|  |  |
| --- | --- |
| Internship Project Title | TCS iON RIO-125: HR Salary Dashboard - Train the Dataset and Predict Salary |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | BK Birla college of Arts, Science & Commerce (Autonomous),kalyan |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 25-10-2023 | 06-12-2023 | | 125 | | Jupyter Notebook | Python 3,Flask |
| Milestone # | 3 | Milestone: | | Create data set, Clean and sanitize dataset, Preprocessing data set, Test and train the dataset, Build the classifier models and fit the data in the model, Dash board creation. | | |

**TABLE OF CONTENTS**

|  |
| --- |
| **Acknowledgement** |
| 1. **Objective** |
| 1. **Introduction** |
| 1. **Internship Activities**    1. Creating the dataset    2. Analyzing the dataset    3. Cleaning and sanitizing    4. Splitting the data set    5. Modeling    6. Hyper Parameter tuning 2. **Result & Conclusion** 3. **Enhancement scope** |
| 1. **Link to code and executable file** |

**ACKNOWLEDGEMENTS**

I extend my sincere gratitude to my mentors, colleagues, and all those who generously shared their guidance and support during this internship. Thank you for your time, insights, and unwavering support, which have been instrumental in my professional development.

I would also like to extend my gratitude to all my friends who offered their valuable suggestions during this project. Their input has greatly enriched my work, and I am thankful for their unwavering support.

**1. OBJECTIVE**

The objective of this project is to build a salary prediction dashboard for human resource management. This dashboard will use machine learning algorithms to predict the salary of job candidates based on factors like their experience, age, and qualifications. This information can help HR managers to make better decisions when hiring candidates for job positions.

**2. INTRODUCTION**

The human resources department often has a lot of job applicants to process and must choose the best candidates for each job. Candidates often consider salary when deciding whether to accept a job offer, so it's important for HR to offer competitive salaries.

In this project, we will use a dataset that includes information on over 1,00,000 job candidates, such as their experience and salary. This dataset is good for our analysis because it has a wide range of job profiles and salaries. We will use this data to build a salary prediction dashboard to help HR managers make better decisions about salaries for job candidates.

I have trained and tested my data using Logistic Regression,Linear Regression, KNN, Decision tree and Random Forest compared them to select the best model.

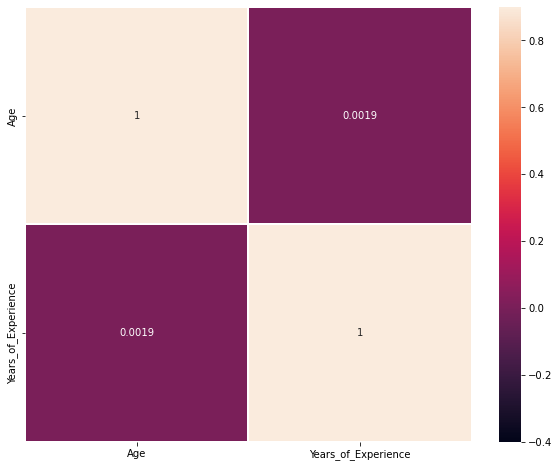
**3. INTERNSHIP ACTIVITIES**

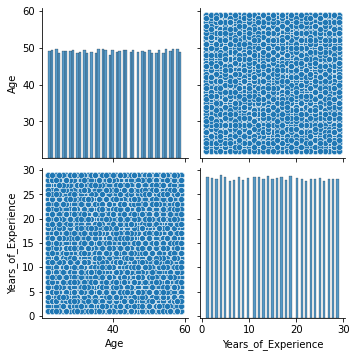
**3.1 Creating the dataset**

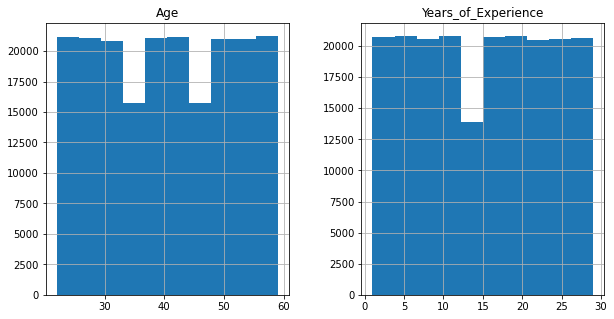
For this project, a dataset with 2,00,000 rows and 4 columns has been selected. The columns in the dataset include Age, Years Of Experience and Salary.

**3.2 Analyzing the dataset**

The results of this analysis are presented below.







**3.3 Cleaning and sanitize the dataset**

In the data set, no null values are present, but ‘?’ present in some columns and are replace with respective mode of each column. Then checked for the outliers before processing, and in our data, outliers are found in some of the columns like age and hours-per-week and are removed. The categorical data has been converted to a numerical form in this step using label encoding.

**3.4 Splitting the data**

The dataset has been split into two parts.

**3.5 Modeling**

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. A classification model tries to draw some conclusions from the input values given for training. It will predict the class labels/categories for the new data.

**A. Logistic regression classifier**

Logistic regression is a machine learning algorithm used for classification tasks. It models the probabilities of possible outcomes of a single trial using a logistic function, making it useful for understanding how independent variables influence a single outcome variable. However, logistic regression only works when the predicted variable is binary, assumes that all predictors are independent of each other, and requires the data to be free of missing values.

**B. K-Nearest Neighbours**

**Neighbors-based classification is a type of machine learning algorithm that falls under the category of lazy learning. It does not attempt to construct an internal model but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point. This algorithm is simple to implement and robust to noisy training data, making it effective when the training data is large. However, determining the value of k can be challenging, and the computation cost is high as it needs to compute the distance of each instance to all the training samples.**

**C. Decision Tree classifier**

A decision tree is a machine learning algorithm that generates a sequence of rules based on the input data's attributes and corresponding classes. It is a popular algorithm because it is easy to understand and visualize, requires minimal data preparation, and can handle both numerical and categorical data.However, decision trees have some limitations. They can create overly complex trees that may not generalize well to new data. Additionally, small variations in the data can result in completely different decision trees being generated, making them unstable.

**D. Random forest classifier**

The Random Forest classifier is a meta-estimator that fits multiple decision trees on various sub-samples of datasets and uses an average to improve the predictive accuracy of the model while controlling over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement. This reduces over-fitting, and in most cases, Random Forest classifiers are more accurate than decision trees.

However, one disadvantage of the Random Forest classifier is that it may not be suitable for real-time prediction, as it can be computationally expensive to build the forest and classify new data. Additionally, implementing a Random Forest classifier may be more difficult than implementing simpler models due to the complexity of the algorithm and its many hyperparameters that need to be tuned.

|  |  |
| --- | --- |
| **MODEL** | **ACCURACY** |
| **LINEAR REGRESSION** | **0.4986** |
| **Logistic regression classifier** | **0.4986** |
| **K-Nearest Neighbours** | **0.4986** |
| **Decision Tree classifier** | **0.4986** |
| **Random forest classifier** | **0.4986** |

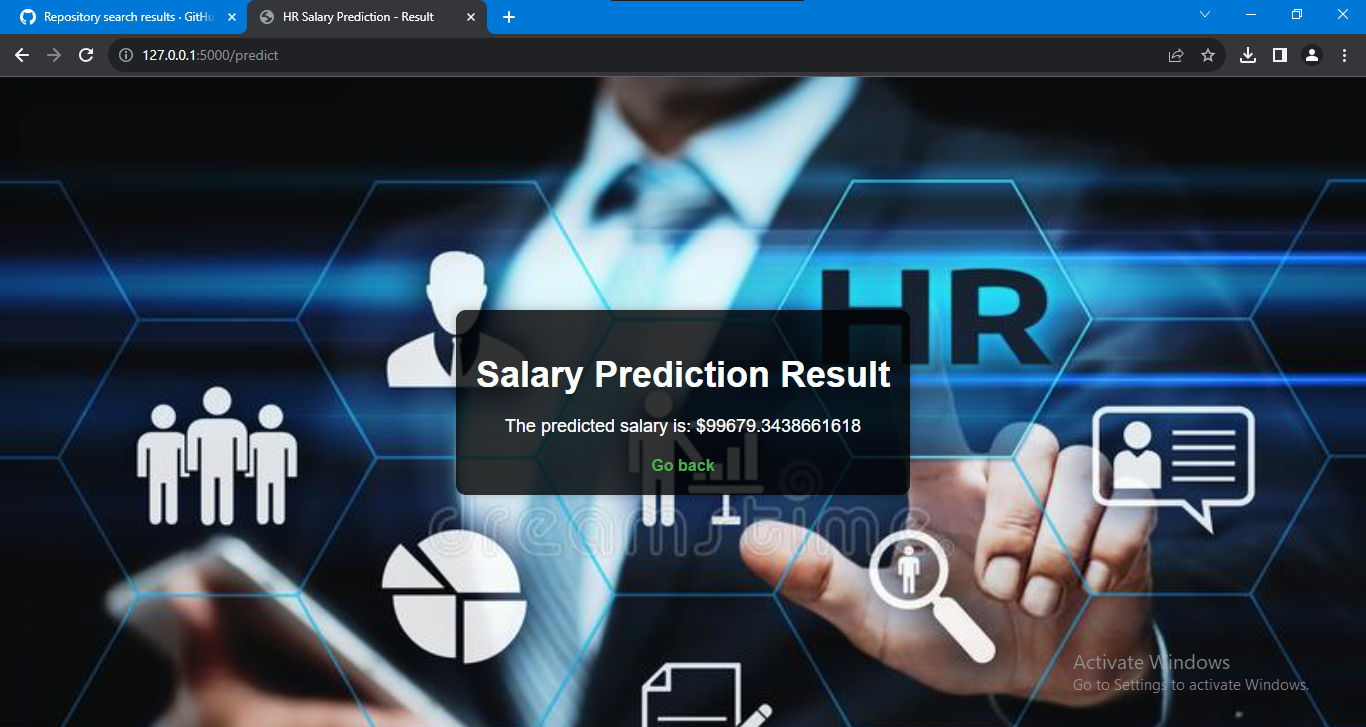
**3.6 Hyper Parameter Tuning**

It is a crucial step to improve accuracy of the forecasted results. Most performance variation can be attributed to just a few hyperparameters; the tunability of an algorithm, hyperparameter, or interacting hyperparameters is a measure of how much performance can be gained by tuning it. In machine learning, a hyperparameter is a parameter whose value is used to control the learning process. The values of other parameters are derived via training.Here I have done fine tuning of Linear regression and decided to use it as model for HR salary prediction.

**4. RESULTS & CONCLUSION**

The dataset for HR salary prediction has 4 features, but only 2 of them were found to be relevant for making accurate predictions. After experimenting with various machine learning models, it was observed that Linear regression model performed the best with the highest accuracy. The Linear regression model was fine-tuned to further improve its performance, and it was ultimately chosen as the most suitable model for this particular dataset.A website is hosted using these features to predict salary.





**5. ENHANCEMENT SCOPE**

This industry project has a broad scope. It involves predicting the salary of an individual using their resume or CV. Some of the Natural language processing techniques will help in developing this website.

**6. LINK TO CODE AND EXECUTABLE FILE**

<https://github.com/NAVINB-07/TCS-Hr-Salary-Prediction-Dashboard>