ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

A Comparative Analysis of Linear Regression Techniques: Evaluating Predictive Accuracy and Model Effectiveness

Madhuparna Das Hait[1]; Priya Das[2]; Washim Akram[3]; Siddhartha Chatterjee[4*]

[1]Department of Information Technology, Haldia Institute of Technology, Haldia, Purba Medinipur – 721657, West Bengal, India

[2]Department of Computer Science Engineering, Bengal College of Engineering and Technology Durgapur – 713212, West Bengal, India

[3]Department of Computer Science Engineering, College of Engineering and Management Kolaghat, KTPP Township, Purba Medinipur – 721171, West Bengal, India

[4]Department of Computer Science and Engineering, College of Engineering and Management Kolaghat, Purba Medinipur – 721171, West Bengal, India

Corresponding Author: Siddhartha Chatterjee[4*]

Publication Date: 2025/07/09

Abstract: The main objective of this research is to determine which of the three methods of regression; Ordinary Least Squares (OLS) regression, Baseline regression, and Polynomial regression offers the most accurate predictive capability and an ability to capture the associations between two variables. Other assessment indicators include R squared and Mean squared error (MSE) while graphical techniques include residual charts. The paper presents a concise review of the linear regression method, the mathematical background of the method, and the procedure for improving the efficiency of the model by selecting relevant features. It discusses the use OLS regression as the fundamental technique of statistical inference and its relative accuracy to other methods. Using regression lines, residual graphs, outliers influence and effects of outliers, the research shows how reliable predictions can be made using such models. This work contributes to the understanding of statistical modelling, giving practicable guidelines for enhancing data analysis techniques for all fields of study, mainly economics, natural science, and social science to enable improved decision-making and enhanced accuracy of the analysis.

Keywords: *Predictive Accuracy, Ordinary Least Squares (OLS), Model Evaluation Metrics.*

How to Cite: Madhuparna Das Hait; Priya Das; Washim Akram; Siddhartha Chatterjee (2025) A Comparative Analysis of Linear Regression Techniques: Evaluating Predictive Accuracy and Model Effectiveness. *International Journal of Innovative Science and Research Technology*, 10(7), 127-139. https://doi.org/10.38124/ijisrt/25jul349

I. INTRODUCTION

Linear regression is a very basic statistical method that is used to determine the relationship between a variable and one or more other variables. The main idea is to find the regression line that is used to bring the difference between the observations and the predictions to the minimum. This is used in predictive statistics to analyze future trends and the trends to make good and sound decisions. Some of the distinct characteristics of the linear regression model include the following the model is easy to understand Thus, linear regression can be used in most domains. It uses a straight-line model that shows that the changes in the independent variables result in a similar magnitude of changes in the dependent variable. This trend can in most cases be modelled using '***Ordinary least square***', by the formulation above which seeks to determine the parameters of the model such that the sum of squared residuals is optimized. Further, the use of linear regression makes it easy the assess relation strength and direction through coefficients, as well as hypothesis testing and identification of the significance of the predictors in the given model.

➢ *Background*

Linear regression is one of the oldest and widely used methods of modelling the behaviour of numerical variables and US rooted in different categories, for example, economics, biology, engineering, and social studies. The method dates back to the nineteenth century agreeing with the twentieth century, and it was named by some mathematicians

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

such as Carl Friedrich Gauss the founder of the method of least squares of the parameters of linear models. It then developed and has since become one of the most commonly utilized methods of statistical learning. This research area is concerned with how to model and forecast the profile of complied systems through straight linearity. Quantitative research provides a mathematical expression of a dependent variable to one or more independent variables that can be used to make conclusions from the data gathered.

In addition, due to the growing amount of data produced in terms of both size and variety, the need for accurate and explicable approaches is critical. Linear regression is equally the most popular type of regression analysis for its simplicity and interpretability regardless of the statistical expertise of its user. Moreover, continuous developments in computational hardware and the data science field complement linear regression by strengthening the model while improving its results. As the purpose of this research is to lay out the underlying reasons for linear regression it is hoped that the work will provide an addition to the current understanding of the field as well as demonstrate the continued significance of linear regression in the face of the current issues and future advancements in statistical modelling.

➢ *Rationale*

Linear regression has also gained popularity in the recent past due to increased computational abilities and vast data accessibility, especially in fields like machine learning, economics, and public health. Even with newer more complicated models, linear regression is still applicable because they are easy to use or interpret. Recent studies are still being made to extend and improve the enhancements of linear regression that incorporate the L1 and L2 approaches and make the non-linear regression models more resistant to outliers making it more relevant in the time of big data and high data and analytical demands.

The reason for this paper on linear regression is to explore its practicality across different disciplines and its importance to society. Being one of the basic types of statistical methods, linear regression helps to analyze the connections between different variables, and, thus, to understand seemingly maze-like systems. A recognition of such relationships is essential to proper decision-making in various fields including economics, health, and the study of the environment.

➢ *Aim and Objectives*

The aim of this research is to compare Ordinary Least Squares (OLS) regression, Baseline regression, and polynomial regression models in assessing accuracy and evaluating the effectiveness of these models using two variables, analyzing their features to determine the best approach for accurate predictions.

➢ *Objectives*

- To justify the effectiveness of the linear regression model through relevant evaluation parameters such as R-squared, Mean Squared Error (MSE) along with graphical performances like residual plots
- To give a comprehensive overview of linear regression, the major ideas behind it, and what mathematical theories apply to this method
- To examine the use and efficiency of Ordinary Least Squares (OLS) regression in estimating model parameters and as a building block of statistical inference.
- To compare the predictive accuracy of Ordinary Least Squares (OLS) regression, Baseline regression, and Polynomial regression models using two variables and assess their effectiveness in capturing relationships within the data.
- To analyze the features and correlations of the selected variables to identify which model provides the most accurate predictions and insights into the underlying patterns in the dataset.

➢ *Research Questions*

- Is there any relation between selected variables and the R-squared and the Mean Squared Error (MSE) and how are they useful in the testing of a linear regression model?
- What are the fundamental concepts and mathematical principles employed in Linear Regression and how do they impact the performance of the model?
- What can be learned from the graph of the regression line and the residuals when discussing the model fit and model prediction on the penguin dataset?
- What are the differences between the different techniques of linear regression?

➢ *Research Significance*

The importance of this work is to contribute to the improvement of knowledge and use of linear regression, as an indispensable, commonly applied statistical procedure across numerous disciplines, such as economics, social and natural sciences. It is with these objectives that this study seeks to establish the methods by which the accuracy and reliability of linear regression models may be better understood through the use of R-squared and Mean Squared Error (MSE), to help practitioners better identify strategies that will be useful for their data analysis. In addition, the study area as to the effect of feature selection on the model performance enhances efficiency by directing the researcher to a set of important predictors. Looking at the graphs of the regression line and residuals will help in arriving at the correct decision as far as evaluating the fit of the model is concerned and studying the impact of outliers brings out the point that while analyzing the data, robust modelling should be used.

Overall, this research contributes to the body of knowledge in statistical modelling, offering practical implications for improving statistical analytics and decision-making processes across various domains.

II. LITERATURE REVIEW

This paper's literature review comprises the existing works and research done on regression models with specific

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

emphasis on linear, Ordinary Least Squares (OLS), as well as polynomial regression methods. These models are basic in predictive modelling and carry out the identification of variables or determination of outputs. Previous studies address antecedent theories on OLS regression, decision-making on parameter estimation, polynomial regression that helps in the determination of non-linear functions, and baseline models for a comparison of the performance. In modulating model performance, feature selection and correlation analysis are also identified in the literature as good practices. R-squared and Mean Squared Error (MSE), were some of the important pointers used in model authenticity checks that include, residual plots. The advantage of comparing the efficiency of the analyzed regression models is to derive increased comprehension of the data features. This literature review is intended to summarize the existing literature and to examine difficulties for further development and successful use of models by synthesizing the existing conclusions.

- *Review of Linear Regression Methodology and Assessment of its uses in Practice*

Montgomery, Peck, and Vining (2021) present Introduction to Linear Regression Analysis, a textbook for students and practitioners, which remains one of the essential sources providing the analysis of the key approaches of linear regression and their application in modern science and engineering, business, and social sciences. The authors provide theory and examples including theoretical concepts and concrete applications thus the book can benefit students or professionals to discover and perform regression models adequately [13].

Starting from primitive principles, linear regression, the author focuses not only on modelling concepts between variables but also emphasizes the link between dependent and independent variables. It explains simple linear regression, where there is only one independent variable which predicts the dependent variable, and then it elaborates multiple linear regressions, in which there is more than one predictor [1,12].

Within this context, the authors discuss the most commonly used approach to parameter estimation, namely Ordinary Least Squares (OLS). OLS strives to fit the best line of regression as it seeks to make the overall sum of squared residuals, or in other words the overall sum of squares between the actual result and the predicted one, as small as possible [2]. They also then investigate commonly used measurements for ranking regression models and their performance such as R Squared, Adjusted R Squared and Mean Squared Error (MSE). These metrics provide information on what proportion of the dependent variable variation is explained by the model and form the basis for model comparison.

Over a third of the book is devoted to model diagnostics and validation, which is rather tens of pages. The authors emphasize the need to test for violations of regression assumptions by use of diagnostic tests such as residual plots. Issues such as multi co linearity, heteroscedasticity and autocorrelation are explored, measures of checking them and dealing with them. For example, they suggest Variance Inflation Factors (VIF) for use in addressing multi co linearity and transformation of variables for use under heteroscedasticity of variance.

In addition to linear models, the book offers methodological information on polynomial regression, which goes beyond linear regression under introducing nonlinear relations. In this section, it is shown how polynomial terms can be added to linear terms to enhance the applicability of the model to more complicated data. Stepwise regression and other methods for feature selection, such as those presented by the authors, also aim at increasing the accuracy of a model by showing which predictors are most important.

According to Montgomery, Peck, and Vining it is noble to compare the baseline models, OLS regression as well as other techniques to identify which regression delivers the best results given the data in question. They also present several current actual cases and datasets, which enables readers to solve linear regression models in realistic settings. More to that, the inclusion of what is in use current software tools like R and Python makes the book more valuable for the current data analysis [3,14].

In summary, this textbook serves as a crucial resource for anyone working with regression models, offering both theoretical understanding and practical skills. It highlights the importance of careful model building, diagnostics, and evaluation, ensuring that models are not only accurate but also interpretable and reliable. The book remains relevant for researchers and practitioners by covering advanced topics like non-linear regression and model selection, alongside a strong foundation in OLS regression.

- *Foundations and Applications of Linear Regression: A Gateway to Advanced Statistical Learning*

James *et al.* (2023) have dedicated a chapter of their book to linear regression from which we explain linear regression as one of the most important algorithms in the supervised learning technique. The authors state that what they have attempted to bring across in their book is the fact that linear regression is indispensable and recognized as a mainstream method for forecasting quantitative target variables even though modern statistical learning methods have emerged. Sometimes, it provides the original knowledge base upon which more complex machine-learning algorithms will be built upon [4,15].

The chapter starts with an explanation of the use of linear regression in analysing relationships between variables and making quantitative predictions. The first approach covered is the Ordinary Least Squares (OLS) which seeks to minimize the sum of squared residuals in order to fit a line. The approach guarantees the model gives accurate estimates by diminishing the gaps in between forecasted and actual values.

Taking Advertising data, the authors provide real examples of how linear regression can be used for solving real-world issues like designing the right strategy for

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

marketing. They illustrate how the sales data can be expressed as a function of advertising cost within various media, and how regression models offer solutions to sales decisions [5]. These practical examples help the reader to grasp the readership of linear regression in industry, as well as in consulting [17].

Moreover, the entire chapter stresses the importance of linear regression as an entry point to the other more advanced statistical models. It is a fundamental statistical method [emerging] from simple linear regression, and many of the current approaches are combining or modifications of the simple linear regression. Understanding linear regression is crucial before making the shift towards more complex machine learning flows [19].

To the same, the authors also point towards the aspect of model assessment using residual plots, R-squared and MSE for establishing correct predictive values and valid insights.

In conclusion, the chapter establishes linear regression not only as a practical tool for predictive modelling but also as a conceptual foundation for learning more advanced statistical techniques.

- *Advances in Linear Regression for Machine Learning and Predictive Analytics*

Linear regression is perhaps one of the simplest and most common algorithms in machine learning helping perform predictive analysis across various industries. The technique of linear regression was established by Sir Francis Galton in 1894; it relates the dependent variable to the independent variable. In that angle, as a mathematical technique, it allows for the estimation of continuously distributed or real variables of relationships within databases [20].

This technique provides an opportunity to express and evaluate the extent of association and dependency between variables more effectively than univariate approaches, such as ANOVA, Chi-square or Fisher's exact tests, especially when many variables or covariates are present. Regression has a considerable advantage over other methods when one is in a position to analyze the effect of an independent variable in light of other confounding factors [6,16]. Partial correlation and also regression kinds of relationships are far more helpful in analysing these relationships and enabling clarification about the complicated and closely intertwined forces in question.

As a prediction technique and a model of statistical learning, linear regression is used in machine learning. It acts as the basis for continuous outcomes prediction by applying a linear relationship between the predictor variables and a target variable. In addition, regression is considered as the basis for other more complex procedures. Linear regression lies at the heart of many contemporary methods and statistical techniques, as well as being found at the root of numerous newly developed machine learning methods [18].

Most recent papers on linear regression are devoted to enhancing the time complexity of the algorithm, as well as the examination of new data and the corresponding performance comparison. Linear regression models have been widely used by researchers in different fields of study, given data sets and algorithms to test the prediction power and robustness of an algorithm. These efforts also investigate feature selection methodologies to enhance performance levels. The review presented in this paper includes an overview of the most commonly used techniques for linear regression during the last five years, the algorithms employed, the databases utilized and the measures of performance [7].

- *The Role of Linear Regression in Machine Learning: Applications, Advancements, and Challenges*

Linear regression is one of the category-defining algorithms in machine learning, commonly used in financial, medical and engineering domains as a predictive model. The method describes the relationships between a set of dependent (target) variables and a set of independent (input) variables from observed data by estimating a linear equation with the least square technique. Both ridge regression and Lasso regression are born in advanced regression for dealing with multi co linearity and over fitting issues after facing today's modern world challenges for regression [24].

Literature on machine learning focuses on applying linear regression for both univariate and multivariate predictions citing the simplicity of the technique compared to another complex model. Linear regression is favoured for the projection of continuous variables as it can be used to forecast outcomes that could be estimated in real numbers including the rate of stocks or health of a patient [8].

In the current studies, linear regression is made as a comparison with other ML techniques such as decision trees, neural nets and ensembles. While there might be the existence of non-linearity in these cases, linear regression can work best when the two variables in a dataset are related linearly. Succeeding reforms have brought decrements within the blend of linear regression using other examinations to improve both, including the use of ridge regressions in minimizing the effect of multi co linearity [9].

It is also apparent in the partial correlation analysis where the method is employed enabling the researchers to hold constant interference which may distort the relationship between predictors and outcomes. This is important especially when several covariates are present such as in the analysis of epidemics [23]. Many works demonstrate that linear regression is still important with newer ML models due to its simplicity, speed to solve and efficiency for small to medium datasets. In conclusion, linear regression has an important position in the machine learning system as an interpreter of the real models along with the analysis.

- *Literature Gap:*

There are quite several gaps in the literature that remains to be explored even with the broad use and recognized importance of linear regression in machine

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

learning. First, although great attention has been paid to the evaluation of linear regression in different fields, few works systematically investigate how it compares with other models in practice especially in datasets with high dimensionality [10]. Second, prior work frequently does not include the investigation of the combination of linear regression with deep learning and ensemble learning, which can help improve predictive performance and model generalization. Furthermore, how the quality of the data and the data pre-processing affected the linear regression model has not garnered much attention, especially in other disciplines such as health and finance. Last but not least, more research seems to be needed on the analysis of users' interpretations and the communication of the linear regression models and the results with stakeholders who may not have technical backgrounds.

- *Theoretical Framework:*

The theoretical basis for linear regression in machine learning is underpinned by the following statutes of statistical theories. Although the topic of regression analysis could fill a whole book, the heart of the standard linear model lies in the Ordinary Least Squares (OLS) theory, which seeks to minimize the sum of squared residuals, defined as the difference between the actual and predicted values to estimate the parameters of a linear model. According to the Gauss-Markov Theorem OLS possesses the property of Being the Best Linear Unbiased Estimators any efficiency in parameter estimation work is guaranteed [21].

In addition, a theory known as Statistical Learning Theory offers information on how to use complexity for generalization and the problem encountered when employing linear regression on high cardinality data sets [22]. This theory gives rise to the idea of Regularization, Lasso (L1 regularization) and Ridge (L2 regularization) being used in handling multi co linearity and also to make the models stronger.

Furthermore, an approach within regression analysis is Bayesian Inference that allows prior beliefs towards the parameters of the model to be balanced. This approach provides the means for quantification of the uncertainty of the predictions made [11]. Finally, the Causal Inference Theory is crucial in understanding the mechanisms by which independent and dependent variables function, so that researchers can implement correlation effects rather than correlations, which increase the value of linear regression results.

- *Summary:*

In this chapter, the theoretical base of linear regression is discussed under the lens of the learning process in the field of machine learning. OLS underpins it, which provides the smallest squared-error estimates; more to that, the OLS model can offer unbiased estimates as the Gauss-Markov Theorem postulates. The chapter focuses on Statistical Learning Theory and the concept of model complexity with the risk of over fitting, new techniques for model stability in many-feature spaces: Lasso and Ridge. Also, it discusses the Bayesian inference that enables prior beliefs and the measure of uncertainty in parameter estimates. Finally, the Causal Inference Theory is discussed in the chapter whereby the authors explain how to establish causality between two variables, which is useful when analyzing findings of linear regression. Together, these theories offer a green view on the principles and uses of linear regression as employed in data analysis studies.

III. METHODOLOGY

This chapter outlines the methodology employed to investigate the effectiveness of linear regression models, specifically focusing on Ordinary Least Squares (OLS), Baseline regression, and polynomial regression. The research aims to evaluate the predictive accuracy of these models using two variables, with a strong emphasis on statistical evaluation parameters such as R-squared and Mean Squared Error (MSE).

The methodology comprises data collection, pre-processing, and analysis procedures designed to ensure the reliability of results. By incorporating graphical performance assessments like residual plots, this study seeks to enhance understanding of model fit and prediction capabilities. Additionally, the methodology will address the impact of feature selection and the handling of outliers, providing a comprehensive framework for evaluating linear regression in diverse applications. Ultimately, this chapter aims to clarify the systematic approach adopted to achieve the research objectives and answer the defined research questions effectively.

- *Research Approach and Design:*

This research employs a quantitative approach, utilizing secondary data to explore ecological sexual dimorphism and environmental variability in Antarctic penguin communities. The primary focus is on three species: Adélie, Chinstrap, and Gentoo penguins. The design is structured to analyze existing datasets collected during extensive field studies conducted from 2007 to 2009 across various Antarctic islands, including Torgersen, Dream, and Biscoe Islands.

The secondary data sources encompass biological measurements of penguins, including bill length, bill depth, flipper length, body mass, sex, and species classification. These measurements were gathered through systematic observations during the breeding seasons, ensuring a comprehensive dataset that reflects the penguins' ecological characteristics and adaptations to environmental conditions.

Data analysis will employ statistical methods to assess the relationships between sexual dimorphism and environmental factors such as temperature, food availability, and habitat type. This quantitative analysis aims to determine if significant differences exist in morphological traits between sexes across the different species and how these traits may correlate with environmental variables.

The research design acknowledges potential limitations associated with secondary data, such as inconsistencies in data collection methods and potential biases in the original studies. However, the utilization of established datasets

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

allows for a broader analysis and minimizes resource expenditure while still providing valuable insights into penguin ecology.

By adopting a structured quantitative approach and relying on secondary data, this study aims to enhance understanding of ecological patterns and sexual dimorphism within penguin communities, contributing to conservation strategies and informing future research in this critical area of ecological study.

➢ *Tools and Techniques:*

This research adopts a quantitative approach to evaluate the effectiveness of various linear regression models—specifically, Ordinary Least Squares (OLS), Baseline regression, and polynomial regression—in predicting outcomes based on two selected variables. The quantitative methodology allows for objective measurement and statistical analysis, facilitating comparisons of model performance. The study employs a comparative research design, focusing on the assessment of predictive accuracy through statistical metrics such as R-squared and Mean Squared Error (MSE). Data will be collected from relevant sources, ensuring a representative sample that captures the relationships among the variables of interest.

The analysis will involve pre-processing steps, including normalization and handling of outliers, to enhance data quality and model robustness. By utilizing graphical methods, such as residual plots, the research aims to visualize model fit and evaluate the assumptions underlying linear regression. This structured approach enables a comprehensive understanding of how each model performs under specific conditions, ultimately contributing to informed decision-making in various disciplines. The findings will provide insights into the applicability and limitations of linear regression models in analyzing complex datasets. Python programming language is used for this implementation.

➢ *Data description:*

- *Dataset Summary and Explanation*

In 2007, a dataset was gathered with data regarding Adelie penguins living in the Torgersen and Biscoe islands. It consists of multiple body parameters of a particular penguin and other factors that could be species and gender and the year when it was observed. The key variables in this dataset are described below:

✓ *Species:*

All data present in the dataset feature Adelie, a most ordinary type of penguin characterized by a black head and white under parts.

✓ *Island:*

The data includes information for two islands: Torgersen and Biscoe which belong to the Palmer Archipelago and are ideal areas for studying penguins.

- *Bill Length (mm) and Bill Depth (mm):*

These measurements are in millimetres and give the length of the penguin's bill and the depth of the groove that runs down its centre. Such fluctuations imply bill dimensions assist researchers in determining species behaviour, feeding patterns and sexual selection.

✓ *Flipper Length (mm):*

This measurement normally involves a record of the flippers in penguin and this is very important in realizing the swimming capacity of the animal. For instance, the attributes of flipper length as presented in the range between 174 mm and 198 mm.

✓ *Body Mass (g):*

Body mass in grams varies between 3200g respectively 4675g. This metric is useful for seeing how penguins are doing and their condition in the various areas where they are situated.

✓ *Sex:*

Some records indicate sex information, including male or female; however, there are many records of missing values noted as "NA." A lot of behavioural patterns, reproductive success and population dynamics can be explained by determining sex.

✓ *Year:*

Every record in the dataset has the year 2007, which means that the given dataset contains information on the penguins of that particular year only.

✓ *Missing Values:*

The darker shaded cells denote rows with missing data, labelled by "NA". This could have been due to issues regarding measurement in the field such as incomplete measurement or human factors.

Altogether, the dataset under analysis may be used as a starting point for studying other biological patterns of penguin populations: for instance, the ratio of males to females and sexual dimorphism in size and the impact of environmental factors.

➢ *Data collection:*

The research followed the secondary approach of data collection which mainly focuses on collecting data from internet sources. The data on penguins in this study was collected from the research titled *"Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)."* The data focuses on three penguin species—Adélie, Chinstrap, and Gentoo—and explores ecological differences, including sexual dimorphism, environmental adaptations, and species-level variability. The data collection follows a **secondary data approach**, meaning the dataset was compiled from previously collected field observations and research. The original researchers gathered detailed biological measurements for individual penguins, including attributes like bill length, bill depth, flipper length, body mass, sex, and species classification. Field studies were conducted across

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

several Antarctic islands, such as Torgersen, Dream, and Biscoe Islands, which represent distinct environments. Secondary data approaches like this involve reusing datasets collected by other researchers or institutions. It saves time and resources, allowing the focus to shift toward analysis rather than data acquisition. However, challenges can arise, such as missing data—evident here, where some measurements and sex classifications were incomplete. In such cases, data pre-processing steps, including imputation or deletion of missing entries, are essential to ensure analytical accuracy. This dataset provides valuable insights into penguin ecology, environmental impacts, and physical variability while demonstrating how secondary data reuse can enhance research efficiency.

➢ *Data analysis:*

- *Procedure for Data Analysis*

Exploratory data analysis starts with descriptive statistics the penguin dataset that have 344 observations and eight variables: The primary attributes of interest being bill length, bill depth, flippers length, and body mass. Some values were omitted or unknown, these cases were treated in a pre-processing step to ensure that all resulting dataset values were not null. Categorical variables were then converted into one-hot encoded columns to make it easy to use the algorithms in machine learning.

IV. FINDINGS AND ANALYSIS

➢ *Data Pre-Processing*

Data inspection starts with data pre-processing, where the data collected is cleaned, and checked if it is relevant or not and in the right format. Here, it is about handling missing values, data cleaning of features, and outliers' removal or managing them to improve on the predictive model. Next, variable selection is carried out to select the most important variables, normally, depending on the correlation analysis and variance thresholds in an attempt to eliminate the multi-collapse problem and over fitting.

➢ *Descriptive Statistics*

Following pre-processing, Exploratory Data Analysis is performed. Descriptive data analysis is used by employing histograms, scatter plots and box plots to understand the nature of correlations. For regression models, scatter plots with the trend line are a good way to look at whether there is a straight-line relationship between dependent and independent variables. Furthermore, the heat map of correlation allows for analyzing multi co linearity which is often a problem that requires either variable to be omitted or transformed.

Visualizations, including histograms, box plots, and violin plots were employed as part of EDA to demonstrate the distribution of body mass and flipper length in the species. The distribution of body mass appeared to be right-skewed with evidence of the peak around 3500-4000 grams whereas flipper length also fell around 190 mm – results that are of a different order to those inside the relationship. The results from a correlation matrix demonstrated a considerable positive relationship between body mass and both FL and BL, which stated that these two variables are good measures for predicting body mass.

The variability in body mass across species is conveyed using box plots, violin plots and scatter plots and Gentoo penguins exhibited the highest body mass and the highest amount of variation. Scatter plots reinforced the positive correlation between flipper length and body mass, differentiating species through colour coding. Overall, this analysis provides insights into the biological characteristics and relationships among different penguin species.

➢ *Description of Dataset*

```
data.describe()
```

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
count	342.000000	342.000000	342.000000	342.000000	344.000000
mean	43.921930	17.151170	200.915205	4201.754386	2008.029070
std	5.459584	1.974793	14.061714	801.954536	0.818356
min	32.100000	13.100000	172.000000	2700.000000	2007.000000
25%	39.225000	15.600000	190.000000	3550.000000	2007.000000
50%	44.450000	17.300000	197.000000	4050.000000	2008.000000
75%	48.500000	18.700000	213.000000	4750.000000	2009.000000
max	59.600000	21.500000	231.000000	6300.000000	2009.000000

Fig 1 Description of Dataset

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

Figure 1 gives the description from which we get quantitative data of measurements of penguins such as bill length, bill depth, flipper length and body mass in several years (2007-2009). These variables are useful while assessing linear regression models as they have strong quantifiable relationship, signal fluctuation and patterns. To the research aim, these features enable a determination of the relationship prediction accuracy of Ordinary Least Squares (OLS), Baseline, and Polynomial regression by using appropriate parameters such as R-squared and Mean Squared Error (MSE). Mean, standard deviation, and range values of the dataset enable pattern analysis for underlying data, correlation identification, and model evaluation on real and different datasets.

➢ *Implementation of Exploratory Data Analysis*

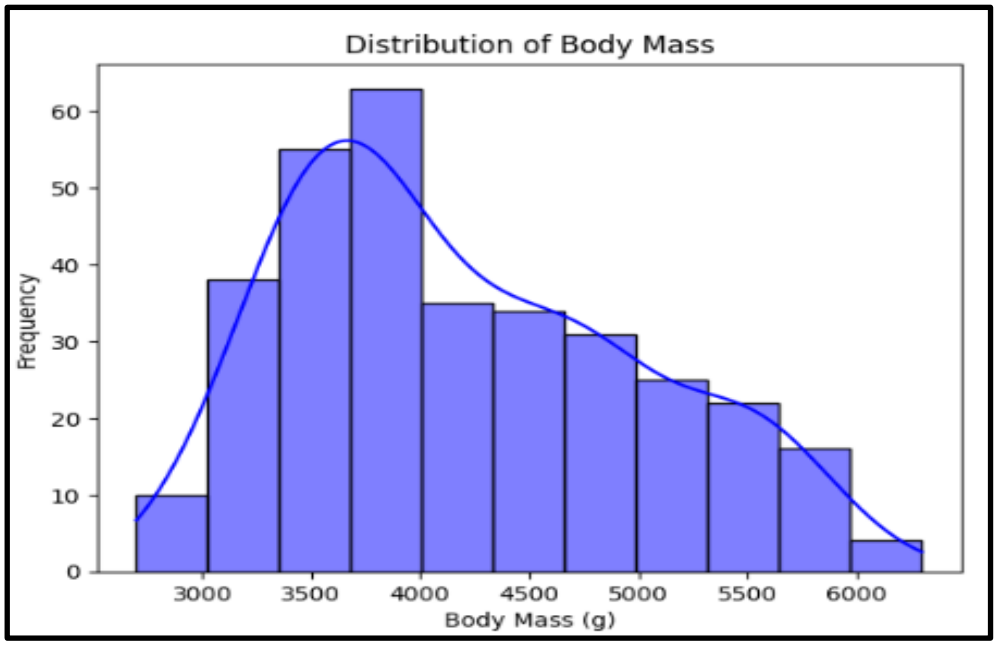

Fig 2 Distribution of Body Mass

Figure 2 shows the distribution of body mass, which appears approximately right-skewed; with the most frequent body mass values centred around 3500-4000 grams. The smooth curve overlay suggests a unimodal distribution, tapering off at higher body mass values above 5000 grams.

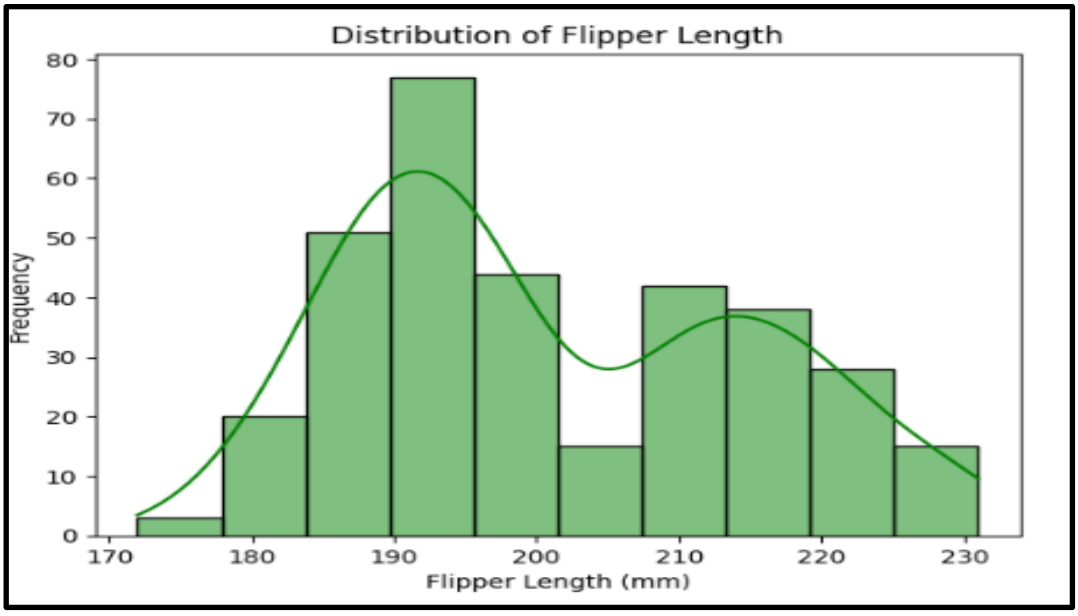

Fig 3 Distribution of Flipper Length

Figure 3 shows a histogram plot of flipper lengths for penguins, with the distribution peaking around 190 mm and following a roughly slightly right-skewed shape. The overlaid density curve suggests multiple small sub-peaks, indicating some variability in flipper length across the population.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

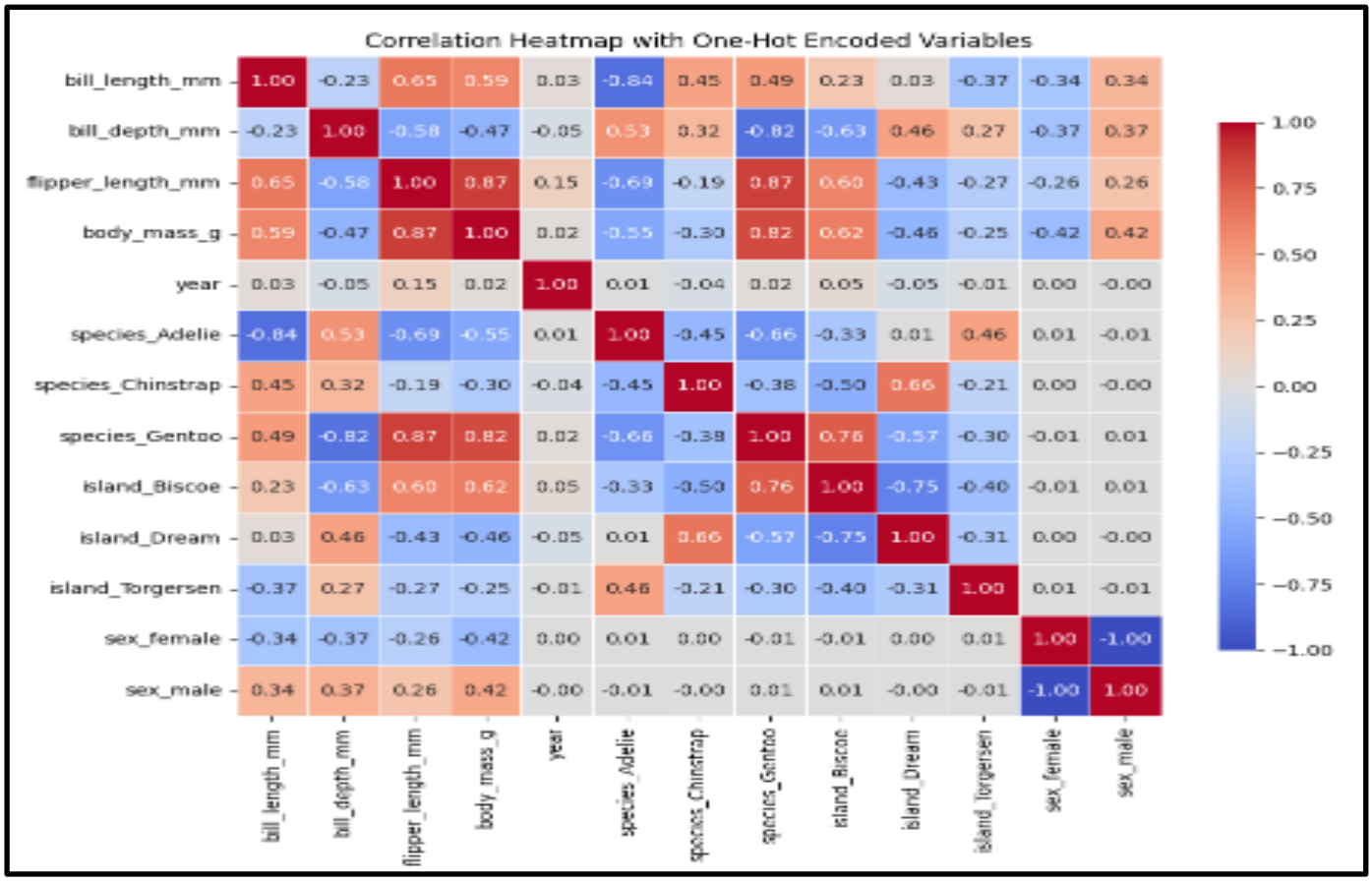

Fig 4 Correlation Heat Map

Figure 4 shows the relationships between penguin features, including one-hot encoded categorical variables. The body mass has a strong positive correlation with flipper length and bill length, suggesting these are important predictors for body mass, while negative correlations appear for species-specific variables, indicating some variance across penguin species.

➢ *Performance of Linear Regression Model*

The linear regression model's performance on predicting body mass using flipper length is summarized by the following metrics:

➢ *Mean Squared Error (Mse):*

135158.65 Indicates the average squared difference between the predicted and actual body mass values. Lower MSE suggests better model accuracy.

➢ *Mean Absolute Error (Mae):*

288.82 Represents the average absolute error in the predictions, providing a straightforward interpretation of the prediction error.

➢ *R-Squared (R^2):*

0.75 shows that approximately 75% of the variance in body mass can be explained by the flipper length feature. This indicates a relatively strong t, though there is room for improvement. These metrics serve as a foundation for assessing the accuracy and performance of the model.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

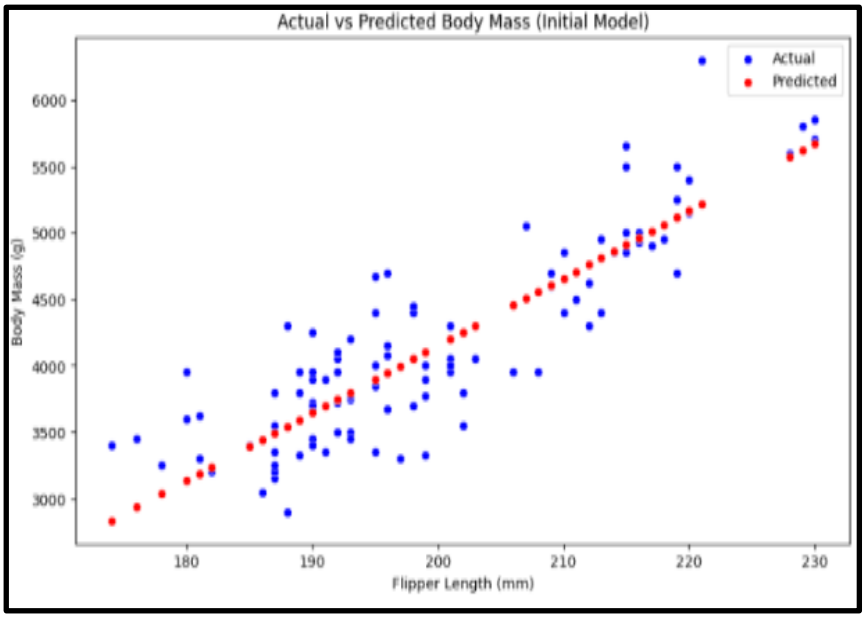

Fig 5 Actual vs. Predicted Body Mass

Figure 5 shows the relationship between the actual and predicted body mass values (in grams) based on flipper length (in mm) for the OLS linear regression model. The model's predictions (in red) closely follow the trend of the actual values (in blue), though there are some deviations, indicating that while the model captures the general linear relationship, there are instances where it under or overestimates body mass.

➢ *Performance of Polynomial Model:*

- *Mean Squared Error (MSE):*

Measures the average squared difference between the actual and predicted values, with a result of 121,892.07. A lower MSE indicates better model performance.

- *Mean Absolute Error (MAE):*

Calculates the average absolute difference between actual and predicted values, with a result of 271.25, showing the model's average prediction error in the same units as the target variable.

- *R-Squared (R^2):*

Indicates the proportion of variance in the target variable explained by the model, with a score of 0.77.

This value shows that the model explains 77% of the variance in the test data, suggesting a fairly good t for the polynomial model. Overall, these metrics imply that the polynomial model performs reasonably well but could still be improved by implementing model tuning.

➢ *Performance using Baseline Model:*

- *Mean Squared Error (MSE):*

548537.91 This high MSE value rejects substantial error in predictions when only using the mean as the predictor.

- *Mean Absolute Error (MAE):*

621.93 Shows the average error in predictions, which is relatively large, indicating poor predictive accuracy.

- *R-Squared (R^2):*

-0.02 A negative R^2 value indicates that the baseline model does worse than a horizontal line at the mean, emphasizing the need for a more sophisticated model. These metrics highlight that the linear regression model, with an R^2 of 0.75, significantly outperforms the baseline.

➢ *Comparison of Regression Predictions with Actual Predictions*

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

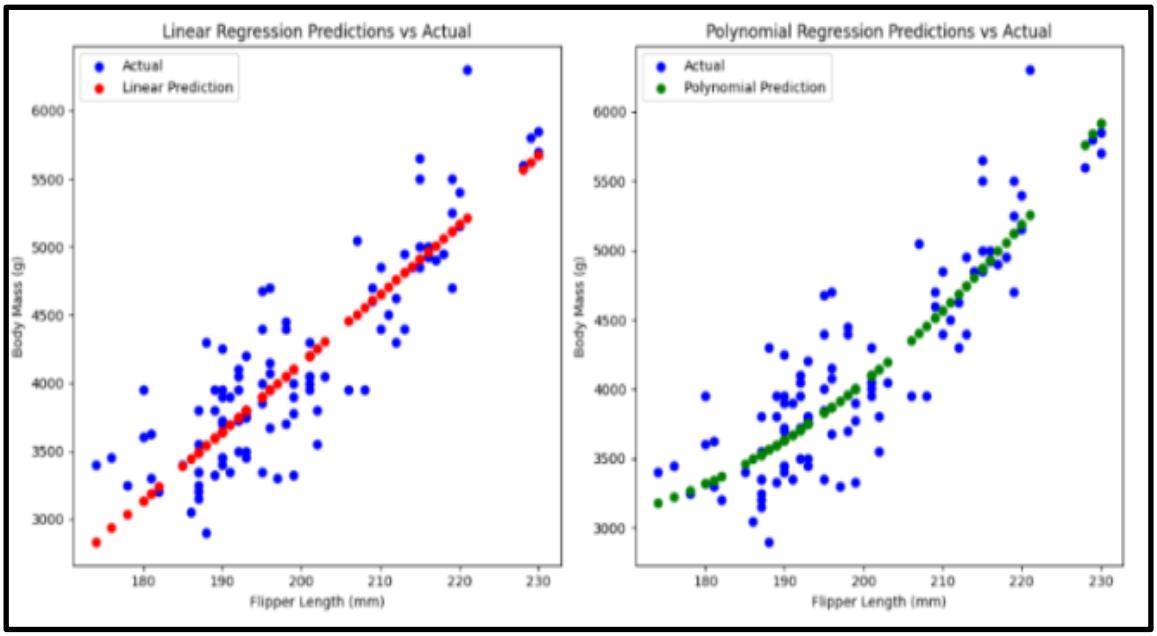

Fig 6 Comparison of Regression Predictions with Actual Predictions

Figure 6 shows the predicted values (in red) closely follow a straight line, capturing the general trend of the actual values (in blue) but missing some of the non-linear patterns in the data. In the polynomial regression plot (right), the predictions (in green) more accurately capture the curve in the data distribution, providing a closer fit to the actual values, especially for extreme flipper lengths.

--- Model Comparison Summary ---

Model	MSE	R^2
Baseline Model	548538	-0.02
Linear Model	135159	0.75
Polynomial Model	121892	0.77

Fig 7 Comparison of Model Summary

Figure 7 shows the Baseline Model has a high MSE (548538) and a negative (R^2), indicating poor fit. The Linear Model significantly improves performance with a lower MSE (135159) and a strong (R^2) of 0.75. The Polynomial Model performs the best, with the lowest MSE (121892) and a slightly higher (R^2) of 0.77, suggesting it explains the data variance most effectively.

The plots show that the Baseline Model has the highest Mean Squared Error (MSE) and the lowest (R^2) value, indicating poor predictive accuracy. In contrast, the Linear and Polynomial Models have significantly lower MSE and higher \(R^2 \) values, suggesting they perform much better at explaining the variance in the data, with the Polynomial Model achieving the lowest MSE.

V. CONCLUSION

This chapter has shown the usefulness of the different regression models that were built to predict the body mass of the penguins using the flipper length and other characteristics. During data cleaning and preliminary inspection, it is possible to identify patterns, for example, density values of the body mass, flipper length, and the length of the bill are positively correlated. The other models considered include Polynomial Regression, of which it had the lowest value of Mean Squared Error, 121,892 and the highest R square value (0.77). Such findings suggest that the utilization of polynomial relations leads to improvement in aspects such as; non-linear relationship prediction. On the flip side, low evidence in support of the Baseline model proves some systems' worth on which more complex models rely for accurate forecasts.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

➢ *Future Scope:*

Future work can include the following:

- *Feature Engineering:*

Expanding the range of included biological and environmental indicators in order to make more accurate predictions.

- *Advanced Models:*

Looking on how Random Forests or Gradient Boosting techniques and so on can be used to provide stronger predictions.

- *Cross-Validation:*

To guarantee the model stability across the different subsets of the data set, We decided to use the k-fold cross validation strategy.

- *Application Expansion:*

Perpetuating the information contained in the models to other animal species or different environmental data for the improvement of general ecological knowledge.

- *Hyperparameter Tuning:*

Tuning coefficients of the model for better performance in Polynomial and other sophisticated models of machine learning.

Such improvements may bring more detailed and diverse capabilities to the applications in the field of ecological and biological research.

➢ *Relation with Sustainability:*

This paper relates to sustainable development in the following ways:

- *Biodiversity Understanding:*

The analysis of penguin features like body mass and flipper length helps monitor species' health, contributing to conservation efforts vital for ecosystem balance.

- *Data-Driven Conservation:*

Regression models provide insights into environmental and biological factors affecting species, enabling data-driven decisions for habitat preservation.

- *Education and Awareness:*

Promotes the use of machine learning and statistical tools in ecological research, fostering interdisciplinary approaches critical for sustainable practices.

- *Sustainable Methodology:*

Encourages minimal environmental disturbance through non-invasive data collection and analysis techniques, aligning with sustainable research principles.

REFERENCES

[1]. Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. Introduction to linear regression analysis. *John Wiley & Sons.*

[2]. James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J., 2023. Linear regression. In An introduction to statistical learning: With applications in python (pp. 69-134). Cham: Springer International Publishing.

[3]. Maulud, D. and Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(2), pp.140-147.

[4]. Kibria, B.G. and Lukman, A.F., 2020. A new ridge-type estimator for the linear regression model: simulations and applications. *Scientifica, 2020(1),* p.9758378.

[5]. Abu-Faraj, M.A., Al-Hyari, A. and Alqadi, Z., 2022. Experimental Analysis of Methods Used to Solve Linear Regression Models. *Computers, Materials & Continua*, 72(3).

[6]. Ottaviani, F.M. and De Marco, A., 2022. Multiple linear regression model for improved project cost forecasting. *Procedia Computer Science*, *196*, pp.808-815.

[7]. Etemadi, S. and Khashei, M., 2021. Etemadi multiple linear regression. *Measurement*, 186, p.110080.

[8]. Shaker Reddy, P.C. and Sureshbabu, A., 2020. An enhanced multiple linear regression model for seasonal rainfall prediction. *International Journal of Sensors Wireless Communications and Control*, *10*(4), pp.473-483.

[9]. Shewa, G.A. and Ugwuowo, F.I., 2023. A new hybrid estimator for linear regression model analysis: Computations and simulations. *Scientific African*, *19*, p.e01441.

[10]. Gupta, A.K., Singh, V., Mathur, P. and Travieso-Gonzalez, C.M., 2021. Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario. *Journal of Interdisciplinary Mathematics*, *24*(1), pp.89-108.

[11]. Arum, K.C., Ugwuowo, F.I., Oranye, H.E., Alakija, T.O., Ugah, T.E. and Asogwa, O.C., 2023. Combating outliers and multicollinearity in linear regression model using robust Kibria-Lukman mixed with principal component estimator, simulation and computation. *Scientific African*, *19*, p.e01566.

[12]. Ghosh, P., Hazra, S., and Chatterjee, S. Future Prospects Analysis in Healthcare Management Using Machine Learning Algorithms. *the International Journal of Engineering and Science Invention (IJESI), ISSN (online)*, 2319-6734.

[13]. Hazra, S., Mahapatra, S., Chatterjee, S., and Pal, D. 2023. Automated Risk Prediction of Liver Disorders Using Machine Learning. In *the proceedings of 1st International conference on Latest Trends on Applied Science, Management, Humanities and Information Technology (SAICON-IC-LTASMHIT-2023) on 19th June* (pp. 301-306).

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul349

[14]. Gon, A., Hazra, S., Chatterjee, S., and Ghosh, A. K. 2023. Application of machine learning algorithms for automatic detection of risk in heart disease. In *Cognitive cardiac rehabilitation using IoT and AI tools* (pp. 166-188). IGI Global.

[15]. Das, S., Chatterjee, S., Sarkar, D., and Dutta, S. 2022. Comparison Based Analysis and Prediction for Earlier Detection of Breast Cancer Using Different Supervised ML Approach. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 3* (pp. 255-267). Singapore: Springer Nature Singapore.

[16]. Das, S., Chatterjee, S., Karani, A. I., and Ghosh, A. K. 2023, November. Stress Detection While Doing Exam Using EEG with Machine Learning Techniques. In *International Conference on Innovations in Data Analytics* (pp. 177-187). Singapore: Springer Nature Singapore.

[17]. Hazra, S. 2024. Pervasive nature of AI in the health care industry: high-performance medicine.

[18]. Sima Das, Siddhartha Chatterjee, Sutapa Bhattacharya, Solanki Mitra, Arpan Adhikary and Nimai Chandra Giri "Movie's-Emotracker: Movie Induced Emotion Detection by using EEG and AI Tools", In the proceedings of the 4th International conference on Communication, Devices and Computing (ICCDC 2023), Springer-LNEE SCOPUS Indexed, DOI: 10.1007/978-981-99-2710-4_46, pp.583-595, vol. 1046 on 28th July, 2023.

[19]. Chatterjee, R., Chatterjee, S., Samanta, S., & Biswas, S. (2024, December). AI Approaches to Investigate EEG Signal Classification for Cognitive Performance Assessment. In *2024 6th International Conference on Computational Intelligence and Networks (CINE)* (pp. 1-7). IEEE.

[20]. Adhikary, A., Das, S., Mondal, R., & Chatterjee, S. (2024, February). Identification of Parkinson's Disease Based on Machine Learning Classifiers. In *International Conference on Emerging Trends in Mathematical Sciences & Computing* (pp. 490-503). Cham: Springer Nature Switzerland.

[21]. Ghosh, P., Dutta, R., Agarwal, N., Chatterjee, S., and Mitra, S. (2023). Social media sentiment analysis on third booster dosage for COVID-19 vaccination: a holistic machine learning approach. *Intelligent Systems and Human Machine Collaboration: Select Proceedings of ICISHMC 2022*, 179-190.

[22]. Rupa Debnath; Rituparna Mondal; Arpita Chakraborty; Siddhartha Chatterjee 2025 Advances in Artificial Intelligence for Lung Cancer Detection and Diagnostic Accuracy: A Comprehensive Review. International Journal of Innovative Science and Research Technology, 10(5), 1579-1586. https://doi.org/10.38124/IJISRT/25may1339

[23]. Nitu Saha; Rituparna Mondal; Arunima Banerjee; Rupa Debnath; Siddhartha Chatterjee; (2025) Advanced Deep Lung Care Net: A Next Generation Framework for Lung Cancer Prediction. International Journal of Innovative Science and Research Technology, 10(6), 2312-2320. https://doi.org/10.38124/ijisrt/25jun1801

[24]. Poushali Das; Washim Akram; Arijita Ghosh; Suman Biswas; Siddhartha Chatterjee (2025) Enhancing Diagnostic Accuracy: Leveraging Continuous pH Surveillance for Immediate Health Evaluation. International Journal of Innovative Science and Research Technology, 10(7), 7-12. https://doi.org/10.38124/ijisrt/25jul123

[25]. Manali Sarkar; Aparajita Das; Sraddha Roy Choudhury; Siddhartha Chatterjee 2025. A* Based Optimized Travel Recommendation System for Smart Mobility. International Journal ofInnovative Science and Research Technology, 10(5), 3185-3193. https://doi.org/10.38124/ijisrt/25may2352

[26]. Hazra, S., Chatterjee, S., Mandal, A., Sarkar, M., Mandal, B.K. 2023. An Analysis of Duckworth-Lewis-Stern Method in the Context of Interrupted Limited over Cricket Matches. In: Chaki, N., Roy, N.D., Debnath, P., Saeed, K. (eds) Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023. ICDAI 2023. Lecture Notes in Networks and Systems, vol 727. Springer, Singapore. https://doi.org/10.1007/978-981-99-3878-0_46