IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

PRODUCT RECOMMENDATION SYSTEM

YESHIV SAHU, PRANSHU SHUKLA , AVINASH KASHYAP, SHUBHAM KUMAR

B. TECH, B. TECH, B. TECH, B. TECH

COMPUTER SCIENCE & ENGINEERING

BHILAI INSTITUTE OF TECHNOLOGY RAIPUR, RAIPUR, CHHATTISGARH, INDIA

ABSTRACT:

Today's consumers are exposed to a wider range of products and information than ever before. This results in rising "consumer" demand, creating a new challenge for retailers to meet client preferences with the appropriate products. A technique to meet this difficulty is the recommender system, which allows for product recommendations that meet consumer needs and expectations while also luring in newcustomers. However, the huge size of the transactional database typical of retail businesses reduces efficiency and quality of product. To address these issues, we employ content-based filtering and data mining techniques in this paper. The recommendation algorithm starts to obtain a similar group of customers using customer data. The algorithm was tested with data from a chain of perfumeries. The experimental finding demonstrates its ability to recognize the consumer category to which a userbelongs and how helpful this is for online purchasing and shopping.

Recommender system are the tools to overcome the problems of demand of customers by fulfilling their needs and challenges of retail industries to recommend the customer the correct product based on the customer's preferences. In this paper we have discussed about data mining techniques that has been used in product recommendation system. The data mining techniques used in product recommendation system are linear regression, content-based filtering, KNN algorithm. Recommendation system profits the customers by making him the suggestions about the items he is doubtless to buy it.

Keywords – Product Recommendation System, Data mining , Content-based filtering, Linear regression KNN algorithm.

1. INTRODUCTION:

***1.1 Customer Behavior*:** Customers participate in cross-category decision-making. In a retail setting, a multi-category decision-making process leads to a shopping basket that contains the assortment of goods that customers buy during a single store visit. Online retailers have a history of being interested in learning the make-up of their customers' market baskets because this information may be used to develop micromarketing campaigns and targeted cross-selling initiatives. Technologies like recommender systems help businesses implement theseplans. To boost the likelihood of cross-selling, build customer loyalty, and fulfill consumers' needs by identifying products they might be interested in, recommender systems are used in e-commerce.

1.2 Data mining: Data mining is a technique for extracting useful information and patterns from large amounts of data. It is a very useful technique and is important as the data is increasing day by day in our daily lives, and to maintain this increasing data and data storage, this technique is useful. It is also known as the "knowledge discovery process," which includes the following steps: data integration, data cleaning, data selection and transformation, data mining, pattern evaluation, and knowledge representation. The goal of this technique is to find unknown patterns and use them to make certain decisions once they are found. For knowledge discovery and recommendation systems, various data mining algorithms and techniques (such as classification, clustering, regression on decision trees, and so on) are used.

1.3 Recommendation System: Recommendation systems are used in hundreds of different services-everywhere from online shopping tomusic/movies and many. These recommendation systems have become the integral part of the web applications. This is mainly because the e- commerce websites have a large variety of product inventory’s available the data which is to be displayed to the user can be overwhelming. As a result the user may face difficulties to search for the product that they may be looking for. At this point the recommendation system’s comes to usage. In order to increase the sale of the product or to increase the efficiency in the market recommender system is needed. This system will require large amount of information in order to make correct decision. The information which is provided to the recommender system must be consistent in nature. The recommender system will take the information and formulate the decision in one of the following two ways- either by the use of collaborative filtering or by the use of content filtering.

1.4 Content-based filtering: Content-based filtering is based on the content or item-item relationship. In this filtering, a relationship is formed between the group of items. When a customer searches for an item from the group of items, based on their interest in the search, similar items are recommended to that customer. Based on the features or attributes associated with the compared items, the similarity of the items is calculated and matched with the customer's search history. For example, if a customer searches on Amazon for Samsung mobile phones, the content-based recommender system will learn the customer's preferences and recommend or display related items based on the customer's search history.

1.5 Regression: Regression is based on supervised learning. Regression creates a goal prediction value based on independent variables. It is mostly used to determine how variables and forecasting relate to one another. Regression models vary depending on the quantity of independent variables they use as well as the type of relationship they take into account between the dependent and independent variables.

1.6 Linear regression: is a statistical method for predictive analysis and works on continuous (real) or numeric variables such as sales, salary, product price, etc. The linear regression is used to find the relationship between the dependent variable and one or more independent variables; hence, it is called the "linear regression." For representing the relationship between the variables, it provides a sloping straight line.

1.7 KNN algorithm: The K-Nearest Neighbor algorithm, which falls under supervised learning, is a simple and effective method for filtering datasets. A labeled dataset is divided into different clusters in this case, with similar items having similar properties. In this algorithm, the number of clusters is denoted by K. This algorithm supports a variety of filtering methods, including Euclidean distance, cosine similarity, Pearson correlation, and others.

2. LITERATURE REVIEW:

In this section we will discuss the related works :

A. Raich, B. Ganguly, and M. Tota (2019) described the relatively simple implementation of machine learning. Clear methods for identifying consumer buying habits and transforming them into actionable strategies. Any information, no matter how minor, for online or other retailers would be greatly appreciated. They talked about using market baskets in various circumstances and using machine learning to make connections.[1]

B. A mining algorithm for frequent patterns in noisy data was put forth by Abboud et al. in 2019. The noise in data is too much for traditional mining algorithms to handle. As a result, two strategies are used: noisy data mining and the C3Ro pattern mining algorithm. C3Ro has some benefits, such as less memory usage and the ability to mine large, noisy databases. Comparisons were made between the experimental results and algorithms like CCSpan and CM-Clasp.[2]

C. A pattern mining approach was put forth by Belhad et al. (2020) to address the hashtag retrieval issue on Twitter. With the aid of temporal and trivial transformation techniques, the algorithm converts thecollection of tweets into a database. Patterns are then found to locate the tweets that users have requested using a similarity search method. The simulated results with huge tweets showed improved run time,memory, and F-measure results.[3]

D. Jyoti and Er. Amandeep Singh Walia (2017) created a recommendation system based on the user's navigational patterns and provideappropriate recommendations to meet the user's current needs by usingK-Nearest Neighbor and data mining techniques.[4]

E. M. A. Ula's research has focused on the application of data mining to theanalysis of market baskets. Here, they discussed their multipleapplications and provided examples. They desired to group the consumers based on how they typically make purchases.[5]

F. Mehrbakhsh Nilashi provided an overview of data mining techniques applied to the creation and execution of recommendation systems. He discussed various methods of data mining used throughout recommendation systems, such as classification, clustering, and prediction methods.[6]

G. Jatin Sharma, Kartikay Sharma et.al (2021) examined the various mechanisms and techniques required for recommender systems to recommend products or items in the domains of fashion and books. Similarly, Deuk Heepark et al. reviewed 210 articles on recommendation systems from 46 different journals to identify trends in the recommender systems.[7]

H. In order to find an anomaly in spatiotemporal data from the flow distribution probability (FDP) databases, Djenouri et al. (2019) postulated the K Nearest Neighbor method. The KL divergence distance measurement is used to examine how similar the sequences are to one another. The KNN algorithm and FDP are used to identify outliers. The suggested method is used to mine a set of traffic flow data to find outliers.[8]

I. The authors of this paper, Abhishek Saxena and Navneet Gaur (2019) developed a recommender system for an e-commerce application that uses a personalised information retrieval method to identify sets of products that will be of interest to users. The main principle of the recommendation stipulates that any product set that commonly occurs together must require that each item listed (or other set of items) take place at least as regularly.[9]

J. A new recommendation system for generating recommendations based on user ratings and personal profiles was proposed by Vignesh Thannimalai and Li Zhang (2021). In addition to incorporating the content-based filtering algorithm with the Nave Bayes classifier, they proposed item-based collaborative filtering to recommend tourist attractions based on user ratings.[10]

K. Freyne and Berkovsky (2010) were able to forecast the performance index for a goal recipe using a content-based algorithm (CB) based on the appropriate information of the ingredients used in those recipes. [11]

L. Umair Javed and Kamran Shaukat (2021) gave a review of the content- based and context-based recommendation systems. They proposed a semantic recommender framework for e-learning methods in a content- based recommendation system, allowing learners to discover and select appropriate learning materials for their field.They also proposed a few ontologies related to the content-based framework, which can improve the functionalities of an instructional system. [12]

M. Using a content-based methodology, Kundan S. Rana proposed a recipe recommender system application to assist users in discovering their favourite foods and their nutritional value. Unit testing was used by the author to confirm the strategy. Letizia utilised content-based filtering to anticipate which pages a user might be fascinated with based on his activities through the sites.[13]

N. Balraj Kumar and Neeraj Sharma proposed a paper on approaches, issues, and challenges in recommender systems. The purpose of this article was to present a comprehensive and thorough review of current recommender systems. The methodology used in this case consists of a search plan and paper selection criteria. The search strategy attempts to retrieve research studies from multiple digital libraries, and the paper setof criteria assists in further filtering to select the most related research tocollect information against the various research questions.[14]

3. METHODOLOGY:

The recommendation system is built by the techniques namely Content based filtering , linear regression ,k nearest neighbor. In the suggested system, users interact with a web portal, providing data that is then recorded in a raw log file after going through a number of preprocessing, data cleaning, and other procedures to separate the useful information fromthe raw data and put it into a structured format.After that, the clean data is used to find patterns and recommend the top n products to each user. The system architecture is depicted below in Fig.3.1.

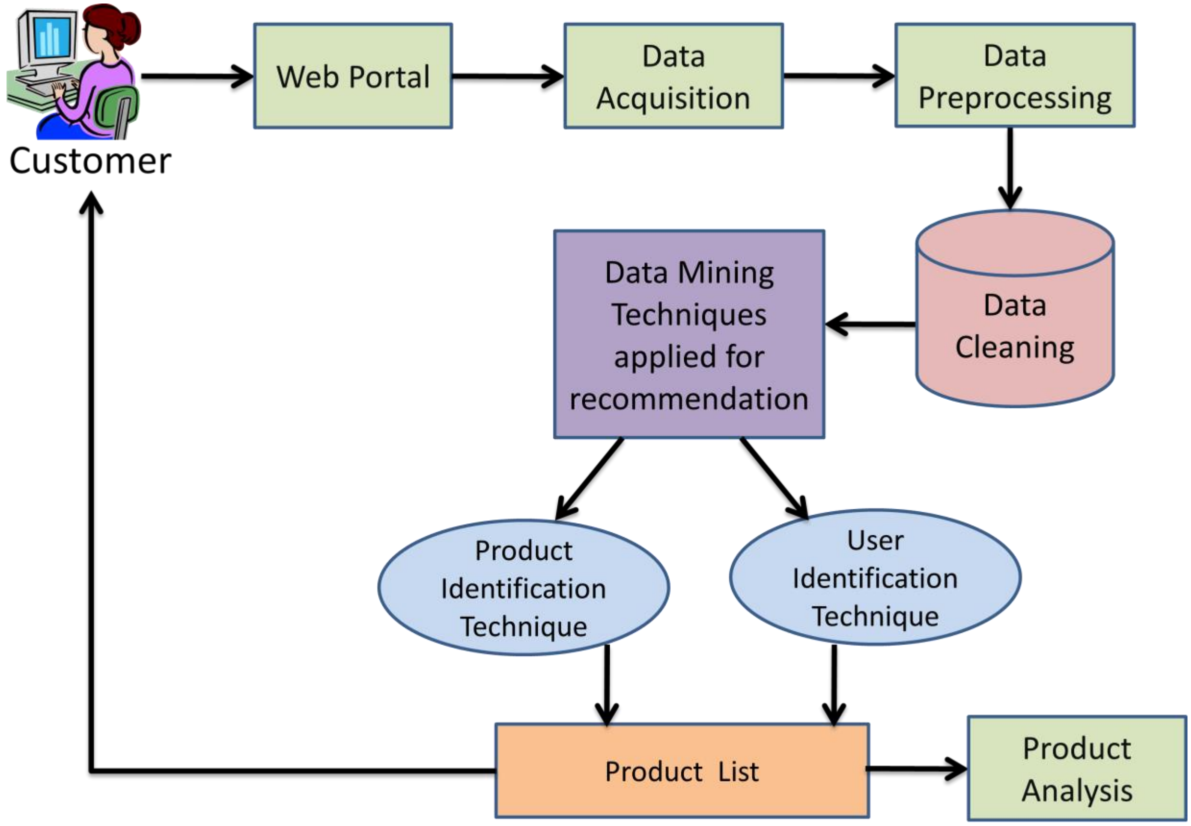

Customer
Web Portal
Data Acquisition
Data Preprocessing
Data Cleaning
Data Mining Techniques applied for recommendation
Product Identification Technique
User Identification Technique
Product List
Product Analysis

Fig.3.1 System Architecture

3.1 Proposed work as follow:

Data acquisition - Extraction of relevant information, data transformation, and loading of the data into the target program are all steps in the data acquisition process.

Data preprocessing - Data preprocessing is a technique which is used to transform the raw data in a useful and efficient format

Data Cleaning- data that is collected during acquisition phase may havenull value, missing data etc so those inaccurate record must be correctedto perform data mining techniques.

Data mining- After data cleaning the data mining techniques are appliedon it .

Here KNN (K Nearest-Neighbor) algorithm is applied on it.

The K-NN working can be explained on the basis of the below steps:

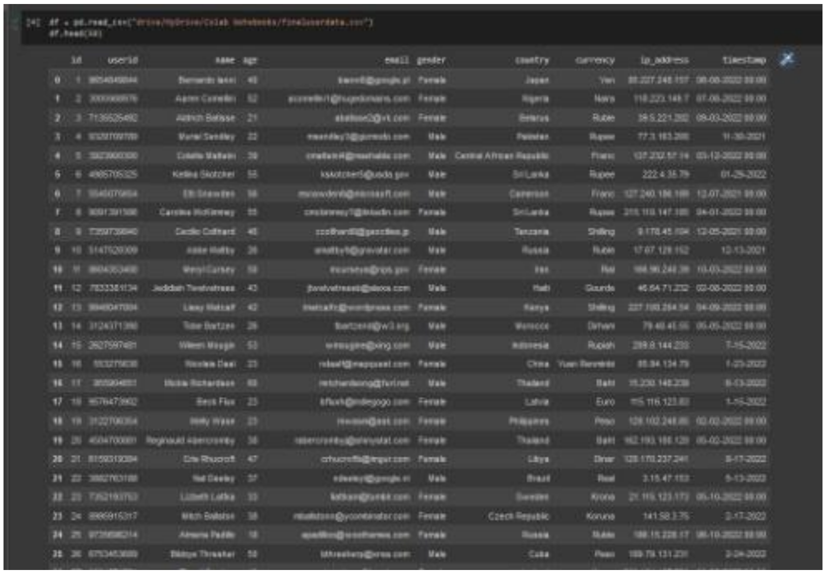
- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the Calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the datapoints in each category.
- **Step-5:** Assign the new data points to that category for which thenumber of the neighbor is maximum.

- **Step-6:** Our model is ready.

4. *RESULT & DISCUSSION:*

Python is used to develop a recommendation analysis, and the id for executing code is anaconda prompt. The online shopping portal provides raw data. We are taking two datasets: one is finaluserdata.csv which contains the userid, name, age, and email, gender, country, currency, ip address, and timestamp are all variables. and the other is productdata.csv which contains the productid, agestart, ageend, and category.

- Userid: It indicates the ID of particular user.
- Name: It provides the name of the user.
- Age: It is the age of the particular user .
- Email: It provides the email used by the particular user.
- Gender: Gender can be male or female.
- Country: Country provides the location of the user.
- Currency: Currency provides the type of currency used by the user.



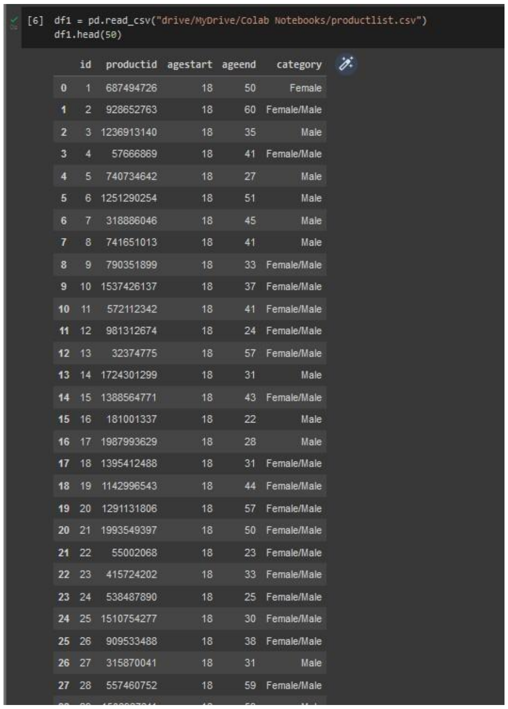
```
[4] df = pd.read_csv("drive/MyDrive/Colab Notebooks/finaluserdata.csv")
    df.head(30)
```

	id	userid	name	age	email	gender	country	currency	ip_address	timestamp
0	1	[illegible]	Bernardo Iassi	40	[illegible]@google.pl	Female	Japan	Yen	88.227.248.157	08-08-2022 00:00
1	2	[illegible]	Aaron Cornellin	52	acornellin1@hugedomains.com	Female	Nigeria	Naira	118.223.148.7	07-08-2022 00:00
2	3	7138625492	Aldrich Batisse	21	abatisse2@vk.com	Female	Belarus	Ruble	39.5.221.202	09-03-2022 00:00
3	4	[illegible]	Muriel Sandley	22	msandley3@gizmodo.com	Male	Pakistan	Rupee	77.3.183.200	11-30-2021
4	5	[illegible]	Colette Maltain	39	cmaltain4@mashable.com	Male	Central African Republic	Franc	137.232.57.14	03-12-2022 00:00
5	6	4886705325	Kellina Skotcher	55	kskotcher5@usda.gov	Male	Sri Lanka	Rupee	222.4.35.79	01-29-2022
6	7	[illegible]	Elli Snowden	58	[illegible]@microsoft.com	Male	Cameroon	Franc	127.240.186.188	12-07-2021 00:00
7	8	[illegible]	Carolina McKenney	55	cmckenney7@linkedin.com	Female	Sri Lanka	Rupee	215.110.147.180	04-01-2022 00:00
8	9	[illegible]	Cecilio Colthard	46	ccolthard8@geocities.jp	Male	Tanzania	Shilling	9.178.45.104	12-05-2021 00:00
9	10	5147520309	Addie Maltby	38	amaltby9@gravatar.com	Male	Russia	Ruble	17.67.128.152	12-13-2021
10	11	[illegible]	Meryl Carsey	59	mcarseya@nps.gov	Female	Iran	Rial	166.96.248.38	10-03-2022 00:00
11	12	7833381134	Jedidiah Twelvetrees	43	jtwelvetreesb@[illegible].com	Male	Haiti	Gourde	46.64.71.232	02-08-2022 00:00
12	13	[illegible]	Lissy Metcalf	42	lmetcalfc@wordpress.com	Female	Kenya	Shilling	[illegible]	04-09-2022 00:00
13	14	3124371386	Tobe Bartzen	26	tbartzend@w3.org	Male	Morocco	Dirham	79.48.45.55	05-05-2022 00:00
14	15	2627597481	Willem Meggie	53	wmeggiee@bing.com	Male	Indonesia	Rupiah	299.8.144.233	7-15-2022
15	16	[illegible]	Nicolas Deal	23	ndealf@mapquest.com	Female	China	Yuan Renminbi	85.94.134.79	1-23-2022
16	17	[illegible]	Mollie Richardson	60	mrichardsong@[illegible].net	Male	Thailand	Baht	15.230.148.238	8-13-2022
17	18	[illegible]	Beck Flux	23	bfluxh@indiegogo.com	Female	Latvia	Euro	115.116.123.83	1-16-2022
18	19	3122708364	[illegible]	23	[illegible]@aol.com	Female	Philippines	Peso	128.102.248.86	02-02-2022 00:00
19	20	[illegible]	Reginauld Abercromby	36	rabercrombyj@shinystat.com	Female	Thailand	Baht	162.193.186.120	05-02-2022 00:00
20	21	[illegible]	Cris Rhucroft	47	crhucroftk@imgur.com	Female	Libya	Dinar	128.170.237.241	8-17-2022
21	22	3882763188	Nel Deeley	37	ndeeleyl@google.ru	Male	Brazil	Real	3.15.47.153	5-13-2022
22	23	7352193753	Lizbeth Lafka	33	llafkam@tumblr.com	Female	Sweden	Krona	21.115.123.173	05-10-2022 00:00
23	24	8869915317	Mitch Balstone	38	mbalstonen@ycombinator.com	Female	Czech Republic	Koruna	141.58.3.75	2-17-2022
24	25	[illegible]	Almeria Paddle	18	apaddleo@[illegible].com	Female	Russia	Ruble	188.15.228.17	06-10-2022 00:00
25	26	[illegible]	Bibbye Thrasher	52	bthrasherp@[illegible].com	Male	Cuba	Peso	169.79.131.231	2-24-2022

Fig. 4.1 user data

The product data set consist of following attributes:

- Productid: It indicates the ID of particular product.
- Agestart and Ageend : It provides the info. about the product that isliked by the users beween the ages mentioned in those attributes.
- Category: Provides the info of user liking the product whether male orfemale.

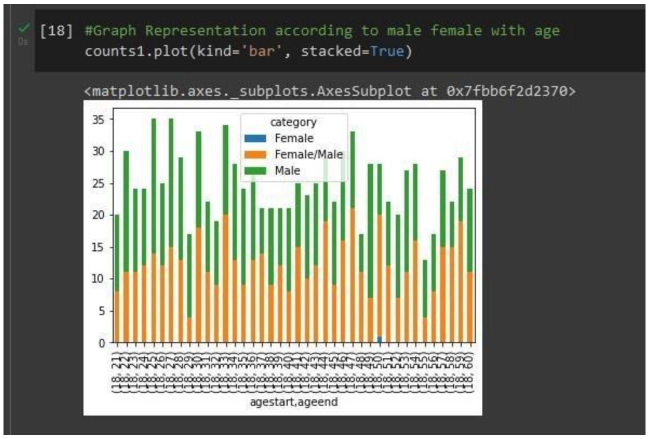


```
[6] df1 = pd.read_csv("drive/MyDrive/Colab Notebooks/productlist.csv")
    df1.head(50)
```

	id	productid	agestart	ageend	category
0	1	687494726	18	50	Female
1	2	928652763	18	60	Female/Male
2	3	1236913140	18	35	Male
3	4	57666869	18	41	Female/Male
4	5	740734642	18	27	Male
5	6	1251290254	18	51	Male
6	7	318886046	18	45	Male
7	8	741651013	18	41	Male
8	9	790351899	18	33	Female/Male
9	10	1537426137	18	37	Female/Male
10	11	572112342	18	41	Female/Male
11	12	981312674	18	24	Female/Male
12	13	32374775	18	57	Female/Male
13	14	1724301299	18	31	Male
14	15	1388564771	18	43	Female/Male
15	16	181001337	18	22	Male
16	17	1987993629	18	28	Male
17	18	1395412488	18	31	Female/Male
18	19	1142996543	18	44	Female/Male
19	20	1291131806	18	57	Female/Male
20	21	1993549397	18	50	Female/Male
21	22	55002068	18	23	Female/Male
22	23	415724202	18	33	Female/Male
23	24	538487890	18	25	Female/Male
24	25	1510754277	18	30	Female/Male
25	26	909533488	18	38	Female/Male
26	27	315870041	18	31	Male
27	28	557460752	18	59	Female/Male

Fig. 4.2 Data set of product list

In ecommerce, there are three product categories: the first is only male used product, the second is female used product, and the third is both male and female used product here we are showing the graph of category of product according to the Agestart and Ageend.

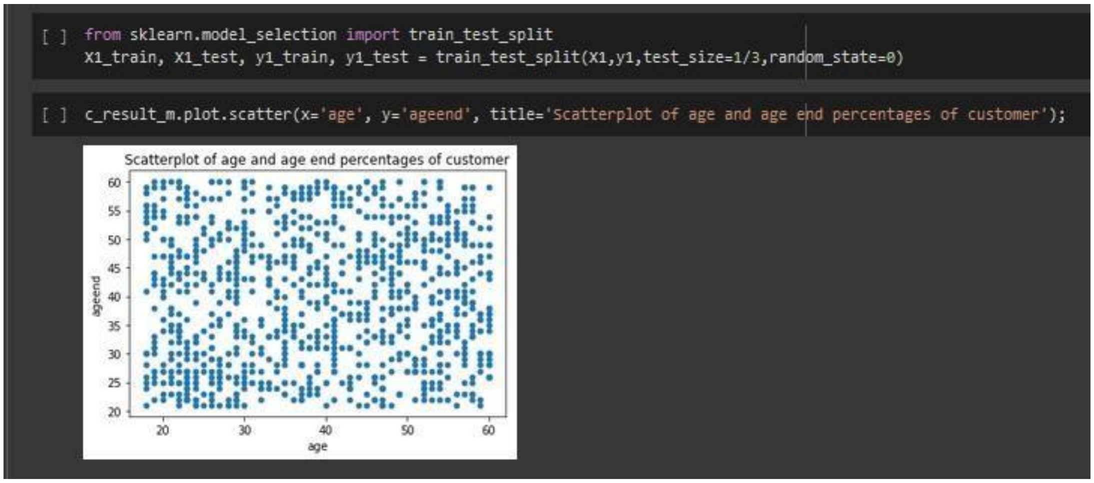


```
[18] #Graph Representation according to male female with age
     counts1.plot(kind='bar', stacked=True)

<matplotlib.axes._subplots.AxesSubplot at 0x7fbb6f2d2370>
```

Fig 4.3: Number of male female in data frame according to age

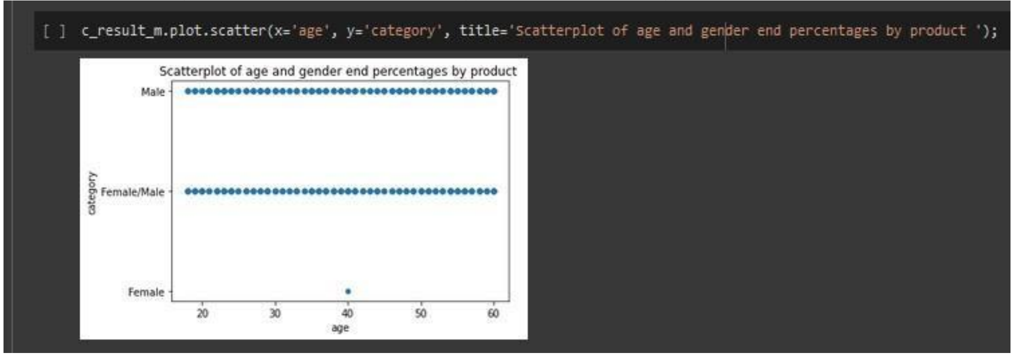
Creating a scatterplot graph with ageend and age to determine the percentage of customers. In this section, we will determine how many people fall into each age group. and another scatter plot graph for age, gender, and percentages by product based on product category and age.



```
[ ]  from sklearn.model_selection import train_test_split
     X1_train, X1_test, y1_train, y1_test = train_test_split(X1,y1,test_size=1/3,random_state=0)
```

```
[ ]  c_result_m.plot.scatter(x='age', y='ageend', title='Scatterplot of age and age end percentages of customer');
```

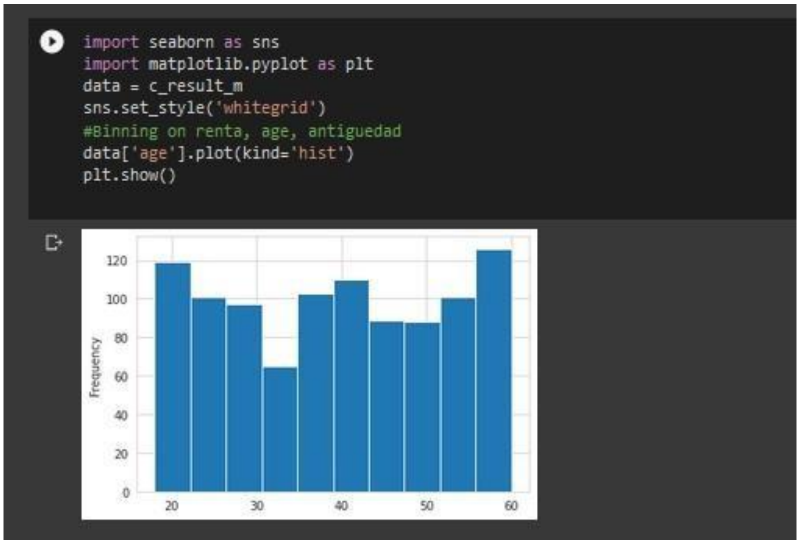
Fig. 4.4: Scatterplot of age end percentages of customer



```
[ ]  c_result_m.plot.scatter(x='age', y='category', title='Scatterplot of age and gender end percentages by product ');
```

Fig. 4.5: Scatterplot of age and gender end percentages by product

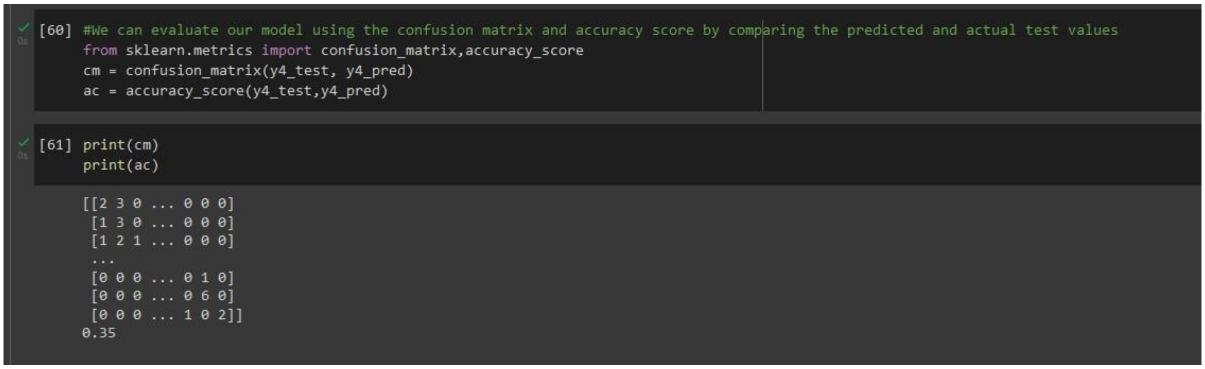
The linear regression algorithm is used to determine the relationship between the dependent and independent variables. The linear regression class expects entries that may contain more than one value, but also a single value, and we also expect the training and test datasets in order to find a regressor. The intercept_value is 38.1166, and the regressor. Coefficient value is 0.02091. We created a graph to representthe outcome of using a linear regression algorithm.



```
import seaborn as sns
import matplotlib.pyplot as plt
data = c_result_m
sns.set_style('whitegrid')
#Binning on renta, age, antiguedad
data['age'].plot(kind='hist')
plt.show()
```

Fig 4.6: graph according to linear regression result

We also use the K-Nearest Neighbor (KNN) algorithm, which is used for both regression and classification. For classification, we choose n neighbor value =5, metric = minkowski, and P=2, and we evaluate our model using the confusion matrix and accuracy score by comparing predicted and actual test values.



```
[60] #We can evaluate our model using the confusion matrix and accuracy score by comparing the predicted and actual test values
     from sklearn.metrics import confusion_matrix,accuracy_score
     cm = confusion_matrix(y4_test, y4_pred)
     ac = accuracy_score(y4_test,y4_pred)

[61] print(cm)
     print(ac)

     [[2 3 0 ... 0 0 0]
      [1 3 0 ... 0 0 0]
      [1 2 1 ... 0 0 0]
      ...
      [0 0 0 ... 0 1 0]
      [0 0 0 ... 0 6 0]
      [0 0 0 ... 1 0 2]]
     0.35
```

Fig 4.7: calculating the accuracy score of model

REFERENCES:

[1]. Raich, B. Ganguly, and M. Tota, "Machine Learning for Market Basket Analysis through," IOSR Journal of Engineering (IOSRJEN), pp. 22-23, 2019.

[2]. Abboud, Y., Brun, A., & Boyer, A. (2019). C3Ro: an efficient mining algorithm of extended-closed contiguous robust sequential patterns in noisy data. *Expert Systems with Applications*, *131*, 172-189.

[3]. Belhadi, A., Djenouri, Y., Lin, J. C. W., Zhang, C., & Cano, A. (2020). Exploring pattern mining algorithms for hashtag retrieval problem. *IEEE Access*, *8*, 10569-10583.

[4]. Jyoti, E., & Walia, E. A. S. (2017). “A Review on Recommendation System and Web Usage Data Mining using K-Nearest Neighbor (KNN) method. *Int. Res. J. Eng. Technol. IRJET*, *4*(4), 2931-2934.

[5]. Ulas, M. A. (2001). Market Basket Analysis For Data Mining. In *Academia. edu*.

[6]. Nilashi, M., bin Ibrahim, O., & Ithnin, N. (2014). Hybrid recommendation approaches for multi-criteria collaborative filtering. *Expert Systems with Applications*, *41*(8), 3879-3900.

[7]. Sharma, J., Sharma, K., Garg, K., & Sharma, A. K. (2021). Product recommendation system a comprehensive review. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012021). IOP Publishing.

[8]. Djenouri, Y., Belhadi, A., Lin, J. C. W., & Cano, A. (2019). Adapted k- nearest neighbors for detecting anomalies on spatio–temporal trafficflow. *IEEE Access*, *7*, 10015-10027.

[9]. Saxena, A., & Gaur, N. K. (2015). Frequent item set based recommendation using apriori. *International Journal of Science, Engineering and Technology Research*, *4*(5).

[10]. Thannimalai, V., & Zhang, L. (2021, December). A Content Based and Collaborative Filtering Recommender System. In *2021 International Conference on Machine Learning and Cybernetics (ICMLC)* (pp. 1-7).IEEE.

[11]. Berkovsky, S., & Freyne, J. (2010, September). Group-based recipe recommendations: analysis of data aggregation strategies. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 111-118).

[12]. Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., & Luo, S. (2021). A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET), 16*(3), 274-306.

[13]. Rana, K. S. (2016). Food recommendation system based on content filtering algorithm. *Bachelor's Degree in Computer Science. Tribhuwan University*.

[14]. Kumar, B., & Sharma, N. (2016). Approaches, issues and challenges in recommender systems: a systematic review. *Indian journal of science and technology*, *9*(47), 1-12.

[15]. Park, D. H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2011). A literature review and classification of recommender systems on academicjournals. *Journal of intelligence and information systems*, *17*(1), 139-152.