IRAQI Academic Scientific Journals

العراقية المجلات الأكاديمية العلمية

Al-Kut University College Journal

ISSN (E) : 2616 - 7808 II ISSN (P) : 2414 - 7419

www.kutcollegejournal.alkutcollege.edu.iq

k.u.c.j.sci@alkutcollege.edu.iq

Vol. 9 , Issue 1 , June 2024

Approaches and Methods for Regression Analysis used in Supervised Machine Learning

Shubar Sahib Jawad AL-KHAIAT [1]

Abstract

Regression analysis is an important and simple technique in machine learning, which entails determining the optimal line that intersects the original data points to determine the strength of the relationship between one or more independent variables and the dependent variable. Multiple linear regression analysis can be used to describe this relationship between variables based on their scores. There are a variety of algorithms used in supervised learning methods. In recent years, a large number of supervised learning methods have been introduced into machine learning. Supervised learning techniques have become a field of scientific research activities and are applied in processing and analyzing various data sets, which is called the regression approach. It has become one of the most critical features of supervised machine learning, with the ability to analyze available data and future predictions. The supervised machine learning method and the regression method of all kinds are the two basic techniques in which we will cover the basic aspects of the topic. Whereas, machine learning is concerned with the field of predictive modeling, reducing model error، or making predictions more accurate than they can be at the expense of interpretability. In machine learning applications, we will use algorithms and replicate their use in various other fields, including statistical science, and how to investigate them. In this field, linear regression has been studied and developed as a model to understand the multiple relationships between variables of inputs and outputs. It is a machine learning algorithm and a statistical algorithm.

Keywords: Machine Learning, Supervised Machine Learning, Simple Linear Regression, Multiple Linear Regression, Polynomial Regression

Affiliation of Author

[1] Institute of Graduate Studies, Department of Statistics, Ondokuz Mayis University, Turkey, Samsun , 55200

[1] shubar.s.jawad@gmail.com

[1] Corresponding Author

Paper Info.

Published: Jun. 2024

مناهج وطرق تحليل الانحدار المستخدمة في التعلم الآلي الخاضع للإشراف

شبّر صاحب جواد الخياط [1]

الخلاصة

تحليل الانحدار هو أحد الأساليب الأساسية في التعلم الآلي ، والذي يستلزم تحديد الخط الأمثل الذي يتقاطع مع نقاط البيانات الأصلية لتأسيس العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة. يمكن استخدام تحليل الانحدار الخطي المتعدد لوصف العلاقة بين متغير استجابة واحد والعديد من المتغيرات المستقلة بناءً على درجاتهم. هناك مجموعة متنوعة من الخوارزميات المستخدمة في طرق التعلم تحت الإشراف. في السنوات الأخيرة ، تم إدخال عدد كبير من أساليب التعلم الخاضع للإشراف في التعلم الآلي. أصبحت تقنيات التعلم الخاضع للإشراف مجالًا لأنشطة البحث العلمي ويتم تطبيقها في معالجة وتحليل مجموعات البيانات المختلفة ، وهو ما يسمى نهج الانحدار. لقد أصبح أحد أهم ميزات التعلم الآلي الخاضع للإشراف ، مع القدرة على تحليل البيانات والتنبؤ بالمستقبل. سنغطي الجوانب الأساسية لتقنيتين أساسيتين: طريقة التعلم الآلي الخاضعة للإشراف وطريقة الانحدار بجميع أنواعها. يهتم التعلم الآلي بمجال النمذجة التنبؤية ، حيث يقلل من خطأ النموذج أو يجعل التنبؤات أكثر دقة ممكنة ، على حساب القابلية للتفسير. في التعلم الآلي التطبيقي ، سنستخدم ونعيد

انتساب الباحث

[1] كلية الدراسات العليا للعلوم، قسم الإحصاء، جامعة أوندوكوز مايس، تركيا، سامسون، 55200

[1] shubar.s.jawad@gmail.com

[1] المؤلف المراسل

معلومات البحث

تأريخ النشر : حزيران 2024

استخدام الخوارزميات من العديد من المجالات المختلفة ، بما في ذلك الإحصائيات واستخدامها لتحقيق هذه الغايات. في حين تم تطوير الانحدار الخطي في مجال الإحصاء ودراسته كنموذج لفهم العلاقة بين المتغيرات العددية للمدخلات والمخرجات ، فقد تم استيراده عن طريق مناهج التعلم الآلي. فهي خوارزمية تعلم الي وخوارزمية احصائية.

الكلمات المفتاحية: التعلم الآلي، التعلم الآلي الخاضع للإشراف، الانحدار الخطي البسيط، الانحدار الخطي المتعدد، الانحدار متعدد الحدود

Introduction

Recently, industries have witnessed the emergence of artificial intelligence and information technology, a field where machines are smart enough to implement many specific features through learning alone. Machine learning (ML) is one of the most important of these areas. Experts liken it to the science that enables computers to learn on their own, without human intervention or informational programming. With so much data, ML is particularly powerful. This big data allows for increasingly accurate and balanced learning while training machine learning machines. These datasets cannot be seen with the naked eye but can be examined by computing devices that use machine learning algorithms. The amount of available data and the ability to process it are increased dramatically, with the advent of big data. The ability of machines to learn and thus obtain accurate, faster, and smarter results has also increased. Machine learning has become particularly well-suited to many problems in which the associations or rules involved may be intuitive, but cannot be easily solved and described by simple logical rules, to specify output values or expected outcomes, but the action that can be taken is based on different, indistinguishable circumstances. Uniquely identified or predicted. Data is a particular problem for some traditional methods, analytics techniques, and manipulations in big data and highly correlated data [1].

Extracting knowledge from data is the technique used by the machine learning system. It is a field that researches artificial intelligence, statistics, and computer science, we can also define it as a statistical learning model or a predictive analytics suite. Machine application has become one of the best learning methods in recent years, and it can be seen everywhere in daily life. In terms of food to be ordered and purchased, recognition of friends in photos, technical recommendations about watching and following movies, dedicated online radio, many websites and advanced smart devices contain machine learning algorithms [2].

The mission of machine learning is to teach computers to do what is natural and beneficial to humans: learning from past scientific experiments. Where these algorithms use information directly from the available data without reference and rely on a previously known formula, and their performance improves better with the increase in the number of samples available for study. Different types of techniques can also be used, including the supervised learning technique, which trains a model on the nature of known input and output data so that it can predict and extrapolate future output, and the unsupervised learning technique, which finds unknown and hidden patterns or underlying structures in Input data and predict it into the receiver [3].

To achieve the effect of prediction as a regression analysis problem that enables humans to find

appropriate solutions and required results from big data through the rapid development of machine learning technology, the results of predictions and inferences with data have become a matter of our daily life. We can use this technology in many areas on a large scale such as weather forecasting, medical diagnosis, economic forecasts, and countries' financial policies. Therefore, the subject of the study was some problems of regression analysis and the way to treat them through the techniques of machine learning algorithms. However, in the real world, there are often very complex internal and external factors in the subjects of regression problems, and different machine learning algorithms affect scalability and predictive performance differently [4].

Items of Research

The details related to the subject of the study included some basic applications and algorithms in supervised machine learning and its different types. Different types of regression methods were identified and applied in the Python program. Some examples include advertising costs for marketing a specific commodity and the dollar value of sales for one of the producing companies, the relationship between the quantity of a specific commodity. And the factors that affect them, such as price and income, price indicators, the cumulative rate, and work experience, as well as the annual number of industrial accidents in a particular factory, and then obtain quick and accurate results in the probability ratios and the transactions and forecasts expected for each example.

Types of Machine learning

Thanks to artificial intelligence, the computer has become more usable and smarter as it enables it to think for itself. One of the most popular sub-fields of artificial intelligence is machine learning, as many researchers believe that the process of intelligence cannot be defined without a learning approach. There are many types of machine learning techniques shown in the figure below: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, evolutionary learning, and deep learning. As shown in figure (1).

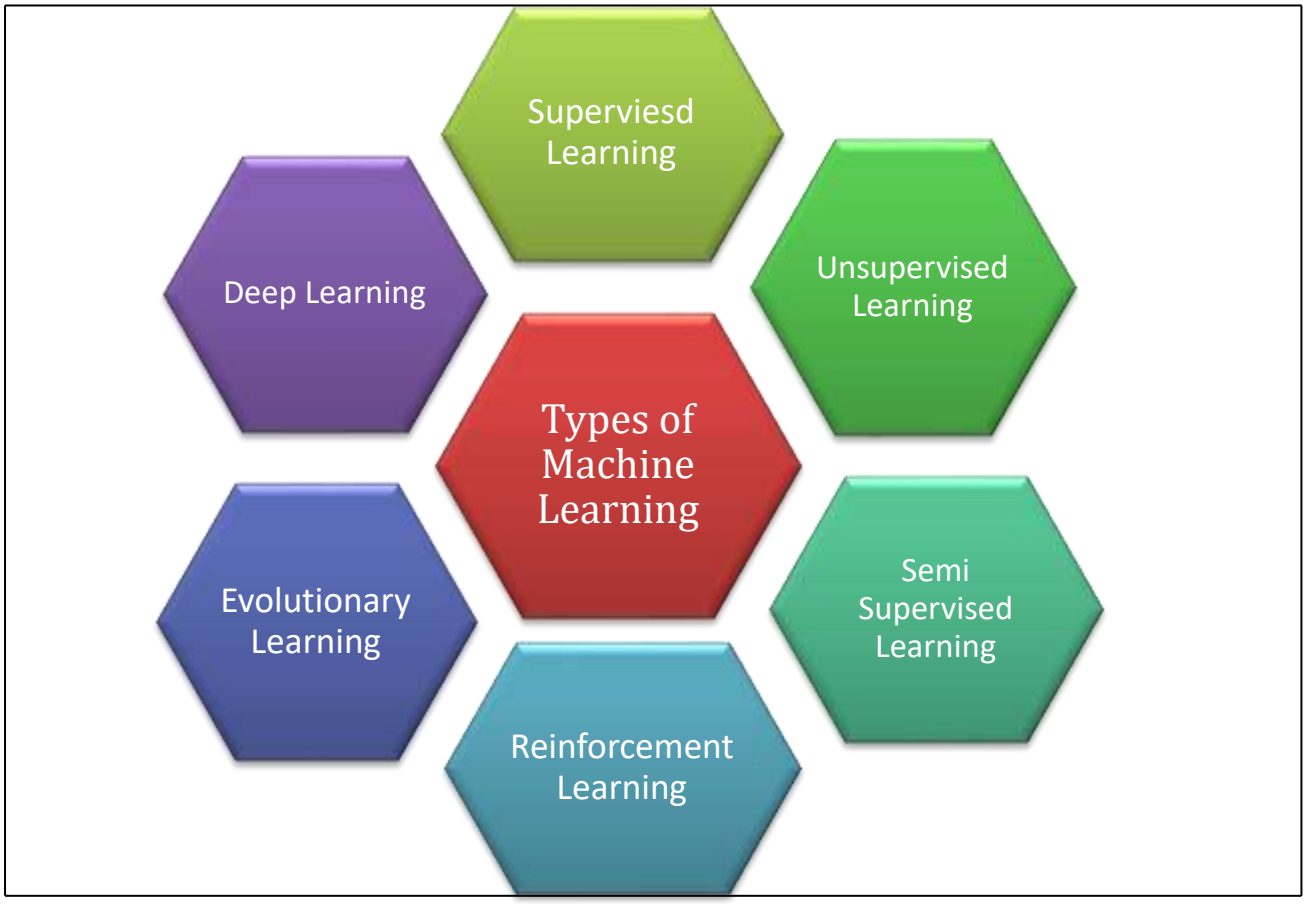

Figure (1): Types of Machine Learning

Supervised Machine Learning

Data mining is one of the most important features and computations in machine learning (ML) applications. Some errors are often discovered during data analysis when trying to find a relationship between variable traits, and this makes it difficult to find appropriate solutions to these problems. Here, machine learning is successfully applied to these problems to improve system efficiency and machine design. With the same set of features, the study data used is represented by the application of algorithms.

When the computer is initialized, trained, and provided with data and output data, by training the computer to recognize links and patterns between data and output so that the model can later predict and recognize new output values, this is called supervised learning, as learning is monitored through output data. Unlike unsupervised learning, where the computer is trained on only feeding data, learning is unsupervised and the output is not given to the data. Through the training process, the computer will build and find connections and common patterns among the data so that it can anticipate and predict new output values for the data. [5].

Two varieties of supervised learning are regression and classification. Classification: giving yes or no predictions, Regression: giving "how much" and "how much" answers [6].

What is Regression Analysis?

The process of representing and estimating the links between the values of the independent

variables and the values of the dependent variable is known as regression analysis. In other words, it means fitting an arithmetic function from a given set of functions to the sample data given the error value of the function. Regression analysis is one of the most important introductory tools for machine literacy used for calming. With regression, we can measure the nature of the study data and try to estimate it through future predictions or keep the original data points. Some real-world exemplifications of retrogression analysis include house price soothsaying, endorsement of the impact of SAT/GRE scores on board approval, soothsaying deals grounded on input parameters, rainfall vaticinators, etc. [7].

The main purpose of regression is to create efficient models for predicting related features of a set of functional variables. Regression problems arise when the output variables are real or continuous (salary, weight, area, etc.). Regression is also defined as a statistical mathematical technique used in applications such as housing, investment, etc. Used to predict the nature of the relationship between a set of independent variables and the value of the dependent variable.

Types of Regression

The study of most regression methods determines the effect of the independent variables on the dependent variable. Different types of regression are used in data science and machine learning. The following are the types of regression:

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression [8].

Linear regression in machine learning

Linear regression is a statistical regression technique used in predictive analytics. This is a straightforward algorithm for showing relationships between continuous variables. It is used to simplify machine learning models and solve regression problems. It is called linear regression because it shows the nature and type of linear relationship between the values of the independent variables (x-axis) and the values of the dependent variable (y-axis). If there is only one input variable (x), this type of regression is called simple linear regression. The second type of regression is called multiple linear regression when there are multiple input variables. The linear regression equation can be mathematically expressed as follows:

$$\boldsymbol{y = a_0 + a_1 * x + \varepsilon} \qquad (\mathbf{1})$$

Here, $\boldsymbol{y}$ = dependent variables (target variables),

$\boldsymbol{x}$ = independent variables (predictor variables),

$\boldsymbol{a_0}$ = regression line intercept (obtainable by setting x = 0),

$\boldsymbol{a_1}$ = the slope of the regression line or coefficient of linear regression,

$\boldsymbol{\varepsilon}$ = random error term (for a good model it would be negligible) [9].

Types of linear regression

Linear regression can be divided into two types of algorithms:

1. **Simple linear regression:** A regression algorithm that models the relationship between a dependent variable and a single independent variable. The relationship exhibited by a simple linear regression model is called a

simple linear regression because it is a linear or sloping straight-line relationship. A simple linear regression algorithm has two main goals:

- A model of the relationship between two variables. For example, income/expenditure ratio, experience, salary, etc.
- Predict new notes. When a single independent variable is used to predict the value of the dependent variable, in the same way, that temperature and annual business investment are used to predict the weather, this linear regression algorithm is called simple linear regression [10].

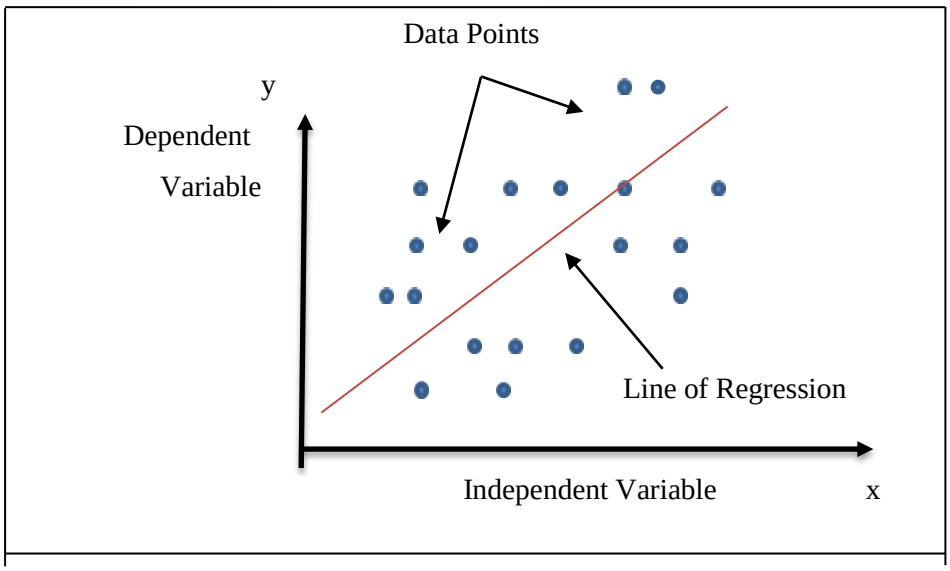

Figure (2): Best Fit Line for a Linear Regression Model

From the figure above, we notice that:

x-axis = independent variable

y-axis = output/dependent variable

Line of Regression = best-fit line for a model

Data Points = study sample data

Material points that fit all problems are plotted as a line of fit called the 'best fitting line'. The purpose of applying the linear regression algorithm is to obtain the best fit line as shown in the above figure [10].

2. **Multiple Linear Regression:** This relationship can be explained by multiple variables by trying to explain the dependent variable by multiple independent variables using multiple regression. The first type is one of the main uses of multiple regression analysis. Determine the dependent variable from multiple independent variables. For example, crop yield quality depends on temperature, precipitation, and other independent variables. The second is to assess the strength of relationships between various variables. For example, see how crop yields change with more precipitation or cooler temperatures.

Multiple regression assumes that each independent variable does not strongly relate to the others. Along with the presumption that each independent variable and each dependent variable are correlated with one another, The result of these linkages is to add a unique regression coefficient to each independent variable, ensuring that the mostsignificant independent factors drive the dependent value. Multiple linear regression should

be employed when several independent factors influence the results of a single dependent variable. When forecasting more complex relationships, this is frequently the case. The multiple regression equation has one y-intercept, several slopes (one for each variable), and multiple slopes. Except for the fact that numerous variables impact the relationship's slope, it is understood as a straightforward linear regression equation [11].

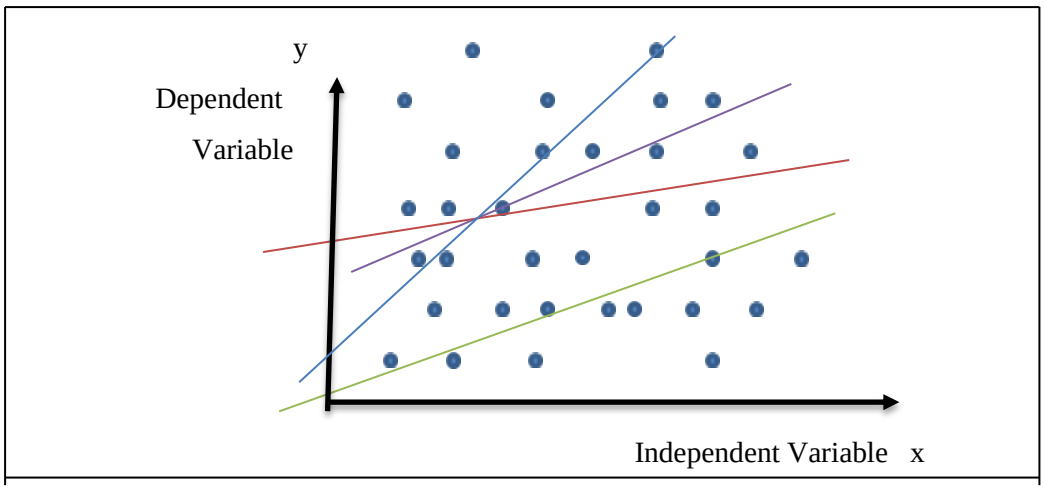

Figure (3): Multiple Linear Regression Model

And the mathematical equation for Multiple Linear regression is as:

$$\boldsymbol{y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + \varepsilon} \quad (2)$$

$\boldsymbol{y}$ = the predicted value of the dependent variable

$\boldsymbol{b_0}$ = the y-intercept.

$\boldsymbol{b_1 x_1}$= the regression coefficient ($\boldsymbol{b_1}$) of the first independent variable ($\boldsymbol{x_1}$).

$\boldsymbol{b_2 x_2}$= the regression coefficient ($\boldsymbol{b_2}$) of the second independent variable ($\boldsymbol{x_2}$).

... = We do the same with many of the independent variables you test.

$\boldsymbol{b_n x_n}$ = the regression coefficient of the last independent variable

ε = the random error

Polynomial Regression: A type of regression analysis known as "polynomial regression" in statistics models the relationship between the independent variable (x) and the dependent variable (y) as a polynomial of degree (n) in (x). The nonlinear relationship between the value of (x) and the conditional mean corresponding to (y), indicated by E (y | x), is the best fit by polynomial regression. The regression function E (y | x) is linear in the unknown parameters that are estimated from the data, which is an issue for polynomial regression even though it fits a non-linear model of data. Because of this, multiple linear regression is a particular instance of polynomial regression [12].

When the relationship between the data is linear, a simple linear regression procedure only works; however, if the data are not linear, linear regression is unable to create the best possible line and fails in such situations. A non-linear relationship is present in some regression models, which is not realistic and does not perform well. To solve this issue, we introduce polynomial regression, which enables the identification of the

curvilinear relationship between the independent and dependent variables [13].

And the mathematical equation for Multiple Linear regression is as:

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n + \varepsilon \quad (3)$$

y = the predicted value of the dependent variable

b_0 = the y-intercept.

$b_1 x$ = the regression coefficient (b_1) of the first independent variable (x).

$b_2 x^2$= the regression coefficient (b_2) of The square of the second independent variable (x).

... = We do the same with many of the independent variables you test.

$b_n x^n$ = the regression coefficient (b_n) of the last independent variable(x^n)

ε = random error term.

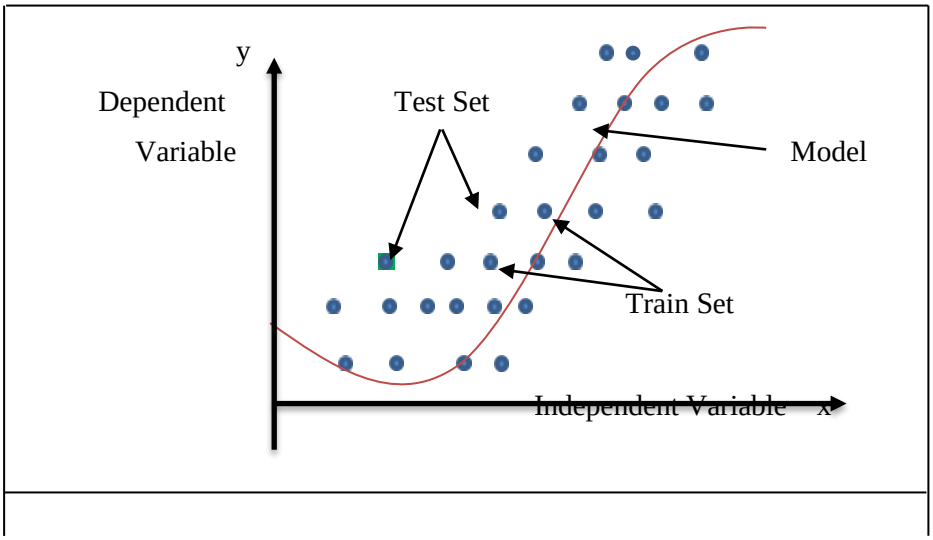

Figure (4): Polynomial Regression Model

There are three types of polynomials:

1. Liner

$$y = b_0 + b_1 x \quad (4)$$

2. Quadratic

$$y = b_0 + b_1 x + b_2 x^2 \quad (5)$$

3. Cubic

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 \quad (6)$$

The cubic polynomial has a degree of 3, the quadratic polynomial has a degree of 2, and the linear polynomial has a degree of 1, as can be shown. The curve more closely matches the data set with higher-degree polynomial equations [14].

Logistic Regression: Applications like machine learning and data mining depend heavily on classification algorithms. Classification issues make up nearly 70% of data science issues. Although there are numerous classification issues, logistic regression is a popular and effective regression technique for addressing binary classification issues. Classification issues like spam detection can be solved using logistic regression. Numerous more instances in the suite include predicting diabetes if a specific client would purchase a specific product or compete with another rival, and whether a user will click on a specific ad link. It is one of the most popular machine learning methods for categorizing two classes, and because it defines and estimates the relationship between a single dependent binary variable and independent variables, it may also

serve as the foundation for any binary classification issue.

A statistical model for forecasting binary categories is called logistic regression. There are only two viable categories because the resulting variable or target is bidirectional. For instance, it can be applied to issues with cancer detection. determines the likelihood that an event will occur and is used as a dependent variable in the probability register. Using the logit function, logistic regression forecasts the likelihood of a binary occurrence [15].

Logistic regression is used to calculate the best-expected weights b_0, b_1, …, b_r when the function p (**x**) is as similar to all of the actual responses as possible y_i, where is the number of observations and i = 1, ..., n. Model training or fitting is the process of selecting the optimal weights based on the feedback that is provided.

The x-function of f (**x**) is the logistic regression function p (**x**):

$$\boldsymbol{p}\,(\boldsymbol{x}) \;=\; \mathbf{1}\,/\,(\mathbf{1} \;+\; \boldsymbol{exp}\,(-\boldsymbol{f}\,(\boldsymbol{x}))) \qquad (\mathbf{7})$$

As a result, p (**x**) frequently approaches 0 or 1. Often The expectation of the chance that the product equals one is represented by the function p (x). As a result, the likelihood that the output is zero is 1 - p (x).

By maximizing the log-likelihood function (LLF) for all observations with i = 1, ..., n, the best weights can typically be found. The equation represents this technique, known as a maximum-likelihood estimation [16].

$$\boldsymbol{LLF} \;=\; \boldsymbol{\Sigma}_i \left(\boldsymbol{y}_i\, \boldsymbol{log}\left(\boldsymbol{p}\,(\boldsymbol{x}_i)\right) + \,(\mathbf{1} - \,\boldsymbol{y}_i)\boldsymbol{log}\left(\mathbf{1} \,-\, \boldsymbol{p}\,(\boldsymbol{x}_i)\right)\right) \qquad (\mathbf{8})$$

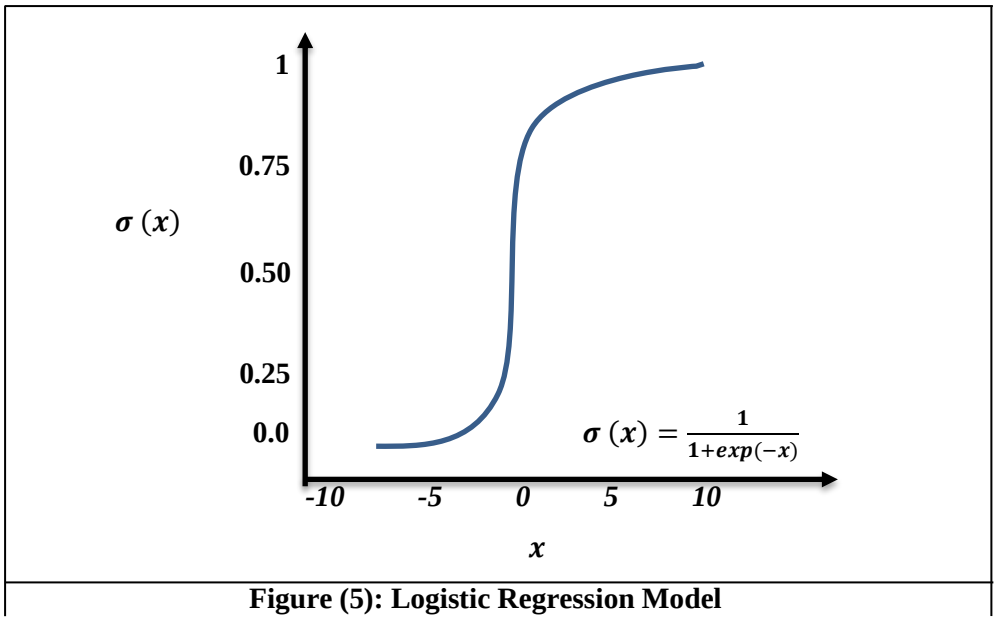

Figure (5): Logistic Regression Model

In the majority of its domains, the sigmoid function has values that are very near to 0 or 1. They can therefore be used in classification models because of this. The picture below displays the natural logarithm (x) form for a few x-variables with x values ranging from 0 to 1 [16].

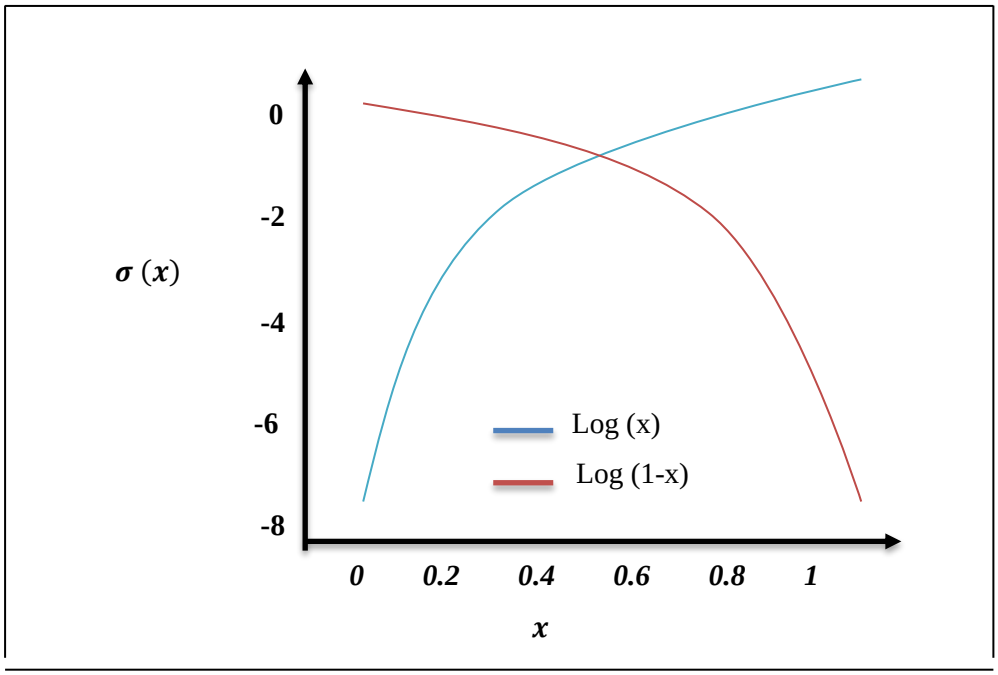

Figure (6): The Natural Logarithm (x) for Some Variables of x

Using Python, we will implement a simple linear regression algorithm and a polynomial model:

Example: a statement for a simple polynomial linear regression using two variables, the independent variable (x) and the dependent variable for a set of random data (y). The objectives of this issue are:

- We are interested in determining whether these two variables are correlated in any way.
- We'll identify the line that fits the data set the best.

How does the dependent variable change by changing the dependent variable? We will implement the two models and compare them using Python.

- Step 1: First, we import statistical packages, functions, and categories to obtain the data of the independent variable and the dependent variable. As shown in the figure below (7):

```
# Step 1: Import Packages, Functions, Classes, and Get Data
import numpy as np
import pandas as pd
import math
import matplotlib.pyplot as plt
import sklearn.metrics as skm
import statsmodels.api as sm
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from scipy import stats
x=np.linspace(1,2,12).reshape(-1,1)
y=np.array([1.5,1.6,1.7,2.37,2.35,2.25,2.3,1.8,1.6,1.79,1.06,0.21])
x,y
```

```
(array([[1.        ],
        [1.09090909],
        [1.18181818],
        [1.27272727],
        [1.36363636],
        [1.45454545],
        [1.54545455],
        [1.63636364],
        [1.72727273],
        [1.81818182],
        [1.90909091],
        [2.        ]]),
 array([1.5 , 1.6 , 1.7 , 2.37, 2.35, 2.25, 2.3 , 1.8 , 1.6 , 1.79, 1.06,
        0.21]))
```

Figure (7): Import Packages, Functions, Classes, and Get Data

- Step 2: A simple linear model was created and how trained for obtaining the values of the coefficients of the variable, the intercept value, the coefficient of determination, the correlation coefficient, and the error percentage (MSE) was calculated, as in Figure 8. where the mean squared error value, MSE = 0.52.

```
# Step 2: Create Linear Model, Train it, and Calculating the Mean Square Error (MSE)
model=LinearRegression().fit(x,y)
slope=model.coef_
print('Slope:',model.coef_)
intercept=model.intercept_
print('Intercept:',model.intercept_)
r2=model.score(x,y)
print('Determination Coefficient:',r2)
r=math.sqrt(model.score(x,y))
print('Correlation Coefficient:',r)
y_pred=model.predict(x)
print('Predict response:',y_pred)
MSE=np.sqrt(skm.mean_squared_error(y,y_pred));
print('MSE:',MSE)
```

```
Slope: [-0.91807692]
Intercept: 3.0879487179487173
Determination Coefficient: 0.23491626044712988
Correlation Coefficient: 0.4846816072919725
Predict response: [2.16987179 2.08641026 2.00294872 1.91948718 1.83602564 1.7525641
 1.66910256 1.58564103 1.50217949 1.41871795 1.33525641 1.25179487]
MSE: 0.51995017430585
```

Figure (8): Create the Model, Train it, and Calculating MSE

Where the values were as follows:

- Slope Value: -0.918
- Intercept Value: 3.088
- Determination Coefficient: 0.2349
- Correlation Coefficient: 0.4847
- Predict Values: [2.17, 2.09, 2.00, 1.92, 1.84, 1.75, 1.67, 1.59, 1.50, 1.425, 1.341, 1.25].
- The error percentage (MSE) = 0.52.

Thus, the simple linear regression equation will be in the following form:

$$y = 3.088 - 0.918 * x \qquad (9)$$

- Step 3: Estimate the data values of the model as shown in the figure below:

```
# Step 3: Evalute the Model
y_pred=model.intercept_+model.coef_*x
print('Predict values:',y_pred)
```

```
Predict values: [[2.16987179]
 [2.08641026]
 [2.00294872]
 [1.91948718]
 [1.83602564]
 [1.7525641 ]
 [1.66910256]
 [1.58564103]
 [1.50217949]
 [1.41871795]
 [1.33525641]
 [1.25179487]]
```

Figure (9): Evaluate the Data Values of the Model

- Step 4: In this step, a simple linear regression model was drawn as in Figure 10, and we notice from the figure that the direction of the regression line was in the negative direction

```
# Step 4: Draw a simple linear regression model
plt.figure(figsize=(10,6))
plt.title('Simple Linear Regression',fontsize=15)
plt.xlabel( 'x',fontsize=15)
plt.ylabel('y',fontsize=15)
plt.scatter(x,y,color='blue')
plt.plot(x,y_pred,color='red')
plt.show()
```

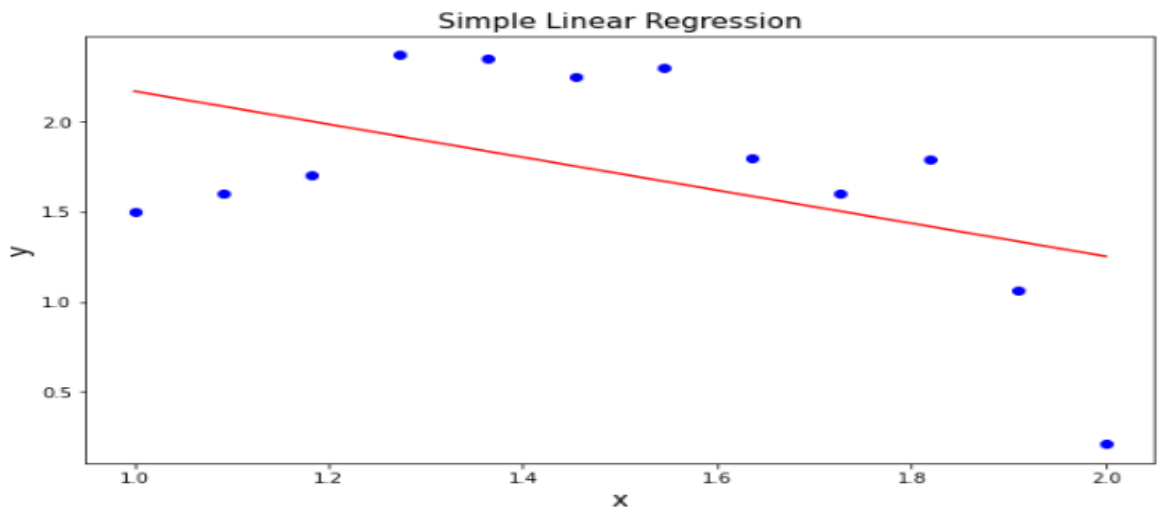

Figure (10): Draw a Simple Linear Regression

- Step 5: In this step, the linear model was transformed into a polynomial model of degree 2, As shown in Figure 11:

```
# step 5: Convert the Linear form to a Polynomial
x_poly=PolynomialFeatures(degree=2).fit_transform(x)
x_poly
```

```
array([[1.        , 1.        , 1.        ],
       [1.        , 1.09090909, 1.19008264],
       [1.        , 1.18181818, 1.39669421],
       [1.        , 1.27272727, 1.61983471],
       [1.        , 1.36363636, 1.85950413],
       [1.        , 1.45454545, 2.11570248],
       [1.        , 1.54545455, 2.38842975],
       [1.        , 1.63636364, 2.67768595],
       [1.        , 1.72727273, 2.98347107],
       [1.        , 1.81818182, 3.30578512],
       [1.        , 1.90909091, 3.6446281 ],
       [1.        , 2.        , 4.        ]])
```

Figure (11): Convert the Linear form to a Polynomial

- Step 6: The following set of values was obtained:
  - Slope Value: [0, 15.68, -5.53]
  - Intercept Value: -8.814
  - Determination Coefficient: 0.89
  - Correlation Coefficient: 0.94
  - Predict Values: [1.33, 1.71, 1.99, 2.18, 2.28, 2.29, 2.20, 2.03, 1.76, 1.40, 0.95, 0.41]
  - and the value of MSE = 0.194 was calculated, which was the lowest value when compared with its other value in the simple linear regression model, which indicates the superiority and quality of the polynomial regression model. As shown in Figure 12:

```
# Step 6: Create Polynomial Model and Calculate MSE
model_poly=LinearRegression().fit(x_poly,y)
poly=PolynomialFeatures(degree=2).fit_transform(x_poly,y)
slope=model_poly.coef_
print('Slope:',model_poly.coef_)
intercept=model_poly.intercept_
print('Intercept:',model_poly.intercept_)
r2=model_poly.score(x_poly,y)
print('Determination Coefficient:',r2)
r=math.sqrt(model_poly.score(x_poly,y))
print('Correlation Coefficient:',r)
y_pred=model_poly.predict(x_poly)
print('Predict response:',y_pred)
MSE=np.sqrt(skm.mean_squared_error(y,y_pred))
print('MSE:',MSE)
```

```
Slope: [ 0.         15.67711538 -5.53173077]
Intercept: -8.813653846153784
Determination Coefficient: 0.8927685833283914
Correlation Coefficient: 0.9448643200631461
Predict response: [1.33173077 1.70543706 1.98770979 2.17854895 2.27795455 2.28592657
 2.20246503 2.02756993 1.76124126 1.40347902 0.95428322 0.41365385]
MSE: 0.19465621030367936
```

Figure (12): Create Polynomial Model and Calculate MSE

Thus, the quadratic regression equation becomes:

$$y = -8.81 + 15.68x - 5.53x^2 \qquad (10)$$

- Step 7: In the last step, we drew a polynomial regression model, and we note the curvature of the regression line and its fit with the data values. This indicates that we obtained the best regression model, and the degree of the model increased. This is shown in Figure 13 below:

```
# Step 7: Draw a Polynomial regression model
plt.figure(figsize=(10,7))
plt.title('Polynomial Regression',fontsize=15)
plt.xlabel( 'x',fontsize=15)
plt.ylabel('y',fontsize=15)
plt.scatter(x,y,color='b')
plt.plot(x,y_pred,color='r')
plt.show()
```

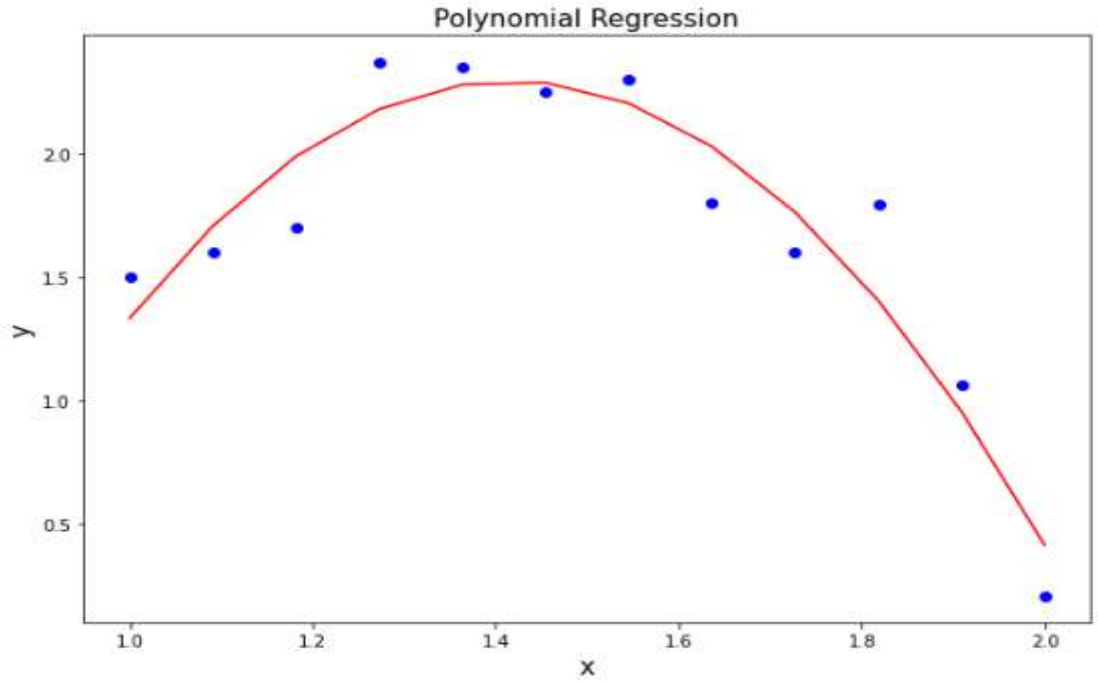

Figure (13): Polynomial Regression Curve Model

Results of Program

Results and interpretive analyses were used to test regression hypotheses and compare the accuracy of different variables. The following graphs and statistics are used as part of the explanatory analysis:

- Scatterplots and scatterplot matrices.
- Regression equations and predictions for new observations.
- Modified coefficients of determination R2 and R2.
- Regression coefficient and intercept coefficient values.

Interpretive analysis should start with the selection of independent variables and the construction of a regression model. One of the main assumptions that must be satisfied is that the model must be linear. It is possible to evaluate the linearity between dependent and independent variables using scatterplot matrices.

After building a regression model, several statistical outputs are available, including the regression equation and R2 value. Once the

regression model is built, we have the necessary outputs, tables, and graphs.

Understanding how each independent variable affects the dependent variable and predicted value is made possible by the regression equation, including the regression coefficient for each independent variable. To assess the relative impact of each independent variable on the dependent variable, the propensity values can be compared; the greater the change in the slope value from zero (whether positive or negative), the greater the effect. By entering values for each independent variable, regression equations can also be used to forecast the value of a dependent variable.

The coefficient of determination, represented by R2, measures the suitability of modeling the regression equation for the actual data points. The value of R2 is a number between 0 and 1. The closer to 1, the more accurate the model. R2 values represent a complete model.

Conclusions

Researchers use several statistical techniques to help understand the nature of the study data. Business and organization leaders can make better judgments through the use of linear regression techniques. Organizations that collect a lot of data may use linear regression to manage reality more effectively rather than relying on intuition and experience. They can turn a lot of raw data into useful knowledge.

In addition, other types of linear regression can be used to highlight patterns and relationships that co-workers may have previously noticed and assumed they already knew. We can detect unique buying patterns on certain days or at certain times, for example, by analyzing sales and purchases of certain items. Entrepreneurs can take advantage of the insights from regression analysis to predict when the demand for their products will be highest.

References

[1] Padala, V. S., Gandhi, K., and Dasari, P. (2019). Machine learning: the new language for applications. IAES International Journal of Artificial Intelligence, 8(4), 411.

[2] Müller, A. C., and Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".

[3] Ciaburro, G. (2017). MATLAB for machine learning. Packt Publishing Ltd.

[4] Huang, J. C., Ko, K. M., Shu, M. H., and Hsu, B. M. (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. Neural Computing and Applications, 32(10), 5461-5469.

[5] Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.

[6] Fatima, M., and Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.

[7] Aqeel, A. (2021) A beginner's Guide to Regression Analysis in machine learning, https://towardsdatascience.com/a-beginners-guide-to-regression-analysis-in-machine-learning-8a828b491bbf.

[8] Waseem, M. (2022, January 5). How to Implement Linear Regression for Machine Learning? Edureka.

https://www.edureka.co/blog/linear-regression-for-machine-learning/.

[9] Gawali, S. (2021) Linear regression algorithm to make predictions easily, Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/

[10] Kanade, V. (2022) What is linear regression? types, equations, examples, and best practices for 2022, What is Linear Regression? Examples & Best Practices | Spiceworks 1. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/.

[11] Blokhin, A. (2022) Linear vs. multiple regression: What's the difference? Investopedia. Investopedia. https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp.

[12] Aden, H. (2020) Polynomial regression: The only introduction you'll need. https://towardsdatascience.com/polynomial-regression-the-only-introduction-youll-need-49a6fb2b86de.

[13] Agrawal, R. (2022) Polynomial regression: What is polynomial regression, Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/all-you-need-to-know-about-polynomial-regression/.

[14] Polynomial regression: Everything you need to know! (2022, August 9). Voxco. https://www.voxco.com/blog/polynomial-regression-everything-you-need-to-know/.

[15] Navlani, A., 2021. [online] Available at: https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python.

[15] Logistic Regression in Python. (2022, September 1). https://realpython.com/logistic-regression-python/.