**DR. AMBEDKAR INSTITUTE OF TECHNOLOGY**

**(An Autonomous Institution, Affiliated to VTU, Belgaum and**

**Aided by Government of Karnataka)**

**Near Jnana Bharathi Campus, Bangalore-560056**



**Department of Information Science & Engineering**

**MINI PROJECT SYNOPSIS**

**ON**

**“R.I.K(Retailer Insights using KNN classification)”**

**BACHELOR OF ENGINEERING**

**IN**

**INFORMATION SCIENCE AND ENGINEERING**

2023-2024

**SUBMITTED BY**

ADITYA G [USN: 1DA21IS004]

YASH SANJAY KAROGAL [USN: 1DA21IS059]

**UNDER THE GUIDANCE OF:**

Drs. S PUSHPALHA

Asst. Professor, Dept. of ISE

Rd. Ambedkar Institute of Technology.

Dr

Dr

d

  Rd. AI

(An Autonomous institution, Aided by Government of Karnataka, Near Jnana Bharathi  Campus, Bengaluru-560056)



**CERTIFICATE**

This is to certify that this project report entitled **“**R.I.K(Retailer’s Insights using KNN classification**”** by Aditya G(1DA21IS004), Yash Karogal (1DA21IS059)  submitted in partial fulfilment of the requirements of the 6th semester seminar for  the degree of Bachelor of Engineering in Information Science & Engineering of  Dr. Ambedkar Institute of Technology, Bengaluru, during the academic year 2022- 23, is a bonafide record of work carried out under my guidance and supervision.

Signature of the Guide Signature of HOD Signature of Principal

Dr. S Pushpalatha Mrs. Vanishree Abhay Dr. M V Vijayakumar Dr.C Nanjundaswamy Asso. Professor Head of the Department Principal  Dr. AIT Dept. of ISE, Dr. AIT Dr.AIT

External viva:

Name of the examiners with date Signature

**INDEX**

**CONTENT PAGE**

INTRODUCTION ………………………………………………………………………………………….. 5

LITERATURE SURVEY …………………………………………………………………………………. 8

OBJECTIVE …………………………………………………………………………………………………… 11

ADVANTAGES AND DISADVANTAGES ………………………………………………………….. 12

EXISTING SYSTEMS ……………………………………………………………………………………. 13

CODE AND REVIEW ………………………………………………………………………………….. 14

REASONS FOR CHOOSING KNN MODEL ………………………………………………………. 18

PROPOSED SYSTEMS ……………………………………………………………………………….. 20

SYSTEM REQUIREMENTS AND SPECIFICATIONS ………………………………………… 22

**ABSTRACT**

Understanding customer behavior is crucial for businesses as it significantly enhances personalization, targeted marketing, and customer experience, leading to increased satisfaction and loyalty. By delving into customer behavior, companies can predict churn, design effective loyalty programs, and guide product development and service improvements. Insights into price sensitivity and dynamic pricing strategies further optimize revenue, while proactive and personalized support boosts satisfaction. Analyzing behavior gives companies a competitive edge, helping them stay ahead of market trends and enabling informed strategic decisions, which reduces risks and enhances outcomes.

Efficient resource allocation becomes possible, allowing businesses to focus on high-value customers and reduce waste. Practical applications of understanding customer behavior include e-commerce personalization, customer segmentation, churn prediction, and dynamic pricing. For instance, a retail business can use customer data to recommend products tailored to individual preferences, optimize inventory management to ensure popular items are always in stock, and design loyalty programs that reward frequent shoppers and encourage repeat business. This data-driven approach not only enhances the customer experience but also drives growth and profitability.

Moreover, understanding customer behavior allows businesses to anticipate market demands and trends, enabling them to innovate and adapt swiftly. For example, companies can identify emerging consumer preferences and adjust their offerings accordingly, ensuring they remain relevant and competitive. Additionally, insights into customer behavior can inform marketing strategies, making campaigns more effective by targeting the right audience with the right message at the right time.

Customer behavior analysis also plays a pivotal role in enhancing customer support. By understanding common issues and pain points, businesses can proactively address these concerns, providing timely and personalized support that boosts overall satisfaction. This proactive approach not only improves customer retention but also builds a positive brand reputation, as customers are more likely to share their positive experiences with others.

In the context of dynamic pricing, understanding how customers respond to price changes allows businesses to implement strategies that maximize revenue without alienating customers. For instance, by analyzing purchase patterns and price sensitivity, companies can adjust prices in real-time to match demand, ensuring optimal sales and profitability.

Overall, understanding customer behavior is not just about improving individual transactions but about building a comprehensive strategy that enhances every aspect of the business. From product development and marketing to customer support and pricing strategies, insights gained from customer behavior analysis empower businesses to make data-driven decisions that drive success. By continuously monitoring and analyzing customer behavior, companies can stay agile, respond to changes swiftly, and maintain a competitive edge in an ever-evolving market.

**INTRODUCTION**

**Predicting customer behavior with KNN (K nearest neighbor)**

K-Nearest Neighbors (KNN) classification is a simple and intuitive machine learning algorithm used for classification tasks. It works by storing all training data instances and, when a new instance needs to be classified, calculating the distance between this new instance and all training instances. The 𝑘*k* nearest neighbors are identified, and the class of the new instance is determined by majority voting among these neighbors. KNN is sensitive to feature scales, so normalization or standardization is important. Choosing the optimal 𝑘*k* and handling computational efficiency for large datasets are key considerations. Despite its simplicity, KNN is effective for various applications, including customer behavior analysis and image recognition.

Machine learning, particularly K-Nearest Neighbors (KNN) classification, can significantly aid in understanding and predicting customer behavior. Here's how KNN can be applied:

**Understanding Customer Behavior with KNN**

1. **Customer Segmentation**:
   * KNN can group customers with similar behaviors and preferences by identifying patterns in their data. For instance, it can segment customers based on purchase history, browsing patterns, or demographic information.
2. **Predicting Customer Purchases**:
   * KNN can predict what products a customer might buy next by analyzing the purchase history of similar customers. If a customer has a purchase pattern similar to others, KNN can suggest products that those similar customers have bought.
3. **Customer Churn Prediction**:
   * By comparing current customer behavior to historical data, KNN can help identify customers at risk of churning. If a customer’s recent behavior is similar to that of customers who have previously churned, the model can flag this customer as a potential churn risk.
4. **Personalized Marketing**:
   * KNN can personalize marketing efforts by recommending products or offers to customers based on the preferences and behaviors of their nearest neighbors. This ensures that marketing efforts are relevant and more likely to be effective.

**Practical Steps to Implement KNN for Customer Behavior Prediction**

1. **Data Collection**:
   * Gather customer data, including purchase history, demographic information, browsing behavior, and any other relevant metrics.
2. **Feature Selection**:
   * Choose the features that best represent customer behavior. This might include total purchase amount, frequency of purchases, types of products bought, and more.
3. **Data Preprocessing**:
   * Normalize or standardize the data to ensure that each feature contributes equally to the distance calculations. Handle missing values and encode categorical variables if necessary.
4. **Choosing the Number of Neighbors (k)**:
   * Experiment with different values of 𝑘*k* to find the optimal number of neighbors. Use cross-validation to ensure the chosen 𝑘*k* provides the best performance.
5. **Distance Metric**:
   * Choose an appropriate distance metric, such as Euclidean distance, to measure the similarity between customers.
6. **Model Training**:
   * Train the KNN model on the preprocessed data. The model will store the training instances and use them to classify new instances.
7. **Making Predictions**:
   * For a new customer or a new behavior instance, calculate the distance to all training instances and identify the 𝑘*k* nearest neighbors. Use the majority class of these neighbors to make a prediction.
8. **Evaluation**:
   * Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score. Ensure the model generalizes well to new, unseen data.

**LITERATURE SURVEY**

To provide a comprehensive understanding of how K-Nearest Neighbors (KNN) classification can be applied to predict customer behavior, a review of relevant literature is essential. Here are key papers and articles that can help you delve into this topic:

1. **"A Survey on Customer Behavior Prediction using Machine Learning Techniques"**
   * **Authors**: Varsha Sharma, Preeti Sharma
   * **Journal**: International Journal of Computer Applications
   * **Abstract**: This paper provides an overview of various machine learning techniques, including KNN, used for predicting customer behavior. It discusses the strengths and limitations of these techniques and compares their performance in different scenarios.
2. **"Customer Churn Prediction Using Improved Balanced Random Forest"**
   * **Authors**: Myoung-jae Park, Il-yeop Lee, and Byoung-ju Choi
   * **Journal**: Journal of the Korea Institute of Information and Communication Engineering
   * **Abstract**: While this paper primarily focuses on random forests, it provides insights into how KNN can be used in combination with other methods to improve customer churn prediction.
3. **"Customer Behavior Analysis using KNN and LSTM"**
   * **Authors**: G. Gupta, R. Kaur
   * **Conference**: 2019 International Conference on Advances in Computing, Communication, and Control (ICAC3)
   * **Abstract**: This paper explores the use of KNN alongside LSTM for analyzing customer behavior. It highlights the effectiveness of hybrid approaches in improving prediction accuracy.
4. **"Comparison of Classification Techniques for Customer Behavior Prediction"**
   * **Authors**: P. Kumar, S. Bhatia
   * **Journal**: International Journal of Computer Science and Information Security
   * **Abstract**: This study compares KNN with other classification techniques such as decision trees, SVM, and neural networks for predicting customer behavior. It provides empirical results and discusses the advantages of KNN in certain contexts.
5. **"Machine Learning Techniques for Customer Behavior Analysis: A Comparative Study"**
   * **Authors**: A. Sharma, B. K. Saini
   * **Journal**: Journal of Information and Computational Science
   * **Abstract**: This paper offers a comparative analysis of various machine learning techniques, including KNN, for customer behavior analysis. It emphasizes the practical applications and performance metrics of these techniques.
6. **"Effective Customer Segmentation Using K-means and KNN in Big Data Environment"**
   * **Authors**: S. Patil, M. Kumar
   * **Journal**: International Journal of Recent Technology and Engineering (IJRTE)
   * **Abstract**: This paper discusses the application of K-means clustering followed by KNN classification for customer segmentation in a big data environment. It demonstrates how this combination can enhance customer behavior prediction.

**OBJECTIVE**

The primary objective of employing K-Nearest Neighbors (KNN) classification for predicting customer behavior is to accurately anticipate future actions based on historical data, fostering personalized interactions, segmentation understanding, and operational efficiency. KNN achieves this by identifying similarities between new and existing customers, facilitating tailored marketing strategies and product recommendations. Additionally, it enables the early detection of anomalies or changes in behavior, empowering businesses to intervene proactively. Overall, KNN serves to enhance decision-making processes, improve customer satisfaction, and drive business growth by leveraging data-driven insights into customer preferences and tendencies.

1. **Accurate Anticipation of Future Actions**: KNN classification enables businesses to predict future customer actions with precision by analyzing historical data. This accuracy is essential for making informed decisions and planning effective strategies.
2. **Fostering Personalized Interactions**: By identifying similarities between new and existing customers, KNN facilitates personalized interactions. This allows businesses to tailor marketing strategies, product recommendations, and communication channels to meet individual customer needs and preferences.
3. **Segmentation Understanding**: KNN aids in customer segmentation by categorizing the customer base into distinct groups based on behavior and preferences. Understanding these segments enables targeted marketing campaigns and product development tailored to specific customer needs.
4. **Operational Efficiency**: KNN's simplicity and ease of implementation contribute to operational efficiency. Businesses can quickly deploy KNN without extensive training phases or complex parameter tuning, allowing for rapid adaptation to new data and efficient decision-making.

**ADVANTAGES**

1. **Simplicity and Ease of Implementation**: KNN is straightforward to understand and implement. It does not require an extensive training phase or parameter tuning, making it accessible even for those with limited machine learning experience. Its simplicity allows for quick deployment and easy adaptation to new data without complex preprocessing or model adjustments.
2. **Non-parametric Nature**: KNN is a non-parametric algorithm, meaning it makes no explicit assumptions about the underlying data distribution. This makes it versatile and effective for a variety of data types and structures, including those with complex, nonlinear relationships. KNN can handle both classification and regression problems with minimal adjustments.
3. **Adaptability to New Data**: KNN classification can easily incorporate new data points without the need to retrain a model from scratch. As new data becomes available, it can be directly included in the existing dataset, and the algorithm can immediately use this data to make predictions. This adaptability makes KNN particularly useful in dynamic environments where data is frequently updated.

**DISADVANTAGE**

1. **Computational Intensity**: KNN can be computationally intensive, especially during the prediction phase. It requires calculating the distance between the query point and all points in the dataset, which can be slow and resource-consuming for large datasets.
2. **Sensitivity to Irrelevant Features**: KNN is sensitive to the presence of irrelevant or redundant features in the dataset. These can distort the distance calculations and negatively impact the accuracy of the model. Proper feature selection and normalization are crucial to mitigate this issue.
3. **Memory Usage**: KNN requires storing the entire dataset for making predictions, leading to high memory usage, particularly with large and high-dimensional datasets. This can make it impractical for applications where storage resources are limited.

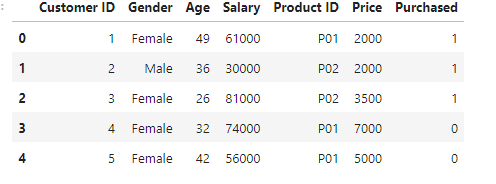
**EXISTING SYSTEM**

1. **Data Quality and Quantity**: High-quality, large datasets are crucial for accurate predictions. Inadequate or poor-quality data can lead to inaccurate results, affecting decision-making and customer experience.
2. **Complexity and Expertise**: Implementing and maintaining these systems require significant technical expertise in machine learning and data science, which can be costly and resource-intensive.
3. **Computational Resources**: Machine learning models, especially those involving large datasets and complex algorithms, demand substantial computational power, which can be expensive and require ongoing investment in hardware and software infrastructure.
4. **Data Privacy and Security**: Handling large volumes of customer data raises concerns about privacy and security. Ensuring compliance with regulations (e.g., GDPR) and protecting sensitive information are critical and challenging tasks.
5. **Bias and Fairness**: Machine learning models can inadvertently perpetuate or amplify biases present in the training data, leading to unfair or discriminatory outcomes. Identifying and mitigating bias is a complex and ongoing challenge.
6. **Integration with Existing Systems**: Integrating advanced machine learning models into existing business processes and systems can be complex and may require significant changes to workflows and infrastructure.
7. **Maintenance and Updates**: Machine learning models need regular updates and maintenance to remain accurate and relevant as customer behavior and market conditions change. This ongoing need for tuning and retraining can be resource-intensive.
8. **Interpretability**: Many machine learning models, particularly deep learning models, can act as "black boxes" with low interpretability, making it difficult for businesses to understand the reasoning behind predictions and build trust in the system's outputs.

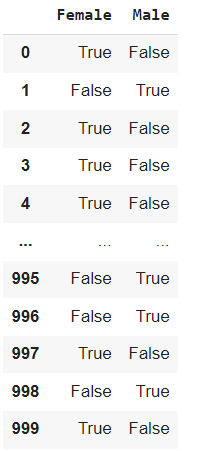
**CODE AND REVIEW**

**Training and Testing**

1. We loaded the dataset from a well-known repository of customer purchases

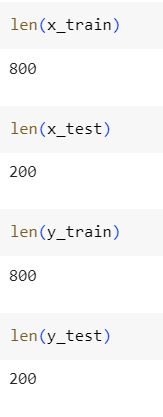
****

1. Onehot encoding for categorical data so it can be understood by the algorithm

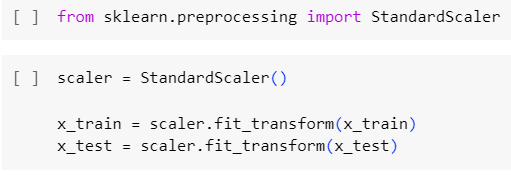


1. In the train and test split we are taking 80- 20 split for improving accuracy

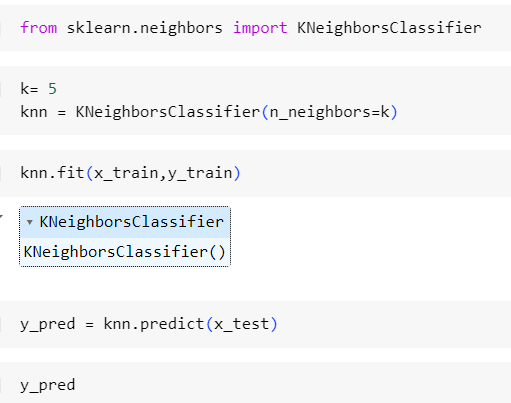


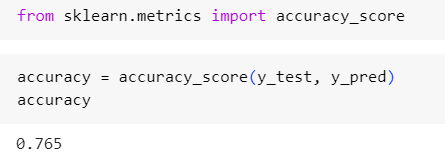


1. We then scale the features equally so no one feature gets biased more and the model can learn from all the features



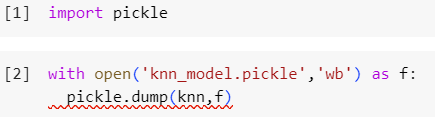
1. Using KNN algorithm to train the model

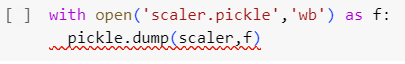


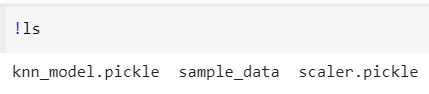


Accuracy of 76.5%

1. Storing the trained model using pickle so we can use it in any environment which is compatible with can run python



****

****

**Predicting**

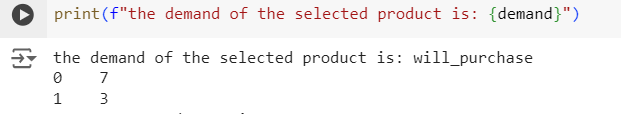
1. We first load the trained model

****

1. Predicting if the customer will buy or not

****

1. Now we will understand the demand of the product

****

7 people will not buy and 3 people will buy

**Reasons for choosing KNN model**

### 1. **Linear Regression / Logistic Regression**

#### Limitations:

* **Assumes Linearity**: These algorithms assume a linear relationship between the input features and the output. Customer behavior data often has complex, non-linear relationships, which these models might not capture effectively.
* **Over-simplification**: They may oversimplify the problem and fail to capture the nuances in customer behavior, leading to lower predictive accuracy.

### 2. **Support Vector Machines (SVM)**

#### Limitations:

* **Scalability**: SVMs can be computationally intensive, especially with large datasets, as they require solving a quadratic optimization problem.
* **Choice of Kernel**: The performance heavily depends on the choice of the kernel and the parameters, which can be difficult to tune for complex, real-world data.
* **Interpretability**: The decision boundary created by SVMs is often not interpretable, making it difficult to understand the customer behavior insights derived from the model.

### 3. **Decision Trees**

#### Limitations:

* **Overfitting**: Decision trees are prone to overfitting, especially when they become too deep and complex. This can lead to poor generalization on unseen data.
* **Instability**: Small changes in the data can lead to completely different tree structures, making the model unstable.

### 4. **Random Forests and Gradient Boosting Machines (GBM)**

#### Limitations:

* **Complexity**: These ensemble methods can create very complex models that are difficult to interpret. Understanding how they make specific predictions about customer behavior can be challenging.
* **Computational Cost**: While more robust than individual decision trees, ensemble methods can be computationally expensive and require more resources for training and prediction.

### 5. **Neural Networks**

#### Limitations:

* **Data Requirements**: Neural networks typically require large amounts of data to perform well. In cases where data is limited, they might not be the best choice.
* **Black Box Nature**: Neural networks are often considered black boxes due to their lack of interpretability. This makes it hard to derive actionable insights from the model regarding customer behavior.
* **Computational Resources**: They require significant computational power and resources, making them less practical for real-time or resource-constrained environments.

### 6. **K-Means Clustering**

#### Limitations:

* **Assumes Spherical Clusters**: K-Means assumes that clusters are spherical and equally sized, which may not be true for customer behavior data.
* **Sensitivity to Initial Centroids**: The algorithm is sensitive to the initial placement of centroids, which can affect the quality of the clustering.
* **Fixed Number of Clusters**: K-Means requires specifying the number of clusters beforehand, which might not always be known and can lead to suboptimal clustering.

**PROPOSED SYSTEM**

1. Data Quality and Quantity: While high-quality data is still important for KNN, it is often less sensitive to data quantity compared to more complex models. KNN can work effectively with smaller datasets, which can simplify data collection and preprocessing requirements.
2. Complexity and Expertise: KNN is straightforward and easy to implement, requiring less technical expertise compared to more complex machine learning algorithms. This reduces the cost and resource burden associated with deploying and maintaining the system.
3. Computational Resources: Although KNN can be computationally intensive during the prediction phase, it does not require a training phase, which saves computational resources upfront. For smaller datasets or applications where real-time predictions are not critical, the computational demands are more manageable.
4. Data Privacy and Security: Since KNN is simple to implement and does not require extensive data preprocessing or transformation, it can be easier to enforce data privacy and security measures. Moreover, because KNN can work with smaller, more relevant datasets, there is less risk of exposing large volumes of sensitive customer data.
5. Bias and Fairness: KNN's non-parametric nature means it does not make strong assumptions about the data, potentially reducing some biases that more complex models might introduce. However, careful feature selection is still necessary to mitigate biases present in the data.
6. **Integration with Existing Systems**: Due to its simplicity, KNN can be more easily integrated into existing systems without requiring significant changes to workflows or infrastructure. It can act as a straightforward addition to enhance existing data analytics capabilities.
7. **Maintenance and Updates**: KNN does not require regular retraining as it is instance-based. This reduces the maintenance burden since the model simply uses the most current data available for making predictions, adapting naturally to changes in customer behavior.
8. **Interpretability**: KNN is highly interpretable because it is easy to understand why a particular prediction was made by examining the nearest neighbors. This transparency helps businesses build trust in the model and make informed decisions based on its outputs.

**SYSTEM REQUIREMENTS SPECIFICATION**

**Hardware required:**

* Computer system(laptop)
* Corei5 with 8 GB of Ram and 2.8 GHz processor speed
* Power supply
* Software required:

**Operating system:**

* Windows (Any operating system)
* Jupiter Notebook, PyCharm
* Collab, 1. TensorFlow
* Python modules like Pandas, Flask framework, Notebook, PyCharm, sklearn, numpy, matplotlib
* Web technology languages and javascript framework like react
* Dataset collected from GitHub, Kaggle, ChatGPT.