**Customer Segmentation Using Data science & Machine Learning**

**A MICRO PROJECT REPORT**

###### ***Submitted by***

**AMBATI VINOD**

**99210041501**

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

IN

**COMPUTER SCIENCE AND ENGINEERING**

****

**SCHOOL OF COMPUTING**

**COMPUTER SCIENCE AND ENGINEERING**

**KALASALINGAM ACADEMY OF RESEARCH**

**AND EDUCATION**

KRISHNANKOIL 626 126

Academic Year 2024-2025

**DECLARATION**

We affirm that the project work titled **“Customer Segmentation using data science & machine learning”** being submitted in partial fulfillment for the award of thedegree of **Bachelor of Technology in Computer Science and Engineering** is the original work carried out by us. It has not formed the part of any other project work submitted for award of any degree or diploma, either in this or any other University.

**AMBATI VINOD - 99210041501**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Signature of the Mentor

Ms. VS. Vetri Selvi

Assistant Professor

Department of Computer Science and Engineering

A person wearing sunglasses

Description automatically generated with low confidence

**BONAFIDE CERTIFICATE**

Certified that this project report **“Customer segmentation using data science & machine learning”** is the bonafide work of “**AMBATI VINOD”** who carried out the project work under my supervision.

**SUPERVISOR HEAD OF THE DEPARTMENT**

Ms. VS. Vetri Selvi Dr. N. Suresh Kumar

Assistant Professor Professor & Head

Computer Science and Engineering Computer Science and Engineering

Kalasalingam Academy of Research Kalasalingam Academy of Research and and Education and Education

Krishnankoil 626126 Krishnankoil 626126

Submitted for the Micro Project Viva-voce examination held on ………….

**Internal Examiner External Examiner**

**ABSTRACT**

Now a days, there is a lot of increase in the number of shopping malls. Due to this there is a huge competition in the business sector. In order to survive in this competition malls can use the machine learning models .These models can be used for the purpose of customer segmentation .Customer segmentation using machine learning can be an easy way to identify the needs of different category of customers of mall . In this project we use K-means clustering to segment the customers in to various categories and visualize the data.

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **ABSTRACT** | | |
| **LIST OF FIGURES** | | |
| **LIST OF ABBERIVATION** | | |
| **CHAPTER I** | **INTRODUCTION……………………………………….………....13** |
| **CHAPTER II** | **LITERATURE REVIEW ………………………………...………22** |

**CHAPTER III PROBLEM DEFINITION……………………………………………………24**

**CHAPTER IV REQUIREMENTS ……………………………………………………………25**

4.2 Hardware requirements

4.3 Software requirements

**CHAPTER V SYSTEM DESIGN…………………………………………………………26**

5.1 Dataflow diagram

**CHAPTER VI PROPOSED APPROACH ……………………....................................32**

**CHAPTER VII MODULE DESCRIPTION………………………………………….37**

**CHAPTER VIII IMPLEMENTATION AND RESULT..………………………..........40**

8.1 Coding

8.2 Output screenshot

8.3 Experiment Result

**CHAPTER IX CONCLUSION AND FUTURE ENHANCEMENT………………...47**

9.1 Conclusion

9.2 Future Enhancement

**CHAPTER X REFERENCES……………………………………………………….....48**

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **FIGURES** | **DETAILS** |
| Figure 1 | Fig-1-Customer segmentation |
| Figure 2 | Fig 2 –Data flow diagram |
| Figure 3 | Fig-3 Flow diagram of K-Means algorithm |
| Figure 4 | Fig 4-Spending score vs Annual income |

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| **Abbreviation** | **Full form** |
| **WCSS** | Within Cluster Sum of Square |
| **DBSCAN** | Density-based spatial clustering of applications with noise |
| **OPTICS** | Ordering Points To Identify the Clustering Structure |

**I.INTRODUCTION**

Machine learning plays a key role in improving the standards of the business. Customer segmentation is the process of dividing the customers into different groups depending upon their similar characteristics. Customer segmentation can be used to target the specific customers who require special attention and also know about the regular customers who frequently visit the mall. Customer segmentation is a powerful marketing tool that helps you understand your customers better, and target them accordingly. It can help you improve your sales, increase customer retention and reduce costs associated with customer acquisition. Statistics show that segmentation makes companies 60% more likely to understand customer challenges and 130% more likely to know their intentions. Segmented email campaigns also have higher open rates, click-through rates and conversions..One of the popular methods used for customer segmentation is K-means clustering. K-means clustering is the unsupervised learning algorithm used to cluster data points into groups based on their similarities. In this project we apply algorithm to a dataset which consists of Customer id, age, gender, annual income and spending\_ score as its attributes. The main goal of this project is to divide the customers into various groups based on their similarities and plan efficient strategies to meet their needs .Overall this project highlights the benefits of customer segmentation and improves the customer experience.



FIG-1-Customer segmentation

**II. LITERATURE REVIEW**

**3.1.Comparative Study of K-means and Hierarchical Clustering Techniques by Dr.Manju Koushik and Mrs . Bhawana Mathur(2016 )-**

* This paper compares the k-means and hierarchical clustering techniques for customer segmentation process by calculating their strengths and weaknesses.
* They prove that K-means clustering technique is more efficient than hierarchical clustering in case of customer segmentation.

**3.2 N. Sharma .“Comparison the various clustering algorithms of weka tools”**

* This paper compares different clustering algorithms.
* The author uses the weka tool for the comparison.
* The comparison includes DBSCAN, EM, Farthest First, OPTICS, and the K-Means algorithms. In this study the benefits of all algorithms are shown.
* Finally they discover that the K-means clustering is the most efficient among all of them.

**3.3. U. Kaymak and M. Setnes, “Extended fuzzy clustering algorithms” .**

* In this paper, authors used the fuzzy algorithm to categorize the datasets into the clusters.
* They also illustrated some of the issues using the fuzzy algorithm technique such as choosing the number of clusters in the data, shape of the clusters etc. T
* This report proposes two extensions of the fuzzy algorithm to overcome the issues. Firstly, the prototypes are extended to hyper volumes whose size is determined automatically from the data being clustered.
* Secondly the similar clusters are combined while clustering in order to make the partition of data.

**III. PROBLEM DEFINITION**

The existing methods for storing the customer data are through paper work and through software in a digital manner. These methods take are time consuming when it comes to the customer segmentation part. This process can be made easy by using the machine learning techniques. This can help in quick and easy customer data analysis.

**IV. SYSTEM REQUIREMENTS:**

4.1 HARDWARE REQUIREMENTS:

* + Processor : Pentium IV
  + RAM : 1 GB
  + Hard Disk : 40 GB
  + Server : XAMPP Server

4.2 SOFTWARE REQUIREMENTS

* + Front End : Html, CSS
  + Middleware : PHP, JavaScript
  + Back End : MYSQL
  + Operating System : Windows 7,8

**V. SYSTEM DESIGN**

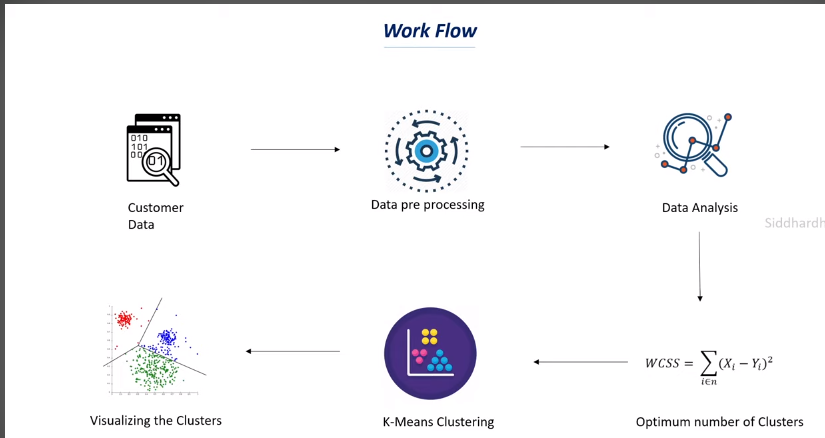
****

Fig 2 –Data flow diagram

**VI.PROPOSED APPROACH**

To overcome the problems of the traditional system we use machine learning techniques. In our project we use K-means clustering for grouping the data into different clusters.

The proposed system consists of following steps-

* Data collection
* Data pre processing
* Data analysis
* Clustering technique
* Data Visualization

**Collection of data**: In the first step we collect the customer data. The data contains Customer id, age, gender ,annual income and spending score as its attributes.

**Data preprocessing**: It refers to transformation, integration and cleaning of data to make it ready for analysis.

**Data Analysis**: It refers to the transformation of collected data into modeled data by the discovery of the useful information from the data.

**Clustering technique:** K-means clustering is applied on the data for the segmentation purpose and the resultant data is visualized in the form of graph.

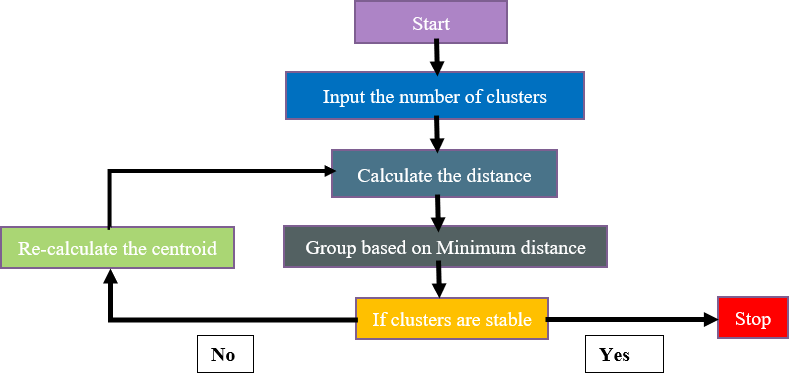
The steps involved in k-means clustering are

Step-1: Specifying the number of clusters to be generated as K.

Step-2: Randomly select K data points and assign each data point to a cluster.

Step-3: Computation of the clusters.

Step-4: Keep iterating the above steps until the clusters are stable without change.

****

**Fig-3 Flow diagram of K-Means algorithm**

**VII. MODULE DESCRIPTION**

**7.1 Importing the libraries:**

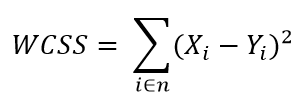
Firstly , we import the numpy, pandas, matplotlib.pyplot , seaborn and sklearn.cluster libraries.

**7.2. Data collection and analysis:**

Secondly, the data from csv file is loaded into the pandas data frame .The rows of the data are displayed and also checked for the missing values.

**7.3. Choosing the number of clusters and plotting the elbow graph:**

We use WCSS (Within-Cluster Sum of Square) method to choose the number of clusters. WCSS is used to find the sum of the squared distance between each point and the cluster centroid. On plotting the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value decreases.



**7.4. Training the K-means clustering model**

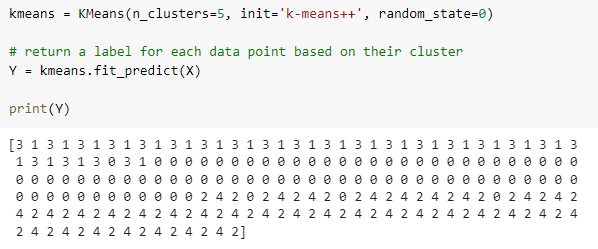


Fig4-Applying k-means algorithm to data

**7.5. Visualization of clusters and Plotting Annual Income vs Spending score**

In this part, we plot all the clusters and their centroids . Finally we get an Annual Income vs Spending score scatter plot.

**VIII. IMPLEMENTATION AND RESULT**

**8.1 Code-**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

customer\_data = pd.read\_csv('/content/Mall\_Customers.csv')

customer\_data.head()

customer\_data.shape

customer\_data.info()

customer\_data.isnull().sum()

X = customer\_data.iloc[:,[3,4]].values

print(X)

wcss = []

for i in range(1,11):

  kmeans = KMeans(n\_clusters=i, init='k-means++', random\_state=42)

  kmeans.fit(X)

  wcss.append(kmeans.inertia\_)

sns.set()

plt.plot(range(1,11), wcss)

plt.title('The Elbow Point Graph')

plt.xlabel('Number of Clusters')

plt.ylabel('WCSS')

plt.show()

kmeans = KMeans(n\_clusters=5, init='k-means++', random\_state=0)

Y = kmeans.fit\_predict(X)

print(Y)

plt.figure(figsize=(8,8))

plt.scatter(X[Y==0,0], X[Y==0,1], s=50, c='green', label='Cluster 1')

plt.scatter(X[Y==1,0], X[Y==1,1], s=50, c='red', label='Cluster 2')

plt.scatter(X[Y==2,0], X[Y==2,1], s=50, c='yellow', label='Cluster 3')

plt.scatter(X[Y==3,0], X[Y==3,1], s=50, c='violet', label='Cluster 4')

plt.scatter(X[Y==4,0], X[Y==4,1], s=50, c='blue', label='Cluster 5')

plt.scatter(kmeans.cluster\_centers\_[:,0], kmeans.cluster\_centers\_[:,1], s=100, c='cyan', label='Centroids')

plt.title('Customer Groups')

plt.xlabel('Annual Income')

plt.ylabel('Spending Score')

plt.show()

**8.2 Screenshot:**

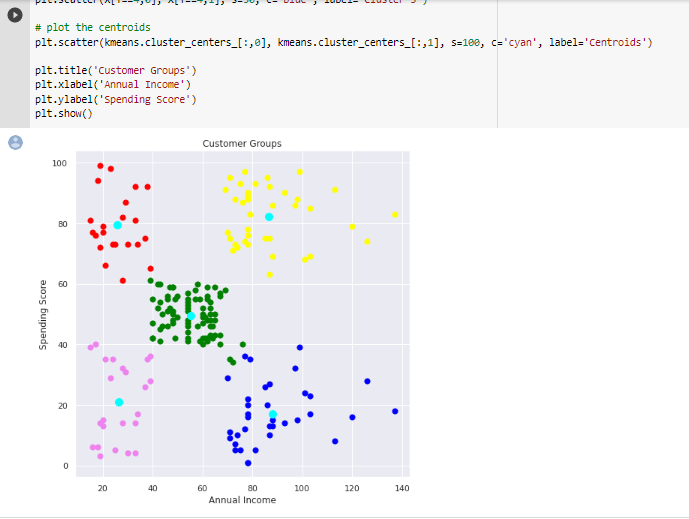
****

Fig 4-spending score vs annual income

**8.3 Experimental Result:**

The graph, the red color clusters denotes the people who has high spending score and low annual income. The violet color denotes the customers with the low annual income and low spending scores. The blue color indicates the people with low spending score and high income .The yellow color indicates the people with high income and high spending score whereas the green denotes the customers with average income and average spending score.

**IX. CONCLUSION AND FUTURE ENHANCEMENT**-

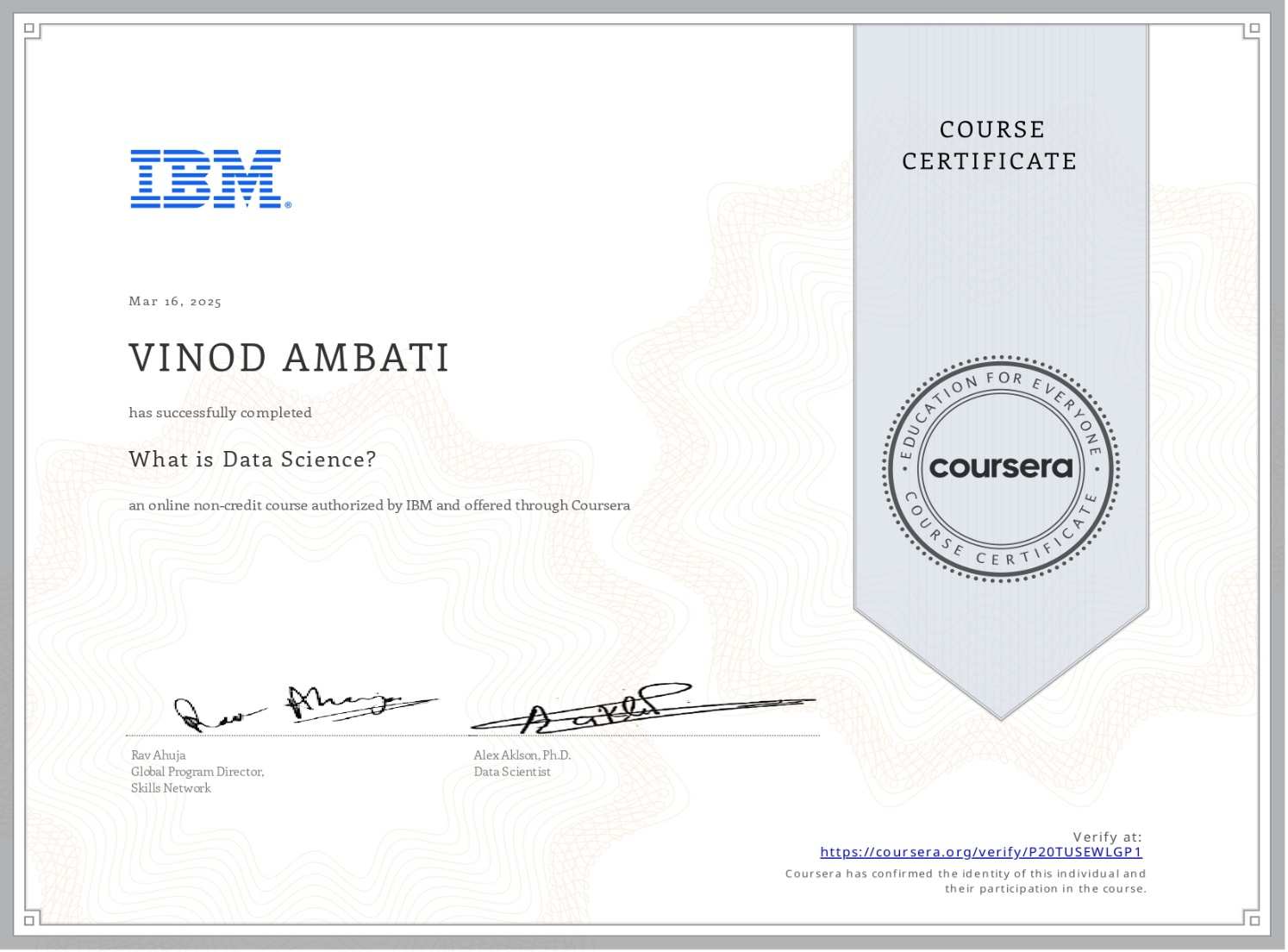
**9.1 Conclusion-**

From the graph the mall owners can analyze the low spending score people and increase the offers for them to attract them towards their mall. They can also easily identify the financial status of their majorly visiting customers and the products to be sold can be easily chosen. This project can make the customer segmentation a very easy task and also helps us to know the customer needs

**9.2 Future Enhancement-**

Deep learning is a new emerging area which exploits artificial intelligence and machine learning to learn features directly from the data, using multiple nonlinear processing layers. Deep learning models can achieve very high accuracy in Customer segmentation.

**Course certification**

****

**X. REFERENCES**

[1]Comparative Study of K-means and Hierarchical Clustering Techniques by Dr.Manju Koushik and Mrs . Bhawana Mathur (2016)

[2]U. Kaymak and M. Setnes, “Extended fuzzy clustering algorithms”

[3]N. Sharma .“Comparison the various clustering algorithms of weka tools”

[4] Bhatnagar, Amit,Ghose, S. (2004), ‘A latent class segmentation analysis of e-shoppers’, Journal of Business Research 57, 758–767.

[5]Vaishali R. Patel and Rupa G. Mehta “Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm”, IJCSI,Year: 2011.



**INTERNAL QUALITY ASSURANCE CELLMICRO PROJECT AUDIT REPORT**

This is to certify that the micro project work entitled “**Customer Segmentation using data science & machine learning**” categorized as an internal project done by AMBATI VINOD of the Department of Computer Science and Engineering, under the guidance of Ms.VS.Vetri Selvi during the Even semester of the academic year 2024 - 2025 are as per the quality guidelines specified by IQAC.

**Quality Grade**

**Deputy Dean (IQAC)**

**Administrative Quality Assurance Dean (IQAC)**

**APPENEDIX**

**(Project Code)**