SUBMITTED TO: INTRNFORTE

ON:20/7/25

Customer Insurance Purchase Prediction Using Machine Learning

BY: ALWIN SABU

**Abstract**

This project aims to predict whether a customer will purchase insurance based on their age and estimated salary. Using a dataset with these two features, we applied various machine learning classification algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest. After comparing their performance, the SVM and Random Forest models gave the most accurate results. This model can help insurance companies better understand customer behavior and make smarter marketing decisions.

**Table of Contents**

* Introduction
* Literature Review
* Problem Statement
* Data Collection and Preprocessing
* Methodology
* Implementation
* Results
* Discussion
* Conclusion
* References
* Appendices

### **1. Introduction**

This project explores how machine learning can be used to improve decision-making in the insurance industry. Specifically, we aim to predict whether a customer will purchase health insurance based on their age and estimated salary. Accurate predictions can help insurance companies better target marketing efforts and understand customer behavior.

**2. Literature Review**

Several studies have shown that classification models such as decision trees and support vector machines are effective in customer segmentation and insurance predictions. While traditional actuarial methods rely on statistical averages, AI-based models offer personalized insights and faster processing.

**3. Problem Statement**

The challenge is to build a model that can classify whether a customer will purchase insurance, given two features: age and estimated salary. The goal is to compare multiple ML models and choose one that balances precision and generalization.

**Assumptions:**

* Only two input features are available
* The dataset is representative
* Personal identifiable information (PII) is excluded

**4. Data Collection and Preprocessing**

* **Data Source:** Dataset extracted from PDF report (provided).
* **Features:**
  + Age
  + Estimated Salary
  + Purchased (target variable)

**Preprocessing Steps:**

* Converted data to CSV
* Standardized features using StandardScaler
* Split into training and test sets using stratification

### **5. Methodology**

The following models were used:

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

### **6. Implementation**

Implementation was done using Python with:

* Pandas, Matplotlib, Seaborn for data handling and visualization
* scikit-learn for preprocessing and model building

Github Link: <https://github.com/Alwin-sabu-2006/Mini-Project>

**7.Results**

Logistic Regression

Accuracy: 0.8375

Precision: 0.8636363636363636

Recall: 0.6551724137931034

F1 Score: 0.7450980392156863

KNN

Accuracy: 0.9

Precision: 0.8620689655172413

Recall: 0.8620689655172413

F1 Score: 0.8620689655172413

SVM

Accuracy: 0.9

Precision: 0.8387096774193549

Recall: 0.896551724137931

F1 Score: 0.8666666666666667

Decision Tree

Accuracy: 0.8625

Precision: 0.7647058823529411

Recall: 0.896551724137931

F1 Score: 0.8253968253968254

Random Forest

Accuracy: 0.9

Precision: 0.8387096774193549

Recall: 0.896551724137931

F1 Score: 0.8666666666666667

Test Case 1: Will Not Purchase

Test Case 2: Will Not Purchase

Test Case 3: Will Purchase

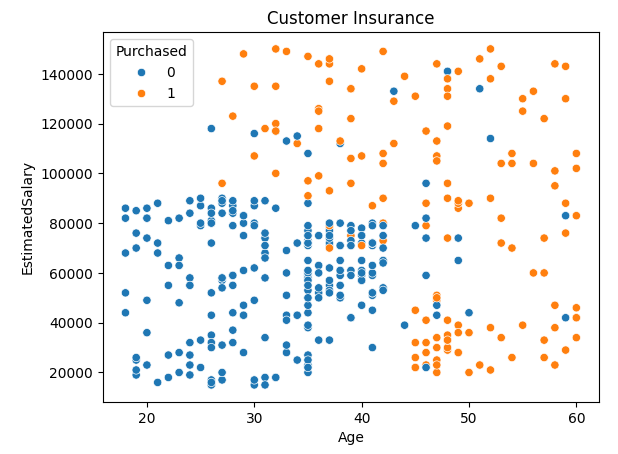
Test Case 4: Will Not Purchase

Test Case 5: Will Not Purchase

Test Case 6: Will Purchase

Test Case 7: Will Purchase

Test Case 8: Will Purchase



**8. Discussion**

* Both SVM and Random Forest performed equally well.
* Salary had a stronger effect on the outcome than age.
* Younger individuals with very high salaries were often classified as buyers.
* Logistic Regression underperformed due to its linear nature.

**Limitations:**

* Only 2 features were available
* Real-world datasets would include more factors

**9. Conclusion**

This study demonstrated the use of various classification models to predict insurance purchase behavior. SVM and Random Forest emerged as the best performers. This model could be used in real-world applications to automate and improve customer targeting in insurance marketing.

**10. References**

* Intrn Forte Internship Videos based on AI
* Google Colab / Jupyter official docs

### **11.Appendices**

* See insurance\_prediction.ipynb for code
* Dataset: insurance\_data.csv
* Use instructions:
  1. Open notebook
  2. Upload dataset
  3. Run all cells