**Project Title** - Customer Insurance Purchases Case Study  
**Name** – S G Sumanth  
**Date** - 5/03/2025  
  
**Abstract-  
Business Goal:** In the role of an Analyst at a Bank Insurance Company, the objective is to leverage a dataset containing user details, excluding personal data such as passwords and account numbers. The focus is on attributes like age and estimated salary. The mission is to develop an AI model capable of predicting whether new customers will purchase insurance based on their age and estimated salary.

**Problem Statement**: The primary challenge is to employ various classification algorithms to construct a comprehensive comparative analysis. The goal is to assess and contrast the performance of these machine learning algorithms. By conducting this study, the aim is to extract valuable insights from the results. The ultimate objective is to compile accuracy metrics for each algorithm and determine the most suitable one for classification. It's essential to identify an algorithm that strikes a balance between precision and generalisation, ensuring it fits the given data without overfitting.

**Approach:**

**Data Preparation**: Begin by preprocessing the provided dataset, ensuring it is ready for training and evaluation. Perform necessary transformations and feature engineering.

**Algorithm Selection**: Choose a variety of classification algorithms, such as Decision Trees, Random Forest, Support Vector Machines (SVM), Logistic Regression, and Neural Networks.

**Model Development**: Implement each chosen algorithm using the preprocessed data. Tune hyperparameters as needed.

**Training and Evaluation:** Split the data into training and testing sets. Train each model on the training set and evaluate their performance on the testing set. Measure accuracy, precision, recall, F1-score, and other relevant metrics.

**Compare and Contrast**: Create a comprehensive comparative analysis of the algorithms' performance. Consider aspects like accuracy, model complexity, training time, and potential for overfitting.

**Insights Generation**: Extract insights from the comparative analysis. Identify patterns and trends in algorithm performance. Understand the trade-offs between accuracy and generalization.

**Select Optimal Algorithm**: Based on the insights gained, select the algorithm that best suits the given data and problem. Emphasize an algorithm that demonstrates high accuracy without compromising on overfitting.

**Results Presentation**: Tabulate the accuracy metrics for each learner in a clear and organized manner. Highlight the strengths and weaknesses of each algorithm.

**Conclusion**: By conducting a thorough compare and contrast study on various classification algorithms, this project aims to identify the optimal model for predicting customer insurance purchases accurately. The final choice of algorithm should provide a balance between accuracy and generalization, ensuring it is well-suited for the provided dataset. This approach aligns with the business goal of enhancing the company's ability to predict customer behavior and make informed decisions in the insurance domain.  
  
**Table Of Contents** -   
  
**1. Introduction-**

The project focuses on Random Forest Classification Algorithm Technique, an ensemble learning technique used for predictive modeling. The objective is to predict whether a user will purchase a product based on their age and estimated salary from the given csv file. The project highlights the importance of machine learning in customer behavior prediction, particularly in targeted advertising and recommendation systems.

**Problem Addressed**: Predicting customer purchase behavior based on demographic features. Improving classification accuracy using ensemble learning.

**Techniques Used:**

Random Forest Classifier for classification. Data Preprocessing including missing value imputation and feature scaling. Performance Evaluation using accuracy metrics.

**2. Literature Review-**

In previous research findings, I have seen that Random Forest Classifiers has demonstrated the effectiveness in handling classification problems with non-linear decision boundaries. The model's strength lies in reducing overfitting while maintaining interpretability which further enhances model’s reliability. Advancements in feature selection, hyperparameter tuning, and ensemble methods have further improved classification accuracy. (future recommendations are open for improvements in the project).

**3. Problem Statement-**

**Objective:** Predict whether a user will purchase a product based on demographic information.

**Assumptions and Limitations:**

The dataset is representative of real-world consumer behavior. Limited feature set (only age and salary) may impact predictive performance.

**4. Data Collection and Preprocessing**

**Dataset Used:** Social\_Network\_Ads.csv

**Features**: Age, EstimatedSalary

**Target Variable**: Purchased

**Preprocessing Steps**:

* Handling missing values: Filled missing salaries with median values.
* Feature Scaling: Standardization using StandardScaler.
* Data Splitting: 80% training, 20% testing using train\_test\_split.

**5. Methodology**

**AI Techniques Used:**

Random Forest Classifier (Ensemble of Decision Trees)

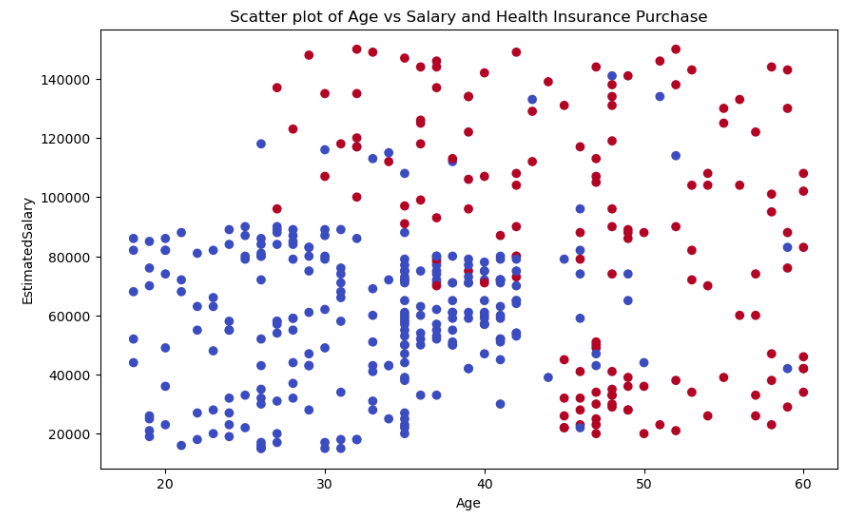
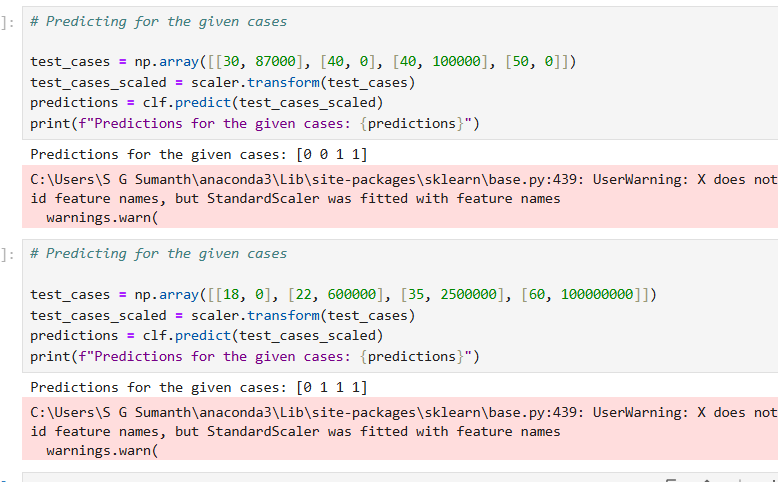
Train-Test Split (to prevent overfitting)

Feature Scaling (StandardScaler for normalization)

**Rationale:**

Random Forest mitigates overfitting by averaging multiple decision trees. It handles both categorical and numerical features effectively. Ensures robustness against missing values and noisy data as far we have seen.

**Results:**

Accuracy Score- 0.9 (90%)  
Visualization- **Graphical Analysis and Predictions-**

Here 1 represents Purchase has been initiated and O represents Purchase has not been done.  
  
According to the hypothesis and assumptions proposed, younger individuals with higher salaries are more likely to purchase health insurance.

Older individuals with higher salaries might be less inclined to purchase health insurance.

Salary might have a stronger impact on insurance purchasing behavior than age.

**Key Learnings From This Study-**  
The study demonstrated how ensemble learning enhances predictive accuracy by reducing overfitting and handling non-linear decision boundaries. Visualization techniques helped in understanding patterns, which is crucial for real-world AI applications.  
  
**Real-Life Applications** like Predicting Loan Default Risk in Banks and Fraud Detection in Online Transactions can be performed.

**References-**

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.