y). Inductive machine learning is the process of learning a set of rules from instances (examples in a training set), or more generally speaking, creating a classifier that can be used to generalize from new instances. The process of applying supervised ML to a real-world problem is described in Figure 1.

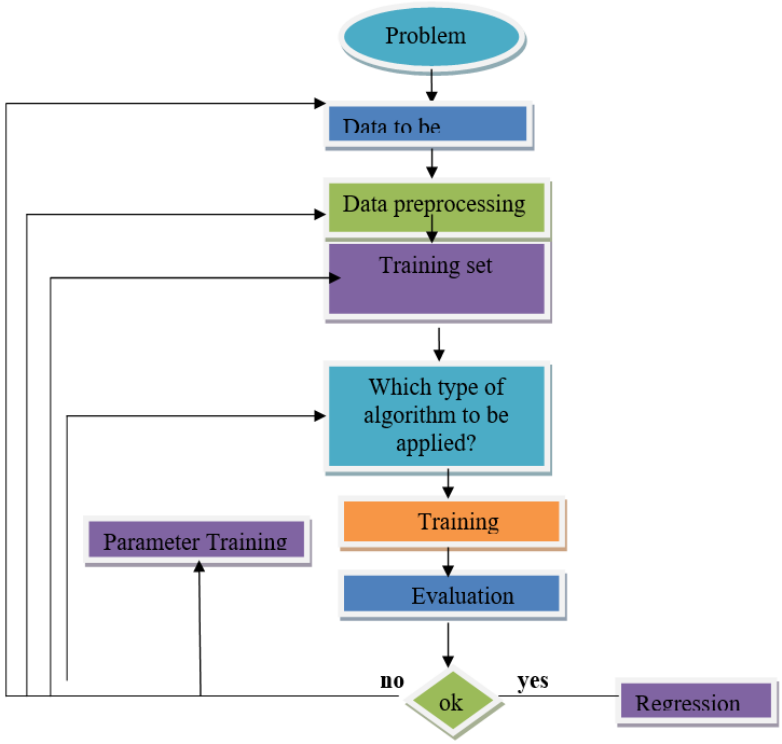

Figure 2. Process of Supervised Machine Learning

The first step of pre-processing data is to collect the meaningful dataset. Then users go through the training, which type of algorithm is applied if there expert is available, then he/she suggest, which type of algorithm is applied. If not then go through the brute force method.[5]. In second steps, the user collects the raw data and applies the algorithm, if desired output is ok then it gives output related to input. it consists of methods for frequently structure a predictive function F:X→Y that maps X to a predict value of Y, given a set of training instance represented by tuples (X_i,Y_i) where Y_i is the output variable and X_i is the vector of input values, which gives the output values according to the datasets. Which types of methods used by users to handle missing values? [6, 7] some steps of supervised learning are:

1) Which types of dataset to be taken?
2) Data collection.
3) Split the training dataset into test data set.
4) Predict the output which comes from the input.
5) Execute the algorithm on the training dataset.
6) If output comes accurate according to giving dataset then our models is accurate.

2.1. Algorithm selection

In this section users check which types of learning algorithm is applied for given dataset. First testing process is judged to be satisfied then classification and regression to be available for routine techniques is used. The classification is based on the prediction accuracy. Different types of techniques are used for data selection. First technique is to split the dataset by using training dataset. Second technique is known as cross-validation; in these techniques to cross the training data set divided into mutual exclusive and finally users check which types of error occur according to our hypothesis. Users can take the training sample dataset with N size, run the two different algorithms in it and estimate the difference between them.[8,9]Supervised learning algorithm is one of the tasks to solve the large scale of data which development based on AI Regression techniques and statistical model. In next section we discuss about the regression techniques which come under the supervised algorithm.

3. REGRESSION METHODS

Regression is a supervised learning algorithm, which finds the correlation between the variables and enables to predict the continuous output based on one or more predictor variable. The equation of regression is:

$$Y=m*X+c \quad \ldots\ldots\ldots\ldots \quad (1)$$

Where, Y=independent variable

M= slope of line

X=dependent variable

it shows a line or curve that passes through all dataset points on the predictors outputs such a way that the vertical distance between data points and regression line is minimum. in regression different types of terminologies are used i.e. Dependent variable, independent variable ,outliers, multicollinearity and under fitting and over fitting [10]. In real world scenario we need some prediction values such as sales prediction, weathercaster prediction etc. for such cases we need to help to predict the continuous or labeled data. Different types of regression techniques are shown in Fig3.

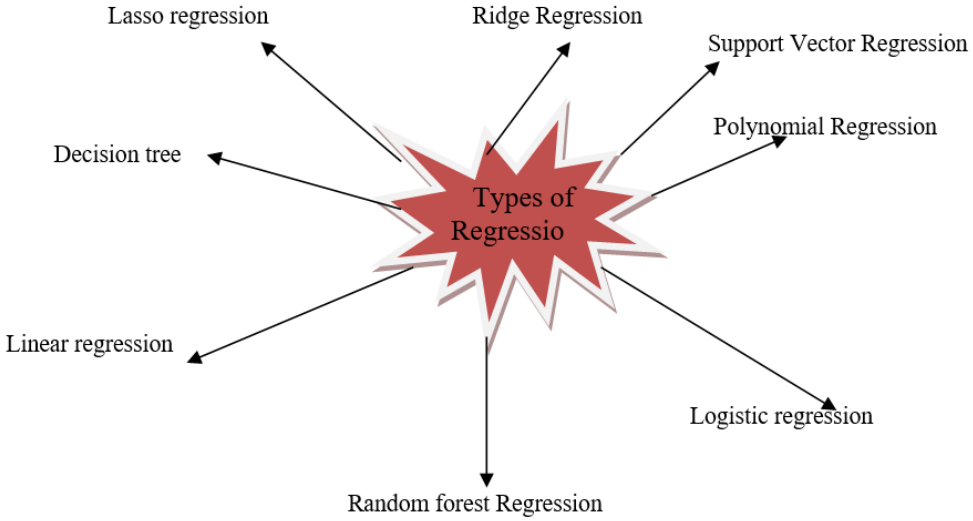

Figure 3. shows the different types of Regression techniques

We knew that regression techniques are provided the relationship between input and output variables. If we have data of multiple variables then we try to find out the variable related to it. For example: we could ask the relationship between people weight, study time, height, test scores etc. so we can explain many transistors the semiconductor industry can pack into circuit.[11] (Moore's Law shown in fig. 4)

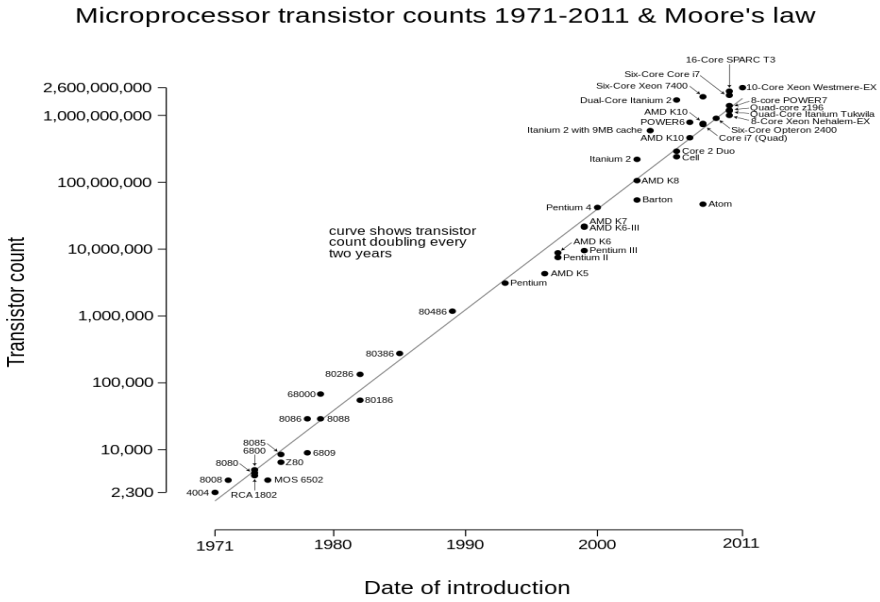

Figure 4. Illustration of Moore's Law ,overtime [11]

3.1. Linear Regression

Linear regression is a statistical regression method, which is used for predictive analysis between single variables. it is very simple algorithm which works on regression and shows relationship between variables. It provides the relationship between dependent variable and independent variable called Linear Regression. if we have one input variable then it is single line regression and if we have multiple values dataset then it is multiple Regression[10]. the equation of Linear regression is:

$$Y=aX+b \text{ …….. (2)}$$

Where Y=independent variable

a&b =linear coefficient

X=dependent variable

3.1.1. Simple Linear regression

This regression provides the linear relationship between dependent and independent variable to predict output in the form of continued values such as sales, age etc. if dependent variable increases on y-axis or independent increases on x-axis then it are called positive linear regression otherwise negative linear regression [10]. it does not mean to show relationship between two variables in straight lines.

$$Y=b_0+b_1 x+€ \text{ …… (3)}$$

Where,

Y=independent variable and X=dependent variable

b_0 , b_1 are parameter and €is error.

3.1.2. Multiple Linear regressions

This regression provides the multiple indepent variables and one dependent variable. It is an extension of linear regression. it contain two variables X_1 and X_2 . Then equation of multiple linear equations is:

$$Y=b_0+b_1 x_1+ b_2 x_2 + b_3 x_{3+} \text{ ….. } +€ \text{ …….(4)}$$

3.1.3. Polynomial Regression

Polynomial regression is a regression algorithm and its provide the relation between dependent and independent variable [12]. Its based on non linear model and its provides the relationship between dependent variable(x) and independent variable(y) as nth degree on polynomial. the linear linear regression is

$$Y=b_0+b_1 x+€ \text{ …… (5)}$$

the multiple linear regression is

$$Y=b_0+b_1 x_1+ b_2 x_2 + b_3 x_{3+} \text{ ….. } + € \text{ …….(6)}$$

Then polynomial regression is

$$Y=b_0+b_1 x_1+ b_2 x_1{}^2 + b_3 x_1{}^3{}_+ \text{ ….. } + b_n x_1{}^n \text{……...(7)}$$

Where dataset are arranged in non linear model where we need the polynomial regression.

3.1.4. Support vector Machine (SVM)

It is very powerful algorithm which comes under linear and non-linear classification as well as regression. it used in variety of tasks such as image classification ,text classification age classification, face detection and gene expression. it adaptable for large dataset and main objective of SVM algorithm is to find the optimal hyperlane in an N-dimensional matrix. it is used to find the closed point between classes.The some terminology of SVM is:

1. **Hyperlane:** it is a decision boundary between to separate data point in different classes.
2. **Support vector**: it is to take decision between hyperlane and margin.
3. **Margin**: it is used to find the distances between support vector and hyperlane.
4. **Kernel**: it is used to map between original input data point into high –value. Common kernel are linear, polynomial, radial basis function and sigmoid.
5. Hinge loss: A typical loss function in SVM is hinge loss.

The SVM kernel function is used to take input of low dimensional and transform into high dimensional space. it convert non-separation problem into separated problems. it is mostly used in non linear separable function. The some equation of kernel are :

$$K(w,b)=w^t x+b \text{ ……Linear equation…….....(7)}$$

$$K(w,x)=(\gamma w^t x+b)^N \text{ ……Polynomil equation……(8)}$$

$$K(w,x)=\exp(-\gamma||x_i-x_j||^n \text{……..Gaussian RBF……} \quad (9)$$

$$K(x_i,x_j)=\tanh(\alpha x_i T x_j+b) \text{…….Sigmoid ……} \quad (10)$$

3.1.5. Decision Tree

It used in interpretable algorithm to predict the models. Its predict the outcomes based on the input data, which suitable for both classification of supervised machine learning i.e. classification and regression. its structured is based on tree where each internal nodes represent the attributes and each leaf represent the decision or prediction. it solves the problem which comes under both regression and classifications[14]. The process of decision tree, where data is split under the each internal nodes criteria based on information gain and gini impurities.

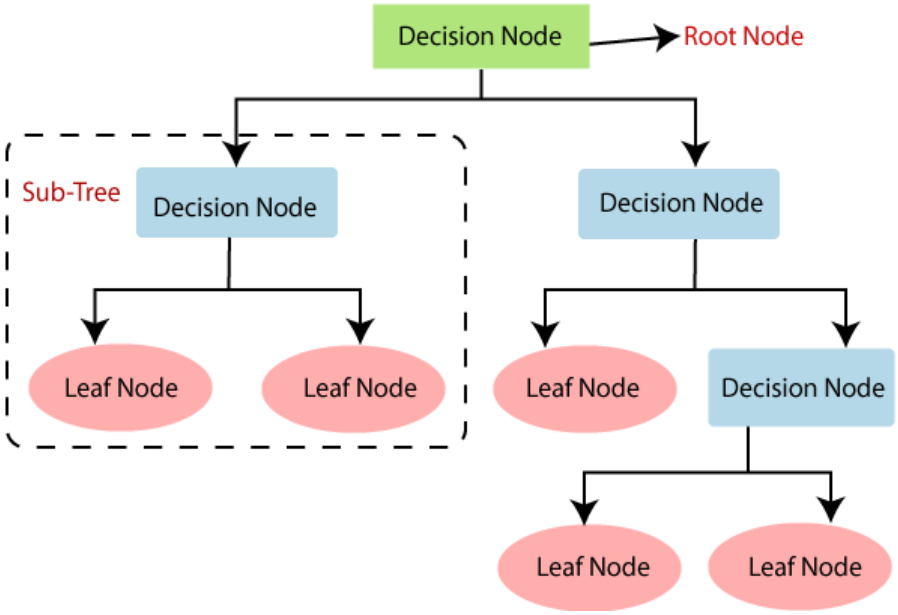

Figure 5. Diagram of Decision Machine Learning [13]

Table 2. Comparison between different supervised learning algorithms according to their features.

Regression models	Linear regression	Multiple regression	Polynomial regression	Support vector machine	Decision tree
Accuracy	xxxx	xxxx	xxxx	xxxx	Xxxx
Simplicity	xxxx	xxxx	xxxx	xxxx	Xxxx
Robust time	xx	xx	xxxx	xxxx	Xxxx
Large dataset	xx	xx	xx	xxxx	Xxxx
Model interpretation	xxxx	xxxx	xx	xxxx	Xxxx
Binary or continuous attributes	continue	continue	Not continue	Continue	not continue

4. REGRESSION APPLICATION IN COMPANY

The real estate group is one of the leading research projects focusing on modern financial side, for its considerable implication on relevant industries and fields such as floor water tank. Constructing a model to predict the price of real estate has been a tough topic, but now with the help of modern machine learning techniques and finding the optimal solution became easily in our supply and with high accuracy. Our company is going to appoint a new applicant. The applicant has give their previous salary slip i.e. 160K per annum and the company owner have to check whether he/she is telling the truth or lie. So we take the dataset from the previous company in which the salaries of the top 10 place are mentioned with their rank. So, we found the **non-linear relationship between the Place levels and the salary**. in this section we solve the problem by apply the linear regression and polynomial regression. We used the linear and polynomial regression to prepare the model that based on the dataset of company.

we tried to apply a supervised machine learning approach to solve the problem discussed above, and we use a linear regression and polynomial regression algorithm to prepare a model support on a training dataset of salary taking to different position in company as showed in table 1.

Problem Description

There is a alanknanda company, which is going to appoint a new applicant. The applicant has give their previous salary slip i.e. 160K per annum and the company owner have to check whether he/she is telling the truth or lie. So we take the dataset from the previous company in which the salaries of the top 10 place are mentioned with their rank. So, we found the **non-linear relationship between the Place levels and the salary**. in this section we

solve the problem by apply the linear regression and polynomial regression. We used the linear and polynomial regression to trained the model that based on the dataset of company.

Table 3. Dataset of company

Position	level	Salary
business	1	40000
junior manager	2	50000
senior manager	3	55000
Manager	4	65000
country manager	5	70000
region manager	6	75000
Partner	7	80000
senior partner	8	85000
C level	9	90000
CEO	10	95000

We perform preprocessing of data steps by using Python, we require importing some predefined Python libraries. These libraries are used to present some particular occupation i.e . numpy, matplotlib.pyplot, pandas. we got to our trained model with the help of Jupyter which is web-based interactive development environment for coding, import data and notebooks. it provides the users interface between client and servers. we take only 20% information of clients and apply the linear regression it gives the [76984.84848485]and polynomial regression with degree 3 i.e. [77712.12121212]

The used Code of linear regression is

```
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd
data=pd.read_excel("C:\\Users\\Admin\\Documents\\shivangi2\\book13.xlsx")  // it takes data from excel
data
…..
…….
……
……
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/5, random_state = 0)
from sklearn.linear_model import LinearRegression   // used sklearn to import the function of it i.e. linear regression.
regressor = LinearRegression()
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
```

output of linear regression is:

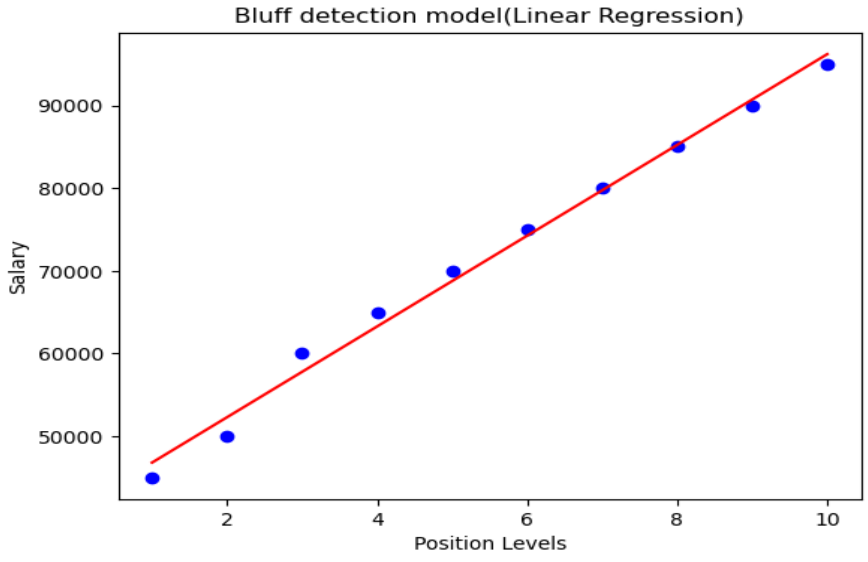

Figure 6. shows the graph between salary and position level by using linear regression.

when we take about polynomial regression,first to used polynomialfeatures class of preprocessing library.then we import the libraby of it.we take the degree 3, it gives more accurate results.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/5, random_state = 0)
from sklearn.linear_model import LinearRegression   // used sklearn to import the function of it i.e. linear regression.
regressor = LinearRegression()
regressor.fit(X_train, y_train)

mtp.scatter(x,y,color="blue")
mtp.plot(x, lin_reg_2.predict(poly_regs.fit_transform(x)), color="red")
mtp.title("Bluff detection model(Polynomial Regression)")
mtp.xlabel("Position Levels")
mtp.ylabel("Salary")
mtp.show()
```

the output of polynomial regression is:

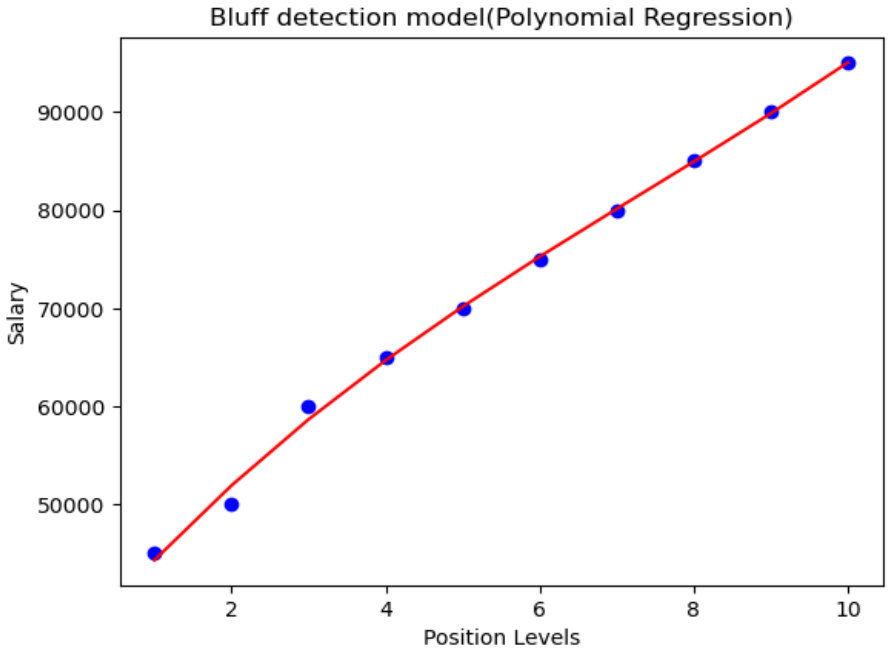

Figure 7. shows the graph between salary and position level by using polynomial regression.

5. CONCLUSION

In this research, we went through the key conception of ML and especially supervised learning algorithm (Regression).in addition we discuss about the linear regression, multi-regression, decision tree, baye's theory and polynomial regression. But in our case we study of real life regression application on water tank company (alanknanda).the whole data splitting, data preprocessing and coded algorithm .this model shows that employee tells our previous salary is accurate or not. We wish that we can simplify the concept of Regression in Supervised machine learning, which is one hottest topic of todays.

REFERENCES:

[1]An, A., Cercone, N. (1999), Discretization of continuous attributes for learning classification rules. Third Pacific-Asia Conference on Methodologies for Knowledge Discovery & Data Mining, 509-514.

[2]Barto, A. G. & Sutton, R. (1997). Introduction to Reinforcement Learning. MIT Press.

[3] Dutton, D. & Conroy, G. (1996), A review of machine learning, Knowledge Engineering Review 12: 341-367

[4]De Mantaras & Armengol E. (1998). Machine learning from examples: Inductive and Lazy methods. Data & Knowledge Engineering 25: 99- 123.

[5]Zhang, S., Zhang, C., Yang, Q. (2002). Data Preparation for Data Mining. Applied Artificial Intelligence, Volume 17, pp. 375 - 381.

[6]Batista, G., & Monard, M.C., (2003), An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence, vol. 17, pp.519-533.

[7] Hodge, V., Austin, J. (2004), A Survey of Outlier Detection Methodologies, Artificial Intelligence Review, Volume 22, Issue 2, pp. 85-126

[8] Dietterich, T. G. (1998), Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10(7) 1895–1924.

[9]Nadeau, C. and Bengio, Y. (2003), Inference for the generalization error. In Machine Learning 52:239– 281.

[10]S. Haidara, **A L Jamil,** Machine Learning – Regression , Content uploaded by Jamil Antone Layous,jan2022.

[11]http://watson.latech.edu/book/future/futureMoores1.html

[12]Ostertagova E, Modelling using polynomial regression, Elsevier, Proceeding Engineering, 2012, 500-506.

[13]https://static.javatpoint.com/tutorial/machine-learning/images/decision-tree-classification-algorithm.png

[14] Kotsiantis.S.B., Supervised Machine Learning: A Review of Classification Techniques, Information 31,(2007) 249-268 249.

Authors

Shivangi Verma, Assistant Professor in ShriRam Institute of Management and technology,Kashipur.i am pursuing PhD from Graphic Era Deemed University and Mtech (CSE) from KIET,Gzb. I have written lots of papers one of them in springer and IEEE.

Chhavi Chaudhary, , Assistant Professor in ShriRam Institute of Management and technology,Kashipur.i am pursuing PhD from TMU, Moradabad.