Predicting Employee Salaries

Youssef Mustafa Sayed, Omar Ashraf Zakaria,

Tarek Muhammed, Bavly Soliman and Hassan Muhammed

Predicting Employee Salaries

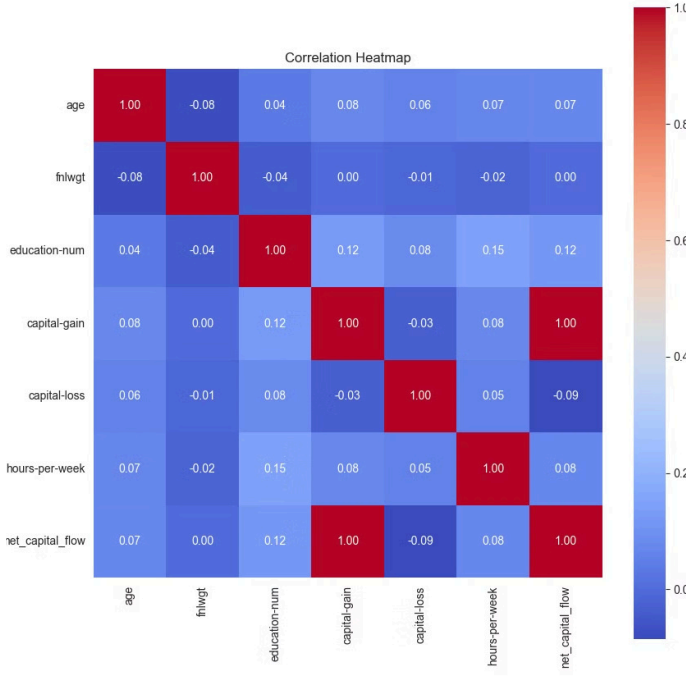
The Goal

This project aims to predict whether employees earn more or less than $50K annually.

The Data

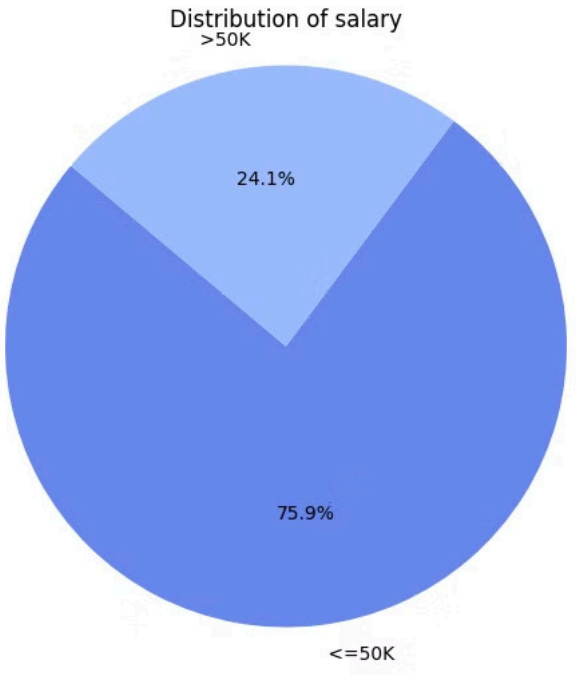
- The data set includes information such as age, education level, experience, job title, and department.
- Uncleaned dataset with many categorical data and not normally distributed
- The missing values were filled by '?'

Visualized EDA :



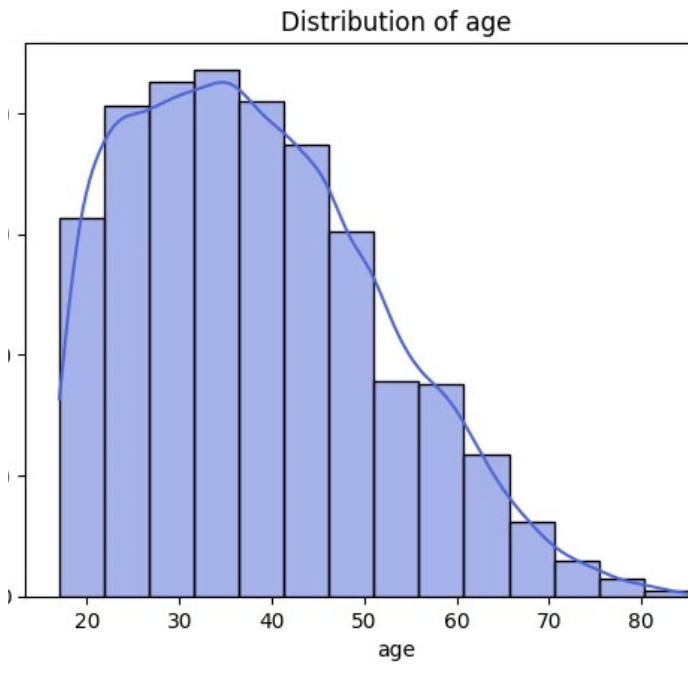
Co-relation Heatmap

To Describe the co-relation between features



Pie Plot

To show the percentage of the distribution of the salary , and it showing clear in-balance problem



Distribution Plot

to show the distribution of the age feature , and it shoes clear right skewing

Phase 1: Data Cleaning and Preprocessing

Handling Missing Values

We addressed missing data points by either removing incomplete rows or imputing values based on statistical techniques.

Outlier Detection and Treatment

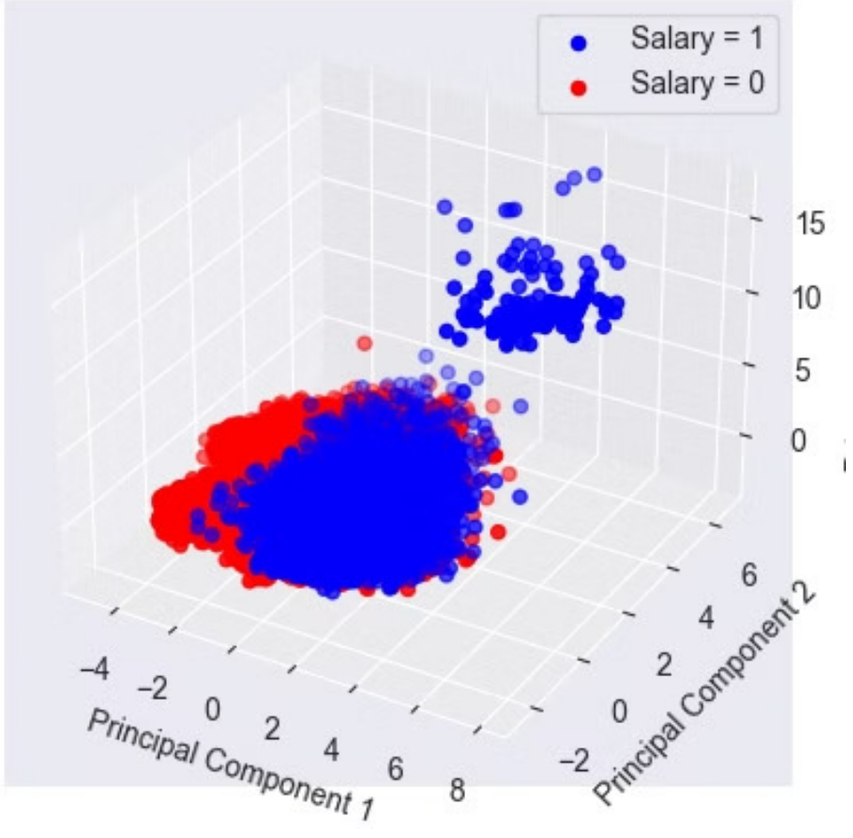
We identified and dealt with outliers, which are extreme data points that can skew the model's accuracy.

Feature Scaling and Encoding

We standardized numerical features and converted categorical data to a numerical format for model compatibility.

Phase 2: Implementing Support Vector Machines (SVM) from Scratch

Plotting the dataset to check if it is Linear Separable



And as it clear from the 3D plot it is not Linear separable

Phase 2: Implementing Support Vector Machines (SVM) from Scratch

Define Kernel

Choose a kernel (linear, polynomial, RBF).

Optimize Hyperparameters

Tune C and kernel parameters.

Train the Model

Fit the SVM to training data.

Make Predictions

Classify new data points.

Phase 3: Model Training and Evaluation

Model Selection

We explored a range of machine learning models, including Logistic Regression, Decision Trees,XGBoost and Random Forests, to find the best fit for our data.

Hyperparameter Tuning

We fine-tuned the hyperparameters of each model to optimize their performance.

Model Evaluation

We evaluated the models using metrics accuracy score to measure their effectiveness.

Random Forest & XGboost Selection

Random Forest Classifier and XGboost emerged as the top-performing models, achieving the highest accuracy between 97% and 98% on the test data and 100% on the training dataset without overfitting and providing robust predictions.