**HR Salary Dashboard - Train the Dataset and Predict Salary**

**Internship Project Report**

Submitted to

**Department of Computer Science**

**Faculty of Science & Technology**

*Under*

**TCS Ion Industry Honour Program**

**Vishwakarma University , Pune**

**(Maharashtra)**

*By*

**Sumeet Dharmendra Choudhary**

*Under the Supervision of*

**Industry Mentor Faculty Mentor**

**Mr. Debashis Roy Prof. Shriprada Chaturbhuj**

**Tata Consultancy Services Vishwakarma University, Pune**

|  |  |
| --- | --- |
| 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 | Vishwakarma University |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 19/06/2024 | 19/07/2024 | 125 | Python | Google Collab / VS Code |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**DECALARATION**

I Sumeet Dharmendra Choudhary (202201767) hereby declare that the work embodied in this project entitled "HR Salary Dashboard - Train the Dataset and Predict Salary", carried out under the supervision of Industry Mentor, Mr. Debashis Roy and Faculty Mentor, Prof. Shriprada Chaturbhuj, Assistant Professor, Faculty of Science and Technology, Vishwakarma University, Pune, is an original work and does not contain any work submitted for the award of any degree in this university or any other university.

Sumeet Dharmendra Choudhary BSc. Computer Science-Big Data Analytics Department of Science and Technology, Vishwakarma University, Pune

**ACKNOWLEDGEMENT**

I would like to extend my heartfelt gratitude to our esteemed academic mentor, Prof. Shriprada Chaturbhuj of Vishwakarma University, and our industry mentor, Mr. Debashis Roy of TCS iON, for their invaluable contributions to the successful completion of my project, "HR Salary Dashboard - Train the Dataset and Predict Salary". I am profoundly grateful for the time and effort they have dedicated throughout the internship. Your insightful guidance and unwavering support have been instrumental to my progress.

I also wish to express my sincere thanks to Vishwakarma University and TCS iON for granting me this exceptional opportunity.

I affirm that this project was undertaken and completed solely by me. It has been submitted to the Department of Computer Science of Vishwakarma University, under the TCS iON Industry Honour Program, during the academic year 2024-2025,

Furthermore, I am deeply appreciative of the support provided by the Head of the Department, Dr. Rajkumar Jagdale.

**CERTIFICATE**

This is to Certify that the Project titled "HR Salary Dashboard - Train the Dataset and Predict Salary" submitted by Sumeet Dharmendra Choudhary, is an original work and has not been previously submitted in part or full for the award of any degree or diploma to this or any other university. The project is submitted to Vishwakarma University, Pune and TCS iON Industry Honour Program, in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in the Subject of Computer Science - Big Data Analytics.

Date:

Prof. Shriprada Chaturbhuj

Faculty Mentor

Dr. Rajkumar Jagdale

Head Of Department

( Computer Science )

**Objective :**

The objective of this project was to develop a robust HR Salary Prediction Dashboard that can predict the salary category of employees based on various features. The goal was to utilize machine learning techniques to train a model and integrate it into an interactive dashboard for user-friendly predictions.

**Introduction / Description of Internship:**

This internship was conducted at TCS iON, where I worked on developing a predictive model for salary classification and creating an interactive dashboard for HR analysis. The project involved data preprocessing, model training, and visualization of predictions through a user-friendly interface.

**Internship Activities:**

* Data collection and preprocessing
* Exploratory data analysis (EDA)
* Model selection and training
* Creation of visualizations for data insights
* Development of an interactive dashboard using Streamlit
* Testing and validation of the predictive model
* Documentation and reporting

**Approach / Methodology:**

1. **Data Collection and Preprocessing:**

* Loaded the dataset and handled missing values.
* Converted categorical variables into numeric form using factorization.
* Scaled the features using StandardScaler.

1. **Exploratory Data Analysis (EDA):**

* Visualized age distribution, working class, education level, marital status, occupation, gender distribution, working hours, and country of origin using histograms, pie charts, and bar plots.

1. **Model Training:**

* Split the dataset into training and test sets.
* Trained a Logistic Regression model on the training set.
* Evaluated the model using precision, recall, and accuracy metrics.

1. **Dashboard Development:**

* Developed an interactive dashboard using Streamlit.
* Integrated model predictions into the dashboard.
* Enabled user inputs for various features to predict salary category.

**Assumptions:**

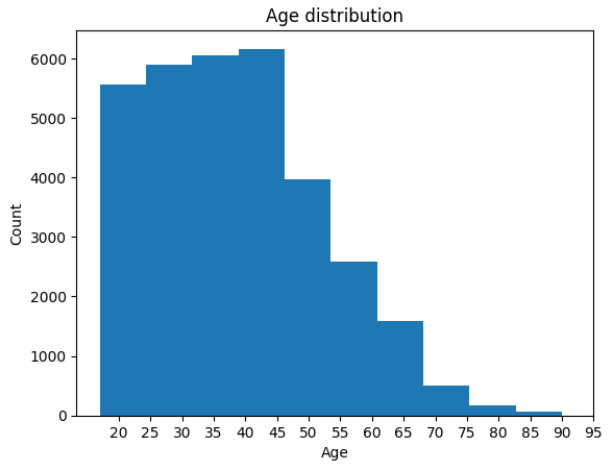
* The dataset is representative of the population and free from significant biases.
* The features provided are sufficient to predict salary categories accurately.
* The logistic regression model is appropriate for this binary classification task.

**Exceptions / Exclusions:**

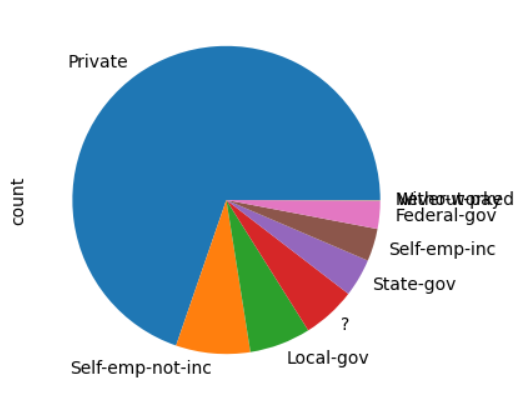
* Data entries with missing or unknown values in critical fields were excluded from the analysis.
* Outliers were not extensively treated but were considered during the EDA.

**Charts, Table, Diagrams:**

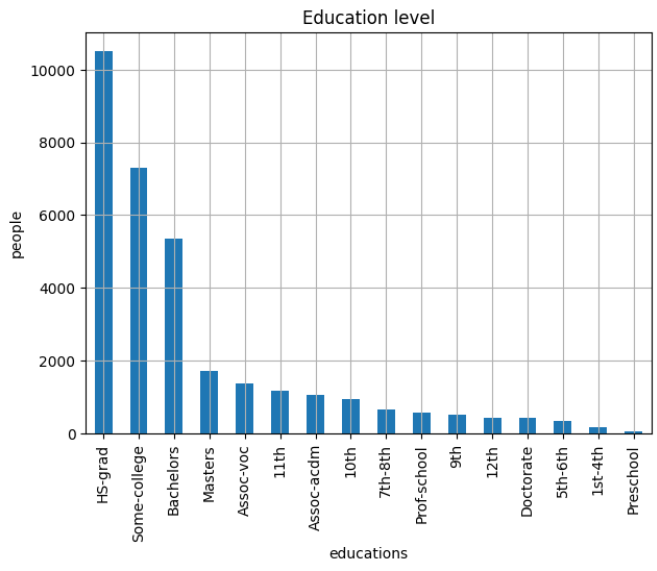
* **Age Distribution:**

****

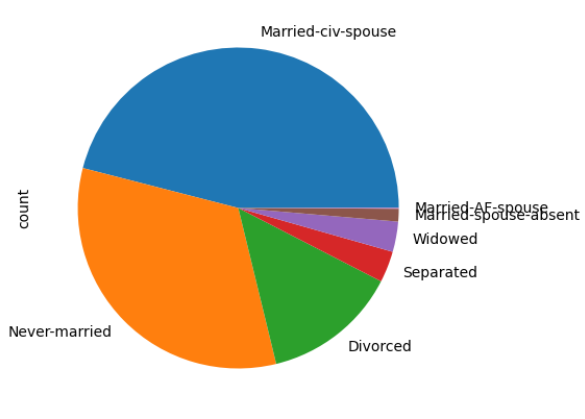
* **Workclass Distribution:**

****

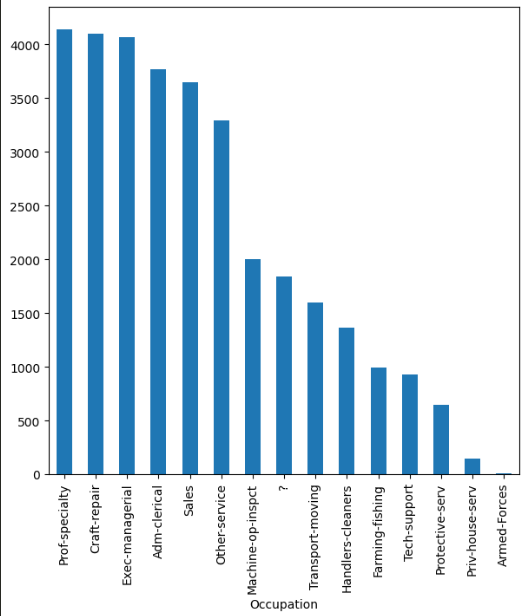
* **Education Level:**

****

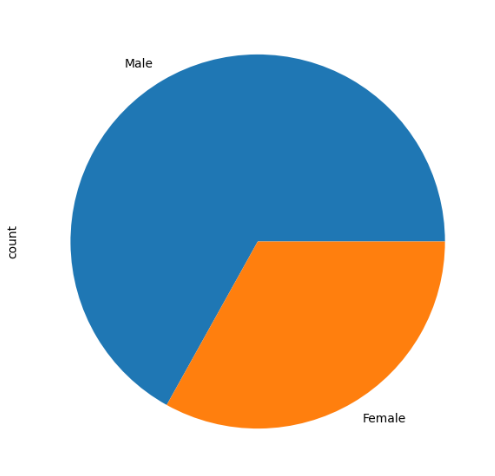
* **Marital Status:**

****

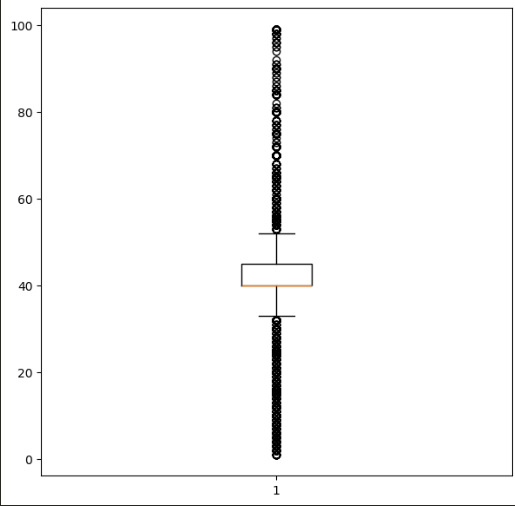
* **Occupation:**

****

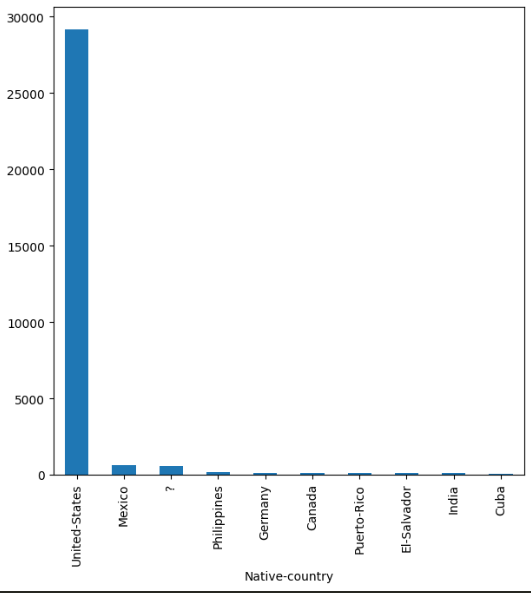
* **Gender Distribution:**

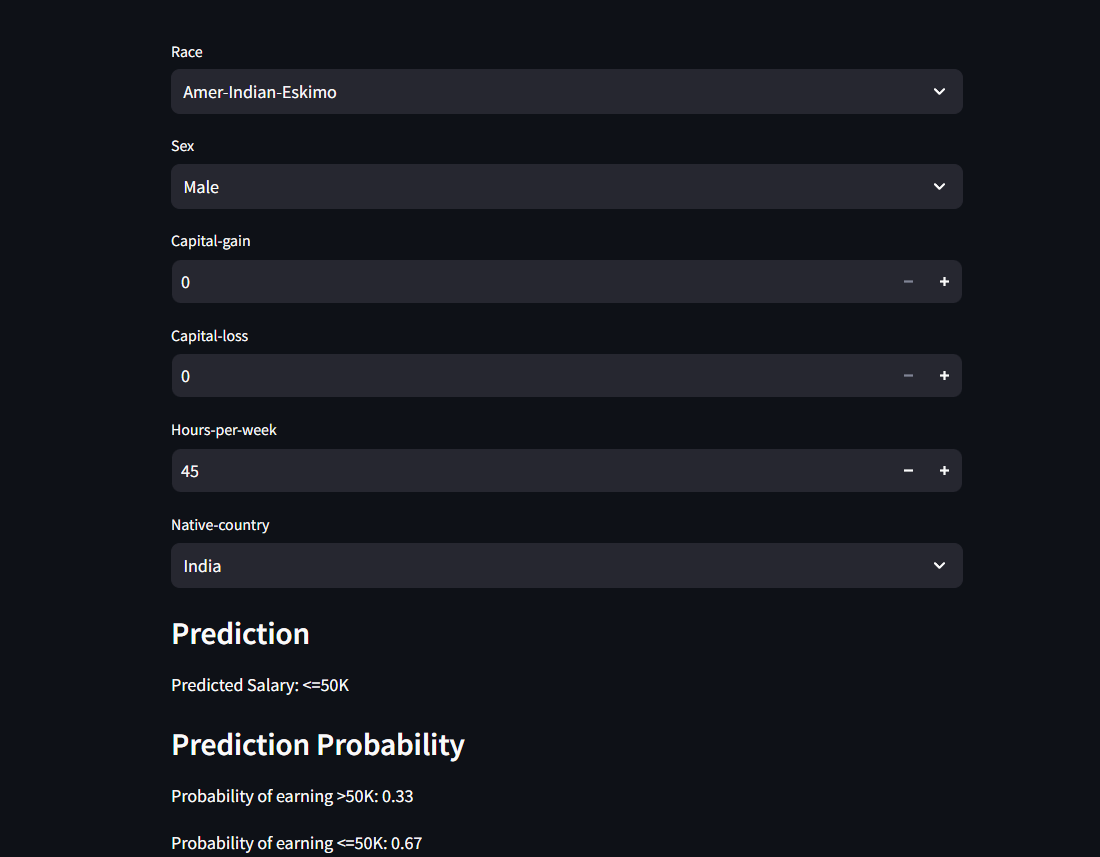
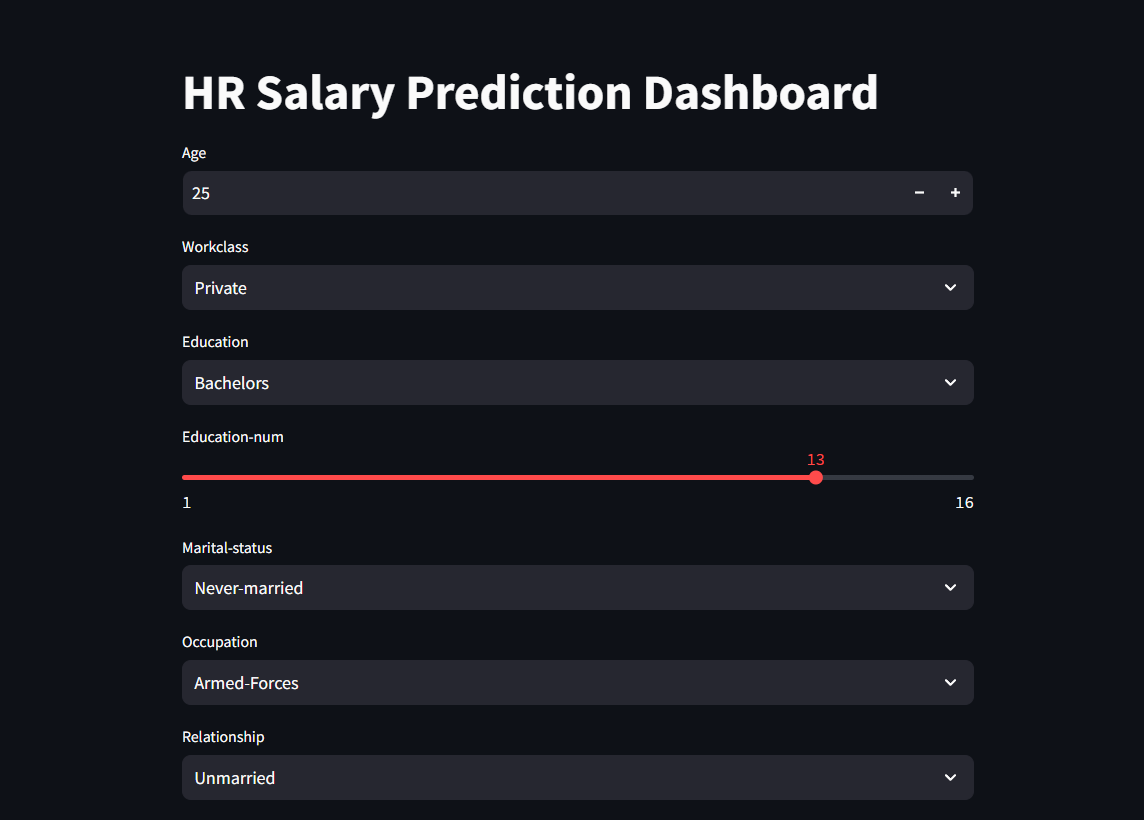
****

* **Working Hours per Week:**

****

* **Country of Origin:**

****

**HR Salary Dashboard :** *Run the Dashboard : streamlit run app.py*****

**Education-num:** These values help in converting the categorical "Education" feature into a numeric form suitable for model training.

|  |  |
| --- | --- |
| **Education** | **Education-num** |
| Preschool | 1 |
| 1st-4th | 2 |
| 5th-6th | 3 |
| 7th-8th | 4 |
| 9th | 5 |
| 10th | 6 |
| 11th | 7 |
| 12th | 8 |
| HS-grad | 9 |
| Some-college | 10 |
| Assoc-acdm | 11 |
| Assoc-voc | 12 |
| Bachelors | 13 |
| Masters | 14 |
| Prof-school | 15 |
| Doctorate | 16 |

**Algorithms:**

**Logistic Regression:** Chosen for its simplicity and effectiveness in binary classification tasks. It provided interpretable results and was suitable for the given dataset.

**Challenges & Opportunities:**

**Challenges:**

* Handling missing data and unknown values in categorical features.
* Scaling and preprocessing data to ensure model accuracy.
* Integrating the model with an interactive dashboard for real-time predictions.

**Opportunities:**

* Improving the model with more advanced algorithms such as Random Forest or Gradient Boosting.
* Enhancing the dashboard with additional features and insights.
* Scaling the solution to handle larger datasets and more complex predictions.

**Risk Vs Reward:**

**Risks:**

* Potential biases in the dataset affecting model accuracy.
* Misinterpretation of predictions by users.
* Technical challenges in integrating and deploying the dashboard.

**Rewards:**

* Providing a valuable tool for HR analytics and decision-making.
* Enhancing predictive capabilities with real-time data inputs.
* Gaining experience in model development, deployment, and user interface design.

**Reflections on the Internship:**

This internship provided me with practical experience in data science and machine learning. I learned how to handle real-world data, develop predictive models, and create user-friendly interfaces. The project reinforced the importance of clear communication and iterative development.

**Recommendations:**

* Incorporate feedback from HR professionals to refine the model and dashboard.
* Explore additional features and more complex models to improve accuracy.
* Conduct regular updates and maintenance to ensure the dashboard remains relevant and accurate.

**Outcome / Conclusion:**

The HR Salary Dashboard successfully predicts salary categories based on user inputs and provides valuable insights into the demographic factors influencing salaries. The project demonstrated the effectiveness of logistic regression and the importance of interactive visualizations in data analysis

**Enhancement Scope:**

* Integrate more sophisticated machine learning models.
* Add more visualizations and analytical features to the dashboard.
* Expand the dataset to include more diverse and recent data.
* Implement user authentication and data security measures.

**Link to code and executable file:**

[**TCS\_iON Github**](https://github.com/sumeet156/TCS-iON-Rio-125)

**Research questions and responses:**

**Research Question 1:** How do demographic factors influence salary distribution?

* Response: Factors such as education level, occupation, and marital status significantly influence salary distribution. Higher education levels and certain occupations tend to have higher salary brackets.

**Research Question 2:** Can logistic regression effectively predict salary categories?

* Response: Logistic regression proved to be an effective model for this binary classification task, achieving satisfactory precision and recall metrics.

**Research Question 3:** What are the key challenges in developing an HR salary prediction model?

* Response: Key challenges include handling missing data, scaling features, and ensuring the model is free from biases. Additionally, integrating the model into an interactive dashboard posed technical challenges.