The role of statistics in machine learning regression models

Bamba Gueye[1,⋆] and Laure Gouba[1,2,⋄]
[1] *African Institute for Mathematical Science (AIMS-Senegal)*,
Km2, Routes de Joal (Centre IRD), Mbour-Thies BP 1418, Senegal
[2] *The Abdus Salam International Centre for Theoretical Physics (ICTP)*
Strada Costiera 11, I-34151 Trieste Italy.
⋆*bamba.gueye@aims-senegal.org*; ⋄ *laure.gouba@aims-senegal.org*

April 17, 2025

Abstract

In this paper, we discuss the role of statistics in simple linear regression, multiple linear regression, and logistic regression. Python has been used to implement the algorithms in these models.

1 Introduction

In today's digital era, the field of machine learning has become increasingly popular. Data plays a crucial role similar to that of oil in the past, and machine learning serves as the driving force behind this data-centric world. The significance of machine learning cannot be emphasized enough given its pivotal role in modern technology. However, many people use the technology or want to become experts in this field without fully comprehending the underlying concept. It's a mistake to see some students solely relying on libraries for learning about machine learning, as it doesn't provide a thorough understanding. A solid grasp of mathematics not only allows us to better comprehend existing algorithms but also empowers us to develop new models and more efficient techniques.

One method for instructing computers to learn from data is through the use of machine learning. Machine learning involves teaching computers how to learn from data, aiming to develop algorithms capable of generating predictions or insights without explicit programming. This field is grounded in mathematical principles [1]. Mathematics provides the necessary basics to build an algorithm for the machine-learning process. We present a simple linear regression model for predicting salaries, a multiple linear regression model for predicting graduate admissions, and a logistic linear regression model for predicting whether a student will pass or fail an exam by using the logistic regression techniques. The role of statistics in these models is discussed.

Statistics is essential for concluding evidence [2]. It involves methods for collecting, presenting, analyzing, and interpreting numerical data. In Machine Learning, statistics play a vital role in handling large amounts of data and are crucial for an organization's progress and success. They are crucial for understanding insights and serve as the foundation for further analysis. Techniques like condensation, summarization, and concluding use methods like central tendencies, dispersion, skewness, kurtosis, correlation, regression, and others.

In this work, we use the Statsmodels library for the analysis of the results of prediction. After the prediction of models, we must have some error between prediction and the actual value in the dataset if we consider that our model are high quality. In linear or multiple regression, the errors between $\hat{y}$ and y are typically measured using the Euclidean norm

$$error = (y - \hat{y})^2. \tag{1}$$

Each prediction comes with an error, so we have **n** errors. The residual sum of squares is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{2}$$

where n is the total number of observations, y_i is the observed value for observation i, $\hat{y}_i$ is the predicted value by the model for observation i.

The goal is to minimize the difference between the predicted and the actual value with the smaller residuals indicating better accuracy. We use the sum of the squares since the residuals can be positive and negative, their sum does not accurately reflect the true error [3].

The paper is organized as follows. In section 2, we present the simple linear regression model, followed by the multiple linear regression model in section 3, and the logistic regression in section 4. The conclusion is given in section 5.

2 Simple linear regression model

Machine learning includes the supervised algorithm known as linear regression. This method performs regression tasks and predicts targets based on the independent variables. The primary focus of linear regression is to identify the relationship between variables and make forecasts, making it applicable in various scenarios [4].

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{3}$$

The linear equation (3) represents a simple linear relationship, where the dependent variable y is expressed as a function of the independent variable x, β_1 is the slope, β_0 is the intercept and ϵ is an error.

2.1 Example of a model to predict salaries

Let's consider an example of using a model to predict salaries with regression techniques in machine learning. We use linear regression to build our model, as it provides the best fit with the training dataset. To do this, we use an experience dataset, apply linear regression, and assess the accuracy and error. Additionally, we intend to use a set of random test cases to observe the predicted salary values [5].

2.1.1 Data loading and preparation

This dataset contains two columns Years of Experience and Salary of some unknown employees. The primary objective is to accurately calculate and predict employees salaries within a specific area using simple linear regression. Viewing the dataset, we obtain the following scatter plot in Figure 1.

	Unnamed: 0	YearsExperience	Salary
0	0	1.2	39344.0
1	1	1.4	46206.0
2	2	1.6	37732.0
3	3	2.1	43526.0
4	4	2.3	39892.0

Figure 1: Salary dataset

After applying data preprocessing techniques to improve the quality of the dataset, such as handling missing data, removing duplicated data, we divide the data into training 80% and test 20% sets. We can see this with the following plot in Figure 2.

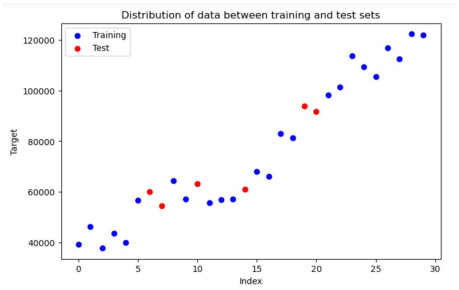

Figure 2: Distribution of data between training and test sets

2.1.2 Model evaluation

We used Python's Statsmodels library, which provides a wide range of algorithms and functionalities for estimating various statistical models. It is an important resource for econometric analysis. Its capabilities include supporting linear regression models, generalized linear models, robust linear models, and time-series analysis [6]. This is shown in Figure 3.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.965
Model:                            OLS   Adj. R-squared:                  0.963
Method:                 Least Squares   F-statistic:                     601.7
Date:                Mon, 27 May 2024   Prob (F-statistic):          1.80e-17
Time:                        14:54:29   Log-Likelihood:                -240.46
No. Observations:                  24   AIC:                             484.9
Df Residuals:                      22   BIC:                             487.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.447e+04   2445.964     10.003      0.000    1.94e+04    2.95e+04
YearsExperience  9436.9135    384.703     24.530      0.000    8639.089    1.02e+04
==============================================================================
Omnibus:                        1.983   Durbin-Watson:                   1.727
Prob(Omnibus):                  0.371   Jarque-Bera (JB):                1.442
Skew:                           0.393   Prob(JB):                        0.486
Kurtosis:                       2.092   Cond. No.                         13.7
==============================================================================
```

Figure 3: Table of Ordinary Least Squares regression results for linear regression model

2.1.3 Interpretation of the ordinary Least Squares regression results

- R-squared and Adjusted R-squared
  - R-squared (0.965): The high coefficient of determination (R-squared) of 96.5% suggests that the independent variable in the model can substantially account for the variation in the dependent variable. This implies that the model exhibits a strong goodness-of-fit to the empirical data under examination.
  - Adjusted R-squared (0.963): The R-squared value is adjusted to account for the number of predictors in the model. It is marginally less than the R-squared value, but still very high, indicating a good fit.
- Analyze of F-statistic and its p-value
  - F-statistic (601.7): This assesses the overall importance of the model. A high F-statistic value suggests that the model is statistically significant.
  - Prob (1.80e-17): This represents the p-value linked with the F-statistic. A very low p-value (considerably less than 0.05) indicates that the model has statistical significance and that the independent variable(s) fit the data well.
- The coefficients (coef) and their standard errors (std err)
  - const (2.447e+04): The intercept represents the predicted value of the dependent variable (Salary) when the independent variable (YearsExperience) is zero. In this case, the intercept is 24.470, which indicates the expected Salary when the individual has no years of experience.
  - YearsExperience (9436.9135): This represents the slope of the regression line, indicating the expected adjustment in Salary for each additional unit increase in YearsExperience. This coefficient implies that an extra year of experience is connected to a rise in Salary of about 9436.91 units.
- The t-statistics and p-values for each coefficient
  - const (t = 10.003, P> $|t|$ = 0.000): The intercept is statistically significant (p-value ¡ 0.05).
  - YearsExperience (t = 24.530, P> $|t|$ = 0.000): This indicates a statistically significant relationship with Salary, as the p-value is less than 0.05.
- Confidence intervals for the coefficients
  - const [1.94e+04, 2.95e+04]: According to the 95% confidence interval for the intercept, we can be 95% certain that the actual intercept value falls between $19,400$ and $29,500$.
  - YearsExperience [8639.089, 1.02e+04]: The 95% confidence interval for the slope indicates that we are confident at a level of 95% that the genuine slope value is between 8639.089 and $10,200$.

2.1.4 Model building and training

We construct the model using the simple linear regression formula, then perform a fonction to fit the model and minimize the parameters, and finally, make predictions on all test data. We can see the following result in Figure 4.

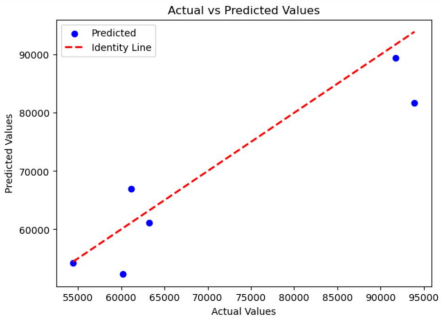

Figure 4: Curve of prediction on test data

2.2 Role of statistics in simple linear regression

Standard errors (std err): The statistical properties of the least squares estimators for the regression parameters β_0 and β_1 can be readily detailed. The model $Y = \beta_0 + \beta_1 x + \epsilon$ assumes that the error term ϵ is a random variable with an expected value of zero and a variance of σ^2. We will examine the bias[1] and variance[2] characteristics of the least squares estimators β_0 and β_1.

Let's begin with β_1. As β_1 is a linear combination of the observed value Y_i, we can apply properties of expectation to demonstrate that the anticipated value of β_1 is

$$E[\hat{\beta}_1] = \beta_1. \tag{4}$$

Same thing for the intercept

$$E[\hat{\beta}_0] = \beta_0. \tag{5}$$

To show that these estimators are unbiased, we need to demonstrate that

$$E[\hat{\beta}_1] = \beta_1; \quad \text{and} \quad E[\hat{\beta}_0] = \beta_0. \tag{6}$$

1. Expectation of $\hat{\beta}_1$: Let's start with $\hat{\beta}_1$. By definition of $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{7}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{8}$$

$$x_i : \text{The individual values of the independent variable.} \tag{9}$$

$$\bar{x} : \text{The mean of the independent variable values.} \tag{10}$$

$$S_{xx} : \text{The sum of the squares of the deviations from the mean of } x \tag{11}$$

It used to estimate and evaluate the regression coefficients. Substituting y_i with its expression in the regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{12}$$

we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})((\beta_0 + \beta_1 x_i + \epsilon_i) - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{13}$$

[1]Bias is the difference between the model's prediction and the correct outcome, with a preference for a certain direction [8]

[2]variance to the stability in performance on future data[8]

Note that

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}, \tag{14}$$

where $\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$. Substituting this expression for $\bar{y}$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})[(\beta_0 + \beta_1 x_i + \epsilon_i) - (\beta_0 + \beta_1 \bar{x} + \bar{\epsilon})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \tag{15}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})[\beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{16}$$

Breaking down this sum,

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{17}$$

Now let's calculate the expectation of $\hat{\beta}_1$

$$E[\hat{\beta}_1] = \beta_1 + E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \tag{18}$$

Given that ϵ_i are independent random variables with zero expectation ($E[\epsilon_i] = 0$),

$$E\left[\sum_{i=1}^{n}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})\right] = \sum_{i=1}^{n}(x_i - \bar{x})E[\epsilon_i - \bar{\epsilon}] = 0. \tag{19}$$

Therefore

$$E[\hat{\beta}_1] = \beta_1. \tag{20}$$

2. Expectation of $\hat{\beta}_0$: For $\hat{\beta}_0$, we have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{21}$$

Taking the expectation

$$E[\hat{\beta}_0] = E[\bar{y} - \hat{\beta}_1 \bar{x}], \tag{22}$$

where $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$ and, using the linearity of expectation, we have

$$E[\hat{\beta}_0] = E[\beta_0 + \beta_1 \bar{x} + \bar{\epsilon} - \hat{\beta}_1 \bar{x}], \tag{23}$$

$$E[\hat{\beta}_0] = \beta_0 + \beta_1 \bar{x} + E[\bar{\epsilon}] - E[\hat{\beta}_1 \bar{x}]. \tag{24}$$

Since $E[\bar{\epsilon}] = 0$ and $E[\hat{\beta}_1] = \beta_1$,

$$E[\hat{\beta}_0] = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}; \tag{25}$$

$$E[\hat{\beta}_0] = \beta_0. \tag{26}$$

3. Now, let's examine the variance of β_1 and β_0: Given our assumption that $V(\epsilon i) = \sigma^2$, it can be concluded that V = σ^2.

The model is given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{27}$$

where $\epsilon_i \sim NID(0, \sigma^2)$.

Estimator $\hat{\beta}_1$:

The ordinary least squares (OLS) estimator for β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}}, \tag{28}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Variance of $\hat{\beta}_1$:

To find the variance of $\hat{\beta}_1$, we use the property of the variance of weighted sums of errors

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{S_{xx}} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{S_{xx}}. \tag{29}$$

Since $E[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, we have

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{S_{xx}}\right).$$

As ϵ_i are independent and identically distributed (i.i.d.)

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2}{S_{xx}^2} = \frac{\sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

Estimator $\hat{\beta}_0$:

The ordinary least squares (OLS) estimator for β_0 is given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{30}$$

Variance of $\hat{\beta}_0$:

To find the variance of $\hat{\beta}_0$, we use the definition

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y} - \hat{\beta}_1\bar{x}). \tag{31}$$

Since $\bar{y}$ and $\hat{\beta}_1\bar{x}$ are linearly combined

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y}) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\mathrm{Cov}(\bar{y}, \hat{\beta}_1). \tag{32}$$

Using $\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n}$ and $\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0$ (because $\bar{y}$ and $\hat{\beta}_1$ are uncorrelated), we obtain

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}. \tag{33}$$

In the context of simple linear regression, the estimated standard error for the slope and intercept are calculated as follows

$$se(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}}. \tag{34}$$

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}. \tag{35}$$

The t-statistics for each coefficient: An essential aspect of evaluating the effectiveness of a linear regression model involves conducting statistical tests on the model parameters and creating specific confidence intervals. This section covers hypothesis testing[3] in simple linear regression, while methods for constructing confidence intervals. To test hypotheses regarding the slope and intercept of the regression model, we need to also assume that the error component in the model, denoted as e, follows a normal distribution[7].

Let us examine the hypothesis that the slope of the model, denoted as $\beta_{1,0}$, is equal to a specific constant value. The relevant hypotheses are:

$$H_0 : \beta_1 = \beta_{1,0}; \qquad H_1 : \beta_1 \neq \beta_{1,0}. \tag{36}$$

The test statistic for the slope, denoted T_0, is given by

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}, \tag{37}$$

where the standard error of the slope is denoted as $\hat{\sigma}_\beta$. Under the null hypothesis that the slope parameter β_1 is equal to a specified value $\beta_{1,0}$, the test statistic T_0 follows a t-distribution with $(n-2)$ degrees of freedom[4]. We would reject $H_0 : \beta_1 = \beta_{1,0}$ if $|t_0| > t_{\alpha/2,n-2}$, where t_0 is computed from the above equation.

Analogous procedures may be employed to investigate hypotheses concerning the regression intercept. To test

$$H_0 : \beta_0 = \beta_{0,0} \qquad H_1 : \beta_0 \neq \beta_{0,0}. \tag{38}$$

We would use the statistic T_0 given by

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} \tag{39}$$

can be rejected if the calculated value of the test statistic, t_0, exceeds the critical value $t_{\alpha/2,n-2}$. Additionally, the null hypothesis should be rejected when the calculated value of the test statistic, t_0, is greater than $t_{\alpha/2,n-2}$. It is important to note that the denominator of the test statistic equation (4.3.19) represents the standard error of the intercept.

A highly significant instance of the hypotheses presented in Equation 4.3.16 is:

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0 \tag{40}$$

The hypotheses pertain to the significance of regression. If we fail to reject the null hypothesis $H_0 : \beta_1 = 0$, it means that we are concluding that there is no linear correlation between the variables x and Y.

Confidence intervals on the slope and intercept: further to calculating the point estimates for the slope and intercept, it is also possible to derive confidence interval

[3]hypothesis test is decision making process[7]

[4]"Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample" [10]

estimates for these parameters. The range of these confidence intervals provides a measure of the overall reliability of the regression line. If the error terms ϵ_i in the regression model follow a normal distribution and are independent.

This leads to the following definition of $100(1-\alpha)\%$. [5].

Assuming the observations follow a normal distribution and are statistically independent, a $100(1-\alpha)\%$ confidence interval for the slope coefficient, β_1, in a simple linear regression model is given by

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\sigma^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\sigma^2}{S_{xx}}}. \tag{41}$$

Similarly, a $100(1-\alpha)\%$ confidence interval on the intercept β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}. \tag{42}$$

F-statistic: It is a method to evaluate the significance of the model. The significance of regression can be tested using a technique known as analysis of variance[6]. This method involves breaking down the total variability in the response variable into distinct components, serving as the foundation for conducting the test. The formula for analysis of variance is expressed as:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2, \tag{43}$$

where

- y_i: The actual value of the i-th observation.
- $\bar{y}$: The mean of all observed values y, calculated as $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- $\hat{y}_i$: The predicted value of the i-th observation.
- $\sum_{i=1}^{n}(y_i-\bar{y})^2$: The total sum of squares (TSS), which measures the total variation in the observed values.
- $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$: The residual sum of squares (RSS), which quantifies the variation in the observed values not accounted for by the model.
- $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$: The regression sum of squares, which quantifies the variation in the predicted values explained by the model. Equation 4.3.10 expresses the relationship between the total sum of squares, the regression sum of squares, and the error sum of squares by [7]:

$$\text{SST} = \text{SSR} + \text{SSE}. \tag{44}$$

R-squared and Adjusted R-squared: The coefficient of determination R-squared is computed by dividing SSR by SST. A value nearing 1 indicates that the variable in the equation has a stronger capability to explain Y, and signifies that the model fits the data well [3].

$$R^2 = \frac{\text{SS}_R}{\text{SS}_T} = 1 - \frac{\text{SS}_R}{\text{SS}_E}. \tag{45}$$

The coefficient judges the adequacy of a regression model.

[5] A confidence interval provides a range of values depending on $\hat{\lambda}$(estimate) such that the probability of λ(parameter) being within the interval is $1-\alpha$ [11]

[6] "Analysis of variance (ANOVA) is a statistical test used to evaluate the difference between the means of more than two groups" [12]

3 Multiple linear regression model

Multiple linear regression is a statistical technique utilized to predict the value of a dependent variable by employing multiple independent variables. The purpose of MLR is to construct a model that represents the linear correlation between the independent variables x and dependent variable y, which will then be examined [9].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_n x_n. \tag{46}$$

3.1 Practical example of predicting graduate admissions

We can take an example like the Prediction of Graduate Admissions. Many students struggle with choosing graduate programs due to a lack of knowledge about university rankings and misleading advice, leading to missed admissions and wasted resources. Our goal is to help students connect with their preferred university by thoroughly evaluating their profiles, ensuring accurate assessments without overestimating or underestimating their potential.

3.1.1 Data loading and preparation

During the preparation of our manuscript, the dataset has been downloaded over 400 times and viewed more than 2000 times. It includes parameters that are meticulously evaluated by the admissions committee, such as GRE scores, TOEFL scores, Undergraduate GPA, University rating, research, Statement of Purpose, and Letter of Recommendation[16].

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

Figure 5: graduate dataset

Viewing the dataset in Figure 5, we obtain the following plot in Figure 6.

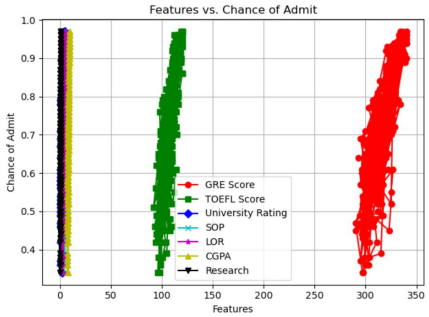

Figure 6: Features vs Chance of Admit

Dividing the data into training 80% and test 20% sets, we have the following plot in Figure 7.

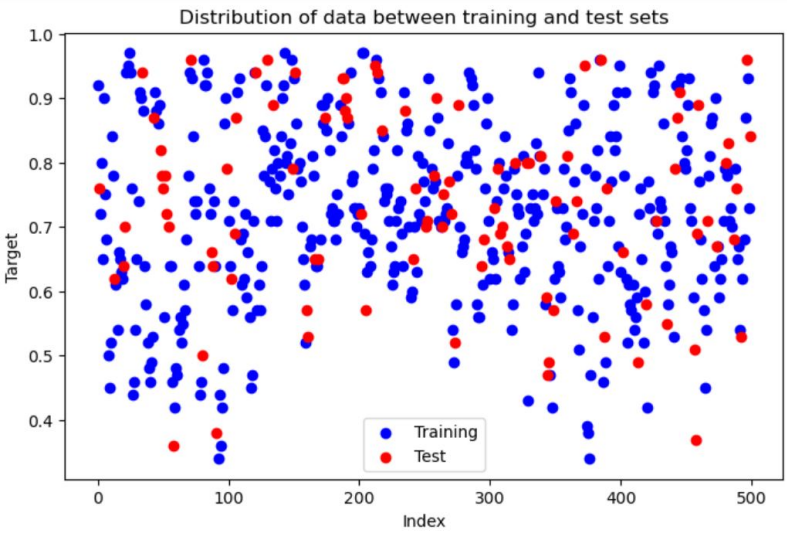

Figure 7: Distribution of data between training and test sets

3.1.2 Model evaluation

We have the result on prediction of test data in Figure 8.

OLS Regression Results

Dep. Variable:	Chance of Admit	R-squared:	0.817
Model:	OLS	Adj. R-squared:	0.813
Method:	Least Squares	F-statistic:	218.4
Date:	Mon, 09 Sep 2024	Prob (F-statistic):	3.85e-139
Time:	23:19:30	Log-Likelihood:	558.59
No. Observations:	400	AIC:	-1099.
Df Residuals:	391	BIC:	-1063.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.2539	0.119	-10.519	0.000	-1.488	-1.020
Serial No.	0.0001	2.16e-05	4.892	0.000	6.31e-05	0.000
GRE Score	0.0016	0.001	2.753	0.006	0.000	0.003
TOEFL Score	0.0031	0.001	3.020	0.003	0.001	0.005
University Rating	0.0043	0.004	0.996	0.320	-0.004	0.013
SOP	0.0057	0.005	1.100	0.272	-0.004	0.016
LOR	0.0147	0.005	3.059	0.002	0.005	0.024
CGPA	0.1198	0.011	10.497	0.000	0.097	0.142
Research	0.0261	0.007	3.509	0.001	0.011	0.041

Omnibus:	63.379	Durbin-Watson:	2.101
Prob(Omnibus):	0.000	Jarque-Bera (JB):	107.853
Skew:	-0.934	Prob(JB):	3.80e-24
Kurtosis:	4.727	Cond. No.	1.67e+04

Figure 8: Result on prediction of test data

3.1.3 Interpretation of the ordinary Least Squares regression results

- R-squared and Adjusted R-squared
  - R-squared (0.817): The R-squared value of 0.817 implies that the independent variables in the model account for approximately 81.7% of the observed variation in the dependent variable, indicating a strong model fit.
  - Adjusted R-squared (0.813): This value, at 0.813, takes into account the number of predictors and remains high, further supporting a good fit for the model.
- F-statistic and its p-value
  - F-statistic (218.4): It evaluates the overall significance of the model, with a high value indicating statistical significance.
  - Prob (F-statistic) ($3.85e-139$): It represents the p-value associated with the F-statistic; an extremely low p-value (much less than 0.05) signifies that the model is statistically significant, suggesting that the independent variables collectively strongly predict the dependent variable.
- Coefficients (coef) and their standard errors (std err)
  - const (-1.2539): The intercept value represents the predicted Chance of Admit when all other independent variables in the regression model are held constant at zero. This intercept has a negative value, which might not have a meaningful real-world interpretation in this context.
  - GRE Score (0.0016): For each one-point increase in GRE Score, there is an associated 0.16% increase in the Chance of Admit while holding all other variables constant- indicated by its statistically significant coefficient with a p-value of 0.006.
  - TOEFL Score (0.0031): Similarly, a one-unit rise in TOEFL Score is associated with a 0.31% increase in the probability of admission while keeping all other factors constant; this coefficient is also statistically significant, with a p-value of 0.003.
  - University Rating (0.0043): The coefficient is not statistically significant with a p-value of 0.320, indicating that University Rating might not have a meaningful impact on the Chance of Admit.
  - SOP (0.0057): The coefficient is not statistically significant with a p-value of 0.272, indicating that SOP might not have a meaningful impact on the Chance of Admit.
  - LOR (0.0147): A one-unit rise in LOR is linked to a 1.47% increase in the probability of admission while keeping all other factors constant. This coefficient demonstrates statistical significance with a p-value of 0.002.
  - CGPA (0.1198): A one-unit rise in CGPA is linked to an 11.98% increase in the Chance of Admit while keeping all other factors constant. This coefficient demonstrates statistical significance with a p-value of 0.000.
  - **Research (0.0261)**: Participation in research activities is associated with a 2.61% higher Chance of Admission when controlling for other factors. This finding is statistically significant with a p-value of 0.001.

- The t-statistics and p-values for each coefficient
  - const (t = -10.519,(P> $|t|$ = 0.000): The intercept is statistically significant (p-value ¡ 0.05) indicating that the constant term is significantly different from zero. This implies that the baseline *Chance of Admit* when all other predictors are zero is meaningful and not due to random chance.
  - Serial No. (t = 4.892, P> $|t|$ = 0.000): The coefficient for *Serial No.* a statistically significant (P-value ¡ 0.05), indicating that the serial number has a non-zero effect on the *Chance of Admit*. However, the effect size is very small and likely not practically significant.
  - GRE Score (t = 2.753, P> $|t|$ = 0.006): The coefficient for *GRE Score* a statistically significant (P-value < 0.05), indicating that GRE Scores have a significant positive effect on the *Chance of Admit*.
  - TOEFL Score (t = 3.020, P> $|t|$ = 0.003): The coefficient for *TOEFL Score* a statistically significant (P-value < 0.05), suggesting that higher TOEFL Scores are associated with a higher *Chance of Admit*.
  - University Rating: (t = 0.996, P> $|t|$ = 0.320): The coefficient for *University Rating* is not statistically significant (P-value ¿ 0.05), indicating that university rating does not have a significant effect on the *Chance of Admit*.
  - SOP: (t = 1.100, P> $|t|$ = 0.272): The coefficient for *SOP* is not statistically significant (P-value ¿ 0.05), suggesting that the statement of purpose does not significantly affect the *Chance of Admit*.
  - LOR: (t = 3.059, P> $|t|$ = 0.002): The coefficient for *LOR* is statistically significant (P-value ¡ 0.05), indicating that stronger letters of recommendation are associated with a higher *Chance of Admit*.
  - CGPA: (t = 10.497, P> $|t|$ = 0.000): The coefficient for *CGPA* is highly significant (P-value ¡ 0.05), suggesting that higher cumulative GPA is strongly associated with a higher *Chance of Admit*.
  - Research: (t = 3.509, P> $|t|$ = 0.001): The coefficient for *Research* is statistically significant (P-value ¡ 0.05), indicating that having research experience is associated with a higher *Chance of Admit*.

3.1.4 Model building and training

We construct the model using the multiple linear regression formula, then perform a function to fit the model and minimize the parameters, and finally make predictions on all test data. We can see the following result on the curve on prediction of test data in Figure 9.

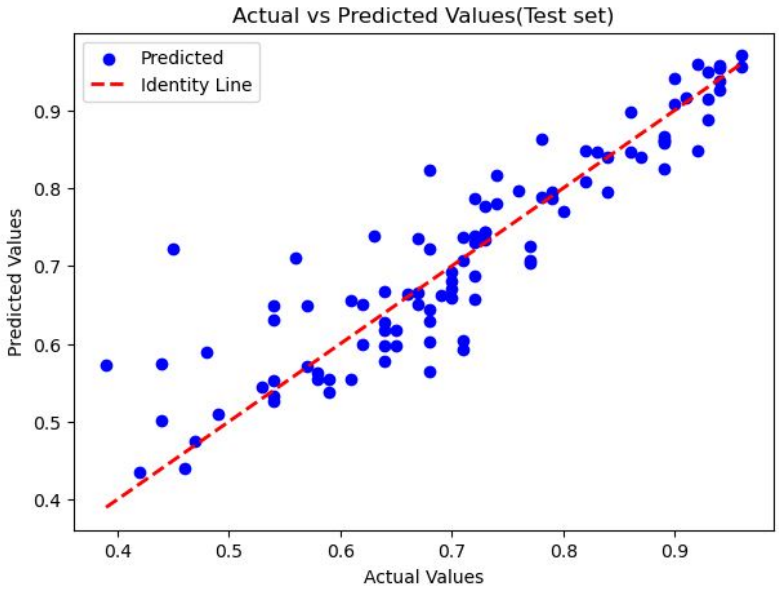

Figure 9: Curve on prediction of test data

3.2 Role of statistics in multiple linear regression

Standard errors(std err): To calculate the standard error we must compute the estimation of variance. The estimator of the variance of the error terms, $\hat{\sigma}^2$, is given by

$$\hat{\sigma}^2 = \frac{\text{SSE}}{(n-p)}. \tag{47}$$

In this formula

- n is the number of observations.
- p is the number of model parameters (including the intercept).

then determine the matrix C_{jj} is the j-th component of the matrix $(X'X)^{-1}$. Once we have c we can extract the diagonal component. Finally the standard errors of the least squares estimator β is

$$\hat{\beta}_j = \sqrt{\sigma^2 \left[(X^T X)^{-1}\right]_{jj}}. \tag{48}$$

The F-statistics for each coefficient: In this part, we outline several significant hypothesis-testing methods. Just like in the case of simple linear regression, hypothesis testing necessitates that the error terms ϵ_i in the regression model follow a normal and independent distribution with an average of zero and variance σ^2. The significance test for regression examines whether there is a linear association between the response variable y and a subset of the predictor variables $x1, x2, ..., xk$. It evaluates the following hypotheses:

$$H_0 \quad : \beta_1 = \beta_2 = \cdots = \beta_k = 0. \tag{49}$$
$$H_1 \quad : \beta_j \neq 0 \text{ for at least one } j. \tag{50}$$

The rejection of the null hypothesis suggests that at least one of the predictor variables $x_1, x_2, \ldots, x_k$ makes a significant contribution to the model. If $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is indeed true, SSR/σ^2 follows a chi-square distribution with k degrees of freedom. The test statistic for ANOVA.

$$F_0 = \frac{\text{SS}_R/k}{\text{SS}_E/n-p}. \tag{51}$$

We must reject the null hypothesis if the calculated test statistic value in Equation (51), denoted as f_0, exceeds $f_{\alpha,k,n-p}$.

The t-statistic: Hypothesis testing could be employed to assess the relative significance of each regressor variable within the regression model. This may involve considering the inclusion of extra variables or the removal of existing ones to enhance the model's effectiveness. One essential hypothesis test involves determining if an individual regression coefficient, denoted as β_j, equals a specified value $\beta_j 0$.

$$H_0 : \beta_j = \beta_{j,0}; \qquad H_1 : \beta_j \neq \beta_{j,0}. \tag{52}$$

The test statistic for this hypothesis is

$$T_0 = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_{j,0}}{\text{se}(\hat{\beta}_j)}. \tag{53}$$

The diagonal element C_{jj} of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponds to $\hat{\beta}_j$. The denominator of Equation 4.3.29 represents the standard error of the regression coefficient $\hat{\beta}_j$. The null hypothesis $H_0 : \beta_j = \beta_{j,0}$ is rejected if the absolute value of the test statistic t_0 exceeds the critical value $t_{\alpha/2,n-p}$. This is considered a partial or marginal test, as the estimated regression coefficient $\hat{\beta}_j$ depends on all other predictor variables x_i included in the model [7].

R-squared and Adjusted R-squared: The coefficient of determination, denoted as R-squared, can be employed as a comprehensive metric to assess the model's goodness of fit. From a statistical standpoint, this value can be computed as:

$$R^2 = \frac{\text{SS}_R}{\text{SS}_T} = 1 - \frac{\text{SS}_R}{\text{SS}_E}. \tag{54}$$

4 Logistic linear regression model

Logistic regression is a statistical model used for classification tasks, where the goal is to predict a categorical or discrete outcome variable. It estimates the probability of the dependent variable B, which is a binary or dichotomous class, based on the independent variables A by determining $P(A|B)$. The model classifies the binary class label B using the following equations [17]:

$$P(B = 1|A) = \frac{1}{1 + \exp\left(W_0 + \sum_{i=1}^{n} W_i A_i\right)}. \tag{55}$$

$$P(B = 0|A) = \frac{\exp\left(W_0 + \sum_{i=1}^{n} W_i A_i\right)}{1 + \exp\left(W_0 + \sum_{i=1}^{n} W_i A_i\right)}. \tag{56}$$

4.1 Practical example

We can take an example like some information about students and aims to predict whether a student will pass or fail based on several characteristics. The columns in the dataset are as follows

1. Student Id: An identifier unique to each student.
2. Gender: The gender of the student, where 'M' stands for male and 'F' stands for female.
3. Age: The age of the student in years.

4. Final mark: The final grade obtained by the student in the exam or course.

5. Pass: A binary indicator where 1 means the student passed and 0 means the student failed.

The goal is to use this information to predict the probability of a student passing or failing the exam or course by using logistic regression techniques, we can model the relationship between the students' characteristics (age and final mark) and their success or failure, as represented by the 'pass' column.

4.1.1 Data loading and preparation

	Student Id	Gender	Age	Final Mark	Pass
0	2024640MN	F	21	8	0
1	2024070VF	F	22	20	1
2	2024052UN	F	19	4	0
3	2024168BF	M	24	3	0
4	2024281MA	F	23	18	1

Figure 10: dataset students

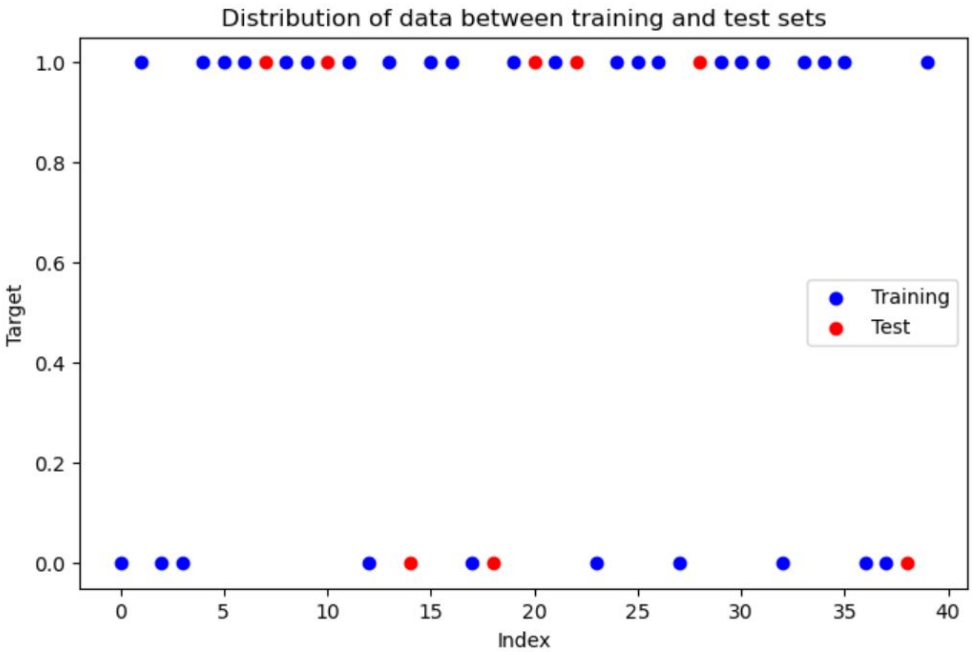

Figure 11: train test split data between training and test sets

4.1.2 Model building and training

It is the same process as the last model.

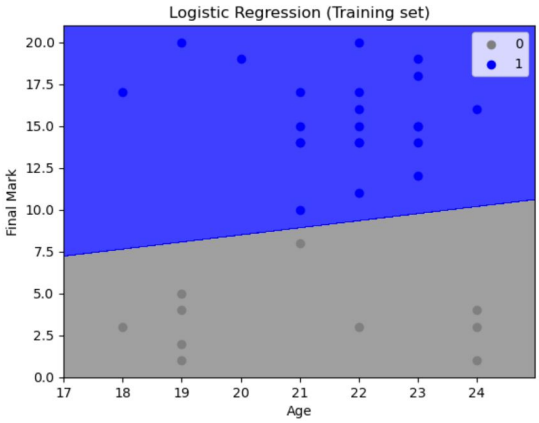

Figure 12: Curve of prediction on training data

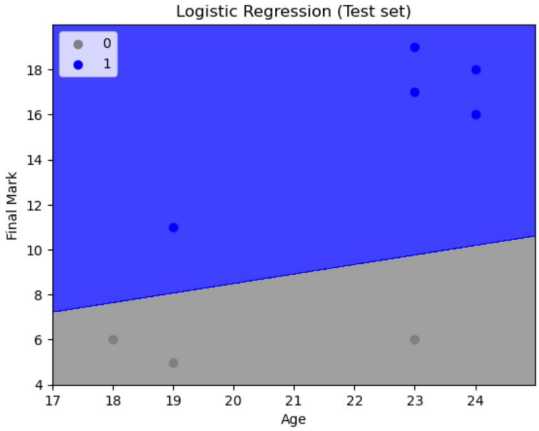

Figure 13: Curve of prediction on test data

4.1.3 Model evaluation

We can see the following figure below

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	5
accuracy			1.00	8
macro avg	1.00	1.00	1.00	8
weighted avg	1.00	1.00	1.00	8

Figure 14: Table of classification report for the logistic regression model

4.1.4 Interpretation

- Precision The precision metric represents the proportion of true position predictions among the total number of positive predictions made by the model. A precision of 1.00 for classes 0 and 1 means that all positive predictions in your model are correct.

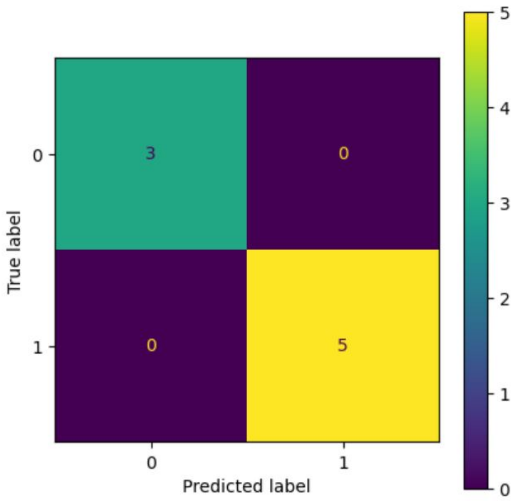

Figure 15: Table of confusion matrix for the logistic regression model

- Recall The recall measure reflects the fraction of true positive predictions out of the total number of actual positive instances. A recall of 1.00 for classes 0 and 1 means that our model has correctly identified all positive instances in each class.
- F1-Score The f1-Score represents the harmonic mean of precision and recall. An f1-Score of 1.00 for classes 0 and 1 means that our model performs perfectly in terms of precision and recall.
- Support The support values represent the number of true examples belonging to each class. In your case, there are 3 instances of class 0 and 5 instances of class 1.
- Accuracy: Accuracy is the proportion of correct predictions out of all predictions. An accuracy of 1.00 means that all predictions in our model are correct.

4.2 Role of statistics in logistic regression

A binary classifier's confusion matrix is depicted in Figure 16. It shows the actual values labeled as True and False, along with their predictions as Positive and Negative. The performance of a classification model is evaluated based on the values of true positive, true negative, false positive, and false negative entries in the confusion matrix.

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 16: Confusion matrix for the binary classification problem

TP (True Positive): A true positive in the confusion matrix occurs when a positive result is accurately predicted.

FP (False Positive): In the confusion matrix, a false positive arises when the model incorrectly predicts a positive outcome, yet the actual outcome is negative. This situation corresponds to Type 1 Error and can be likened to an unwelcome stroke of luck.

FN (False Negative): A false negative occurs in the confusion matrix when a negative result is wrongly predicted as positive. This situation is commonly referred to as a Type 2 Error and is considered equally detrimental as a Type 1 Error.

TN (True Negative): When a negative prediction aligns with the actual outcome, it is considered a True Negative in the confusion matrix. This is demonstrated in the binary classification results depicted in Figure 16 [20].

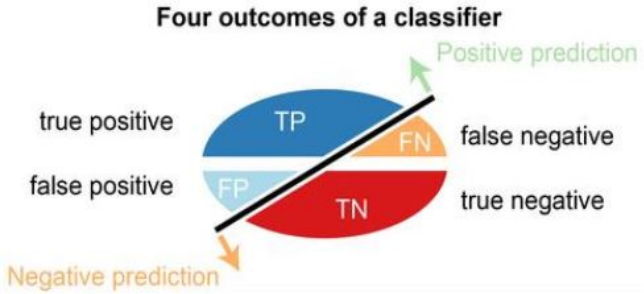

Figure 17: Representation of the test set classification results in elliptical form for four binary outcomes.

The model's accuracy is calculated by summing the number of correctly predicted positive and negative cases (TP + TN) and then dividing that sum by the total number of data sets (P + N). The highest possible accuracy score is 1.0, while the lowest is 0.00 [21].

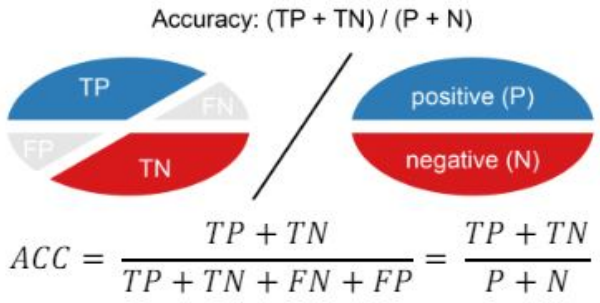

Figure 18: Two ellipses demonstrate the method of accuracy calculation

The sensitivity metric, also referred to as recall or true positive rate, is calculated by dividing the number of correct positive predictions by the total number of actual positive cases. It is referred to as Sensitivity or Recall. A perfect TP Rate is represented by 1.0, while the lowest rate is denoted by 0.0 [21].

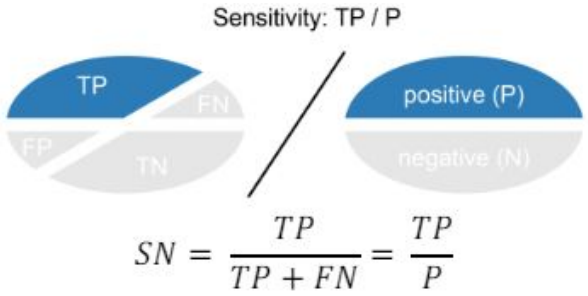

Figure 19: Two ellipses demonstrate the method of accuracy calculation

Precision is determined by the ratio of correctly predicted positive cases to the total number of predicted positive cases (TP + FP). The highest achievable accuracy is 1.0, while the lowest is 0.0 [21].

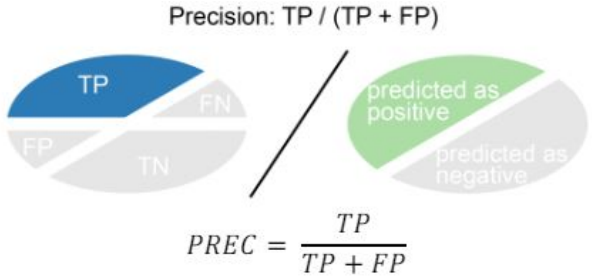

Figure 20: Two ellipses demonstrate the method of precision calculation

The F-Measure, also known as the F-score, quantifies the test's accuracy and is computed using precision and recall in the following formula [21].

$$\text{F-Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{57}$$

5 Conclusion

Statistics plays a crucial role in the field of machine learning. In this work, we studied three models: simple linear regression, a multiple linear regression model, and the logistic regression model. The evidence role of statistics in machine learning is highlighted. The fundamentals of machine learning are deeply rooted in mathematical concepts. It is applied to aid real estate companies in predicting apartment prices, and banks in detecting anomalies and classifying handwritten digit recognition (postal codes) among other uses. This is accomplished through a blend of mathematics and extensive programming skills, particularly using Python. In future work, we will explore other more complex algorithms, fostering innovation and pushing the limits of what machine learning can accomplish. For instance, we can consider exploring the multinomial logistic regression, ridge and lasso regression, and Generalized Linear Models (GLMs) algorithm to understand its mathematical principles, potential enhancements, and efficacies. Why not create our algorithm for machine learning?

Acknowledgments

The authors would like to thank Dr. Douanla Alotse Yaulande for useful discussions. This work has been supported by AIMS Senegal.

References

[1] Dr. Suresh Dara et al. Role of mathematics in machine learning. International Research Journal of Modernization in Engineering Technology and Science, 04(04), April 2022.

[2] Md. Kosher. A combination of mathematics, statistics, and machine learning to detect fraud. In National Mathematics Conference, 2020.

[3] Yilu Wu. Linear regression in machine learning. In Proceedings Volume 12163, International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021), page 121634T, Univ. of Birmingham (United Kingdom), 2022. SPIE. Event: International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021), 2021, Nanjing, China.

[4] Virender and Abhishek Pratap Singh. Applications of mathematics in machine learning. https://naac.iem.edu.in/wp-content/uploads/2024/02/Minor-Project- 7A-Avilash-Sengupta.pdf, 2021. Accessed: 29 May 2024.

[5] ARPAN BISWAS. Salary prediction using machine learning. 2022.

[6] Shukrulloev Bektosh and Murtazaev Misliddin. Using python in the analysis of econometric models. Innovations in Exact Science - Aniq fanlarda innovatsiyalar, 2023. Senior teachers of TMC Institute.

[7] Douglas C Montgomery and George C Runger. Applied Statistics and Probability for Engineers. Wiley, 6th edition, 2013.

[8] Mehmet Koçak et al. Must-have qualities of ai. Balkan Medical Journal, 40(1):7–12, 2023.

[9] Dastan Hussen Maulud and Adnan Mohsin Abdulazeez. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(2):140-147, 2020.

[10] Akhilesh Ganti. Degrees of freedom in statistics explained: Formula and example. 2024. Updated February 28, 2024. Reviewed by Erika Rasure. Fact checked by Suzanne Kvilhaug.

[11] Nathaniel E. Helwig. Confidence intervals, October 17 2020. Copyright © by NEH.

[12] Will Kenton. What is analysis of variance (anova)? https://www.investopedia.com/what-is-anova-4586741, 2024. Updated April 18, 2024. Reviewed by Erika Rasure. Fact checked by Timothy Li. See "Definition of ANOVA" section.

[13] James Chen. Normal distribution: What it is, uses, and formula. https://www.investopedia.com/terms/n/normaldistribution.asp, 2024. Updated March 13, 2024. Reviewed by Khadija Khartit. Fact-checked by Suzanne Kvilhaug.

[14] Damodar N. Gujarati and Dawn C. Porter. Basic Econometrics. McGraw-Hill/Irwin, New York, NY, 5th edition, 2009.

[15] Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. The American Statistician, 70(2):129-133, 2016.

[16] Mohan S Acharya, Asfia Armaan, and Aneeta S Antony. A comparison of regression models for prediction of graduate admissions. In Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019), 2019.

[17] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, and Nitat Ninchawee. Predictiv analytics for chornic kidney disease using machine learning techniques. In The 2016 Management and Innovation Technology International Conference (MITiCON), 2016.

[18] Sanjeev Arora. Mathematics of machine learning: An introduction. In Proceedings of the International Congress of Mathematics, pages 377-390, 2019.

[19] Daniel Jurafsky and James H. Martin. Speech and Language Processing. Pearson, 2023. Draft of February 3, 2024. Chapter 5: Logistic Regression.

[20] T. Saito and M. Rehmeismeier. Basic evaluation measures from the confusion matrix. WordPress, 2017.

[21] Zeljko D Vujovic. A case study of the application of weka software to solve the problem of liver inflammation. Archives of Clinical and Experimental Surgery, 10(10):01–13, 2021.

[22] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, New York, NY, 2006.