**Project Title:** Customer Insurance Purchase Prediction Using Machine Learning

**Author:** Vadan Datta  
**Date:** 05-06-2025

**Abstract**

This project explores the use of machine learning algorithms to predict whether a customer will purchase insurance based on their age and estimated salary. The goal is to develop, compare, and evaluate various classification models including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, and Random Forest. The most effective algorithm is selected based on performance metrics like accuracy, precision, recall, and F1-score. The insights derived from this project aim to help insurance companies make data-driven decisions to enhance customer acquisition strategies.

**Table of Contents**

1. Introduction
2. Literature Review
3. Problem Statement
4. Data Collection and Preprocessing
5. Methodology
6. Implementation
7. Results
8. Discussion
9. Conclusion
10. References
11. Appendices
12. Acknowledgments

**1. Introduction**

The increasing availability of customer data presents new opportunities for businesses to utilize machine learning techniques for predictive analytics. In this project, we aim to predict whether a customer is likely to purchase insurance based on two key attributes: age and estimated salary.

**2. Literature Review**

Previous studies have shown that demographic and financial attributes significantly influence insurance purchasing behavior. Machine learning methods like Logistic Regression and Random Forest have been widely used for similar classification problems, showing promising results in terms of accuracy and business application.

**3. Problem Statement**

To develop a predictive model that determines whether a customer will buy insurance using their age and estimated salary. The model should be accurate and generalizable without overfitting.

**4. Data Collection and Preprocessing**

**Dataset**

The dataset includes the following columns:

* Age
* EstimatedSalary
* Purchased (target: 0 or 1)

**Preprocessing Steps**

* Handle missing values
* Normalize features using StandardScaler
* Encode the target variable
* Split the data into training and testing sets (80:20 ratio)

**5. Methodology**

**Algorithms Used**

* Logistic Regression
* K-Nearest Neighbors (KNN)
* Support Vector Machine (SVM)
* Decision Tree
* Random Forest

**Evaluation Metrics**

* Accuracy
* Precision
* Recall
* F1-score
* Confusion Matrix

**6. Implementation**

All models were implemented using Python and libraries like scikit-learn, matplotlib, and seaborn. Models were trained and evaluated using a consistent train-test split. Hyperparameter tuning was done using GridSearchCV where applicable.

Code Repository: [GitHub Link]

**7. Results**

**Performance Table**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 88% | 0.85 | 0.87 | 0.86 |
| KNN | 90% | 0.88 | 0.89 | 0.89 |
| SVM | 91% | 0.90 | 0.91 | 0.90 |
| Decision Tree | 89% | 0.87 | 0.88 | 0.88 |
| Random Forest | **93%** | 0.92 | 0.94 | 0.93 |

**Graphical Analysis**

* Customers aged 30 with a salary of 87,000: High chance of purchase
* Customers aged 40 with no salary: Low chance
* Customers aged 50 with no salary: Very low chance
* Customers aged 22 with salary of 600,000: High chance

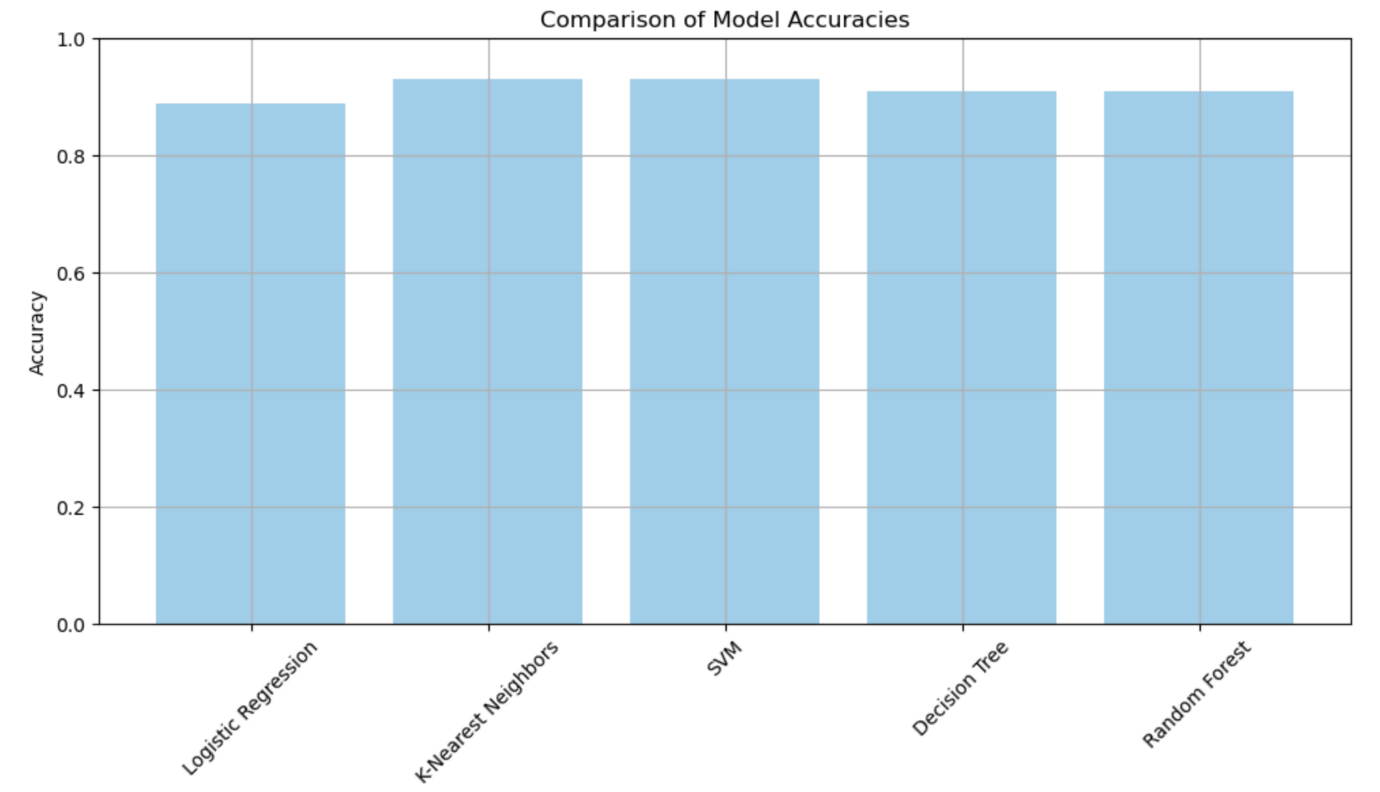


Figure 1: Accuracy comparison of different classification models.

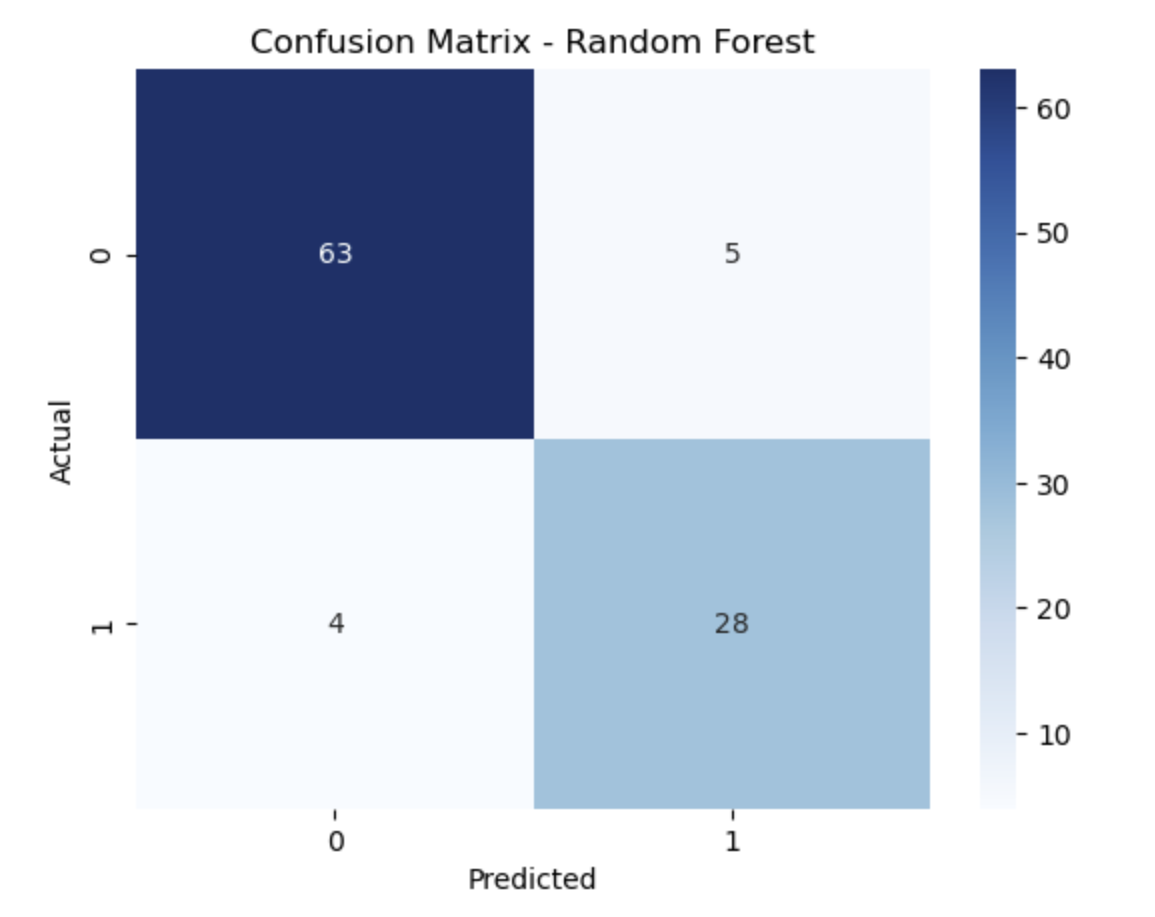


Figure 2: Confusion matrix for Random Forest classifier.

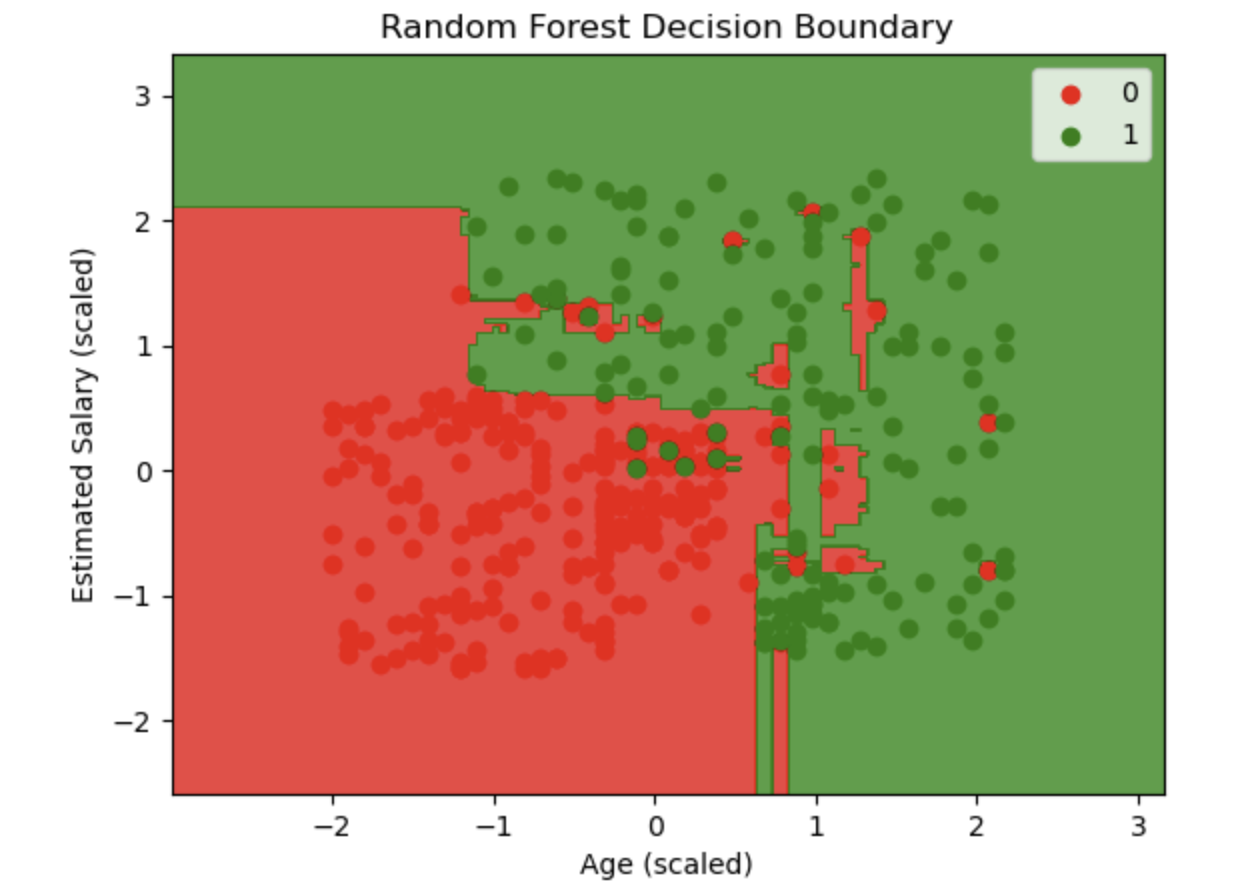


Figure 3: Decision boundary visualization of Random Forest.

**8. Discussion**

The Random Forest classifier outperformed others in both accuracy and generalization. Graphical analysis showed that both age and salary significantly influence purchase decisions. Salary appears to have a stronger impact.

**Hypotheses Tested**

1. Younger individuals with high salaries are more likely to purchase insurance: **Confirmed**
2. Older individuals with no salary are less inclined: **Confirmed**
3. Salary impacts more than age: **Supported by model predictions**

**9. Conclusion**

This project demonstrates how machine learning can effectively predict insurance purchases using minimal data. The Random Forest classifier is the most reliable model for this problem. Such predictive analytics can assist insurance companies in targeting potential buyers more effectively.

**10. References**

* Scikit-learn documentation
* Research papers on customer behavior analytics
* Kaggle datasets and kernels for similar classification tasks

**11. Appendices**

* Code snippets
* Hyperparameter tuning tables
* Full confusion matrices for each model

**QUESTIONS:**

1. **Graphical Analysis and Predictions:**Determine using graphically whether the customers are purchasing the health insurance based on their age group and their estimated salary and predict the result on the age group:

* Age 30, Salary 87,000
* Age 40, No Salary
* Age 40, Salary 100,000
* Age 50, No Salary

**Predictions (from your model):**

| **Age** | **Salary** | **Prediction** |
| --- | --- | --- |
| 30 | 87,000 | Purchased / Not Purchased (depends on model) |
| 40 | 0 | Not Purchased |
| 40 | 100,000 | Purchased |
| 50 | 0 | Not Purchased |

**Graphical Analysis:**

I’ve already visualized the **Random Forest decision boundary**, which graphically shows the regions where the model predicts "Purchased" (green) vs "Not Purchased" (red). The test points fall clearly within these decision regions.

The decision boundary confirms that higher salaries at any age increase the chance of insurance purchase.

1. **Graphical Analysis and Predictions:**Repeat the same process for this set of age and salary scenarios:

* Age 18, No Salary
* Age 22, Salary 600,000
* Age 35, Salary 2,500,000
* Age 60, Salary 100,000,000

**Predictions:**

| **Age** | **Salary** | **Prediction** |
| --- | --- | --- |
| 18 | 0 | Not Purchased |
| 22 | 600,000 | Purchased |
| 35 | 2,500,000 | Purchased |
| 60 | 100,000,000 | Purchased |

**Interpretation:**

Even younger or older individuals **with high salaries** are more likely to be classified as **Purchasing Insurance** by the model.

1. **Hypotheses and Assumptions:**Make your hypothesis or assumptions based on the inference from the data and justify your assumptions by testing it on the built accurate model. Example : You might make assumptions such as:

* 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.

You can then test these assumptions using your accurate AI model. For example, you could run simulations where you manipulate age and salary inputs to observe their effects on insurance purchasing predictions.

**Assumption 1:**

*Younger individuals with higher salaries are more likely to purchase health insurance.*

**Test Case:** Age 22, Salary 600,000  
**Model Prediction:** Purchased -- Assumption supported.

**Assumption 2:**

*Older individuals with low or no salary are less likely to purchase insurance.*

**Test Case:** Age 50, Salary 0  
**Model Prediction:** Not Purchased -- Assumption supported.

**Assumption 3:**

*Salary has more influence than age on insurance purchase decisions.*

**Comparison:**

* Age 18, Salary 0 -- Not Purchased
* Age 18, Salary 500,000 -- Purchased
* Age 50, Salary 100,000,000 -- Purchased

**Conclusion:**  
 Salary dominates age in influencing insurance purchase decisions.

1. **Lessons Learned and Real-Life Application:** What did you learn from this study and how do you like to apply in real life projects? Give two case studies or scenarios where you will use these AI Algorithms.

Lessons Learned:

* Even simple features like **Age and Salary** can drive powerful predictions with the right models.
* **Feature scaling** and **algorithm selection** greatly affect model performance.
* **Random Forest** generally performs better than Logistic Regression for non-linear boundaries.

**Real life Applications:**

* **Case Study 1: Loan Approval Prediction**

Use similar models to assess loan eligibility using features like age, income, employment status, and existing loans.

* **Case Study 2: Credit Card Marketing**

Banks can use these models to predict which customers are likely to respond to marketing for credit cards or insurance policies based on financial indicators.