**Title Page:**

**Project Title:**

**Customer Insurance**

**Purchases Case Study**

**Shreya A**

**22-08-2024**

**Customer Insurance Purchases Case Study**

**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:**

1. **Data Preparation: Begin by preprocessing the provided dataset, ensuring it is ready for training and evaluation. Perform necessary transformations and feature engineering.**
2. **Algorithm Selection: Choose a variety of classification algorithms, such as Decision Trees, Random Forest, Support Vector Machines (SVM), Logistic Regression, and Neural Networks.**
3. **Model Development: Implement each chosen algorithm using the preprocessed data. Tune hyperparameters as needed.**
4. **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.**
5. **Compare and Contrast: Create a comprehensive comparative analysis of the algorithms' performance. Consider aspects like accuracy, model complexity, training time, and potential for overfitting.**
6. **Insights Generation: Extract insights from the comparative analysis. Identify patterns and trends in algorithm performance. Understand the trade-offs between accuracy and generalization.**
7. **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.**
8. **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 behaviour and make informed decisions in the insurance domain.**

**ML Classification Algorithms:**

1. **Logistic regression**
2. **KNN**
3. **Support Vector Machine**
4. **Decision Trees**
5. **Random Forest (Ensemble learning) Classifiers**

**LINK TO DATASET:**

<https://drive.google.com/file/d/1wOrVrq30W3bl1st4cvnt5bvn5UWEK4Ab/view?usp=drive_link>

**Abstract:**

This study compares five classification algorithms for predicting customer insurance purchases, identifying Random Forest as the optimal model due to its superior balance of accuracy and generalization.

**Table of Contents:**

List of sections and subsections in our report, along with clickable links for easy navigation.

**1. Introduction:**

This project aims to enhance customer insurance purchase predictions by evaluating five AI classification algorithms, addressing the challenge of selecting the most accurate and generalizable model. The methodologies include Logistic Regression, KNN, SVM, Decision Trees, and Random Forest to optimize decision-making processes.

**2. Literature Review:**

This review explores recent advancements in AI-driven classification methods, focusing on their application in customer behaviour prediction, while addressing challenges like model selection, accuracy, and overfitting in real-world datasets. Prior research highlights the effectiveness of ensemble methods and the importance of algorithm choice for optimizing predictive accuracy.

**3. Problem Statement:**

This project addresses the challenge of accurately predicting customer insurance purchases using AI models, considering potential limitations like data quality and model overfitting. Assumptions include the availability of relevant features and a balanced dataset.

**4. Data Collection and Preprocessing:**

The project utilizes a dataset of customer information, obtained through company records, which was pre-processed by handling missing values, encoding categorical variables, and normalizing numerical features for model readiness.

**5. Methodology:**

The methodology involves applying machine learning algorithms—Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Trees, and Random Forest—to predict customer insurance purchases, chosen for their balance of complexity and interpretability, with specific parameter tuning for optimal performance.

**6. Implementation:**

The AI model was implemented using Python libraries like scikit-learn, with each algorithm trained and evaluated on a dataset split into training and testing sets, and performance metrics calculated to compare effectiveness.  
[HariniReddy07/AI--IntrnForte (github.com)](https://github.com/HariniReddy07/AI--IntrnForte/tree/main)

**7. Results:**

The results, presented in tables and charts, show the performance metrics of each algorithm including accuracy, precision, recall, and F1 score, with Random Forest achieving the highest overall performance compared to baseline methods.

**8. Discussion:**

The discussion highlights Random Forest as the top-performing algorithm, with insights into its balanced accuracy and efficiency. Unexpectedly lower performance from KNN prompts a revaluation of its parameter settings, while strengths and limitations of each approach are analysed for future improvements.

**9. Conclusion:**

The project identifies Random Forest as the most effective model for predicting customer insurance purchases, demonstrating high accuracy and robustness. Future research could explore hyperparameter tuning and additional algorithms to further enhance predictive performance and generalization.

**10. References:**

Here are two examples of reference links formatted in APA style:

1. Book:

- Smith, J. (2020). \*Introduction to Machine Learning\*. Springer. [Link to book] (https://www.springer.com/gp/book/9783030341252)

2. Journal Article:

- Doe, A., & Brown, B. (2021). Advancements in Neural Networks: A Comprehensive Review. \*Journal of Artificial Intelligence Research\*, 56(3), 123-145. [Link to article] (https://www.jair.org/index.php/jair/article/view/12345)

**11. Appendices:**

Appendices should provide supplementary details such as additional graphs, technical specifications, and code snippets. Include instructions for reproducing the results to support transparency and reproducibility.

**12. Acknowledgment:**

Acknowledge any individuals, institutions, or resources that provided support or contributions to our project. This could include mentors, colleagues, and funding organizations.

**Questions to be solved:**  
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  
**Answer:**

A screen shot of a graph

Description automatically generated

Graphical Analysis and Predictions

Data Visualization:

Scatter plot: Plotting age against estimated salary, coloured by purchase status.

Density plot: Visualizing the distribution of age and salary for purchasers and non-purchasers.

Observations and Predictions:

Age: There seems to be a slight trend towards older individuals purchasing insurance, especially in the higher age brackets.

Salary: Higher salaries appear to correlate with a higher likelihood of purchasing insurance.

Predictions:

Age 30, Salary 87,000: Likely to purchase.

Age 40, No Salary: Unlikely to purchase.

Age 40, Salary 100,000: Likely to purchase.

Age 50, No Salary: Unlikely to purchase.

2. 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   
**Answer:**  
Graphical Analysis and Predictions (Extreme Scenarios)

Observations and Predictions:

Extreme scenarios: The model might break down for very high or very low values of age or salary, as these data points might be outliers.

Predictions:

Age 18, No Salary: Unlikely to purchase.

Age 22, Salary 600,000: Likely to purchase.

Age 35, Salary 2,500,000: Likely to purchase.

Age 60, Salary 100,000,000: Likely to purchase.

3. 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.

**Answer:**  
Hypotheses and Assumptions

Hypotheses:

Age: Older individuals are more likely to purchase insurance due to increased health concerns.

Salary: Higher-income individuals can afford insurance premiums and may prioritize health coverage.

Interaction: The combination of age and salary might influence purchase decisions.

Assumptions:

The data is representative of the target population.

There are no significant missing or erroneous data points.

The relationship between age, salary, and purchase is linear.

Testing Assumptions:

Correlation analysis: Calculate correlations between age, salary, and purchase.

Statistical tests: Conduct hypothesis tests to assess the significance of relationships.

Model building: Develop a machine learning model to predict purchases based on age and salary.

4. 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.  
**Answer:**

Age and salary can be significant factors in predicting health insurance purchases.

Extreme values might require careful consideration and potential adjustments to the model.

Interaction effects between age and salary could be explored.

Real-Life Applications:

Insurance companies: Personalize marketing campaigns based on customer demographics.

Healthcare providers: Identify potential high-risk patients for targeted outreach.

Government agencies: Develop policies to promote affordable health insurance coverage.

Case Studies:

Targeted marketing: An insurance company could offer tailored health insurance plans to individuals in specific age groups or income brackets, based on their predicted likelihood of purchasing.

Risk assessment: A healthcare provider could use the model to identify patients at high risk of not having insurance and offer them resources or assistance in obtaining coverage.

**Thank you.**