Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

|  |  |
| --- | --- |
| Internship Project Title | Rio 210: Salary Prediction Dashboard for HRs |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Himdweep Walia |
| Name of the Institute | Amity University Online |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 15-10-2024 | 23-01-2025 | 210 Hours | Jupyter Notebook | Python |

**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

1. **ACKNOWLEDGEMENTS –**

I am conveying my sincere gratitude towards my industry mentor, Himdweep Walia ma’am, for helping me throughout this project till now and providing me this wonderful platform to complete this project. I am also thankful for answering my queries at every phase of the project.

I would like to express my heartfelt gratitude to TCS iON for providing me with the opportunity to be a part of their esteemed organization as an intern. The experience I gained during my internship has been invaluable, and I am truly grateful for the support and guidance extended to me throughout this journey.

1. **OBJECTIVES –**

The objective of this project is to build a salary prediction dashboard for human resource management. The dashboard is made using various machine learning algorithms in order to predict the salary of the candidates based on parameters like age, gender, number of work hours, education and qualifications. It is crucial for the Human Resource managers to make better decisions for the recruitment of employees in their organizations.

The primary objective of this project is to develop a comprehensive Salary Prediction Dashboard that harnesses the power of data analytics and machine learning methodologies. This advanced tool is designed to forecast job candidates' salaries by analyzing a multitude of variables. By providing a detailed analysis of potential salary outcomes, the dashboard will help HR professionals align their compensation strategies with market standards, ensuring competitive and fair salary offers.

1. **INTRODUCTION / DESCRIPTION OF DASHBOARD –**

Salary Prediction is a vital aspect of modern workforce management bridging the gap between employer expectations and employee compensation. The human resources department frequently receives a large number of job applications and must select the best candidates for each opening. Candidates frequently consider salary when deciding whether to accept a job offer hence, it is critical for HR to provide competitive compensation.

In this project, I made use of a dataset (**salary\_data.csv**) containing information about over 32,000 job candidates, based on their job profiles, experience, education and region. This dataset is useful for our analysis because it includes a variety of candidates belonging to different races, experience, gender and salaries based on total work hours and total experience in the industry.

We will use this information to create a salary prediction dashboard to assist HR managers in making more informed salary decisions for job candidates. There are only two target salary categories in the dataset: salary less than or equal to 50,000 and salary more than 50,000. As a result, we derive a prediction model with the help of a variety of methods for binary classification like logistic regression, SVM and random forest classifiers. I trained and tested my data with Logistic Regression, KNN, Decision Tree, SVM, and Random Forest and compared them to determine the best model with the most accuracy.

1. **INTERNSHIP ACTIVITIES –**
2. Reviewed welcome kit videos to thoroughly grasp the project scope.
3. Prepared for RIO pre-assessment to ensure readiness for evaluation.
4. Successfully completed the RIO pre-assessment to gauge understanding and proficiency.
5. Reviewed the day-wise plan to align the tasks with objectives.
6. Engaged with project and industry reference materials for deeper insight.
7. Attended webinars and jotted down important pointers to enrich my knowledge base with new insights.
8. Actively contributed to digital discussion room interactions for enhanced learning.
9. Utilized educational resources like blogs, books and YouTube videos to deepen the comprehension of project topics.
10. Created a GitHub account for efficient file management and collaboration.
11. Selected a dataset that met project criteria for analysis and modeling.
12. Maintained transparency and accountability through regular progress documentation.
13. Ensured dataset adequacy via thorough verification process.
14. Acquired data cleaning skills through articles and tutorials.
15. Enhanced dataset quality through effective cleaning and sanitization procedures.
16. Conducted exploratory data analysis to uncover valuable insights and trends.
17. Engaged in self-learning via instructional videos on model training.
18. Explored logistic regression and KNN classifiers for model training.
19. Worked on HTML code to develop the front-end interface of the Salary Prediction Dashboard.
20. Created the Salary Prediction Dashboard interface using VS Code environment and tested it.
21. Made a few modifications in the code wherever required.
22. Managed and collaborated on the project through the GitHub repository.
23. **APPROACH/METHODOLOGY –**

* 1. **DATA COLLECTION –**

I started the data collection process by utilizing some easily accessible sources such as Kaggle or the UCI Machine Learning Repository. I carefully selected a dataset which was relevant to my project's objectives and areas of interest. When I first accessed the dataset, I went over any accompanying documentation or metadata in order to gain a good understanding of its structure and variables. This ensures that I am well-prepared to carry out the subsequent data analysis and interpretation tasks effectively.

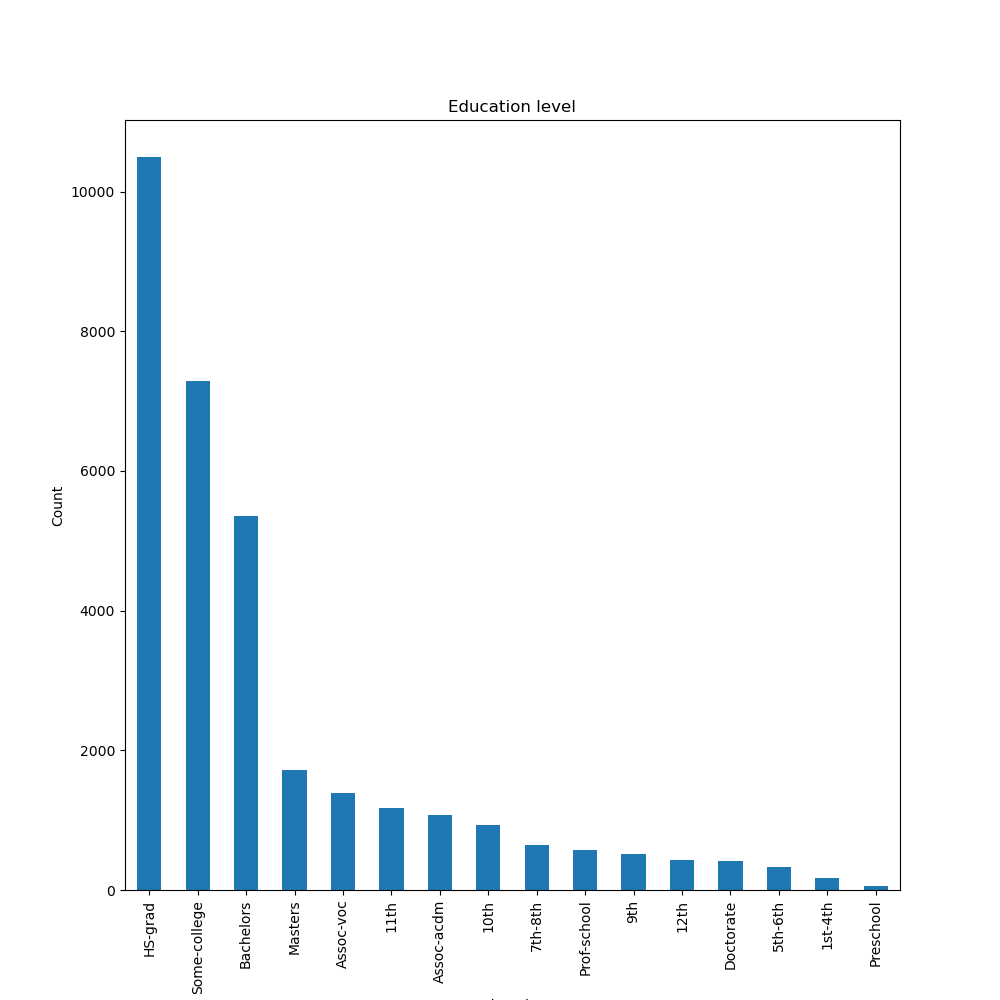
* 1. **DATA INTERPRETATION –**

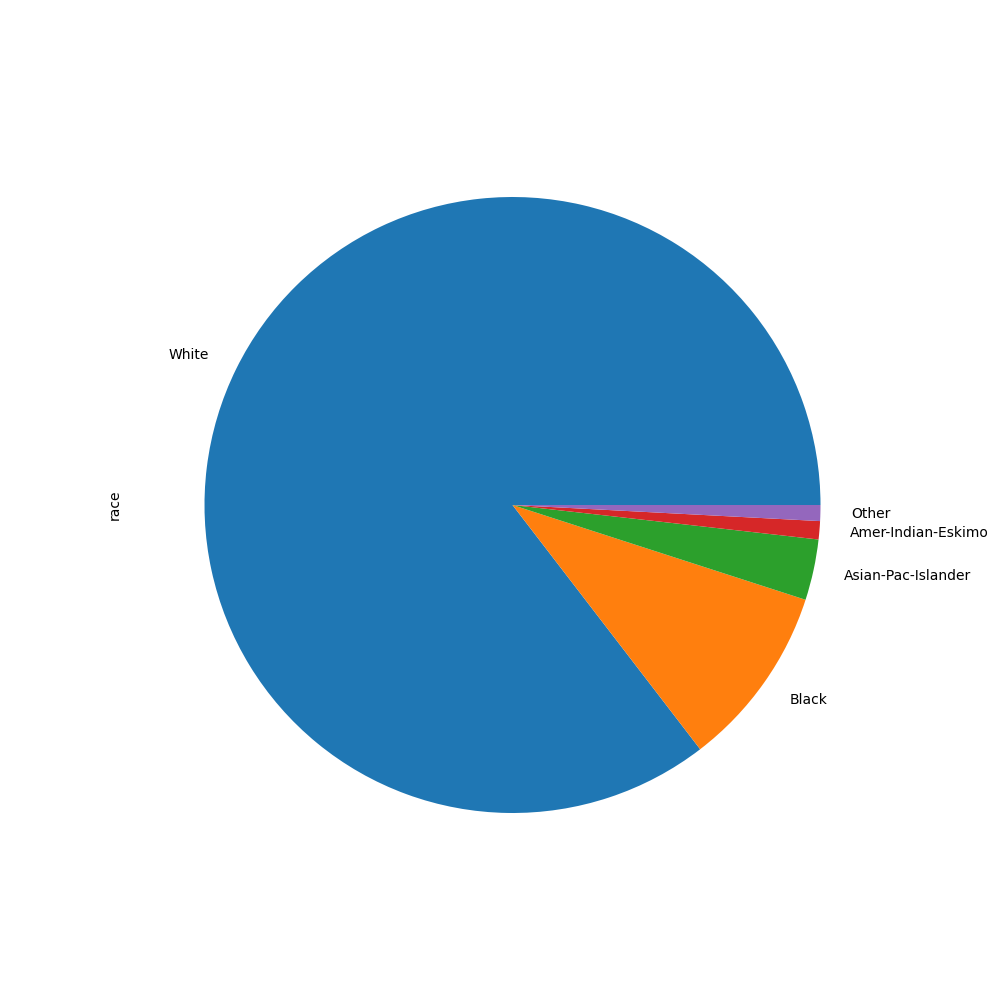
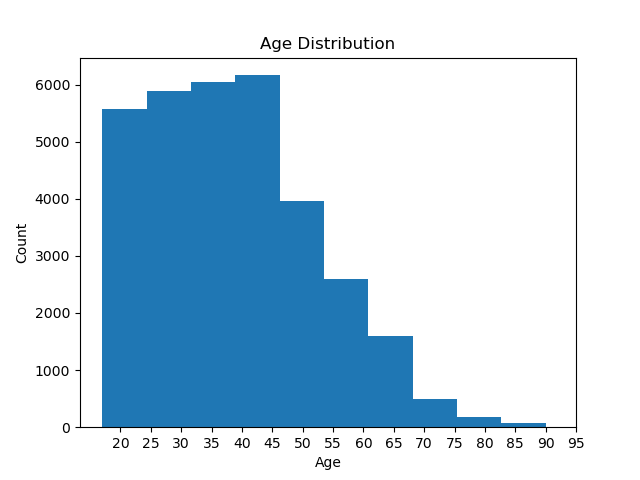
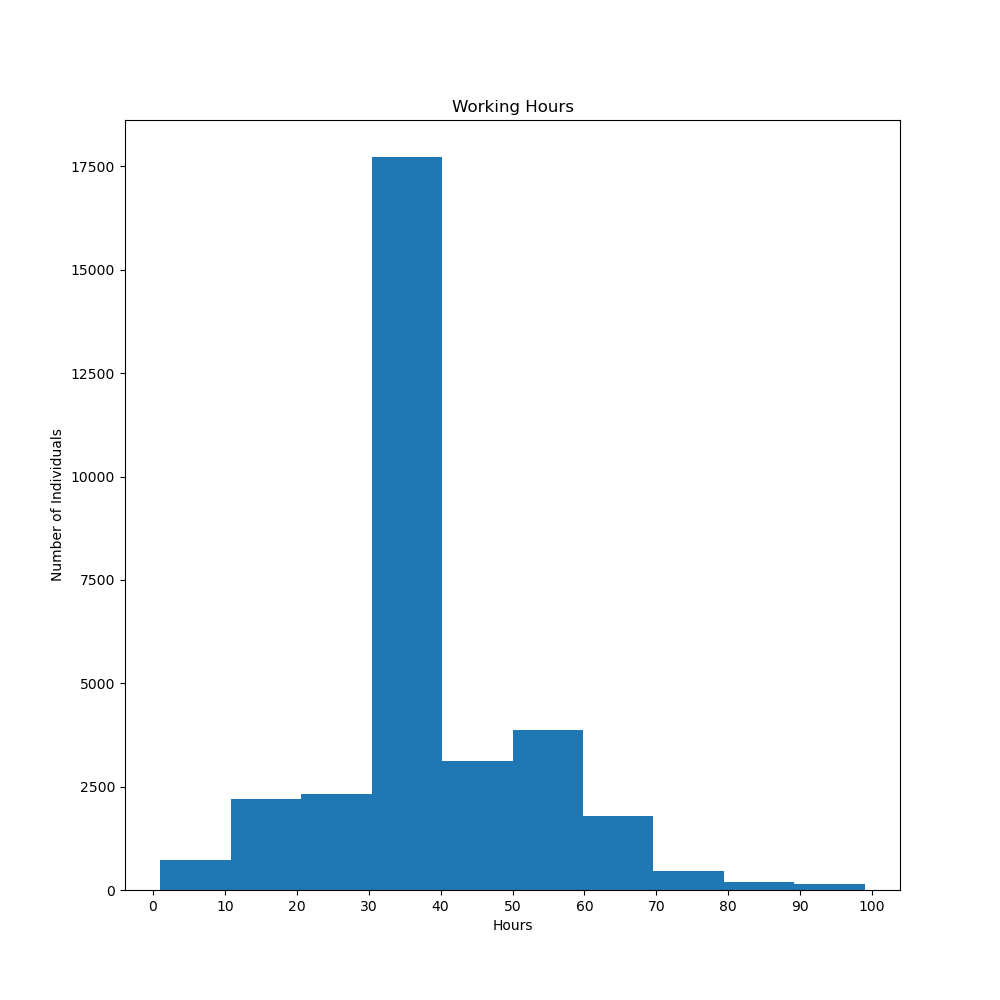
In data interpretation, I begin by delving into the dataset's structure, employing basic Python commands or functions, such as Pandas, to load the dataset and conduct a thorough examination of its layout. Subsequently, I identified the key variables in the dataset, while carefully selecting those which are relevant to my analysis or project objectives. I checked for the null values or missing values from the dataset in order to sanitize it. I made sure that there are no special characters present in the dataset. Following this, I began visualizing the data distribution, creating simple visual representations such as histograms for numerical variables and bar charts for categorical variables. I ensured that the column wise bar plots and pie charts are displayed. These visualizations illuminate the dataset's inherent distributions, which facilitates a better understanding of its underlying characteristics and trends. After which I incorporated the machine learning algorithms like logistic regression, random forest classifiers, k-nearest neighbors, support vector classifiers, and decision tree classifiers to find out the model with most accuracy. I summarized these models in a tabular form and performed hyperparameter tuning on it to further improve its accuracy from 0.80 to 0.82 on the random forest classifier model. I created a Flask application and created a dashboard for salary prediction wherein the user will input the data based on which the predicted output will be displayed and created a GitHub repository to save and share all my files.

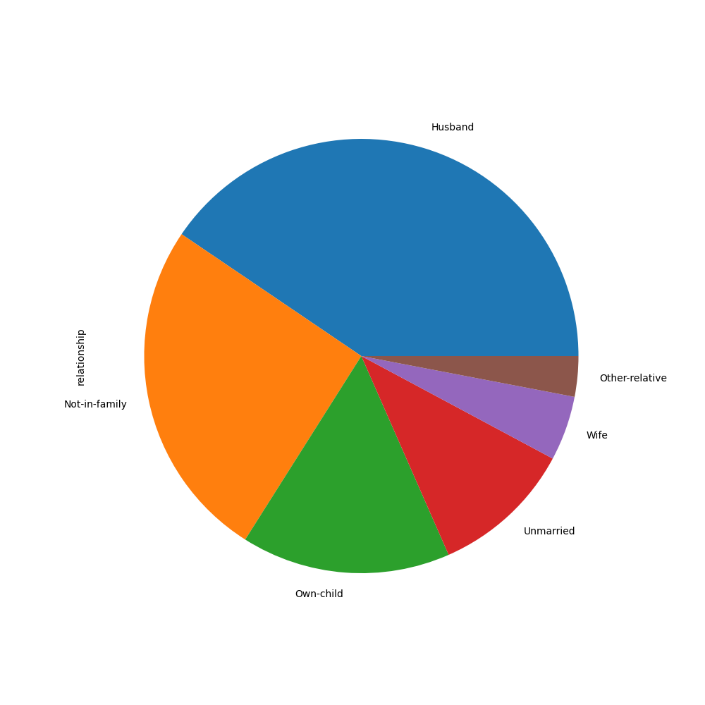
1. **CHARTS/ TABLES/ DIAGRAMS –**

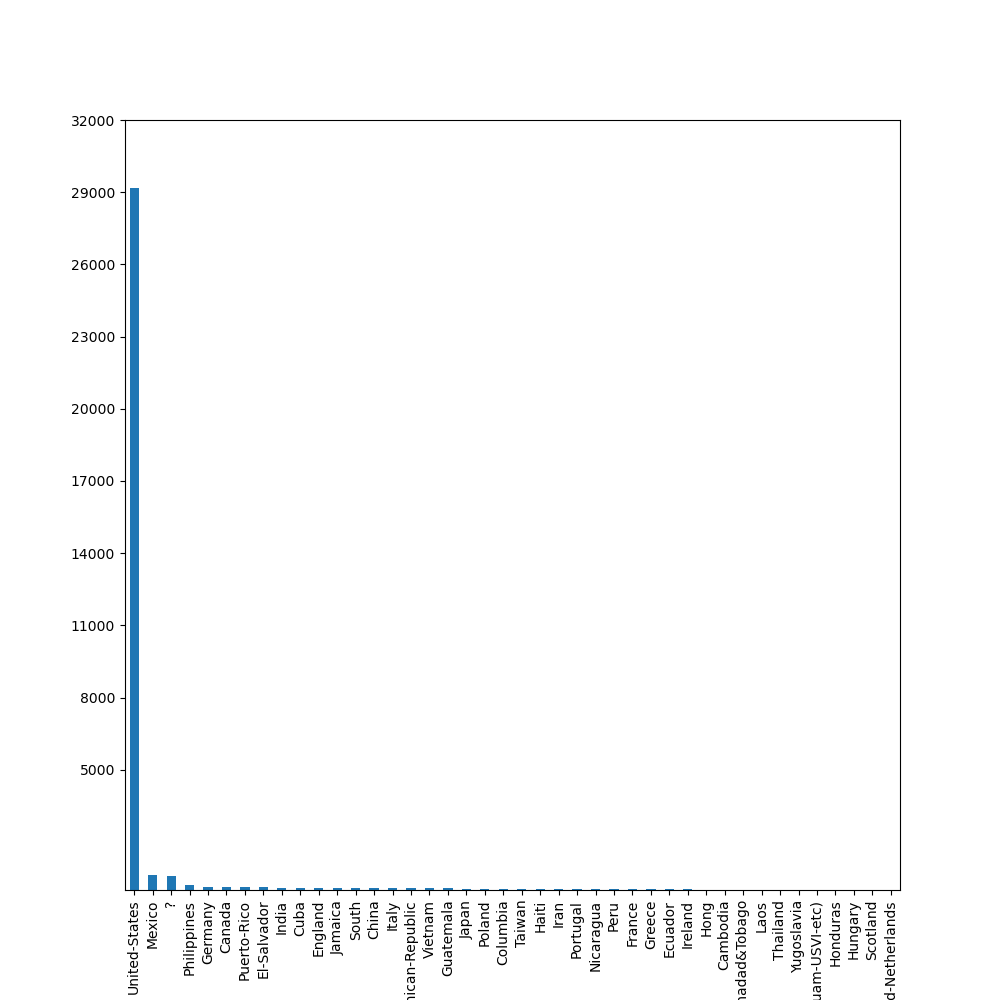
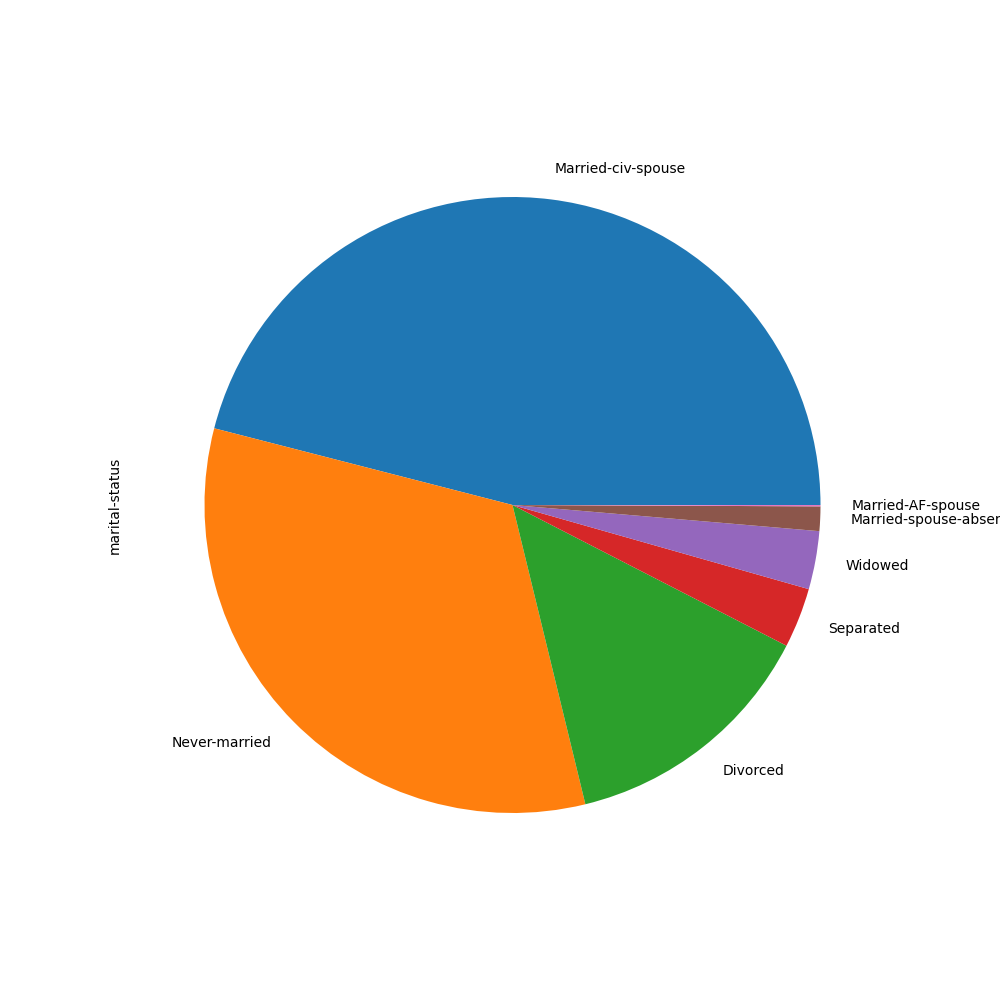
Following are the bar plots and charts according to the column names in the dataset salary\_data.csv

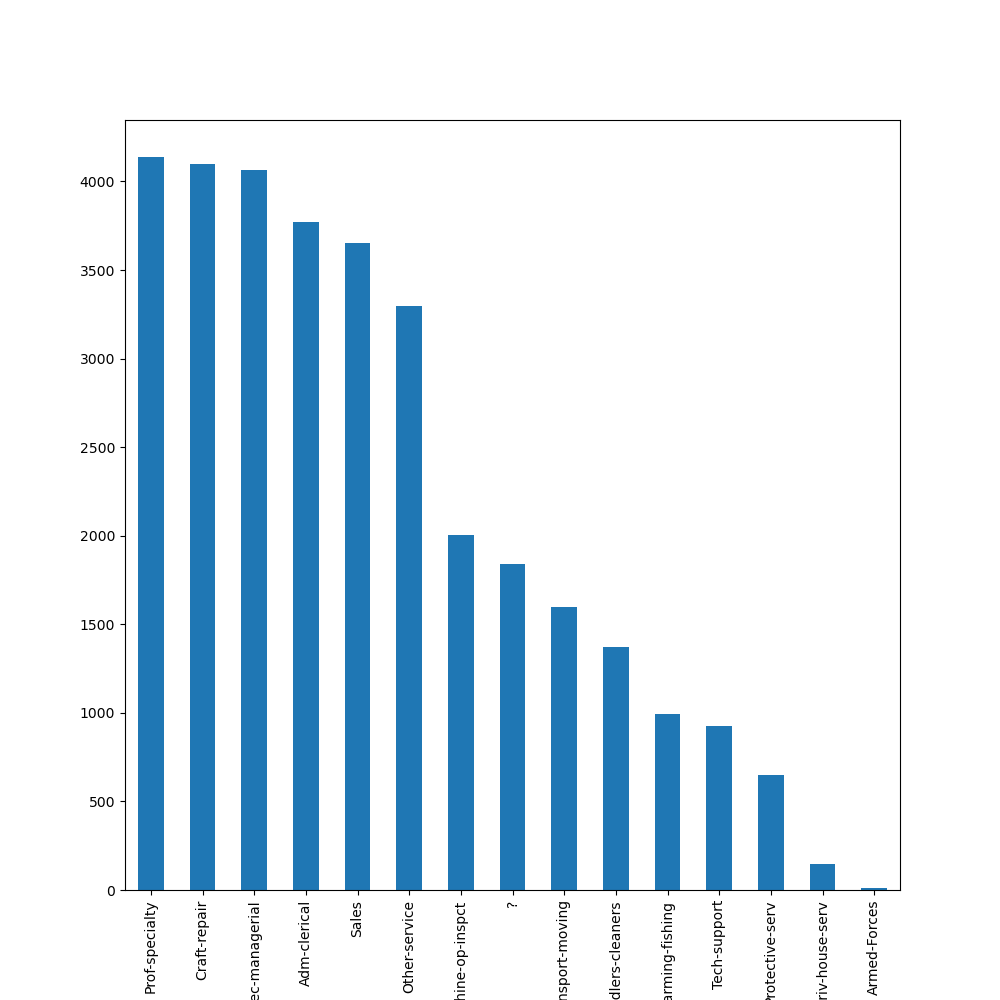
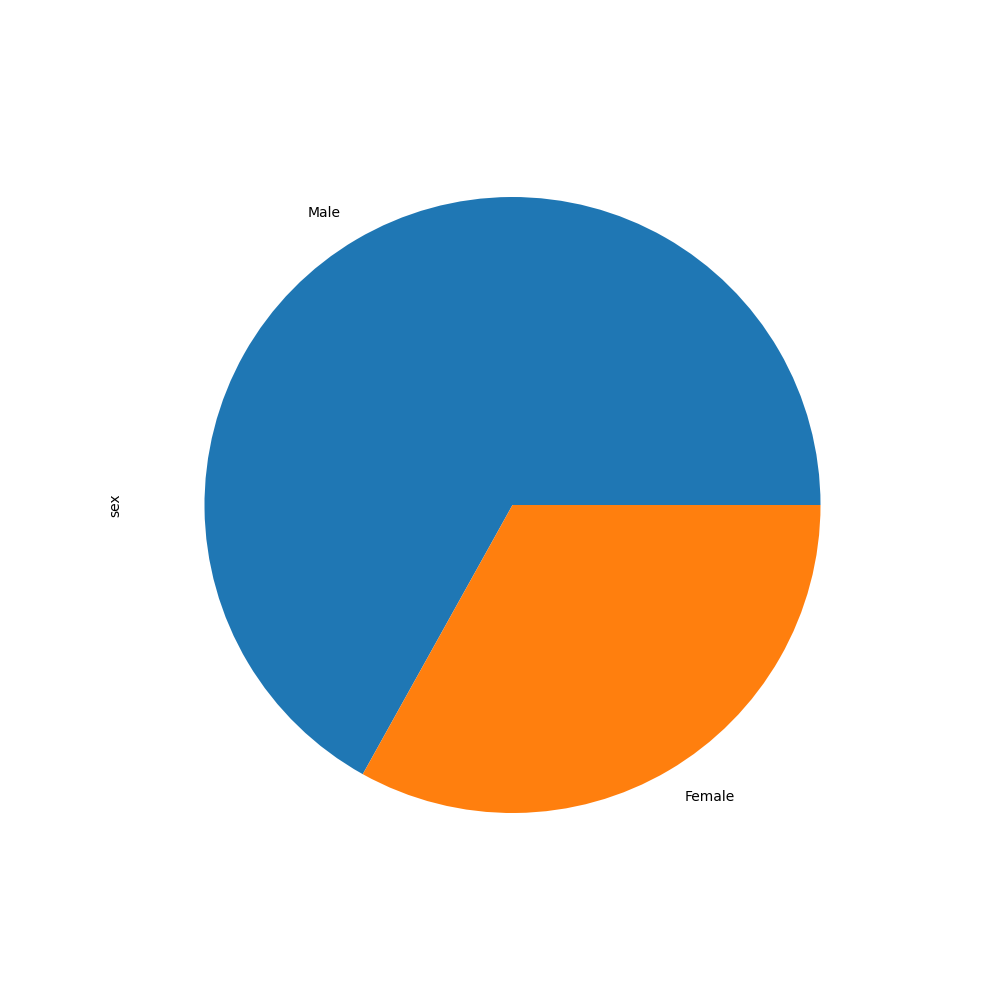
1. Age Distribution
2. Education Level
3. Race
4. Relationship
5. Hours per week
6. Marital Status
7. Native Country
8. Sex
9. Occupation
10. Work Class

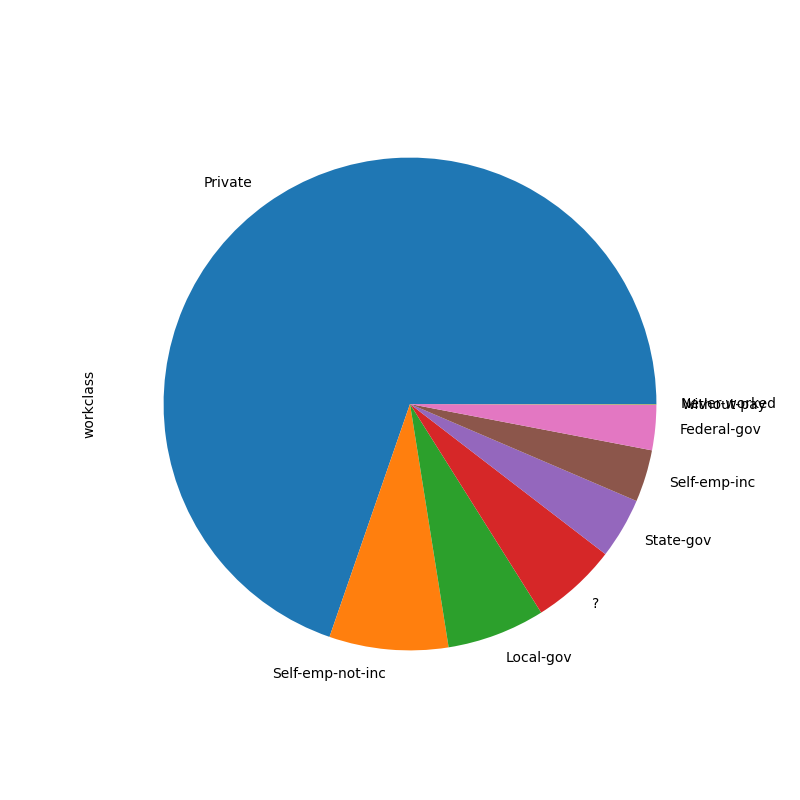
****

****

****

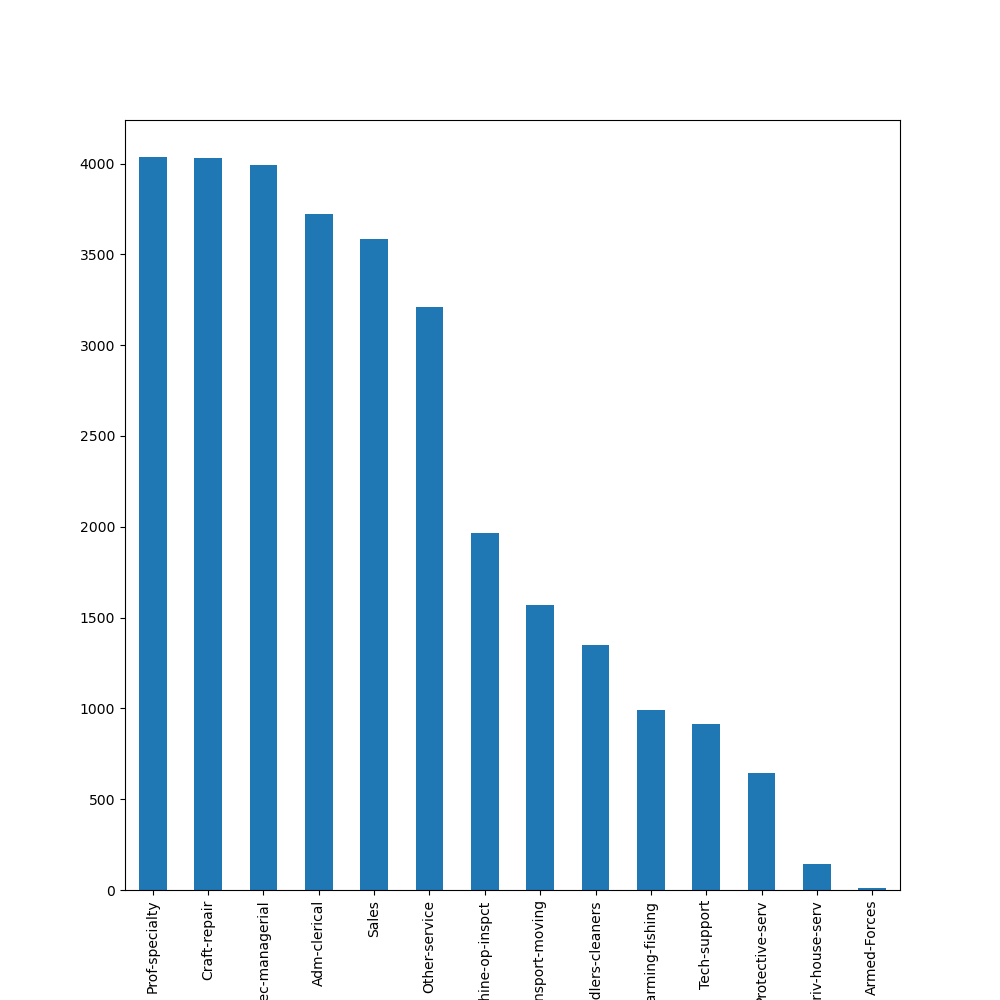
****

****

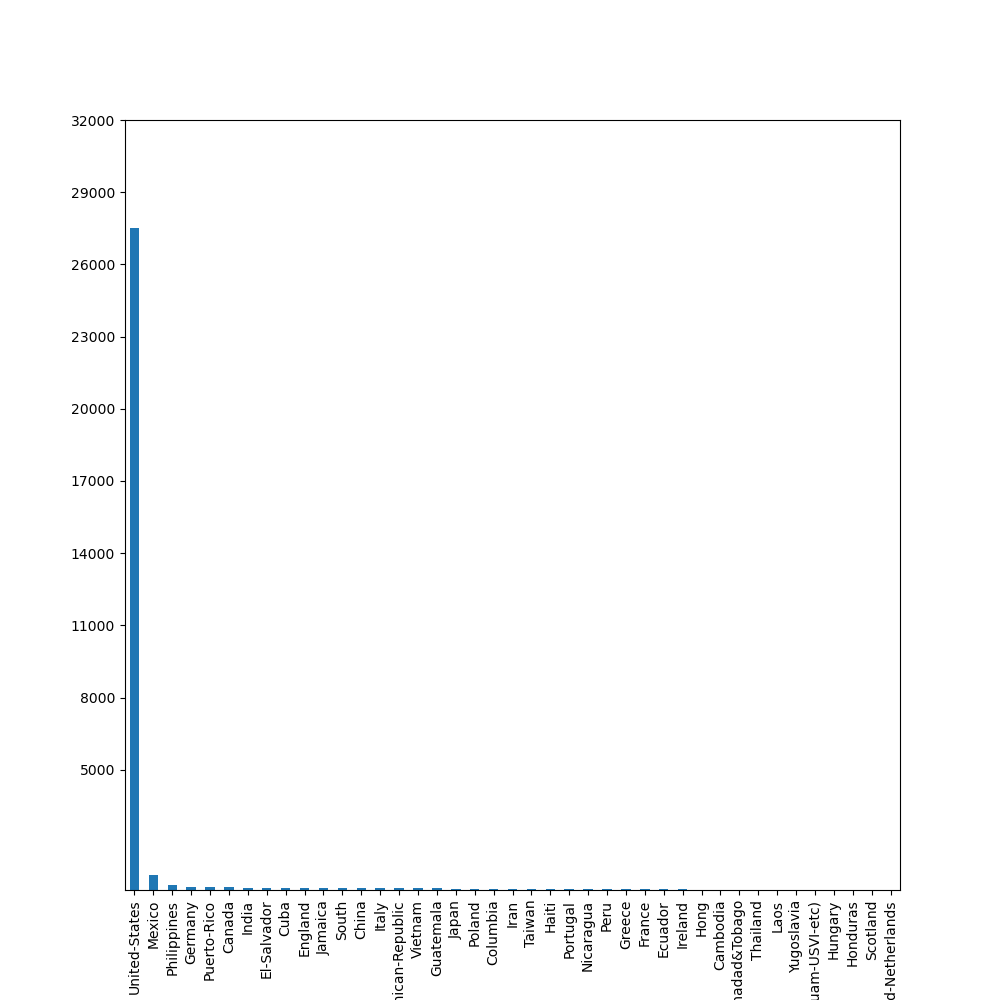
****

**The occupation, work class and native country columns were found with a special character “?” After checking if columns containing “?” are eliminated or not. Furthermore, we have updated those records. After verifying that those columns are updated, we have visualized the columns by the names –**

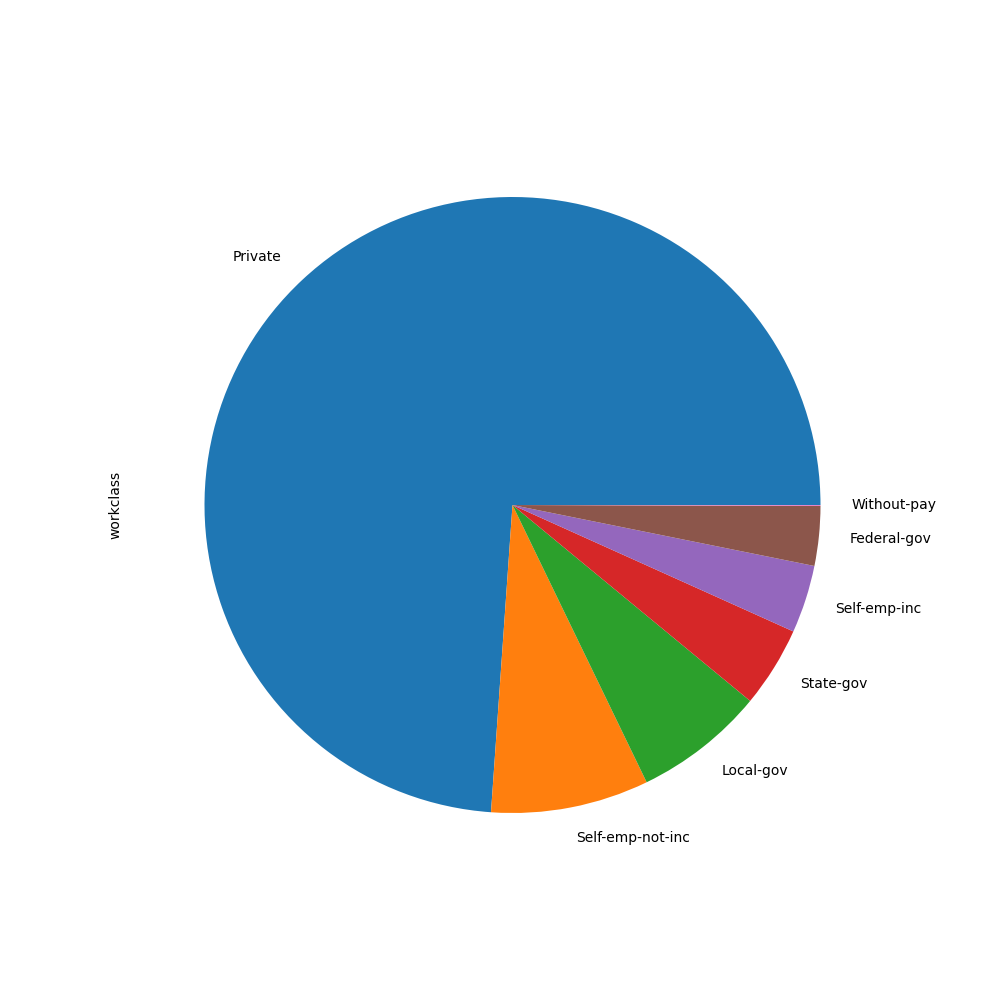
1. **Updated occupation**

****

1. **Updated Native Country**

****

1. **Updated Work Class**

****

**The Model Summary Table is visualized as follows:**

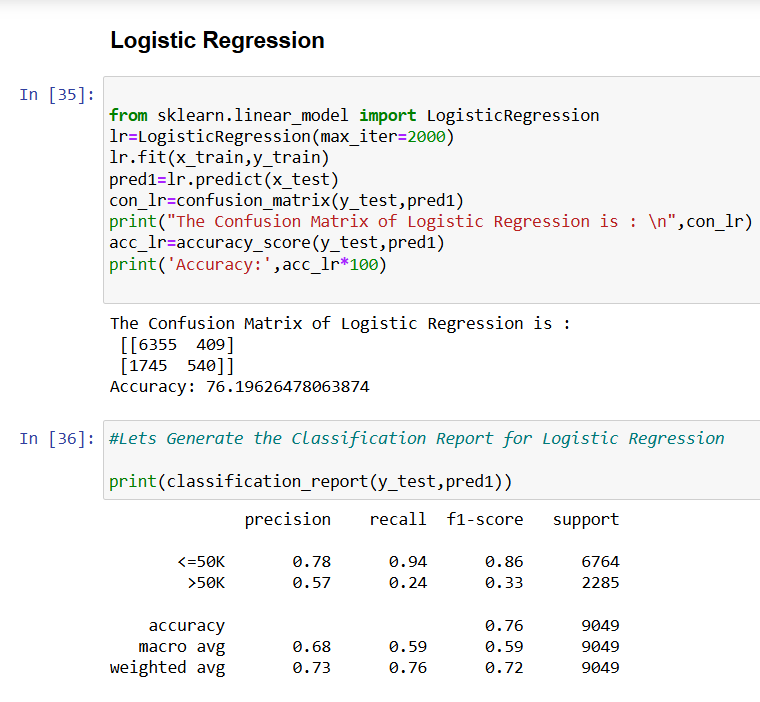
|  |  |
| --- | --- |
| **MODEL** | **ACCURACY SCORE** |
| **Logistic Regression** | **76.19** |
| **Random Forest Classifier** | **80.59** |
| **Decision Tree Classifier** | **76.69** |
| **Support Vector Machines** | **75.10** |
| **K – Nearest Neighbor** | **80.10** |

1. **ALGORITHMS –**

Machine Learning algorithms classify the data based on models which have been developed and make predictions based on these models. It is a piece of code that help people explore analyze and find meaning in complex data sets.

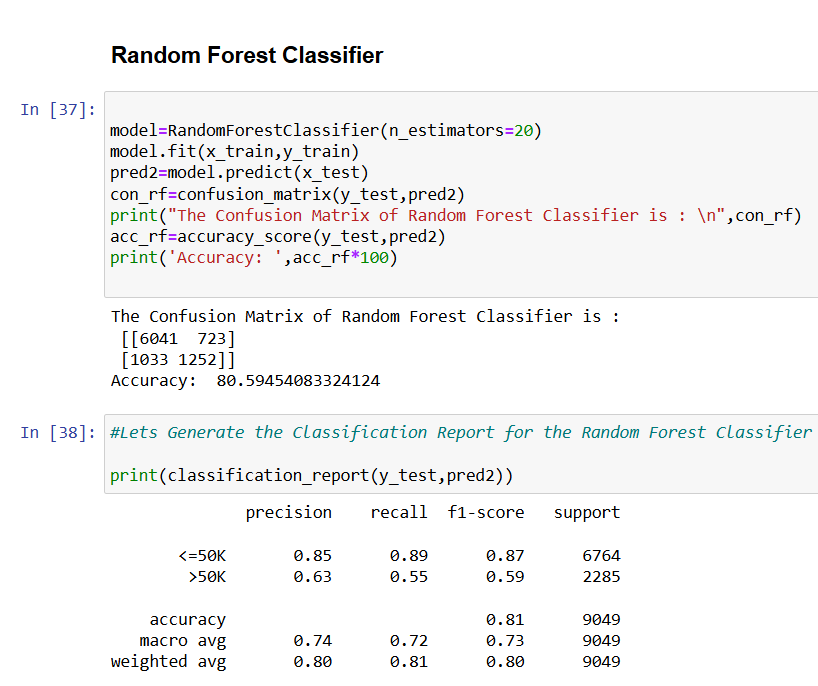
* 1. **LOGISTIC REGRESSION** –

Logistic regression is a statistical supervised learning algorithm used for classification tasks where a goal is to predict the probability that an instance belongs to a given class or not. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The dependent variable must be categorical in nature and the independent variable should not have multi co-linearity. Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. Logistic regression measures how relevant or appropriate an independent/predictor variable is (coefficient size) and also reveals the direction of their relationship or association (positive or negative).



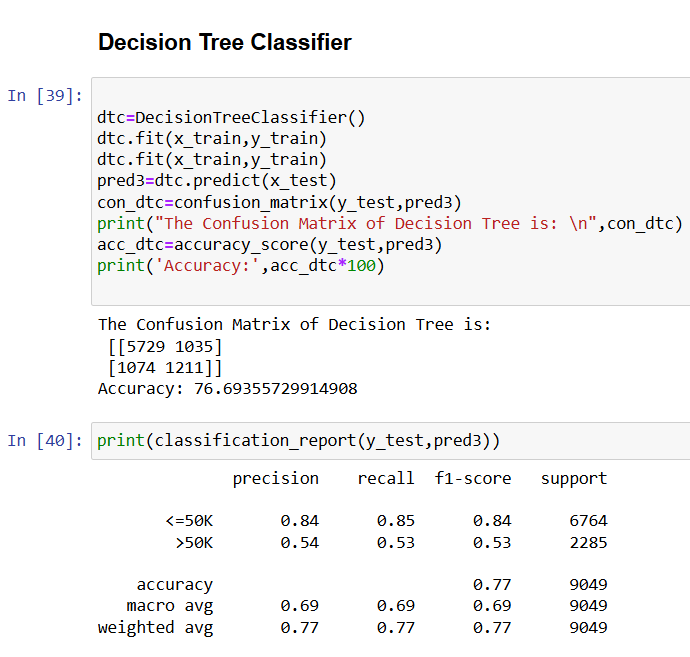
* 1. **RANDOM FOREST CLASSIFIER** –

A Random Forest Classifier is a collection of decision trees that work together to make predictions. Random Forest builds multiple decision trees using random samples of the data. Each tree is trained on a different subset of the data which makes each tree unique. When creating each tree the algorithm randomly selects a subset of features or variables to split the data rather than using all available features at a time. This adds diversity to the trees. Each decision tree in the forest makes a prediction based on the data it was trained on. When making final prediction random forest combines the results from all the trees. Random forest is capable of handling large datasets with high dimensionality. It also enhances the accuracy of model and prevents the overfitting issue. Random forest is a machine learning algorithm that creates an ensemble of multiple decision trees to reach a singular, more accurate prediction or result.



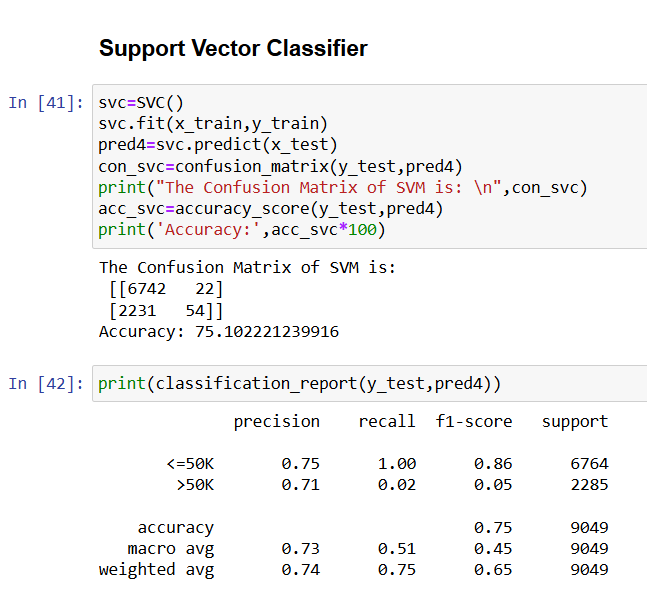
* 1. **DECISION TREE CLASSIFIER** –

A Decision Tree Classifier is a type of non-parametric algorithm that uses a tree-like structure to classify instances based on their feature variables. It is a supervised learning algorithm that is used for classification and regression modeling. A decision tree is a graphical representation of different options for solving a problem and show how different factors are related. It has a hierarchical tree structure starts with one main question at the top called a node which further branches out into different possible outcomes. They are straightforward and easy to understand and can be visualized like a flow-chart. It is capable of capturing non-linear relationships between features and target variables. CART algorithm which stands for Classification and Regression Tree algorithm is used in order to build a tree. It can contain categorical data as well as numerical data. Decision trees consist of root nodes, branches, leaf nodes and internal nodes. Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree.



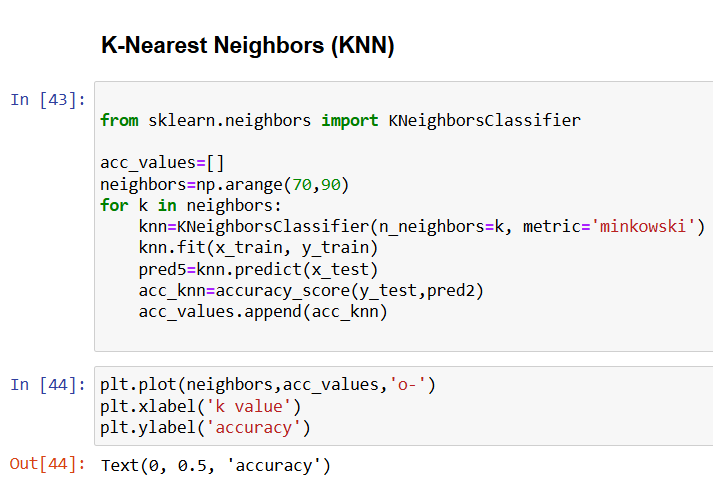
* 1. **SUPPORT VECTOR MACHINES** –

Support Vector Machine is a supervised machine learning algorithm used for both linear and non-linear classification, as well as outlier detection and regression. The primary objective of the SVM algorithm is to identify the optimal hyperplane in an N-dimensional space that can effectively separate data points into different classes in the feature space. It ensures that the margin between the closest points of different classes, known as support vectors is maximized. SVMs are particularly effective because they focus on finding the maximum separating hyperplane between the different classes in the target feature, making them robust for both binary and multiclass classification. They are effective in high dimensional spaces and still effective in cases where number of dimensions is greater than the number of samples. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. SVMs can work well even when the training dataset is too small. It can implicitly handle non-linear separable data by using kernel functions. They are memory efficient and less sensitive to outliers.



* 1. **K-NEAREST NEIGHBORS** –

The K-Nearest Neighbour algorithm is based on supervised learning technique. It is a non-parametric algorithm which does not make any assumption on underlying data. In the **k-Nearest Neighbours (k-NN)** algorithm **k** is just a number that tells the algorithm how many nearby points (neighbours) to look at when it makes a decision. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. K-NN, or the k-nearest neighbour algorithm, is a machine learning algorithm that uses proximity to compare one data point with a set of data it was trained on and has memorized to make predictions. KNN is easy to implement because of how simple and accurate it is. It only needs a few hyperparameters – a k value and a distance metric. Additionally, the KNN algorithm requires no training time because it stores training data and its computational power is used only when making predictions.



1. **CHALLENGES AND OPPORTUNITIES –**

**8.1 CHALLENGES –**

The first challenge is ensuring a higher data quality while determining the most appropriate model amongst several choices, and also balancing the complexities of the model are all significant challenges. These issues can skew results and lead to inaccurate predictions. For instance, if a dataset contains erroneous salary figures or incomplete job descriptions, it can significantly affect the model's performance. Addressing and tackling the dataset biases and imbalances, specifically those related to salary distribution, adds to the complexity. Ensuring that the model generalizes well to unseen data is another significant challenge. A model that performs well on training data may fail to predict salaries accurately for new job postings due to differences in the underlying data distribution Additionally, navigating deployment complexities across multiple platforms is important. Finally, the design of the dashboard itself poses challenges. It must present complex data in an intuitive and user-friendly manner, allowing users to easily interpret the predictions and insights provided.  
  
**8.2 OPPORTUNITIES –**

Advanced algorithms like ensemble methods and deep learning can improve accuracy. By utilizing salary prediction dashboards, companies can refine their recruitment strategies. These tools can analyze job market trends and provide insights into competitive salary ranges, helping organizations attract top talent while remaining budget-conscious. Modern salary prediction dashboards can be tailored to specific industries or job roles, providing more accurate predictions. Using domain knowledge for feature engineering can help capture subtle influences on predictions. Including external datasets enriches the models. Enhancing model interpretability increases user trust. Setting up feedback loops for continuous improvement and working with industry experts encourage innovation and learning. Employers can leverage salary prediction dashboards to make informed decisions regarding compensation packages. By analyzing historical salary data and trends, organizations can adjust their pay structures to remain competitive and equitable, thus improving employee satisfaction and retention.

1. **RISK VS REWARD –**

A key component of our project development of the Salary Prediction Dashboard involves finding a balance between risk and return. The **risk/reward ratio** is a critical metric that can be applied to salary predictions. It compares the expected benefits of accurate salary predictions against the potential risks involved in relying on those predictions.

* 1. **RISKS –**
  2. **Data Quality -** There are significant risks associated with issues like minimizing biases in the dataset, choosing the best models, and guaranteeing data quality. Poor data can lead to misleading predictions, which can have financial implications for businesses and individuals alike.
  3. **Model Selection**: Different machine learning models have varying levels of accuracy and complexity. Choosing a model that is too simplistic may not capture the nuances of salary determinants, while overly complex models may overfit the data.
  4. **Market Variability**: Salaries can fluctuate based on market conditions, industry trends, and economic factors. A model that does not account for these variables may produce unreliable predictions.
  5. **REWARDS –**

1. **Informed Decision-Making**: A well-designed salary prediction dashboard can provide valuable insights, helping employers set competitive salaries and aiding employees in salary negotiations.
2. **Efficiency**: Automating salary predictions can save time and resources for HR departments, allowing them to focus on strategic initiatives rather than manual calculations.
3. **Competitive Advantage**: Organizations that leverage predictive analytics can gain a competitive edge in attracting and retaining talent by offering data-driven compensation packages.

Overcoming these obstacles, however, offers encouraging benefits. While using domain knowledge for feature engineering captures subtle implications on predictions, investigating more complex algorithms such as ensemble methods and deep learning may result in increased accuracy. The models are improved by incorporating external datasets, and improving interpretability builds user trust—a critical factor in HR decision-making. Establishing feedback loops and working with industry experts guarantee innovation and continual progress despite deployment challenges. Significant benefits, such as improved forecasting capacities and empowered decision-making in hiring procedures, are promised by effective risk management.

1. **REFLECTIONS ON THE INTERNSHIP –**

When I think back on my internship experience, it has been a life-changing adventure full with invaluable knowledge and opportunities for growth and improvement. As I've been working on the project, I've learned more about machine learning and data analytics techniques, especially as they relate to creating the Salary Prediction Dashboard. My tenacity and problem-solving abilities have improved greatly as a result of the difficulties I've faced. My educational experience has been enhanced by working with classmates and industry professionals, who have provided a variety of viewpoints and ideas. I've been able to apply my academic knowledge to real-world situations thanks to the internship, which has improved my confidence and practical skills. When I look forward to the future, I am appreciative of the chance to work on worthwhile initiatives and eager to keep developing my data-driven decision-making abilities. In addition to giving me technical know-how, this internship has stoked my curiosity and driven me to keep learning and developing in this ever-evolving area.

1. **RECOMMENDATIONS –**

To enhance our project, we suggest focusing on particular sectors or employment roles in order to improve our project and tailoring wage projections appropriately. It is crucial to put sophisticated data handling strategies into practice, especially when it comes to more skilfully handling outliers and missing numbers. The accuracy of our model could be considerably increased by experimenting with different algorithms and adjusting their parameters. By implementing these tactics, we hope to create a Salary Prediction Dashboard that is more accurate and useful, better meeting the particular requirements and subtleties of the target market or job positions.

1. **OUTCOME / CONCLUSION –**

As we arrive at the project report milestone, substantial progress has been achieved in Interim Project 3. I was successfully able to create a dataset, clean the dataset and sanitize it. I have also trained the dataset and predicted the salary based on the training set. I have improved the predictions in order to make it accurate with the help of hyperparameter tuning of the best model i.e. random forest classifier model. I made various charts and plots to visualize the data. I incorporated various machine learning algorithms and also used tables to summarize the data. I used flask to create a dashboard using vs code. This internship project allowed me to learn about the process, familiarize myself with new concepts. This phase has been enlightening, deepening my understanding across various domains and enhancing proficiency in project-related topics. The journey thus far has instilled confidence in our ability to meet project objectives. Looking ahead, I am eager to continue this momentum and contribute further to the project's success. I’m extremely grateful for all the learning opportunities given to me by TCS iON.

1. **ENHANCEMENT SCOPE**

The proposed enhancements bring a substantial improvement to the quality and effectiveness of our project, particularly the Salary Prediction Dashboard. Exploring advanced algorithms such as ensemble methods and deep learning holds the promise of significantly enhancing the accuracy and reliability of salary estimations. Leveraging domain-specific knowledge for feature engineering ensures a more nuanced understanding of the factors influencing predictions, thereby refining the dashboard's predictive capabilities to align closely with real-world salary expectations. Integrating external datasets will enrich our models with comprehensive information, further enhancing accuracy. Moreover, prioritizing model interpretability fosters trust among HR professionals, crucial for informed decision-making during recruitment processes. Establishing feedback loops and collaborating with industry experts not only promote innovation but also ensure continuous improvement of our solution, making it a vital asset in HR management. These enhancements position our project as a robust and indispensable tool for empowering HR managers with data-driven insights in salary negotiations and talent acquisition.

1. **CORRECTNESS OF SOLUTION**

The solution aligns closely with the specified requirements outlined in the project scope. By focusing on developing the Salary Prediction Dashboard using data analytics and machine learning methodologies, I address the primary objective of the project effectively. The dashboard's predictive capabilities, which consider variables such as experience, age, and qualifications to forecast job candidates' salaries, directly meet the project's goals.

The inclusion of advanced algorithms, meticulous data preprocessing methodologies, and thorough dataset verification processes ensures the accuracy and reliability of the solution. These measures adhere to the project requirements and demonstrate a commitment to delivering a high-quality product.

1. **LINK TO CODE AND EXECUTABLE FILE –**

**GOOGLE COLAB FILE LINK –** [**https://colab.research.google.com/drive/17cplBHi5Z1b9KkkVSydFnciqBr1UEHMI**](https://colab.research.google.com/drive/17cplBHi5Z1b9KkkVSydFnciqBr1UEHMI)

**GITHUB REPOSITORY LINK –**

[**https://github.com/KhushiUpadhye01/Salary-Prediction-Dashboard-for-HRs**](https://github.com/KhushiUpadhye01/Salary-Prediction-Dashboard-for-HRs)