International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-13 Issue-2, May 2023

Predictive Insights: using Machine Learning to Determine Your Future Salary

M. Saraswathi, J. Akhila, K. Sireesha

Abstract: ***Knowing one's expected salary can be a crucial consideration when deciding whether to change careers or seek higher education in today's fiercely competitive work market. Accurate salary forecasts can give important information about the earning potential of various professions because there are so many students graduating each year and workers looking to switch sectors. In order to forecast a salary range, this paper suggests a computerized method that considers a person's country, level of education, number of years of experience, and area of specialization. This kind of system has obvious benefits because it gives individuals and groups the power to decide wisely about job prospects, wage negotiations, and employee retention. The system's data can be used by researchers, academic institutions, and policymakers to evaluate labor market trends and reach informed decisions. The reliability and correctness of the system's data, the forecasting models employed, and the regularity of system maintenance and updates will all have an impact on these factors. However, it is a promising area for further research and development due to the benefits of having a reliable technique for estimating salaries***

Keywords: ***Machine learning, Prediction, Regression, Supervised learning.***

I. INTRODUCTION

The assumptions we make about the future are predictions. Forecasts may, but not always, be based on prior information and experience. It is often impossible to ascertain specific data about the future, as future events are not always guaranteed. However, projections can help us plan for possible developments. A salary prediction model is a machine learning algorithm used to predict an individual's salary based on job title, years of experience, and other factors. A salary prediction machine learning model is used to predict an individual's future salary. The model is trained on historical data about salary and other factors that can affect an individual's salary. This model can be used to predict an individual's salary based on current salary, title, company, location, and other factors.

This model uses country, education level, experience, and specialty as inputs to predict an individual's salary. This will help the employee in the following ways:

- Helping to see the growth in any field
- With the help of machine learning it can easily produce a graph.
- Easy to estimate the salary between the x-y axis.
- Users can give any point to get a salary through the program.
- The salary of the employees can be observed in a particular field according to their qualifications.

II. LITERATURE REVIEW

The system compares the employee's years of experience and annual salary. The algorithms used are Random Forest, Decision Tree Regression, and Support Vector Regression. This paper predicts salaries for active employees who are about to change domains. The system lags behind when it comes to choosing the right features. An overview of the most popular machine learning techniques for solving classification, regression, and clustering. The strengths and weaknesses of these algorithms were investigated, and different algorithms were compared (where possible) in terms of performance, learning rate, etc. Additionally, examples of using these algorithms in practice were covered. [1] The predictive model suggested by the Salary Prediction System uses a seven-characteristic decision tree approach to improve student motivation. In addition, the results of the system are not only the expected annual income but also the third-highest annual income among graduates who have something in common with the user. To test the efficiency of the system, they set up an experiment using his 13,541 records of real data from graduates. The overall accuracy result is 41.39%. Instead of predicting wage changes over time, we want to be able to predict spatial wages across all economic activities and occupations. As a result, previous national surveys were used to determine the median annual income, which was then incorporated into the forecast model. As a result, wage projections for related occupations or other occupations within the same economic sector are projected across all economic sectors. However, the new method allows additional disruptive factors to be taken into account when estimating salaries. With just a small amount of survey data, a systematic process enables accurate salary forecasts. Independent variables are populated in the predictive model based on theoretical and statistical grounds that consider both organizational and occupational characteristics. [2]

Manuscript received on 25 March 2023 | Revised Manuscript received on 05 April 2023 | Manuscript Accepted on 15 May 2023 | Manuscript published on 30 May 2023.

***Correspondence Author(s)
Dr. M. Saraswathi, Assistant Professor, Department of Computer Science Engineering, Sri Chandrasekarendra Saraswathi Viswa Maha Vidyalaya, Enathur (Tamil Nadu), India. E-mail: msaraswathi@kanchiuniv.ac.in, ORCID ID: ORCID ID: https://orcid.org/0009-0006-0642-6899
J. Akhila, B.E, 4th Year Student, Department of Computer Science Engineering, Sri Chandrasekarendra Saraswathi Viswa Maha Vidyalaya, Enathur (Tamil Nadu), India. E-mail: akhilajallavaram0807@gmail.com, ORCID ID: https://orcid.org/0009-0008-1698-1516
K. Sireesha*, B.E, 4th Year Student, Department of Computer Science Engineering, Sri Chandrasekarendra Saraswathi Viswa Maha Vidyalaya, Enathur (Tamil Nadu), India. E-mail: sireesha.scsvmv@gmail.com, ORCID ID: https://orcid.org/0009-0002-5582-6814

Salary Predictor System for Thailand Labour Workforce using Deep Learning used Deep learning techniques to construct a model which predicts the monthly salary of job seekers in Thailand solving a regression problem that is a numerical outcome that is effective. We used five-month personal profile data from well-known job search websites for the analysis. As a result, the Deep learning model has strong performance whether in accuracy or process time by RMSE 0.774 x 104 and only 17 seconds for runtime.[3]
This paper is aimed at providing better assistance to school students regarding the salary that they will aspect after completing their course. Through their mode, they have tried to provide a system for salary prediction where a data processing technique is employed. In this paper, they have compared the profile of the student with the graduated student using data mining techniques. They have also performed an experiment on student datasets using 10-fold-cross-validation. K-nearest neighbor, Decision Tree, and Naïve Bayes algorithms are used [4][5]

III. PROBLEM STATEMENT

The main reason why employees switch companies nowadays is the salary of that employee.
Employees keep switching companies to get the expected salary and it leads to loss of the company to overcome this loss we came up with an idea of what if the employee gets the desired/expected salary from the Company or Organization. In this Competitive world, everyone has higher expectations and goals. But we cannot randomly provide everyone with their expected salary there should be a system that should measure the ability of the Employee for the Expected salary. We cannot decide the exact salary but we can predict it by using certain data sets. A prediction is an assumption about a future event.

IV. PROPOSED SYSTEM

Our model predicts the salary of a person based on his country, education level, experience, and specialization. Most salary prediction models are prioritizing only
experienced employees and those who are trying to switch domains from their present domain. Our system concerns not only experienced people but also the people who are unsure about what salary they will get based on their specializations. The system is user-friendly and it predicts the salary of the employees. We are going to use a decision tree regression model for the prediction. We are trying to achieve better accuracy than the existing models.
We experimented with different data sets before using the present data set. We used linear regression algorithm, decision tree regression algorithm, and random forest regression algorithm on the experimented data and did not get satisfactory results. Then we decided to use a decision tree algorithm on the present used data set.
Advantages:

- User friendly
- Less complexity
- High performance

V. SYSTEM ARCHITECTURE

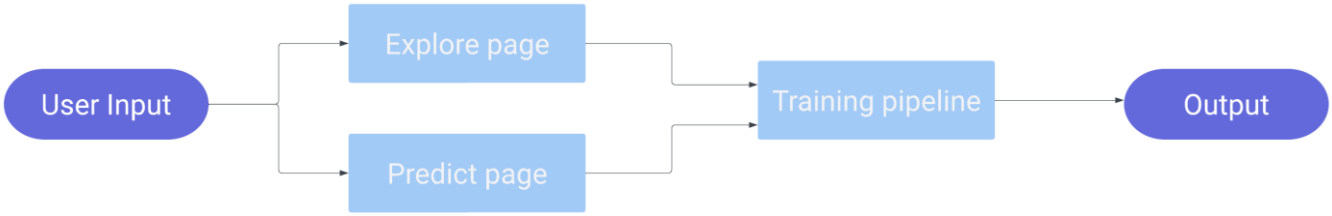

Fig 1. System Architecture

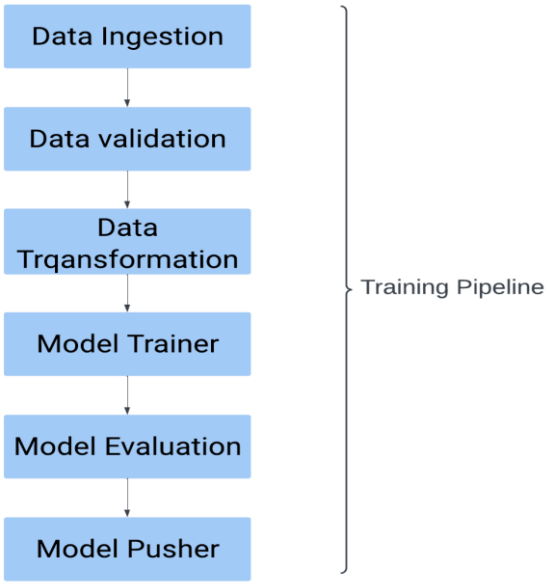

Fig 2. Training Pipeline

A. Modules

User input:

This model seeks the user's country, education level, experience, and specialization as the input.

Training pipeline:

Data ingestion:

In this step loading the data from the database and splitting the data into train tests according to the requirement of the project.

Data validation:

In this step required data analysis has been performed using various data visualization techniques through which we identify the missing data percentage, outlier data, correlation between the features, etc.

Data transformation:

In this step, the identified missing values, outliers, and everything can be resolved using some preprocessing techniques, and data is transformed according to the problem statement.

Model Training:

In this step, we trained the model using linear regression, random forest regression, and decision tree algorithms. The decision tree algorithm has given better results.

Model evaluation:

Model evaluation is performed using mean squared error, root means squared error and r2 score.

Model Pusher:

The trained model is made user accessible using an open-source framework called streamlit.

Output:

The output will be the predicted salary of a person. The output will be shown using the streamlit library.

VI. IMPLEMENTATION WORK

We first gathered the correct data set that is required for training the model. Then we performed data preprocessing steps. Here we split the dataset into the training set and test set. After splitting we trained the model with different algorithms like linear regression, decision tree, and random forest and fit the training set and test set. We got the best accuracy score for the decision tree algorithm. We deployed it using the streamlit library in python.

A. Algorithms

Decision Tree Algorithm:

To predict the salary of a person we have used "Decision Tree Regression".

A decision-making tool called a decision tree can be used to represent decisions and all potential outcomes, including outcomes, input costs, and benefits. A group of supervised learning algorithms includes decision tree algorithms. Works on categorized continuous output variables. A branch/edge represents the result of a node, where a node has either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]

Decision tree regression trains a model in the form of a tree to predict future dates and observe element properties to produce useful continuous results. A continuous output indicates the lack of a discrete output. H. An output that is not simply represented by a discrete, well-known number or set of values.

Discrete output example:

A weather prediction model that predicts whether or not there'll be rain on a particular day.

Continuous output example:

A profit prediction model states the probable profit that can be generated from the sale of a product. Here, continuous values are predicted with the help of a decision tree regression model.

Steps:

Step 1: Import the required libraries.
Step 2: Initialize and print the Dataset.
Step 3: Select all the rows and column 1 from the dataset to "X".
Step 4: Select all of the rows and column 2 from the dataset to "y".
Step 5: Fit the decision tree regressor to the dataset
Step 6: Predicting a new value
Step 7: Visualising the result
Step 8: The tree is finally exported and shown in the tree structure.

Without initially considering entropy, it is challenging to understand information gain. Information theory gave rise to the idea of entropy, which quantifies the impurity of sample values. The following formula serves as its definition, where:

Entropy formula:

$$\text{Entropy}(S) = -\Sigma(c \varepsilon C)\ p(c)\log 2$$

- S represents the data set where entropy is calculated
- c represents the classes in the set, S
- p(c) represents the proportion of data points that belong to class c to the number of total data points in the set, S

VII. RESULT AND DISCUSSION

Predict Page:

On this page, the user has to provide his details like his country, education level, specialization, and experience and it will predict the approximate salary of that person.

The salary mostly depends on the experience of the person. If he is more experienced then his salary will be more and vice versa. It will also depend on his major and country.

The following below screenshot fig 1shows the options that can be chosen to predict the salary:

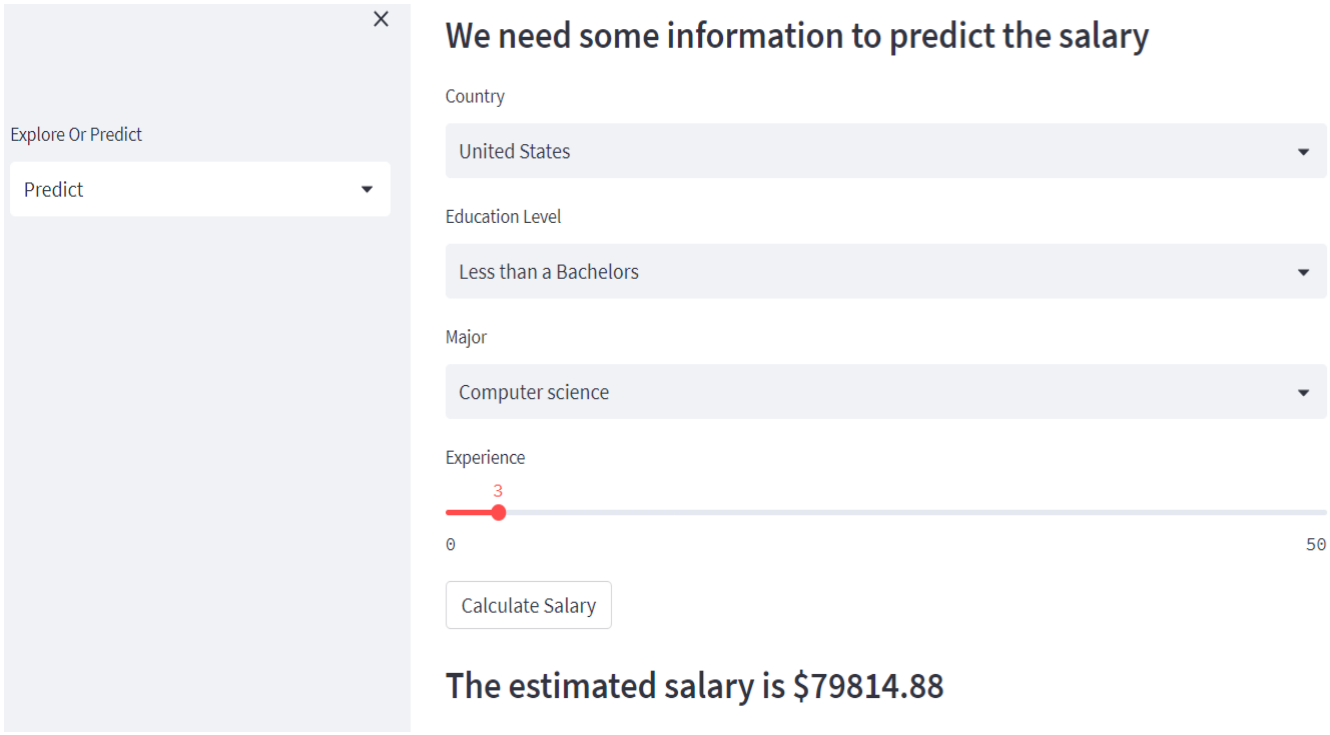

Fig 3. Predict page

The following image shows the predicted salary:

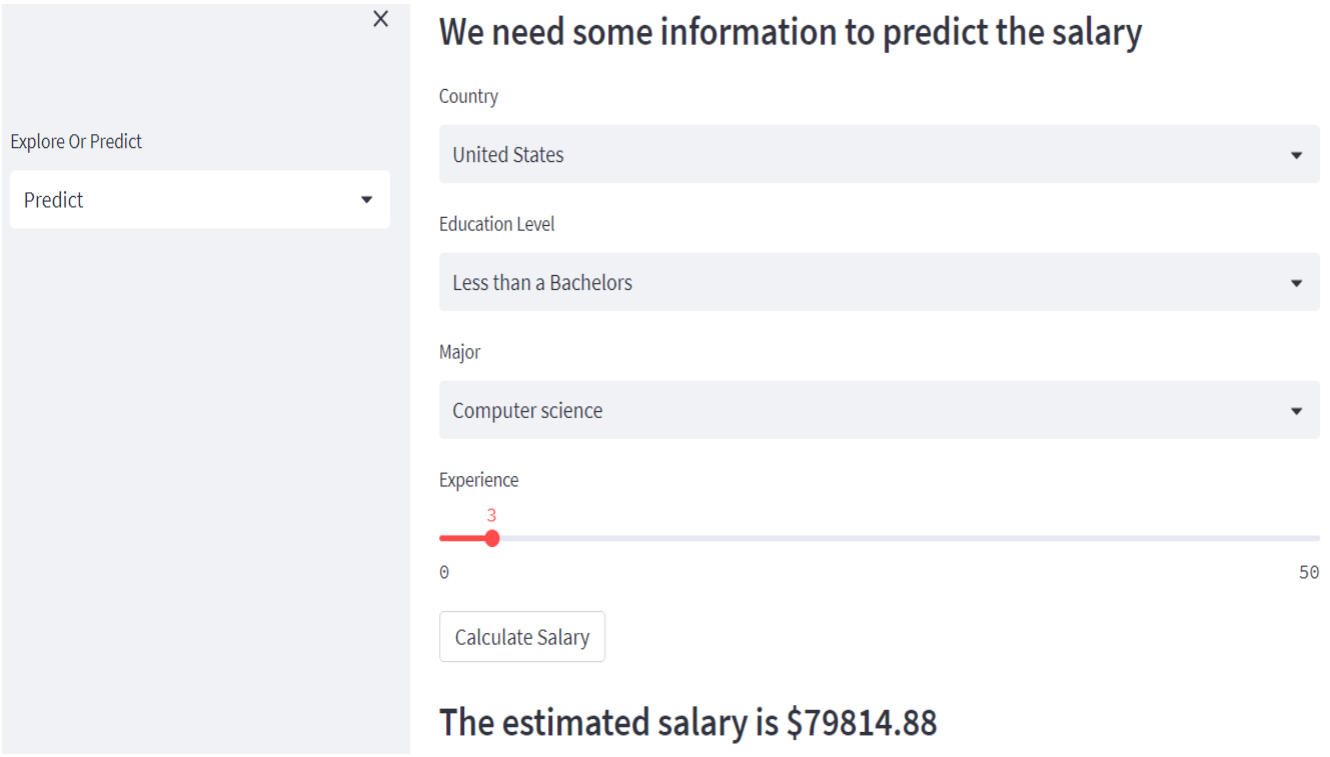

Fig 4. Predict Page

Explore Page:

Data analysis based on features:

The following is the pie chart showing the data that came from each country.

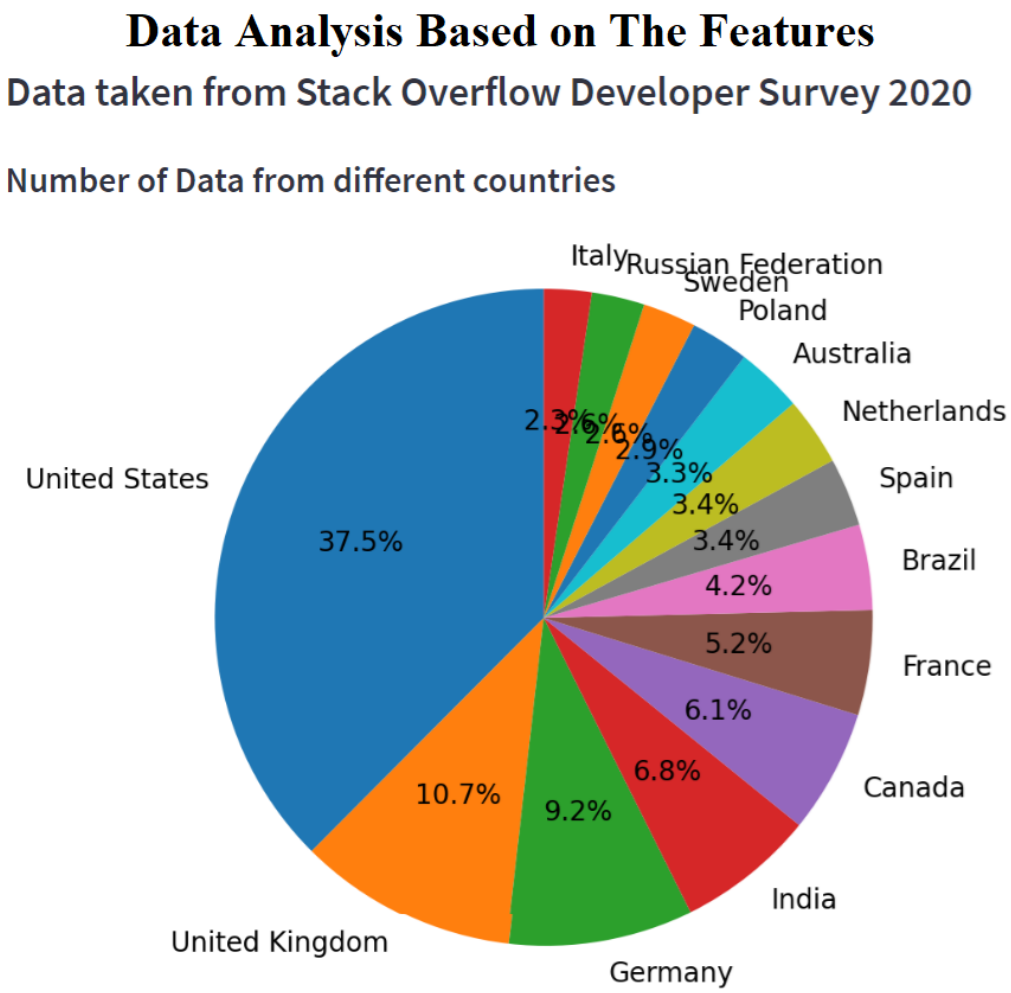

Fig 5. Data analysis based on the features

Mean salary based on country:
The following picture shows the mean salary based on country:

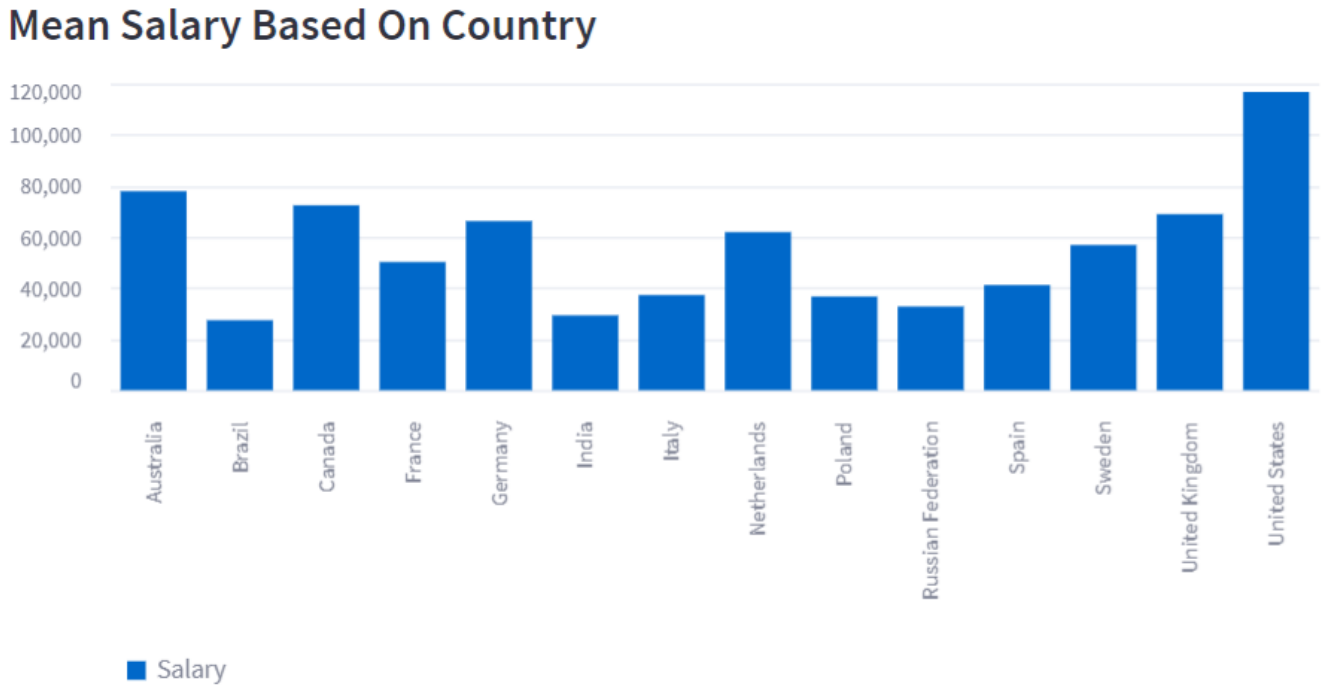

Fig 6. Mean Salary based on country

Mean salary based on experience:
The following line graph depicts the mean salary based on experience:

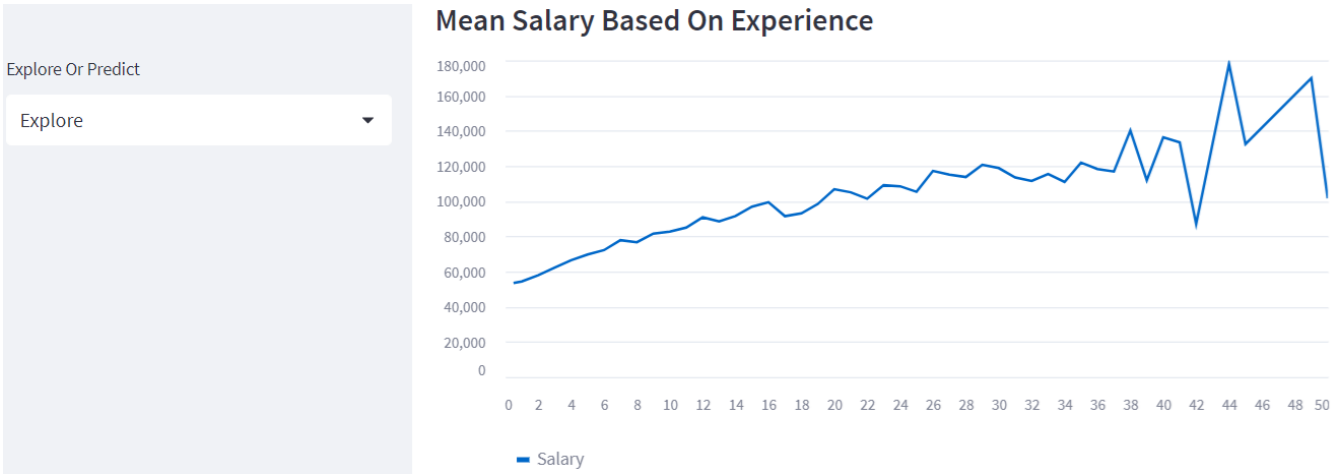

Fig 7. Mean Salary based on experience

Mean salary based on major:

The following graph shows the mean salary based on major:

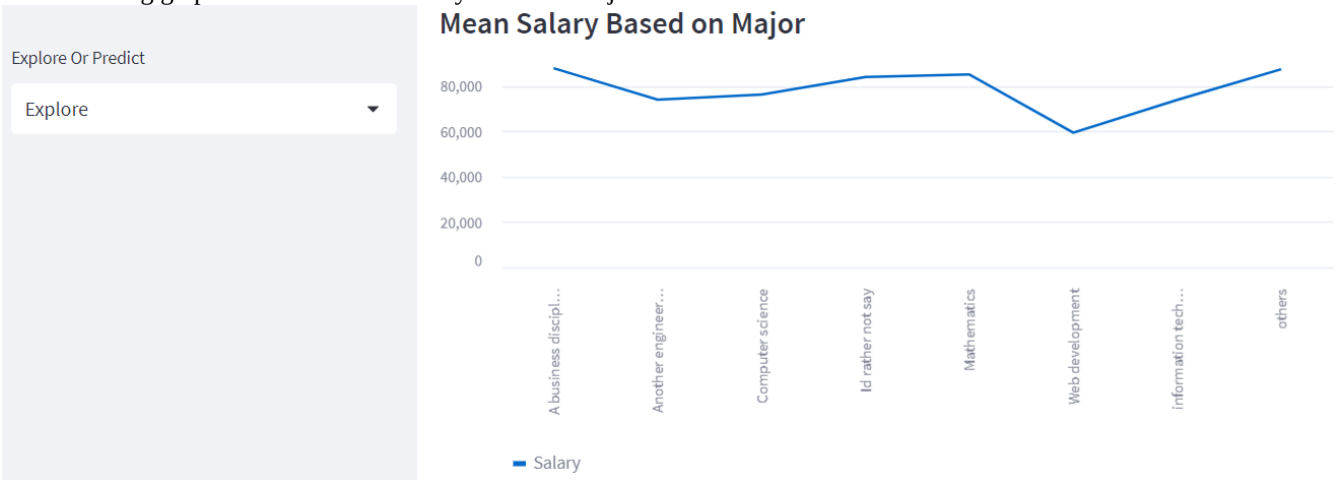

Fig 8. Mean Salary Based on Major

VIII. CONCLUSION

Our system consists of predicting employee salaries. For prediction, we trained a model using a decision tree algorithm. We also ran an experiment using an employee record with 40 records. We used logistic regression, decision trees, and random forest algorithms. I concluded that decision trees give the best results. The output of the system is calculated in an optimal way by comparing it with other algorithms using common metrics such as classification accuracy, F1 score, ROC curve, and precision-recall curve. We only compared algorithms for simple two-attribute models.

As of now, we predicted the salary of an employee by just taking the person's country, education level, experience, and specialization as input. We are using streamlit for hosting it in local systems.

FUTURE ENHANCEMENT

In future work, we would like to deploy it on Heroku and generate an URL so that anyone with that URL can use our model. And try to save and reuse the trained model. We also want to increase the accuracy by training the model with more records.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive any financial support/funding.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. Krishna Gopal, Ashish Singh, Harsh Kumar, Dr. Shrddha Sagar "Salary prediction using Machine Learning" – 2021
2. Prof. D. M. Lothe1, Prakash Tiwari2, Nikhil Patil3 , Sanjana Patil4 , Vishwajeet Patil5," Salary prediction using ML",IJASRET-2021

3. J. Y. Kuo, H. C. Lin, H. T. Chung, P. F. Wang and B. Lei. "Building student course performance prediction model based on deep learning,"-2021
4. Singh, P.; Pattanaik, F. Unequal Reward for Equal Work? Understanding Women's Work and Wage Discrimination in India Through the Meniscus of Social Hierarchy. Contemp. Voice Dalit 2020. [CrossRef]
5. 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

AUTHORS PROFILE

Dr. M. Saraswathi is an Assistant Professor of the Computer Science Engineering Department from Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Deemed to be University, Enathur, Kanchipuram. The area of interest lies in software engineering and cloud computing domain. I served my teaching on OOAD, C, C++, Computer Networks, Data warehousing and Mining, Dot net technologies, Web technology and its lab, Data Structures, and algorithms, Computer system architecture, Visual Programming, SCI Lab, SDL lab using CASE tools. I have attended 5 conferences. I have also published in 8 journals. And also, a member of IACSIT and IAENG.

J. Akhila B.E, 4th Year, Computer Science Engineering Department from Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Deemed to be University, Enathur, Kanchipuram. My area of interest lies in Data science and machine learning and completed training in data science and worked with various machine learning algorithms and performed some research in NLP, currently working as a machine learning intern at Quixy, Hyderabad.

K. Sireesha B.E, 4th Year, Computer Science Engineering Department from Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Deemed to be University, Enathur, Kanchipuram. My area of interest is data science and data analytics. I have done a course on data science where I worked on a project called Dance forms. I have learned Python and its libraries, Statistics, C, and SQL. I am currently spending time increasing my knowledge further in data science.