EMPLOYEE SALARY ANALYSIS AND PREDICTION USING MACHINE LEARNING

|  |  |  |  |
| --- | --- | --- | --- |
| M.Suneetha | K. Narasimha Charyulu | A. Madhava Rao | U. Sai Kumar |
| Department of Computer Science and Engineering | Department of Computer Science and Engineering | Department of Computer Science and Engineering | Department of Computer Science and Engineering |
| Narasaraopeta Engineering College | Narasaraopeta Engineering College | Narasaraopeta Engineering College | Narasaraopeta Engineering College |
| Narasaraopeta | Narasaraopeta | Narasaraopeta | Narasaraopeta |
| [msuneetha973@gmail.com](mailto:msuneetha973@gmail.com) | [kanjarlanarashimhacharyulu@gmail.com](mailto:kanjarlanarashimhacharyulu@gmail.com) |  |  |

# Abstract

Accurately estimating employee pay is essential for both companies and job seekers in today's competitive employment market. This paper suggests a technique that makes use of machine learning algorithms to forecast employee pay. The dataset includes a number of characteristics, including industry, region, employment function, years of experience, and education level. Different machine learning models, such as gradient boosting methods, decision trees, random forests, and linear regression, are trained and evaluated using these features. Several performance indicators are employed to evaluate the models' accuracy and efficacy, The outcomes show how well machine learning approaches work in predicting employee pay with a high level of precision. Employers can use these models to make well-informed judgments about compensation offers, and job seekers can use them to learn what kind of salary range they should expect given their experience and qualifications.

# INTRODUCTION

Precisely forecasting employee pay is a basic duty in the field of human resources management that has major consequences for businesses and employees alike.

Conventional techniques for determining salaries frequently rely on opinionated evaluations or accepted practices in the field, which could introduce bias or incorrect information. On the other hand data-driven approaches to salary prediction are now available in machine learning which make decision making processes more accurate and transparent. For this endeavour, regression algorithms are very useful since they let us model the link between a variety of characteristics, including industry, location, years of experience, job function, and education level, and the related salary. Regression algorithms are highly predictive because of their capacity to identify intricate links in the data and reveal patterns and trends that would not be visible using more conventional techniques. We aim to determine the best method for wage prediction in a real-world setting by evaluating the output of several regression models, such as Linear regression, polynomial regression, lasso regression, multivariate linear regression. Using univariate regression analysis, the relationships between a dependent and independent variable are examined. Since multiple algorithms will perform differently for a given data set, our approaches compare each regression model's performance to determine which

algorithm performs best for pay predictions.

# LITERATURE SURVEY

A quick overview of the different machine learning algorithms that are most commonly used to solve problems with classification, regression, and clustering is provided by **Susmita Ray[8]** in her paper . **Sananda Dutta, Airiddha Halder, and Kousik Dasgupta[6]** "Creation of an innovative Prediction Engine to Estimate Appropriate Compensation for an Employment". The dataset supplied by ADZUNA serves as the foundation for our investigation.

**Pornthep Khongchai and Pokpong Songmuang's study[3],** "Improving Students' Motivation to Study using Salary Prediction System," suggested a method yield a predicted wage.

In their study, "Salary Predictor System for Thailand Labour Workforce using Deep Learning," **Phuwadol Viroonluecha and Thongchai Kaewkiriya[9]** employed deep learning techniques to build a model that accurately predicts the monthly salary of job seekers in Thailand by solving a regression problem with a numerical outcome.

Associative rule mining was used by **K. Lakshmi and A. Parkavil[1]** to analyze data in order to determine the extent of

students' knowledge. They used linear regression and association rule mining to determine which group (or cluster) the student knowledge is most closely related to Data mining .

According to research by **Rajveer Singh[7]**, academic achievement in school and college, affiliation with the institution, and college reputation are significant predictors of starting income for entry-level Indian engineering graduates. In their work, **C.-C. Hung and E.-P. Lim[10]** suggested the "Company, Occupation, Company" (COC) model as a means of obtaining objective salaries through the combination of job review and job post data. This algorithm is able to accurately forecast organizations' inflation, competitiveness, and unbiased compensation.

# PROPOSED MODEL

* 1. **Factors considered in Proposed work**

|  |
| --- |
| Age |
| PHD |
| Gender |
| Salary |

# Dataset Description

|  |
| --- |
| **Name:** Employee Salaries |
| **Source:** Kaggle |
| **Attributes:** 4 |
| **Instances:**100 |
| **Type:** Supervised |
| **Training Data:**80% |
| **Testing Data:**20% |

## Data Preproccesing

Preprocessing is the first step in creating a machine learning model. The process entails transforming an unclean dataset into a cleaned one. Because raw data includes noise, missing values, and duplicate values, it is not appropriate for machine learning models. For this reason, preprocessing is required.

## Splitting Data into Training and Test Data

When estimating the effectiveness of machine learning algorithms that can be used to prediction-based algorithms and applications, the train-test split is utilized.This is a quick and simple process that allows us to compare the output of our machine learning model with that of other machines. By default, 30% of the real data is divided into the Test set and 70% of the real data is divided into the Training set. To assess how well our machine learning model works, we must divide a dataset into train and test sets. The most reliable and practical Python machine learning library is called scikit-learn, sometimes known as sklearn. The splitter function train\_test\_split() is available in the model\_selection module of the scikit-learn toolkit.

**TrainingData:**80% **Testing Data:**20%

## Feature Extraction

Dimensionality reduction techniques aim to simplify complex datasets by reducing the number of variables or features while preserving essential information. One common method within dimensionality reduction is feature extraction. By extracting relevant features from raw data, the process becomes more manageable and computationally efficient, particularly when dealing with large datasets characterized by numerous variables The abundance of variables in large datasets poses a significant computational challenge, requiring substantial computer power for processing. Feature extraction addresses this challenge by selecting and combining variables into meaningful features, thereby reducing the data's dimensionality without losing crucial information. This streamlined approach enhances the efficiency of processing large datasets by focusing on extracting the most relevant features for analysis.

## Alogirthms Used in our Model

* + 1. **Linear Regression**

Linear regression is a fundamental and widely utilized machine learning technique for predictive analysis. It serves as a statistical method for modeling the relationship between one or more independent variables .

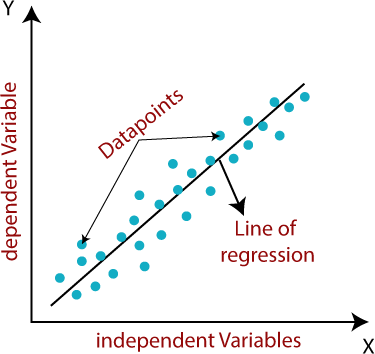


Fig-3.6.1.1 Linear Regression

## Multi Linear Regression

Multiple linear regression indeed extends the concept of simple linear regression by accommodating multiple independent variables to predict a continuous dependent variable. It's a valuable regression technique for modeling complex relationships between the dependent variable and multiple predictors.

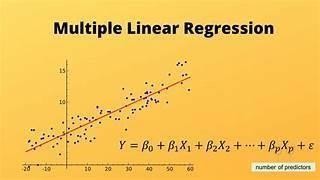


Fig-3.6.2.1 Multi Linear Regression

## 3.6.4Polynomial Regression

In polynomial regression, a polynomial of degree is used to model the relationship between an independent variable(x) and a dependent variable (y) y). Unlike simple linear regression, which assumes a linear relationship between the variables, polynomial regression allows for more flexible modeling of nonlinear relationships. y = b0 + b1 \* 1 + b2 \* 12 + b2 \* 13 +......

bnx1n

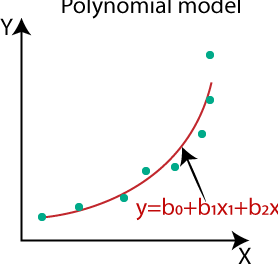


Fig-3.6.4.1 Polynomial Regression

## 3.6.3 Bayesian Ridge Regression

A kind of linear regression called Bayesian regression makes use of Bayesian statistics toestimate a model's unknown parameters. To determine the likelihood of a collection of parameters given observed data, it applies the Bayes theorem. While Bayesian regression makes deeper assumptions aboutthe structure of the data and assigns a prior probability distribution to the parameters, traditional linear regression makes the assumption that the data follows a Gaussian or normal distribution.

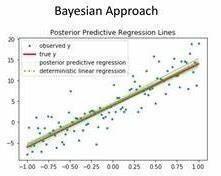


Fig-3.6.5.1 Baysian Ridge Regression

## System Architecture

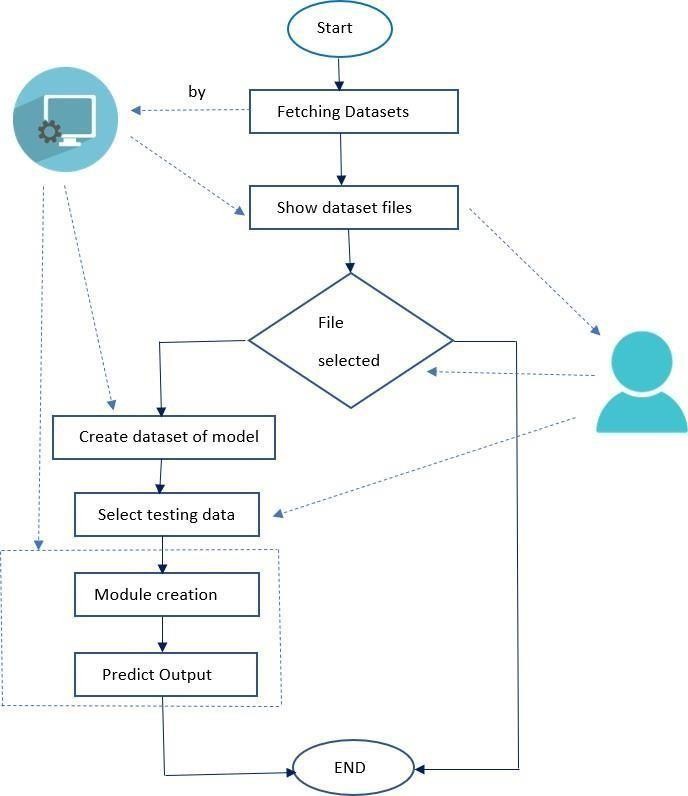


Fig-3.7.1 System Architecture

# RESULTS

## Linear Regression

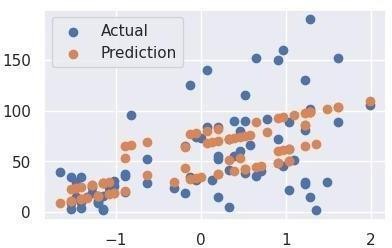


Fig 4.1.1 Actual Prediction plot in linear Regression

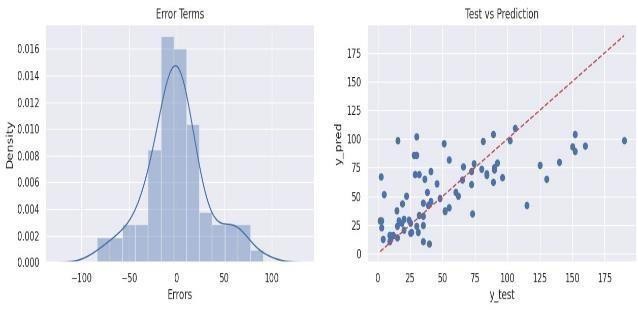


Fig 4.1.2 Test Vs Prediction plot in Linear

* 1. Muli Linear Regression

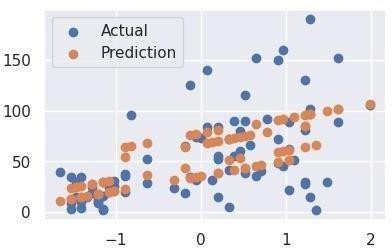


Fig 4.2.1 Actual Prediction plot in Multi Linear Regression

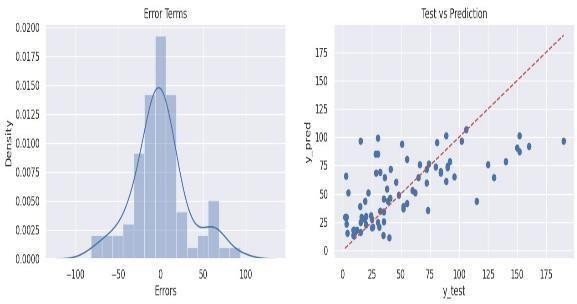


Fig 4.2.2 Test Vs Prediction plot in Multi linearRegression

## 4.3.Polynomial Regression

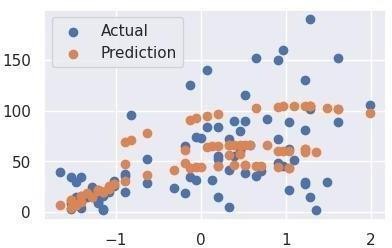


Fig 4.3.1 Actual Prediction plot inPolynomial

Regression

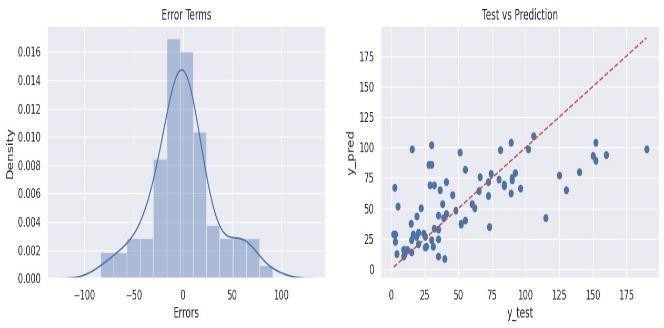


Fig 4.3.2 Test Vs Prediction plot inPolynomial Regression

* 1. Bayesian Ridge Regression

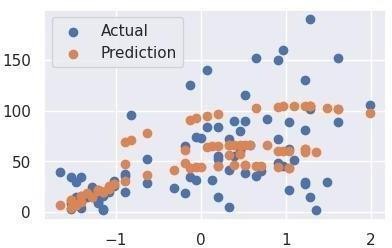


Fig 4.4.1 Actual Prediction plot in Bayesian Ridge Regression

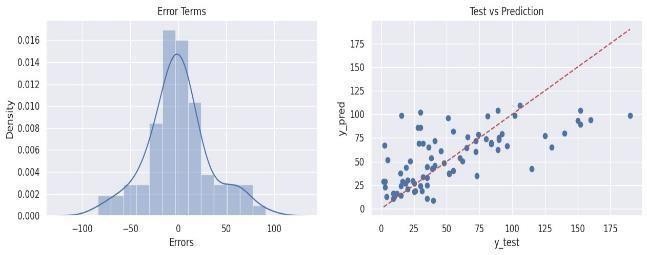


Fig 4.4.2 Test Vs Prediction plot inBayesian Ridge Regression

## Correlation Matrix

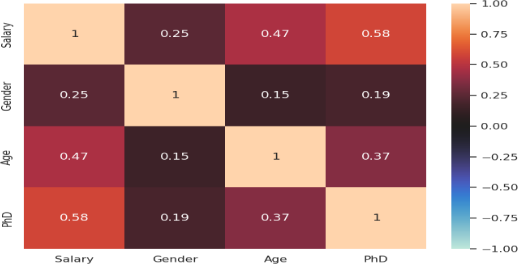


Fig 4.5.1 Correlation Matrix

## Metrics of our Model

* + 1. **For Training**

|  |  |  |
| --- | --- | --- |
| Algorith  m | RMSE Score | MSE Score |
| Multi Linear  Regressi on | 32.3951859892  655 | 1049.44807527  9104 |
| Polyno mial Regressi  on | 31.0618639621  10078 | 964.839392800  6328 |
| Ridge Regressi on | 32.3960250868  2239 | 1049.50244142  60256 |
| Bayesia n Ridge  Regressi on | 31.2011594542  77683 | 973.512351291  262 |

## 3.6.4 For Testing

|  |  |  |
| --- | --- | --- |
| Algorith  m | RMSE Score | MSE Score |
| Multi Linear Regressi on | 29.06554806732  7374 | 844.806084454  118 |
| Polynom ial Regressi  on | 29.41375860566  7862 | 865.169195312  5 |
| Ridge  Regressi on | 29.04832751663  382 | 843.805331513  6254 |
| Bayesia n Ridge Regressi on | 29.24586482282  0437 | 855.320609234  6859 |

1. **CONCLUSION**

For the purpose of employee prediction, Bayesian regression clearly outperforms models after a thorough analysis of linear regression, multi-linear regression, and ridge regression models. Bayesian Ridge regression is the algorithm that consistently shows better performance metrics on the test dataset among those that are being investigated. While preserving competitive predictive accuracy, its regularization approach successfully reduces overfitting. Furthermore, because Bayesian ridge regression features are resilient to multicollinearity—a major problem in employee prediction models—it can handle highly correlated features. The best method for our project is Bayesian ridge regression, which offers trustworthy predictions for outcomes related to employees by balancing bias and variance. Because of this, we can infer with confidence that Bayesian ridge regression is the best option for employee prediction tasks based on our data, opening up a promising new field for applications in the human resource management

## REFERENCES

1. A. Parkavil and A. K. Lakshmi 2017 IEEE International Conference on Smart Computing, Communication, Controls, Energy, and Materials: Predicting students' course knowledge levels using data mining techniques
2. N.A. Rashid, A.M. Shahiri, and A.W. Husain A Review of Data Mining Techniques for Predicting Student Performance .
3. A Survey of Educational Data-Mining Research by Richard A. Huebner, Academic and Business Research Institute, 2013 [4] Using a salary prediction system to increase students' incentive to study, Pornthep Khongchai and Pokpong Songmuang presented at the 13th International Joint Conference on

Computer Science and Software Engineering in 2016.

1. S. Vijayalakshmi Anupama Kumar

M.N. "Inference of Naïve Baye's Technique on Student Assessment Data," Communications in Computer and Information Science book series (volume 270), 2011.

1. Do college students anticipate their future income more accurately than young adults already employed? John Jerrim, Education Economics, Taylor & Francis Journals, vol. 23(2), pages 162-179, 2015.
2. In 2018, Kousik Dasgupta, Sananda Dutta, and Airiddha Halder presented at the Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) on "Design of a novel Prediction Engine for predicting suitable salary for a job."
3. Rajveer Singh, Masters Dissertation: A Regression Analysis of Salary

Determinants in Indian Employment Markets for Entry-Level Engineering Graduates. Institute of Technology, Dublin,2016.

1. Susmita Ray, "A Brief Overview of Machine Learning Algorithms," in 2019 International Conference on Machine Learning, Big Data, Cloud, and Parallel Computing (Com-IT-Con), held in India from February 14–16 .
2. Thongchai Kaewkiriya and Phuwadol Viroonluecha, "Salary Predictor System for Thailand Labour Workforce using Deep Learning" The 18th International Conference on Information and Communication Technologies (ISCIT 2018)
3. Hung, C.C. and E.-P. Lim, "Combining Occupation Salaries from Job Post and Review Data," IEEE Access, Volume 9, 2021