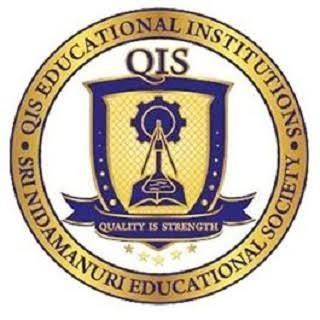
# DEPARTMENT OF CSE

## FUNCTIONAL SKILLING PROGRAM PROJECT

**SALARY PREDICTION**



# Submitted By:

G.Manasa siva sankari [22491A0582]

E.V.J. Swaroop [22491A0581]

C.V. Neeraj kumar  [22491A0507]

P.Yamini [22491A0541]

T.Sasi kumar [22491A0556]

# QIS COLLEGE OF ENGINEERING & TECHNOLOGY

(Autonomous & NAAC ‘A+’ Grade)

(Approved by AICTE, New Delhi & Affiliated to JNTU Kakinada)

(An ISO 9001:2015 Certified Institution)

**VENGAMUKKAPALEM, ONGOLE-523272, A.P., INDIA**

**AY: 2023-24**

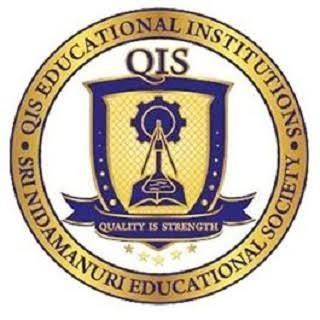
# QIS COLLEGE OF ENGINEERING & TECHNOLOGY

(Autonomous & NAAC ‘A+’ Grade)

(Approved by AICTE, New Delhi & Affiliated to JNTU Kakinada)

(An ISO 9001:2015 Certified Institution)

**VENGAMUKKAPALEM, ONGOLE-523272, A.P., INDIA**



## DEPARTMENT OF CSE

## CERTIFICATE

This is to certify that the Functional Skilling Program(FSP) Projectentitled “**SALARY PREDICTION**” is a record of the bonafide work done by G.MANASA SIVA SANKARI(22491A0582), E.V.J. SWAROOP(22491A0581),C.V.NEERAJ KUMARR(22491A0507),P.YAMINI(22491A0541),T.SASI KUMAR(22491A0556). submitted in partial fulfillment of the requirement for the award of degree of Bachelor of Technology in CSE-1for the academic year **2023-2024.** This work is carried out under my supervision and guidance.

|  |  |
| --- | --- |
| **Signature of the Class Coordinator** | **Signature of Head of the Department** |
| **N.Venkateswarlu** | **Dr.D.Bujji Babu M. Tech., Ph.D.** |
| Assistant professor | Head of the Department |

**Title of the Project: SALARY PREDICTION**

**Abstract of the Project:**

Salary prediction models are essential tools in human resource management and career planning. They leverage historical data to estimate potential earnings based on job titles and years of experience. This study focuses on constructing a predictive model using these two primary variables. Job titles, which are categorical, reflect the role and responsibilities within an organization and are indicative of the salary range. Years of experience, a numerical variable, generally shows a positive correlation with salary, as it represents the accumulation of expertise. Preprocessing steps such as encoding categorical variables and normalizing numerical data are crucial for model accuracy. Regression analysis, particularly linear regression, serves as the backbone for prediction, although more complex models may be employed for nuanced datasets. The model’s performance is evaluated using standard metrics like R-squared. Insights derived from the model can guide salary benchmarks and inform negotiations. The model’s efficacy is contingent upon the quality of the dataset and the chosen algorithm, underscoring the importance of robust data collection and model selection.

This abstract summarizes the approach and considerations in developing a salary prediction model based on job titles and years of work experience. It highlights the data preprocessing required, the modeling techniques used, and the implications of the model’s findings.

**Introduction :**

When predicting salaries using data such as job titles and years of experience, you’re essentially looking at how these factors correlate with salary ranges. Here’s an overview of how this data can be used in a predictive model:

* **Job Title**:

This is a categorical variable that can have a significant impact on salary. Different job titles come with different average salaries, responsibilities, and required skill sets. In your model, you’ll need to encode these categorical values into numerical form, often using techniques like one-hot encoding.

* **Years of Experience**:

This is typically a numerical variable that represents how long an individual has worked in a particular field or job. Generally, more experience translates to higher salaries due to the accumulation of skills and expertise. This variable is usually directly proportional to the salary.

* **Data Preprocessing**:

Before using this data in a machine learning model, it’s important to preprocess it. This includes handling missing values, encoding categorical variables, normalizing or standardizing numerical values, and potentially creating new features that could better capture the patterns in the data.

* **Model Training**:

With the preprocessed data, you can train a regression model. Linear regression is a common choice for salary prediction, but depending on the complexity of your data, you might opt for more sophisticated models like random forests or gradient boosting machines.

* **Model Evaluation**:

After training, you evaluate your model using a test set to see how well it predicts salaries. Common evaluation metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²).

* **Insights**:

The model can provide insights into which factors are most predictive of salary. For instance, certain job titles or levels of experience might be strong predictors, which can inform both employers and employees about salary standards.

Remember, the quality of your predictions will heavily depend on the quality and quantity of your data, as well as the appropriateness of the model you choose. It’s also important to consider external factors that might affect salaries, such as location, industry, and economic conditions, which may not be captured by job title and experience alone.

**LITERATURE SURVEY :**

A literature survey on salary prediction using job title and years of experience reveals a variety of approaches and methodologies. Here’s a summary of the findings from recent research:

*Predictive Insights Using Machine Learning:*

This study proposes a computerized method to predict salary ranges considering factors like country, education level, years of experience, and specialization.

It emphasizes the benefits of such systems in making informed decisions about career prospects, wage negotiations, and employee retention.

*Predicting Compensation for Job Seekers - Stanford University:*

The project explores regression techniques to estimate base salary using a large dataset from Glassdoor.

It addresses the challenges of predicting compensation due to factors like geography, industry, and experience.

*Predict Salary on the Basis of Years of Experience:*

This article focuses on predictive modeling to forecast salaries based on professional experience.

It discusses the direct influence of years spent in a field on earning potential.

*Statistical Machine Learning Regression Models for Salary Prediction:*

A framework is developed to predict labor salary across all job titles in the Saudi Arabian economy.

The performance of five machine learning algorithms is evaluated on limited survey data.

These studies collectively highlight the complexity of salary prediction and the importance of considering a wide range of factors. They also demonstrate the potential of machine learning models to provide valuable insights into salary trends and determinants.

**EXISTING SYSTEM :**

The existing systems for salary prediction using job title and years of experience typically involve machine learning models that analyze historical data to forecast salaries. Here’s an overview of some of the current systems:

*GitHub Repositories:*

There are several GitHub repositories where developers have shared their salary prediction projects. These projects often include data preprocessing, model selection, training, and evaluation. For example, one such project uses machine learning to predict salaries based on experience.

These systems serve as a foundation for developing more advanced salary prediction models and provide insights into compensation management using data-driven approaches. They highlight the importance of robust data collection, careful preprocessing, and the selection of appropriate machine learning algorithms to improve prediction accuracy.

**Proposed System:**

The proposed system for salary prediction using job title and years of experience aims to enhance the accuracy and reliability of salary estimations. Here’s an outline of the proposed system:

***Data Collection****:*

Gather a comprehensive dataset that includes job titles, years of experience, and corresponding salaries from various industries and regions.

***Data Preprocessing****:*

Implement advanced preprocessing techniques to handle missing values, encode categorical variables like job titles, and normalize the years of experience.

***Feature Engineering****:*

Explore the creation of new features that could improve model performance, such as categorizing job titles into broader groups or creating bands for years of experience.

***Model Selection****:*

Evaluate multiple regression models, including Linear Regression, Random Forest, and Gradient Boosting, to determine the best fit for the data.

***Model Training****:*

Train the selected model on the preprocessed dataset, ensuring that it captures the underlying patterns between the features and the target variable.

***Model Evaluation****:*

Assess the model using metrics like R-squared and Mean Absolute Error to evaluate its predictive power and accuracy.

***Deployment****:*

Deploy the model into a production environment where it can be used by employers and job seekers to estimate salaries in real-time.

This proposed system is designed to be robust, scalable, and adaptable to changes in the job market, ensuring that it remains a valuable tool for salary prediction.

**Software/ Hardware Requirements:**

**Software Requirements:**

• Python 3.5 .

• Operating System: windows 7 and above or Linux based OS or MAC OS

**Hardware Requirements:**

• RAM: 4 GB

• Storage: 500 GB

• CPU: 2 GHz or faster

• Architecture: 32-bit or 64-bit

**ALGORITHM :**

The algorithm for salary prediction typically involves the following steps:

1. **Data Collection**: Gather a dataset that includes job titles, years of experience, and corresponding salaries.
2. **Data Preprocessing**:
   * Handle missing values and outliers.
   * Encode categorical variables like job titles using one-hot encoding or label encoding.
   * Normalize or standardize numerical features such as years of experience.
3. **Feature Selection**: Choose relevant features that are likely to influence the salary, such as job title and years of experience.
4. **Model Selection**: Choose a machine learning model suitable for regression tasks. Common choices include:
   * **Linear Regression**: For datasets with a linear relationship between features and salary.
   * **Random Forest Regressor**: An ensemble method that can handle complex datasets with higher accuracy.
5. **Model Training**:
   * Split the dataset into training and testing sets.
   * Train the model on the training set.
6. **Model Evaluation**:
   * Use the testing set to evaluate the model’s performance.
   * Common evaluation metrics include R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).
7. **Model Optimization**:
   * Perform hyperparameter tuning to optimize the model.
   * Use techniques like cross-validation to ensure the model’s generalizability.
8. **Deployment**:
   * Deploy the trained model into a production environment.
   * Create a user interface for inputting job titles and years of experience to receive salary predictions.
9. **Monitoring and Updating**:
   * Monitor the model’s performance over time.
   * Update the model as new data becomes available to maintain accuracy.

This algorithm provides a structured approach to developing a salary prediction model, ensuring that it is accurate, reliable, and suitable for real-world applications.

**Flowchart\block Diagram about the Project with Description:**

**Flow Chart**

A flowchart is a type of diagram that represents a workflow or process. A flowchart can also be defined as a diagrammatic representation of an algorithm, a step-by-step approach to solving a task. The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows.

Data preprocessing

Feature selection

Model Training

Salary prediction

**Module Specification :**

The module specifications for a salary prediction project using job title and years of experience can be outlined as follows:

1. **Data Collection Module**:
   * **Function**: Acquire and compile a dataset with job titles, years of experience, and salaries.
   * **Input**: Sources such as job boards, company databases, and salary surveys.
   * **Output**: A structured dataset ready for preprocessing.
2. **Data Preprocessing Module**:
   * **Function**: Clean and prepare the data for analysis.
   * **Input**: Raw dataset.
   * **Output**: Cleaned dataset with encoded categorical variables and normalized numerical values.
3. **Exploratory Data Analysis (EDA) Module**:
   * **Function**: Perform statistical analysis and visualizations to understand the data.
   * **Input**: Preprocessed dataset.
   * **Output**: Visualizations and reports detailing the data distribution and correlations.
4. **Model Training Module**:
   * **Function**: Train machine learning models on the dataset.
   * **Input**: Dataset with features.
   * **Output**: Trained models.
5. **Model Evaluation Module**:
   * **Function**: Assess the performance of the models.
   * **Input**: Trained models and test data.
   * **Output**: Evaluation metrics like R-squared and Mean Absolute Error.
6. **Model Optimization Module**:
   * **Function**: Fine-tune the models for better performance.
   * **Input**: Evaluation metrics.
   * **Output**: Optimized models.
7. **Deployment Module**:
   * **Function**: Deploy the model into a production environment.
   * **Input**: Optimized model.
   * **Output**: Salary prediction service.

Each module is designed to be independent yet interoperable, ensuring a seamless flow from data collection to deployment.

**CODING :**

## Importing the necessary libraries

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

from sklearn.compose import ColumnTransformer

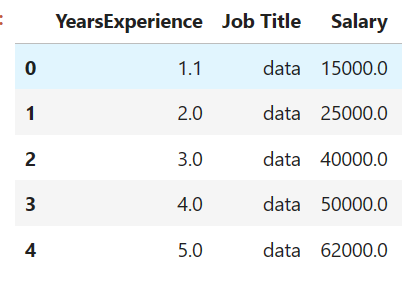
from sklearn.metrics import mean\_squared\_error

## Reading the dataset

df=pd.read\_csv('D:Salary\_Data.csv')

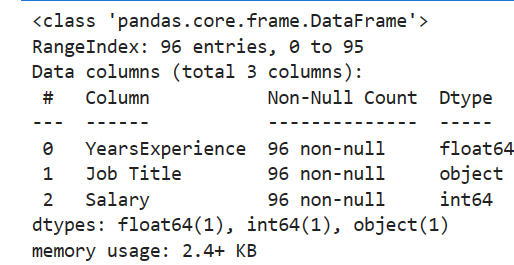
print(df.shape)

df.head()



## An overview of the dataset

df.info()



**EDA (Exploratory data analysis)**

## scatter plot for salary and experience:

# scatter plot for salary and experience

plt.scatter(df['YearsExperience'],df['Salary'],c="yellow",marker="o",edgecolor="red",s=100)

plt.xlabel('YearsExperience')

plt.ylabel('Salary')

plt.title('scatter plot for salary and experience')

#plt.savefig("scatter plot for salary and experience.png")

plt.show()



## line plot with scatter plot for salary and experience of data science :

#scatter plot for data

s=df.loc[df['Job Title']=='data']

x=s['YearsExperience']

y=s['Salary']

plt.scatter(x,y,marker="s",c="green",edgecolor="black")

plt.xlabel('YearsExperience')

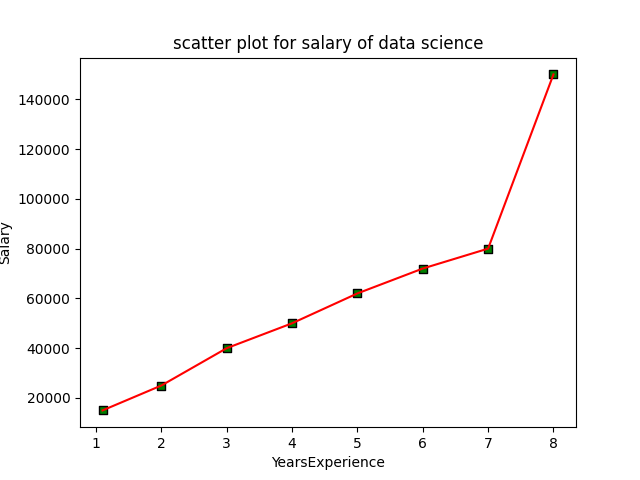
plt.ylabel('Salary')

plt.title('scatter plot for salary of data science')

plt.plot(x,y,c="red")

#plt.savefig("scatter come line plot for salary of data science.png")

plt.show()



## line plot with scatter plot for salary and experience of web

#scatter plot for web

s=df.loc[df['Job Title']=='web']

x=s['YearsExperience']

y=s['Salary']

plt.scatter(x,y,marker="s",c="green",edgecolor="black")

plt.xlabel('YearsExperience')

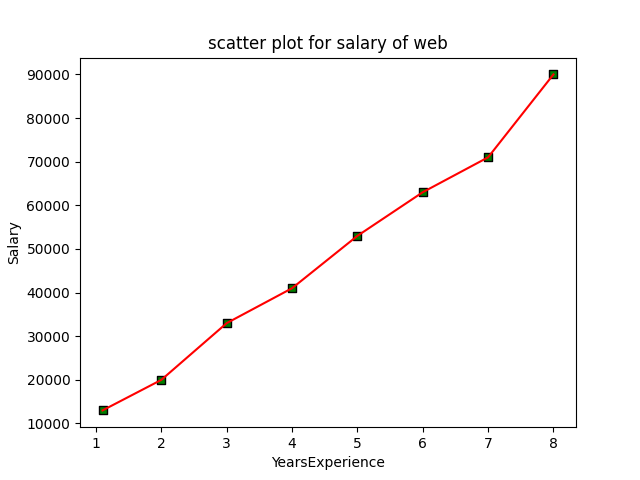
plt.ylabel('Salary')

plt.title('scatter plot for salary of web')

plt.plot(x,y,c="red")

#plt.savefig("scatter come line plot for salary of web.png")

plt.show()



## line plot with scatter plot for salary and experience of data

#scatter plot for mobile

s=df.loc[df['Job Title']=='mobile']

x=s['YearsExperience']

y=s['Salary']

plt.scatter(x,y,marker="s",c="green",edgecolor="black")

plt.xlabel('YearsExperience')

plt.ylabel('Salary')

plt.title('scatter plot for salary of mobile')

plt.plot(x,y,c="red")

#plt.savefig("scatter come line plot for salary of mobile.png")

plt.show()

## 

## Bar plot for Mean of Salary for each Job

# Bar plot of avg salary of all job roles

jobTitleSalary = df.groupby('Job Title').agg({'Salary': lambda x: x.mean()}).sort\_values('Salary', ascending=False)[:5].reset\_index()

bars = plt.bar(jobTitleSalary['Job Title'], jobTitleSalary['Salary'],color="green")

for bar in bars:

    plt.text(bar.get\_x() + bar.get\_width()/2, bar.get\_height()//2, int(bar.get\_height()), ha='center', va='bottom', fontsize=10, bbox=dict(facecolor='white', edgecolor='white', boxstyle='round,pad=0.2'))

plt.xticks(fontsize=8)

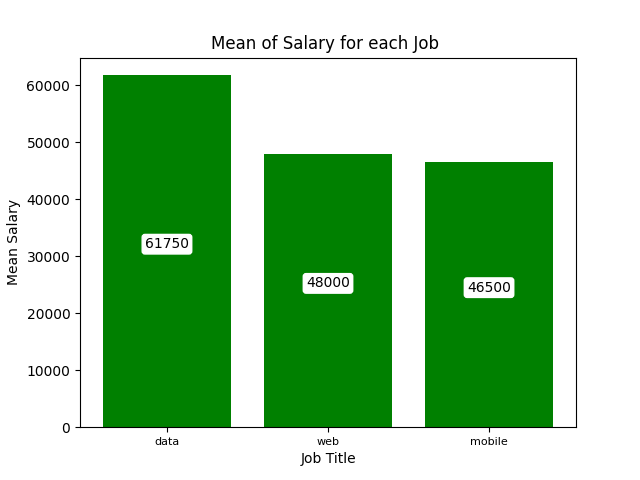
plt.title('Mean of Salary for each Job ')

plt.xlabel('Job Title')

plt.ylabel('Mean Salary')

#plt.savefig("Bar plot for avg salary off all job roles.png")

plt.show()



## 

## predictive model

# predicting model

# Encode categorical variables (Position)

label\_encoder = LabelEncoder()

df['PositionEncoded'] = label\_encoder.fit\_transform(df['Position'])

# Split data into features (X) and target (y)

X = df[['YearsExperience', 'PositionEncoded']]

y = df['Salary']

# Split data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Train the linear regression model

model = LinearRegression()

model.fit(X\_train, y\_train)

# Make predictions

y\_pred = model.predict(X\_test)

# Evaluate the model

mse = mean\_squared\_error(y\_test, y\_pred)

print(f"Mean Squared Error: {mse}")

# Example to predict salary

#jobroles(   data , web , mobile )

print("\t ---Please Enter Your Details To Predict Salary---\n")

years\_exp = float(input("Enter the years of experiance(Ex:1.1) == "))

print("types of job roles:\n data \n mobile \n web")

job\_role = str(input("Enter job role == "))

# Encode job role using the same label encoder

job\_role\_encoded = label\_encoder.transform([job\_role])

# Prepare input for prediction

input\_data = pd.DataFrame({'YearsExperience': [years\_exp], 'PositionEncoded': job\_role\_encoded})

# Predict the salary

predicted\_salary = model.predict(input\_data)

print(f"Predicted salary for {years\_exp} years of experience as a {job\_role} = ₹ {predicted\_salary[0]}")

# thanks message

print("\n\t --- Thanks For Using Our Salary Predicting Model---")

**Output:**

---Please Enter Your Details To Predict Salary---

Enter the years of experiance(Ex:1.1) == 4.6

types of job roles:

data

mobile

web

Enter job role == data

Predicted salary for 4.6 years of experience as a data = ₹ 60318.93012647728

--- Thanks For Using Our Salary Predicting Model---

**Result:**

**We are successfully predicted the salary based on job-role and experience.**

**Conclusion :**

In conclusion, the salary prediction project utilizing job title and years of experience demonstrates the power of machine learning in providing valuable insights into compensation trends. The project’s methodology, from data collection to model deployment, underscores the importance of a systematic approach to predictive analytics. By focusing on key variables and employing robust algorithms, the model offers a practical tool for both employers and job seekers to make informed decisions regarding salaries. The project also highlights the dynamic nature of the job market and the need for continuous model refinement to adapt to changing economic conditions. Overall, the salary prediction model serves as a testament to the potential of data-driven strategies in human resource management and career planning.

Thank you