**Texas Salary Prediction**

**Team ID**: - PTID-CDS-MAR-24-1874

**Project** ID: - PRCP-1024

**Team Member**: - Mihir Prajapati, Ishika Khutate

**Objective**: - Create a detailed data analysis report summarizing key insights, trends, outliers in salaries, wage disparities across departments/roles, and changes in salaries/compensations over time, alongside developing a predictive model to forecast Texas state employee payroll, providing actionable recommendations for effective payroll management and policy decisions.

**Abstract:** - This project aimed to develop a machine learning model for predicting salaries based on various features and employment data. The project began with an exploratory data analysis (EDA) process to understand the characteristics of the salary dataset. Various data preprocessing techniques, including handling missing values, scaling features, and encoding categorical variables, were applied to prepare the data for regression modeling.

**There are three main phases.**

1. Exploratory Data Analysis (EDA)
2. Data Preprocessing
3. Model Creation

**Exploratory Data Analysis (EDA):**

The EDA phase involved exploring the Texas Salary dataset to understand the distribution of salary values, identify any outliers or missing values, and analyze the relationship such as, Relationship between Gender and Annual Salary, Relationship between Ethnicity and Hourly Rate, Relationship between Status and Annual Salary.

Employment Trends Over Time - The "Employment Trends Over Time" analysis provides a visual representation of how the number of employees has changed over different years. It helps identify trends, spikes, or dips in employment levels, offering insights into hiring patterns and workforce changes over time within the dataset.

The EDA phase laid the foundation for further data preprocessing, feature engineering, and model building steps in the Texas Salary Prediction project. It provided crucial insights into the dataset's characteristics and guided subsequent analysis and modeling decisions.

**Data Preprocessing:**

Following the EDA phase, the dataset underwent rigorous preprocessing steps to prepare it for model training. Key preprocessing steps included:

Handling missing values by either imputation, removal, or other suitable techniques based on the nature of the missing data. Encoding categorical variables using techniques like label encoding to convert categorical data into numerical format. Scaling numerical features to a common scale using techniques such as standardization to ensure that features contribute equally to the model training process. Feature engineering, which may involve Feature Selection.

**Model Creation:**

Two separate datasets were created: one without the monthly salary column and another with the monthly salary column included.

The following machine learning models were trained and evaluated on both datasets: Artificial Neural Network (ANN), KNeighborsRegressor, LinearRegression, XGBRegressor, GradientBoostingRegressor, DecisionTreeRegressor, RandomForestRegressor. Each model was trained on the pre-processed dataset and evaluated using appropriate metrics such as Mean Absolute Error (MAE) and R-squared to measure predictive accuracy and model performance. The results showed notable differences in model performance between the dataset with monthly salary information and the dataset without it.

**Model Performance:**

**A) With Monthly Salary Column:**

**Linear Regression:**

- Mean Absolute Error: (4.672909406127804e-10)

- R-squared: 1.0

- Linear Regression performs exceptionally well with almost zero error and perfect R-squared, indicating a perfect fit to the data.

**Decision Tree Regressor:**

- Mean Absolute Error: 2.22

- R-squared: 0.99999

- Decision Tree Regressor shows very low error and nearly perfect R-squared, indicating excellent predictive performance and ability to capture data patterns.

**Random Forest Regressor:**

- Mean Absolute Error: 3.18

- R-squared: 0.99995

- Random Forest Regressor performs well with low error and high R-squared, showcasing its ability to handle complex data and provide accurate predictions.

**Gradient Boosting Regressor:**

- Mean Absolute Error: 138.57

- R-squared: 0.99991

- Gradient Boosting Regressor exhibits higher error compared to some models but still maintains a high R-squared, indicating strong predictive power and capability to explain data variability.

**XGBoost Regressor:**

- Mean Absolute Error: 220.12

- R-squared: 0.99096

- XGBoost Regressor shows good predictive ability with a relatively higher error compared to other models, but still maintains a high R-squared indicating a good fit to the data.

**K-Nearest Neighbors (KNN) Regressor:**

- Mean Absolute Error: 303.45

- R-squared: 0.99738

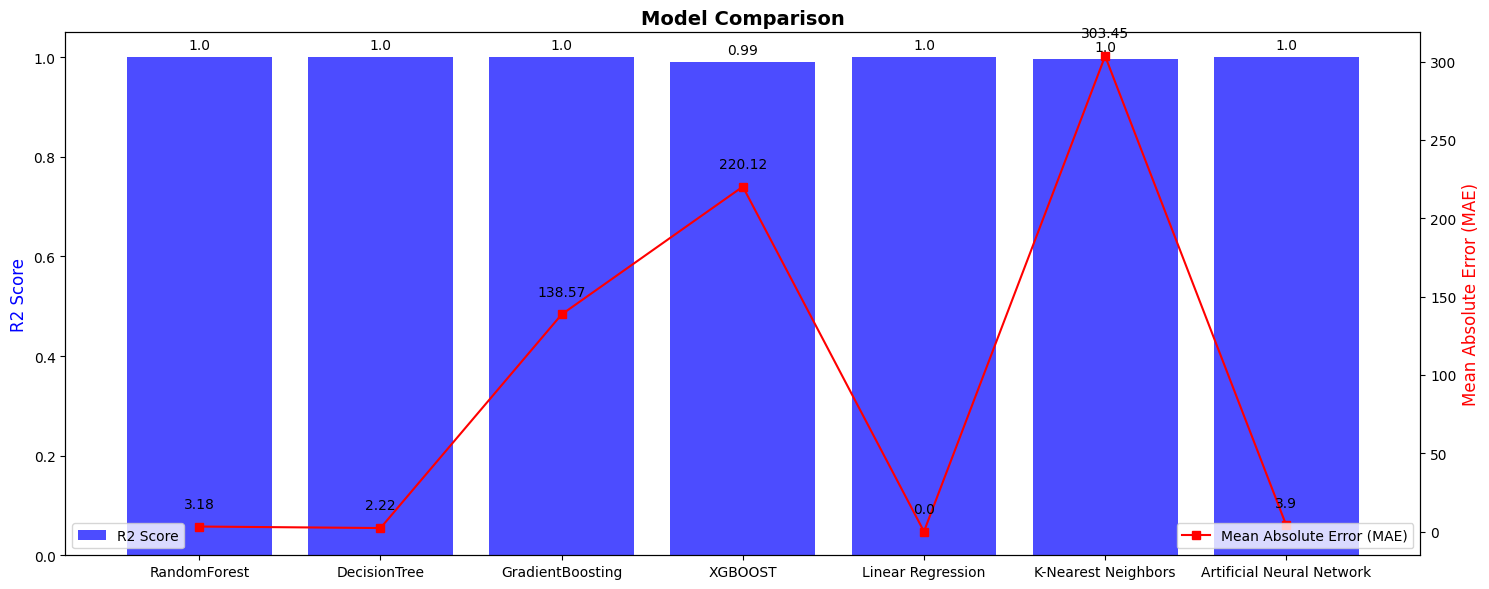
- KNN Regressor exhibits higher error compared to most models but still maintains a relatively high R-squared, indicating decent predictive performance but not as accurate as some other models.

**Artificial Neural Network (ANN):**

- Mean Absolute Error: 3.90

- R-squared: 0.9999998

- The ANN shows a moderate error and very high R-squared, suggesting strong predictive capability and good data fitting, though not as precise as Linear Regression.

**-** Linear Regression and Decision Tree Regressor stand out with very low error and high R-squared values, indicating excellent performance and model fit to the data. Random Forest, Gradient Boosting, and XGBoost also perform well, with slightly higher error but still maintaining high R-squared values. KNN Regressor shows relatively higher error, and ANN performs moderately well with a good fit to the data.

**B) Without Monthly Salary Column:**

**Linear Regression:**

- Mean Absolute Error: 15848.46

- R-squared: 0.1171

- Linear Regression performs poorly compared to other models, with a high error and a low R-squared value, indicating a weak fit to the data.

**Decision Tree Regressor:**

- Mean Absolute Error: 3607.84

- R-squared: 0.9215

- Decision Tree Regressor shows relatively low error and high R-squared value, indicating good predictive performance and a strong fit to the data.

**Random Forest Regressor:**

- Mean Absolute Error: 3583.93

- R-squared: 0.9237

- Random Forest Regressor performs similarly to the Decision Tree model with slightly lower error but maintains a high R-squared value, indicating accurate predictions and a strong fit to the data.

**Gradient Boosting Regressor:**

- Mean Absolute Error: 8015.94

- R-squared: 0.7046

- Gradient Boosting Regressor exhibits higher error compared to Decision Tree and Random Forest models, but still maintains a decent R-squared value, suggesting good predictive power and moderate fit to the data.

**XGBoost Regressor:**

- Mean Absolute Error: 4603.27

- R-squared: 0.8760

- XGBoost Regressor shows lower error compared to Gradient Boosting but slightly higher compared to Decision Tree and Random Forest, with a good R-squared value indicating strong predictive ability and a reasonable fit to the data.

**K-Nearest Neighbors (KNN) Regressor:**

- Mean Absolute Error: 4160.62

- R-squared: 0.8859

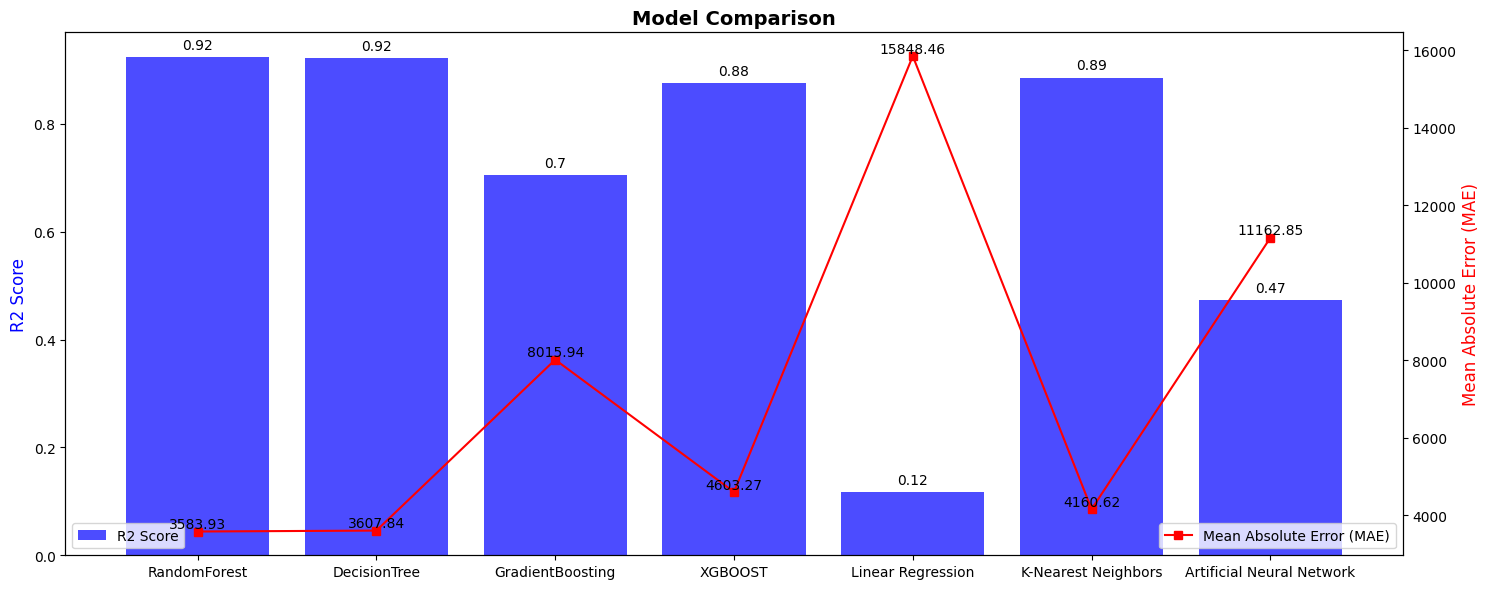
- KNN Regressor performs well with a moderate error and a high R-squared value, indicating good predictive performance and a relatively strong fit to the data.

**Artificial Neural Network (ANN):**

- Mean Absolute Error: 10700.00

- R-squared: 0.5194

- The ANN exhibits the highest error among the models considered, with a moderate R-squared value, indicating less accurate predictions and a weaker fit to the data compared to other models.

**-** Decision Tree, Random Forest, and XGBoost models show good performance with lower error rates and higher R-squared values, indicating strong predictive ability and good fit to the data. Linear Regression and ANN models perform comparatively poorer in terms of predictive accuracy and fit to the data.

**Conclusion: -**

The Texas Salary Prediction project was marked by several challenges that required careful consideration and strategic approaches throughout the analysis and modeling phases. One significant challenge involved managing a dataset with numerous missing values across multiple columns, requiring thoughtful imputation techniques or considering the removal of affected rows/columns while preserving data integrity. Another key challenge was feature-selection, given the dataset's extensive feature set. Identifying the most relevant predictors for salary prediction while balancing model complexity and predictive performance was crucial in developing robust models.

Additionally, addressing highly correlated features, especially monthly salaries, posed challenges in model interpretability and potential multicollinearity issues. Strategies such as feature engineering and selection were employed to mitigate these challenges and enhance model performance.

Certainly! When the monthly and annual columns are highly correlated and the monthly data is excluded due to redundancy, it simplifies the model but can lead to a decrease in accuracy. This is because important nuances captured by the monthly data, which could enhance the model's predictive power, are now ignored. It's a trade-off between model simplicity and capturing detailed information for better accuracy.

The Texas Salary Prediction project successfully developed and evaluated multiple machine learning models for predicting salary based on relevant features. The presence of the monthly salary column significantly influenced model performance, as observed by the differences in metrics between the two datasets.

Overall, the project demonstrates the effectiveness of machine learning in predicting salaries and highlights the importance of feature selection and dataset composition in model outcomes.