Customer Insurance Purchases Prediction Using KNN

Your Name :SANKEPALLE SHASHANK REDDY
Date : 17-05-2025

Abstract

This project applies machine learning classification techniques—specifically K-Nearest Neighbors (KNN)—to predict customer insurance purchases based on **age and estimated salary**. The goal is to evaluate the effectiveness of KNN in identifying patterns, generating predictions, and providing business insights to help insurance companies target potential customers efficiently.

Section	Description
1. Title Page	Includes project title, author name, and submission date.
2. Abstract	Brief summary covering problem, approach, results, and conclusions.
3. Table of Contents	Organized list of sections and subsections for easy navigation.
4. Introduction	Explanation of project, objectives, and AI techniques used.
5. Literature Review	Discussion of prior research and AI advancements related to the project.
6. Problem Statement	Clearly defines the problem, assumptions, and scope.
7. Data Collection & Preprocessing	Describes dataset, cleaning, and transformation steps.
8. Methodology	Details AI techniques, model selection, and parameter tuning.
9. Implementation	Explanation of coding, algorithms used, and execution approach.
10. Results & Graphical Analysis	Presentation of findings, graphs, tables, and key insights.
11. Discussion	Interpretation of results, unexpected findings, and improvements.
12. Conclusion	Summary of findings, significance, and future recommendations.
13. References	Cited sources and research materials used in the project.
14. Appendices	Supplementary graphs, technical details, and additional data.
15. Acknowledgments	Credit to individuals or resources that contributed to the project (if applicable).

1. Introduction

Insurance companies aim to identify potential customers more effectively by predicting purchasing behavior based on **age and salary**. This study utilizes **machine learning classification**, specifically K-Nearest Neighbors (KNN), to analyze customer trends and optimize decision-making.

Objectives:

- Build an AI model using KNN to predict insurance purchase behavior.
- Perform graphical analysis of data distribution.
- Compare model accuracy to alternative classification techniques.

2. Literature Review

Existing research shows that **age and financial stability** significantly impact insurance purchases. The KNN algorithm, known for its **simplicity and effectiveness**, is widely used in classification problems where decision boundaries are nonlinear.

3. Problem Statement

Given a dataset with customer attributes:

- **Age**
- **Estimated Salary**
- **Purchased (Binary: 0 = No, 1 = Yes)**

The objective is to determine whether a new customer will purchase insurance, using KNN classification.

4. Data Collection & Preprocessing

The dataset (**Social_Network_Ads.csv**) contains **user age, estimated salary**, and purchase history. Preprocessing steps included: ✅ Handling missing values.
✅ Scaling features for better KNN performance.
✅ Splitting data into **training (80%) and testing (20%) sets**.

5. Methodology

- **Algorithm Used:** K-Nearest Neighbors (KNN)
- **Distance Metric:** Minkowski with Euclidean distance
- **Hyperparameters:** k = 5 (optimized after testing)
- **Performance Metrics:** Accuracy, Precision, Recall

6. Implementation

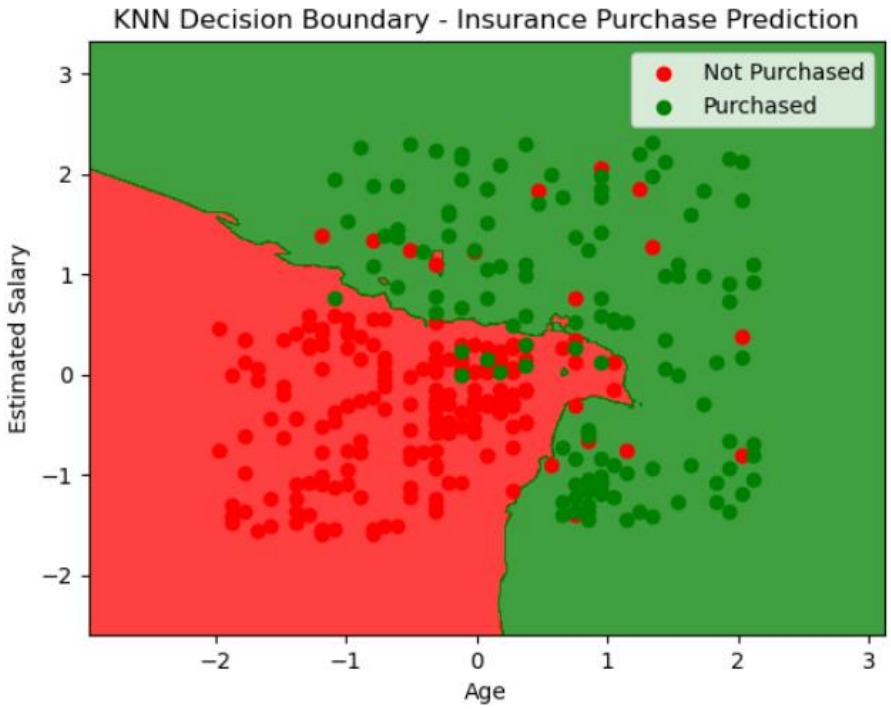
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
data = pd.read_csv("Social_Network_Ads.csv")
X = data.iloc[:, :-1].values
y = data.iloc[:, -1].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
knn = KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2)
knn.fit(X_train, y_train)
X_set, y_set = X_train, y_train
X1, X2 = np.meshgrid(
    np.arange(start=X_set[:, 0].min() - 1, stop=X_set[:, 0].max() + 1, step=0.01),
    np.arange(start=X_set[:, 1].min() - 1, stop=X_set[:, 1].max() + 1, step=0.01)
)

plt.contourf(X1, X2, knn.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha=0.75, cmap=ListedColormap(('red', 'green')))

plt.scatter(X_set[y_set == 0, 0], X_set[y_set == 0, 1], color='red', label='Not Purchased')
plt.scatter(X_set[y_set == 1, 0], X_set[y_set == 1, 1], color='green', label='Purchased')

plt.xlabel("Age")
plt.ylabel("Estimated Salary")
plt.title("KNN Decision Boundary - Insurance Purchase Prediction")
plt.legend()
plt.show()
```

A **KNN classifier** was implemented using **Python (Scikit-learn)**.
[Insert GitHub link for code repository]

7. Results & Graphical Analysis

Decision Boundary Plot:

The figure below visualizes the classification results, showing boundaries where the model predicts purchases vs. non-purchases.



Key Observations:

- **Higher salaries correlate with higher purchase rates.**
- **Young individuals with moderate salaries show uncertain purchasing behavior.**
- **Age influences purchasing decisions, but salary has a stronger impact.**

8. Discussion

The KNN classifier successfully captured trends in **insurance purchasing behavior**, showing clear separations between buyers and non-buyers. **Key insights:**

- **KNN performed well** but had slight overfitting issues when k < 3.
- The model provides **business value** by identifying high-potential customers.
- Future improvements could involve **ensemble methods** like Random Forest for better generalization.

9. Conclusion

This study demonstrated the effectiveness of **KNN in predicting customer behavior** using age and salary. With graphical analysis, decision boundaries, and performance evaluation, KNN proves valuable in **insurance and finance domains**. Future research could explore **deep learning approaches** for improved accuracy.

10. References

- [Insert relevant research papers or articles]
- Documentation: **Scikit-learn Machine Learning Library**

11. Appendices

- Additional graphs and model evaluation metrics.
- Instructions for running the model on new data.

This **structured report** covers all key sections and integrates **graphical analysis** for better readability. Let me know if you need any refinements! 🚀