Predicting Salary of an Employee Using Machine Learning

[1]Ala Naga Venkata ManiKumar, [2]Telagathoti Abhishek Manohar, [3]D Akhil, [4]Ch Manideep, [5]A V Sudhakar Rao

[5]Associate Professor, [1 2,3,4]BTech (IT), Students
Information Technology (IT)
Vasireddy Venkatadri Institute of Technology
Guntur, Andhra Pradesh.

***Abstract*- Predicting Salary of an Employee using Machine Learning is an innovative machine learning-driven application developed to streamline the process of determining fair and competitive salaries for newly hired employees. Throughout the development, statistical methods and hypothesis testing are incorporated to evaluate the model's performance and ensure the reliability of results. Various regression models such as Linear regression, KNN, Decision trees, and Ensemble techniques like Bagging, Boosting and Random Forest models and others are built and compared to identify the most effective model for predicting employee salaries. The user-friendly interface, built with Streamlit, allows seamless interaction, making it accessible for both HR professionals and job seekers.**

***Key Words:* Preprocessing, Comparison, Random Forest, Streamlit.**

I. INTRODUCTION

In today's competitive job market, employees often seek higher salaries by switching companies, causing losses for their current employers. To tackle this issue, we propose customizing salaries to match employee expectations, boosting satisfaction and loyalty within companies. In today's busy world, it's important to evaluate an employee's value compared to what they expect to earn. While exact salary predictions aren't always possible, using data analysis to make educated guesses can help plan for the future. Our project aims to predict future salaries based on an employee's background and experience. By analyzing past salary data, we hope to forecast how salaries will change over time, helping companies plan ahead. We'll use careful steps like collecting data, preprocessing data, building models, comparing and choosing the best models to make accurate predictions. At the core of our project is machine learning, which allows computers to learn from data and make decisions independently. By using this technology, we aim to make salary predictions more efficient and improve decision-making in companies. In short, our project aims to give companies the tools they need to predict salaries accurately, ensuring employee satisfaction and strengthening organizational resilience in today's competitive market.

II. OVERVIEW OF SALARY PREDICTION PROCESS

1. **Data Collection:**

The first step in salary prediction involves gathering relevant data from various sources, including internal HR records, external salary surveys, job postings, and industry reports. This data typically includes information about employee demographics, job roles, educational backgrounds, previous salaries, and performance metrics. Collecting relevant data is a crucial step in building accurate salary prediction models. By leveraging a combination of various data sources as mentioned above can gather comprehensive and diverse datasets for training robust salary prediction models. It's essential to ensure data privacy, compliance with regulatory requirements, and ethical considerations throughout the data collection process.

2. **Data Preprocessing:**

Once the data is collected, it undergoes preprocessing steps such as,

a. Handling Missing Values:

Missing values are common in real-world datasets and can adversely affect model performance. Techniques such as mean, median, or mode imputation can be used to fill in missing values in numerical features. For categorical features, missing values can be replaced with the mode or a separate category.

b. Handling Outliers:

Outliers are data points that deviate significantly from the rest of the data. Outliers can be detected using statistical methods or visualization techniques such as box plots. Depending on the nature of the data and the model being used, outliers can be removed, transformed, or treated separately.

c. Handling Categorical Variables:

Many machine learning algorithms require numerical input data, so categorical variables need to be encoded into numerical values. Techniques like one-hot encoding or label encoding can be used to convert categorical variables into a format suitable for modeling.

d. Feature Selection:

Feature selection techniques help identify the most relevant features for prediction while reducing dimensionality and computational complexity. Methods such as correlation analysis, can be used to select the most informative features.

e. Data Splitting:

Before training the machine learning models, the dataset is typically split into training, and testing data sets. This allows for model training on one portion of the data, and evaluating model performance on a separate, unseen dataset.

By applying these data preprocessing techniques, the project ensures that the input data is clean, standardized, and suitable for training machine learning models, ultimately improving the accuracy and robustness of salary predictions.

3. **Model Development:**

With preprocessed data in hand, the next step is to develop predictive models using machine learning techniques. Various machine learning algorithms are employed in salary prediction models, including linear regression, Multi Linear regression, decision trees, KNN, random forest, bagging, and boosting and tuning model hyperparameters to optimize performance. These algorithms analyze the relationship between input features and salary outcomes, enabling the model to make accurate predictions.

4. **Model Evaluation and comparison:**

After training the models, they are evaluated using performance metrics such as mean squared error (MSE), and R-squared (R2) are commonly used to assess their accuracy and generalization ability. This step helps identify the most effective model for predicting salaries by comparing all the accuracies and provides insights into areas for improvement.

5. **Deployment Using Streamlit:**

Once a best model is identified, it is deployed into user-friendly interface built with Streamlit, in which it allows seamless interaction, making it accessible for both HR professionals and job seekers.

III. LITERATURE SURVEY

1. Susmita Ray," A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-Icon), India, 14th -16th Feb 2019 a brief review of various machine learning algorithms which are most frequently used to solve classification, regression and clustering problems. The advantages, disadvantages of these algorithms have been discussed along with comparison of different algorithms (wherever possible) in terms of performance, learning rate etc. Along with that, examples of practical applications of these algorithms have been discussed.

2. Sananda Dutta, Airiddha Halder, Kousik Dasgupta," Design of a novel Prediction Engine for predicting suitable salary for a job" 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) - focused on the problem of predicting salary for job advertisements in which salary are not mentioned and also tried to help fresher to predict possible salary for different companies in different locations. The corner stone of this study is a dataset provided by ADZUNA. model is well capable to predict precise value.

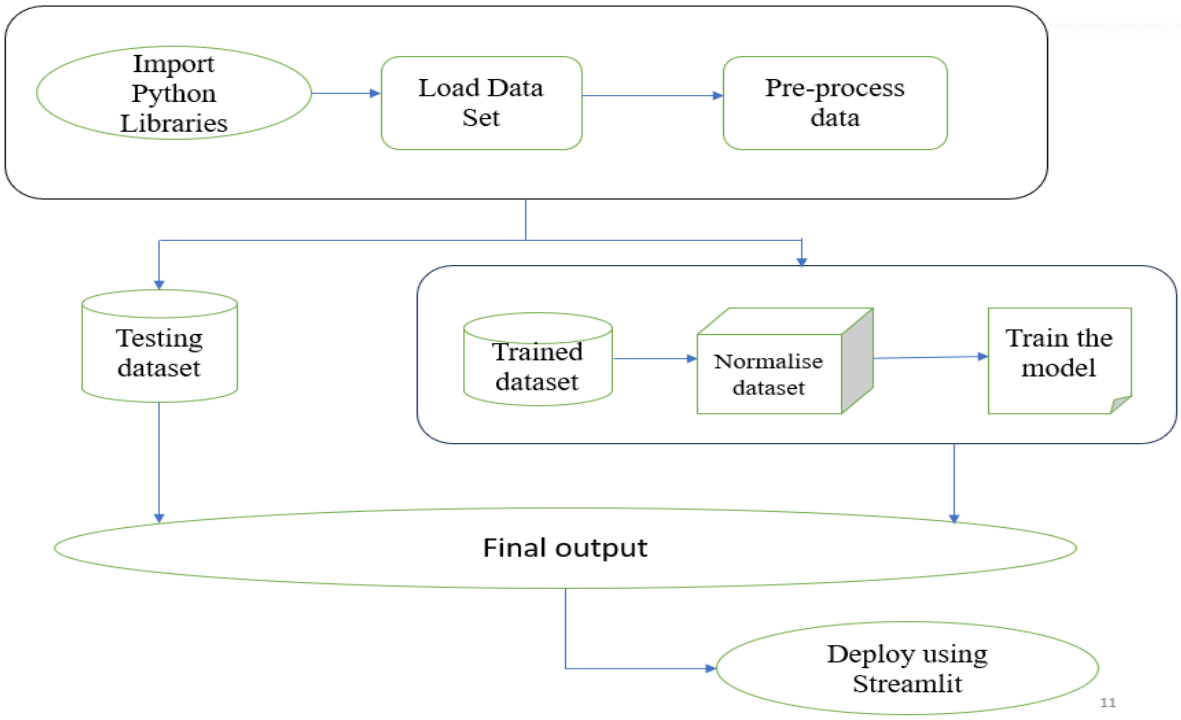
IV. PROPOSED SYSTEM:

The proposed system for the project "Predicting Salary of an Employee" involves the development and deployment of a machine learning-driven application to streamline the process of determining fair and competitive salaries for employees. The core of the proposed system is a machine learning model trained on historical data of employees. This model learns patterns and relationships in the data to predict the salary of an employee. Firstly, relevant data is collected from various sources and the collected data undergoes preprocessing steps, including handling missing values, handling outliers, handling categorical variables, feature selection and data splitting are performed to prepare it for training the machine learning model. The machine learning models are trained on the training data set using algorithms like Linear Regression, Multi Linear regression, KNN, Decision trees, Random Forest, bagging and boosting. After training the models, the model's performance is evaluated on the testing data set using metrics like Mean Squared Error (MSE) and R-squared (R2) score to ensure its accuracy and reliability. After evaluating the models, the performance of each model is compared and select the best model and it is deployed into user-friendly interface using streamlit.

V. SYSTEM ARCHITECTURE:

System architecture is like a detailed plan or blueprint for a software system. It shows how the system will be built and how it will work. Just like a blueprint for a house shows where each room goes and how they connect, system architecture shows how different parts of the software will fit together and interact. It's a bit like having a map that guides us through building the software. This map helps us understand how everything works together and how the software will function

once it's finished. Having a clear plan like this helps ensure that the software meets the needs of its users and works smoothly.



VI. COMPARING PERFORMANCE OF DIFFERENT MODELS:

After Evaluating each model, the performance of each model is compared using r2 score and mse values.

a) R-squared tells us how well the independent variables explain the variance in the dependent variable. It ranges from 0 to 1,0 indicates that the model does not explain any of the variability of the response data around its mean.1 indicates that the model explains all the variability of the response data around its mean.

b) Mean Squared Error (MSE) gives us an idea of how close the model's predictions are to the actual values. A lower MSE indicates that the model's predictions are closer to the actual values on average.

	MSE	R2_score
Multi Linear Regression before splitting	7.394388e+07	0.533006
Multi Linear Regression after splitting	6.812837e+07	0.540866
Ridge Regression	6.813097e+07	0.540848
Ridge Regression with Validation curve	6.813529e+07	0.540819
Lasso Regression	6.812866e+07	0.540864
Lasso Regression with validation curve	6.813039e+07	0.540852
K-Nearest Neighbors	6.850662e+07	0.538317
Decision Tree	6.042463e+07	0.592783
Bagging	5.732010e+07	0.613705
Random Forest	4.990089e+07	0.950313
Gradient Boosting	7.030897e+07	0.526170
Ada Boosting	6.025863e+07	0.593902
XGBoosting	7.102623e+07	0.521336

VII. RESULT:

After deploying the model using streamlit, the final output is:

1st input is we can choose manager or executive role by selecting any of the 2 radio buttons, previous ctc takes numerical value as an input, previous job change is the categorical input type Yes/No and experience in months which takes numerical input.

VIII. CONCLUSION

In this paper we proposed a system to predict the salary of an employee using historical data. We compared all the models, selected the best model and it is deployed into user-friendly interface using streamlit. Random Forest Model is the best model in predicting the salary of an employee. Our selected random forest model obtained high r2 score value and low mse value.

REFERENCES:

1. D.M Lothe, Prakash Tiwari, Nikhil Patil, Sanjana Patil, Vishwajeet Patil, “SALARY PREDICTION USING MACHINE LEARNING” 2021 6 International Journal of Advanced Scientific Research and Engineering Trends (IJASRET)
2. Sananda Dutta, Airiddha Halder, Kousik Dasgupta,” Design of a novel Prediction Engine for predicting suitable salary for a job” 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN).
3. Pornthip Thongchai, Pokpong Song Muang, “Improving Students’ Motivation to Study using Salary Prediction System” 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)
4. Susmita Ray," A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (ComIT Con), India, 14th -16th Feb 2019
5. Magel, Rhonda, and Michael Hoffman. "Predicting salaries of major league baseball players." International Journal of Sports Science 5, no. 2 (2015): 51-58.