Travel Insurance Prediction Report

This project predicts whether a customer will buy travel insurance based on demographic and travel-related features using Logistic Regression.

Dataset Overview

The dataset contains customer demographic and travel-related information. Target variable: TravelInsurance (1 = bought insurance, 0 = did not buy). Features include: Age, Annual Income, Family Members, Chronic Diseases, Employment Type, Graduate or Not, Frequent Flyer, Ever Travelled Abroad.

Methodology

1. Data Cleaning: Removed irrelevant column, checked missing values.
2. Exploratory Data Analysis: Examined distribution of features and target variable.
3. Preprocessing: Encoded categorical variables and scaled numeric ones.
4. Model Building: Split data into 80% training and 20% testing sets. Trained Logistic Regression model.
5. Evaluation: Measured accuracy (~80%) and analyzed confusion matrix.
6. Deployment: Saved model with joblib and created a Streamlit web app for prediction.

Results

The Logistic Regression model achieved around 80% accuracy. Confusion matrix showed more correct predictions for customers not buying insurance. Precision, Recall, and F1-score were reasonable but could be improved with techniques such as SMOTE or trying other models like Random Forest or XGBoost.

Challenges & Future Work

• Handle class imbalance for better recall on minority class.
• Compare with more complex models (Random Forest, XGBoost).
• Deploy app on cloud platforms such as Streamlit Cloud or Heroku.
• Add interpretability features like SHAP values or feature importance charts.

Conclusion

This project demonstrates an end-to-end ML pipeline: data preprocessing, EDA, logistic regression modeling, evaluation, and deployment via Streamlit. It shows how ML can be applied to real-world insurance business problems.