**Project Report: Health Insurance Customer Interest Prediction**

**Introduction**

This report outlines the end-to-end process, solution architecture, and methodology used for predicting customer interest in health insurance. The goal is to determine whether a customer is interested in health insurance based on various demographic and historical data.

**Data Overview**

The dataset consists of customer information with features such as gender, age, driving license status, region code, vehicle information, and annual premium. The target variable is "Response," indicating customer interest in health insurance.

**Data Description**

**id:** Unique ID for the customer

**Gender:** Gender of the customer

**Age:** Age of the customer

**Driving\_License:** 0: Customer does not have DL, 1: Customer already has DL

**Region\_Code:** Unique code for the region of the customer

**Previously\_Insured:** 1: Customer already has Vehicle Insurance, 0: Customer doesn't have Vehicle Insurance

**Vehicle\_Age:** Age of the Vehicle

**Vehicle\_Damage:** 1: Customer got their vehicle damaged in the past, 0: Customer didn't get their vehicle damaged in the past.

**Annual\_Premium:** The amount the customer needs to pay as premium in the year

**Policy\_Sales\_Channel:** Anonymized Code for the channel of outreaching to the customer

**Vintage:** Number of days the customer has been associated with the company

**Response:** 1: Customer is interested, 0: Customer is not interested

**Data Exploration and Analysis**

**Data Preprocessing:**

* Imported necessary packages.
* Extracted and checked the information of the data.
* Checked for null values (no null values found).
* Obtained basic statistics of the data.
* Checked the data types.

**Exploratory Data Analysis (EDA):**

* Dropped the 'id' column as it is unique and not needed for analysis.
* Explored the distribution of categorical variables (Response, Gender, Driving\_License, Region\_Code, etc.).
* Utilized visualizations (count plots, correlation heatmap, pair plots) to gain insights into the relationships between variables.

**Feature Engineering:**

* Encoded categorical variables using label encoding.

**Model Development:**

* Utilized four different algorithms for prediction: Logistic Regression, Decision Tree, Random Forest, and K-nearest Neighbors (KNN).
* Standardized the data using Standard Scaler.
* Evaluated the models based on training accuracy, testing accuracy, precision score, and F1-score.

**Model Evaluation Results**

| Algorithm Name | Training Accuracy | Testing Accuracy | Precision Score | F1-score |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.876564 | 0.880927 | 0.000000 | 0.000000 |
| Decision Tree | 0.999905 | 0.824972 | 0.282686 | 0.293716 |
| Random Forest | 0.876564 | 0.880927 | 0.000000 | 0.000000 |
| K-nearest Neighbours | 0.878368 | 0.877804 | 0.382178 | 0.076542 |

**Model Selection**

Considering precision score and overall performance, K-nearest Neighbors (KNN) is chosen as the ideal model for predicting customer interest in health insurance.

**Model Deployment**

* Saved the KNN model using joblib for future predictions.
* Prediction on New Data
* Loaded the saved KNN model.
* Extracted and preprocessed new data for predictions.
* Made predictions on the new data to determine customer interest.

**Conclusion**

This project successfully explores, analyzes, and predicts customer interest in health insurance based on provided data. The chosen K-nearest Neighbors model demonstrates promising results for predicting customer behavior.

The end-to-end process includes data exploration, feature engineering, model development, evaluation, selection, deployment, and prediction on new data. This comprehensive approach ensures the model's practical applicability in real-world scenarios.