**Project Title:**  
Predicting Customer Insurance Purchases Using Classification Algorithms

**Your Name:**  
Jathin Lalith. D

**Date:**  
29 June 2025

**Abstract**

The main aim of the project is to predict whether a new person with a certain age and salary is likely to purchase insurance or not. Many different types of classification algorithms—including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Random Forests—are implemented to develop predictive models. The project involved preprocessing the data, training each model, and evaluating their performance using metrics such as accuracy, precision, recall, and F1-score. By comparing these algorithms, the project aims to determine which yields the highest predictive accuracy. The results provide insights into which machine learning technique is most suitable for predicting insurance purchasing behavior in this context.

**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 project focuses on predicting whether customers will buy health insurance based on age and estimated salary. The performance of the model is crucial as it helps improve marketing strategies and decision making.

The problem addressed is how to classify new customers effectively using limited input features while maintaining high accuracy and performance. The main goal is to build and compare different models and select the best and most suitable one for prediction.

For this task, various machine learning models were used, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Random Forest Classifiers. Data preprocessing was performed using NumPy and pandas.

**2. Literature Review**

Many studies have examined predicting customer insurance purchases using statistical and machine learning techniques. These studies show that classification methods—especially Random Forests—are effective for this purpose. Achieving good predictive performance without overfitting requires careful feature selection, parameter tuning, and ensuring model generalization to new data.

**3. Problem Statement**

The project aims to predict if a customer will buy insurance based on age and estimated salary, building a model to classify whether a customer will purchase insurance.

**Assumptions:**

* The data provided is accurate and sourced from real customers.
* There are no missing values in the dataset.
* Age and salary are the main factors influencing insurance purchase.

**Limitations:**

* The model excludes other factors that might influence purchase decisions.
* The limited number of features may reduce prediction accuracy; even with 90% accuracy, some wrong predictions may occur.

**4. Data Collection and Preprocessing**

The dataset was obtained from the Bank Insurance Company at:  
<https://drive.google.com/file/d/1wOrVrq30W3bl1st4cvnt5bvn5UWEK4Ab/view?usp=drive_link>

**Fields:**

* Age
* Estimated Salary
* Purchased

**Preprocessing Steps:**

* Checked for missing values and inconsistent data
* Handled missing values (if any)
* Split data into training and testing sets
* Applied standardization on Age and Estimated Salary

**5. Methodology**

**Algorithms Used:**

* Logistic Regression
* K-Nearest Neighbors
* Support Vector Machine
* Decision Tree
* Random Forest

**Rationale:**  
These algorithms cover both linear and nonlinear classifiers.

**Evaluation Metrics:**

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

**6. Implementation**

* Split data into training and test sets.
* Implemented models using libraries such as:
  + numpy
  + pandas
  + sklearn
  + matplotlib

**7. Results**

| **Algorithm** | **Accuracy (%)** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 89 | 0.89 | 0.75 | 0.81 |
| K-Nearest Neighbors | 93 | 0.88 | 0.91 | 0.89 |
| Support Vector Machine | 90 | 0.92 | 0.75 | 0.83 |
| Decision Tree | 91 | 0.83 | 0.91 | 0.87 |
| Random Forest | 91 | 0.85 | 0.88 | 0.86 |

**Baseline Accuracy:**  
If 68 out of 100 samples are "no purchase," the baseline accuracy would be 68%.

**8. Discussion**

**Insights:**

* KNN achieved the highest accuracy (93%).
* Decision Tree and Random Forest also performed well (91%), handling complex patterns effectively.
* Logistic Regression and SVM had high precision but lower recall.

**Unexpected Outcomes:**

* KNN outperforming all others was unexpected.
* Logistic Regression was less accurate at identifying actual buyers.

**Strengths:**

* Multiple models tested and compared.
* Results are easy to interpret.
* All models outperformed baseline guessing.

**Limitations:**

* Only two features were used.
* KNN may be slower on very large datasets.

**9. Conclusion**

**Key Findings:**

* KNN had the highest accuracy (93%), indicating similar customers tend to behave alike.
* Decision Tree and Random Forest also showed strong performance (91%).

These models can help improve marketing strategies, sales, and decision-making.

**Future Improvements:**

* Incorporate more features such as purchase history, health status, or family details.
* Use cross-validation to validate model stability.
* Explore other algorithms like Neural Networks for larger datasets.

**10. References**

* scikit-learn documentation. Retrieved from <https://scikit-learn.org/stable/documentation.html>
* Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

**11. Appendices**

**Technical Details**

**Environment:**

* Python 3.x
* scikit-learn
* NumPy
* pandas
* matplotlib / seaborn for visualization

**Data Preprocessing:**

* Used StandardScaler to normalize age and salary.
* Missing salary values imputed using median.

**Sample Code:**

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

**To Reproduce Results:**

* Install packages:  
  pip install scikit-learn pandas matplotlib
* Clone repository:  
  git clone https://github.com/Jathin-24/Project\_1  
  cd Project\_1
* Run main script:  
  python main.py

**12. Questions**

**1. Graphical Analysis and Predictions (K-NN):**

| **Age** | **Salary** | **Prediction** |
| --- | --- | --- |
| 30 | 87,000 | Not Purchased |
| 40 | No Salary | Not Purchased |
| 40 | 100,000 | Purchased |
| 50 | No Salary | Purchased |

**2. Graphical Analysis and Predictions (K-NN):**

| **Age** | **Salary** | **Prediction** |
| --- | --- | --- |
| 18 | No Salary | Not Purchased |
| 22 | 600,000 | Purchased |
| 35 | 2,500,000 | Purchased |
| 60 | 100,000,000 | Purchased |

**3. Hypotheses and Assumptions:**

* **Hypothesis 1:** Younger individuals with higher salaries are more likely to purchase health insurance.  
  **Test:** Used KNN to simulate predictions for ages under 30 with salaries above ₹500,000.  
  **Result:** High probability of purchase predicted.
* **Hypothesis 2:** Older individuals with higher salaries might be less inclined to purchase health insurance.  
  **Test:** Input ages over 50 with salaries above ₹1,000,000.  
  **Result:** Predicted purchase probabilities were lower (60–70%).
* **Hypothesis 3:** Salary has a stronger impact on purchasing than age.  
  **Test:** Kept age constant (35 years) and varied salary from ₹100,000 to ₹2,500,000.  
  **Result:** Purchase probability increased significantly with salary, confirming salary’s stronger influence.

**4. Lessons Learned and Real-Life Applications:**

* Machine learning models can accurately predict customer behavior using few features.
* KNN performed well, showing simple models can be powerful with clean data.

**Real-Life Applications:**

* Targeted insurance marketing
* Credit card upselling