"Dissemination of Education for Knowledge, Science and Culture"
-Shikshanmaharshi Dr. Bapuji Salunkhe

(स्वायत्त) कोल्हापूर

VIVEKANAND COLLEGE

KOLHAPUR

(Empowered Autonomous)

DEPARTMENT OF STATISTICS

A PROJECT REPORT
on

"Discovering a Hotel Selection Preference of Youth: A Statistical Approach"

Submitted by

Miss. Gurav Sanjivani Sadanand
Miss. Chougale Swati Pandit

in partial fulfillment for the award of the degree of

MASTER OF SCIENCE

in

STATISTICS
2023-24

CERTIFICATE

This is to Certify that,

Sr. No.	Name	Roll No.
1	Miss. Gurav Sanjivani Sadanand	1407
2	Miss. Chougale Swati Pandit	1403

Have satisfactorily completed the project work on **"Discovering a Hotel Selection Preference of Youth: A Statistical Approach"** as a part of practical evaluation course for **M.Sc. II**, prescribed by the Department of Statistics, ***Vivekanand College, Kolhapur (Empowered Autonomous)*** in the academic year **2023-24.**

This project has been completed under our guidance and supervision. To the best of our knowledge and belief, the matter presented in this project report is original and has not been submitted elsewhere for any other purpose.

Date: 29/05/2024

Place: Kolhapur

Pawar A.A.

Project Guide
(Mr. Pawar A.A.)

30/05/24

Examiner

Head
(Mrs. Shinde V.C.)

HEAD
DEPARTMENT OF STATISTICS
VIVEKANAND COLLEGE, KOLHAPUR
(EMPOWERED AUTONOMOUS)

ACKNOWLEDGEMENT

As we approach the culmination of this enormous endeavor, it brings immense satisfaction to reflect upon and express gratitude for the collective efforts and support from all those who enthusiastically contributed. Together, we have transformed what once seemed like a far-off aspiration for industrial training into reality.

We are thankful to Principal Dr. R.R. Kumbhar for his kind cooperation and valuable support. We are also thankful to Mrs. Shinde V.C., Head, Department of Statistics, Vivekanand College, Kolhapur (Empowered Autonomous). We are indeed grateful to OJT Coordinator Mr. Pawar A.A., Assistant Professor, Department of Statistics, Vivekanand College, Kolhapur (Empowered Autonomous) for his kind and valuable support.

We are extremely thankful to guide Mr. Pawar A.A., Assistant Professor, Department of Statistics, Vivekanand College, Kolhapur (Empowered Autonomous) for his valuable guidance and mentorship throughout this project work given to us during the study.

We are also thankful to all the teaching and non-teaching staff of our department for their direct and indirect support. We also thanks to all the members of Next Level Design Academy.

Additionally, we sincerely thanks to our parents, friends and colleagues for their direct and indirect contributions.

Yours Sincerely,

M.Sc.II

Department of Statistics

INDEX

Sr. No.	Content	Page No.
1	Introduction	5
2	Objectives	6
3	Data Collection & Methodology	7
4	Graphical Representation	10
5	Statistical Analysis	15
6	Conclusion	23
7	Reference	24
8	Scope of the study	25
9	Appendix	30

INTRODUCTION

The hotel industry has been one of the most competitive industries, especially in the 21st century. For this reason, enchanting customer loyalty is one of the key aspects of enhancing growing businesses in the industry as well as ensuring business continuity.

The motivation behind this project lies in understanding the evolving landscape of the hotel industry. We seek to uncover the factors that drive customer choice and preference, guiding hotels to enhance their services. To find actionable insight for the hotels, enabling them to tailor their offerings to better meet the diverse needs of guests. By understanding customer expectations, the project aims to contribute the improvement of overall guest satisfaction and the success of hospitality establishments.

For instance, it identifies factors such as location, price, and traveller's hotel sections. Other criteria of interest to travellers are location, facilities of guest rooms, staff facilities. Recently identified three attribute's facilities and breakfast recently identified three attributes of the low-priced hotel segment that are more valuable in terms of improving consumer satisfaction. Cleanliness, silence, and air conditioning demonstrated that expectations of hotels are influenced by personal factors such as gender, purpose, nationality, culture, and the private domain of hospitality.

The hospitality industry is dynamic and constantly evolving to meet the diverse needs and preferences of guests. In this context, our project aims to delve into the intricacies of hotel management and customer preference. The data collected can help you to figure out what others in the hotel industry are doing, and how to become better than them in terms of services and experience.

OBJECTIVES

- ✓ To explore the customer preferences.
- ✓ To explore the preference of booking platforms.
- ✓ To understand the importance of Amenities and Factors preference.

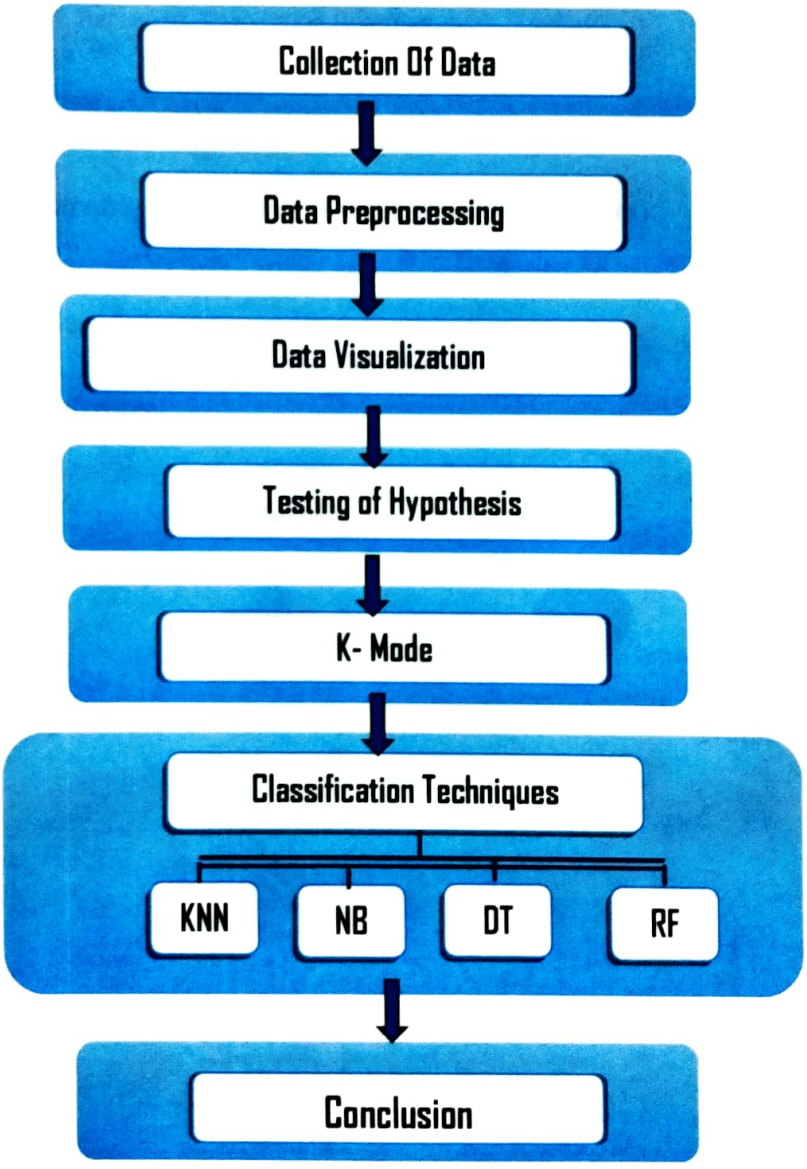
DATA COLLECTION &METHODOLOGY

We are employing a diverse approach combining survey design to collect data from a representative sample. Statistical analysis and data mining technique's will then uncover insights, guiding our understanding of customer preference and shaping practical strategies for the local hotel industry. The data collection process includes a well-structed questionnaire.

This project not only contributes to the body of knowledge in hotel management but also provides actionable insights for hospitality establishments. By understanding customer preference hotels can tailor their offerings to their guests. Hotel can customize what they offer to match exactly what their guests want and need.

Analyzing the results to draw meaningful choices about the factors influencing hotel choices and customer expectations.

METHODOLOGY: FLOW CHART



STATISTICAL TOOLS

Exploratory Data Analysis:

- Radar Chart, Bar charts, Correlation Heatmap
- Scree plots
- Chi-square test

Machine Learning Algorithms (Data Mining Classifiers):

- K-Nearest Neighbour, Naïve Bayes, DecisionTree, Random Forest.
- K-Mode Clustering

Statistical Software:

MS-Excel

Python

SPSS

R software

GRAPHICAL REPRESENTATION

Amenities Preference:

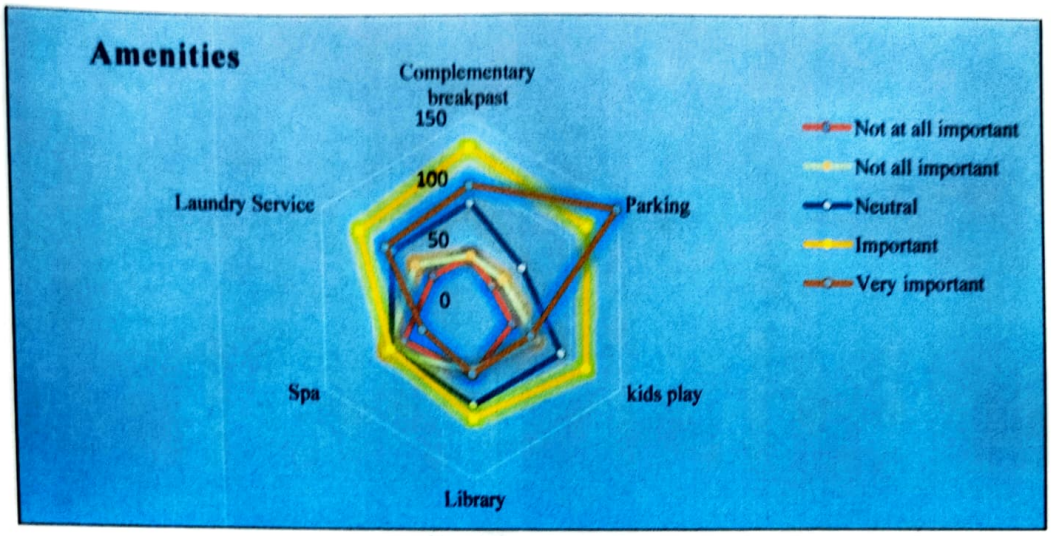


Fig 1.1: Amenities

Conclusion: From the Fig 1.1 we conclude that, most of the customers prefers Parking & Complementary breakfast among the amenities while choosing a hotel for stay.

Room Features preference:

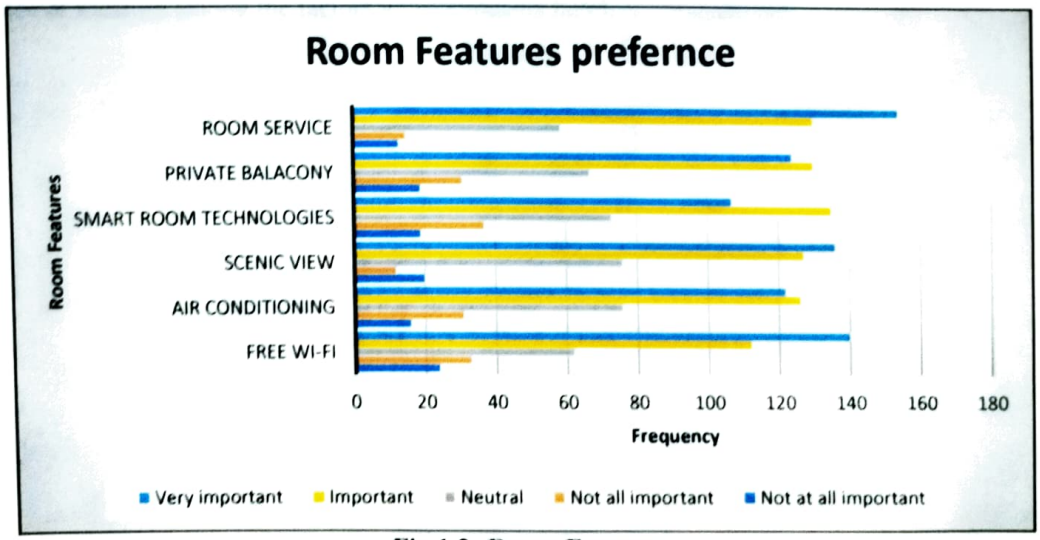


Fig 1.2: Room Features

Conclusion: From the Fig 1.2 we conclude that, most of the customers prefers room service and free wi-fi among room features.

Factor Preference:

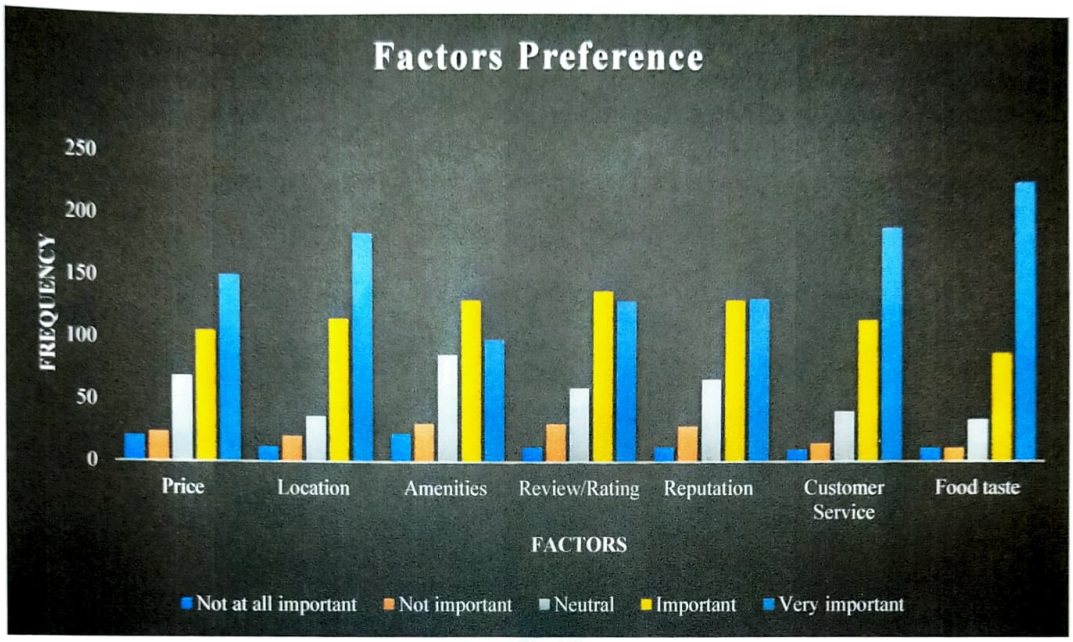


Fig 1.3: Factors

Conclusion: From the above Fig., we see that most of the customers prefers food taste, customer service & location among the factors while choosing hotel.

CORRELATION HEATMAP

Correlation Heatmap is visual representation of the correlation between different variables in the dataset. By examining the heatmap, we can quickly identify patterns & relationship between variables. This helps us understand which variables are closely related & may influence each other.

The following map shows the correlation between different amenities which is constructed using Python software.

Amenities:

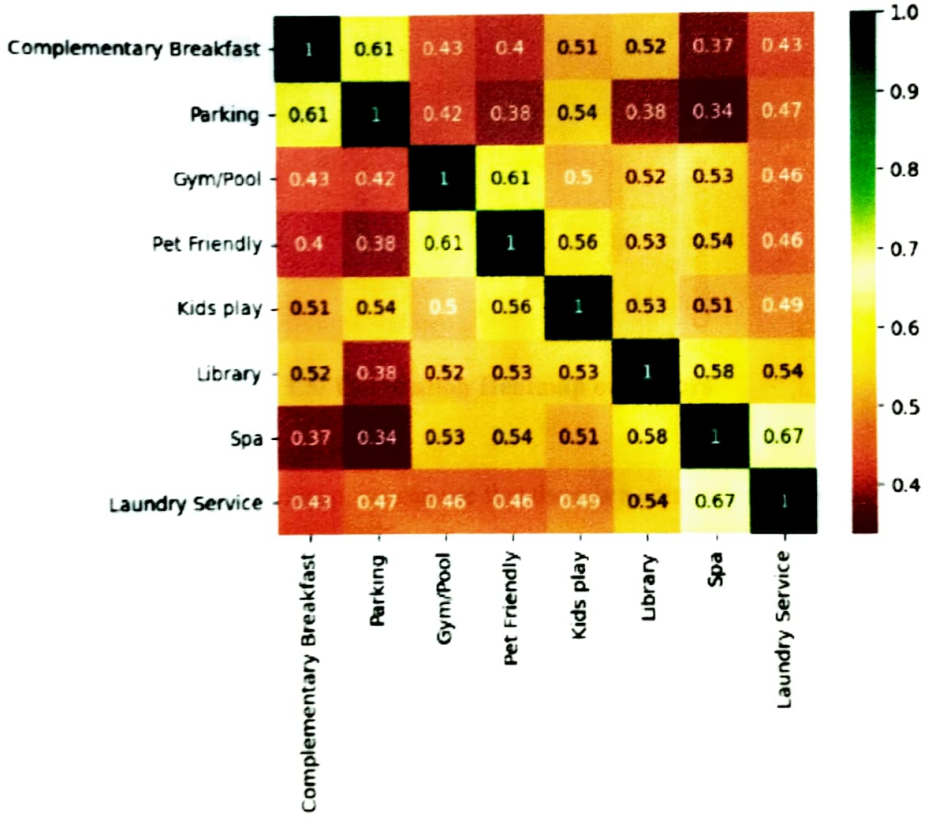


Fig 1.4: Correlation Heatmap of Amenities

Conclusion: From the above heatmap, we can see that the Parking & Complementary breakfast, Gym /Pool & Pet-friendly, Laundry service & Spa are closely related and may influence each other.

Factors:

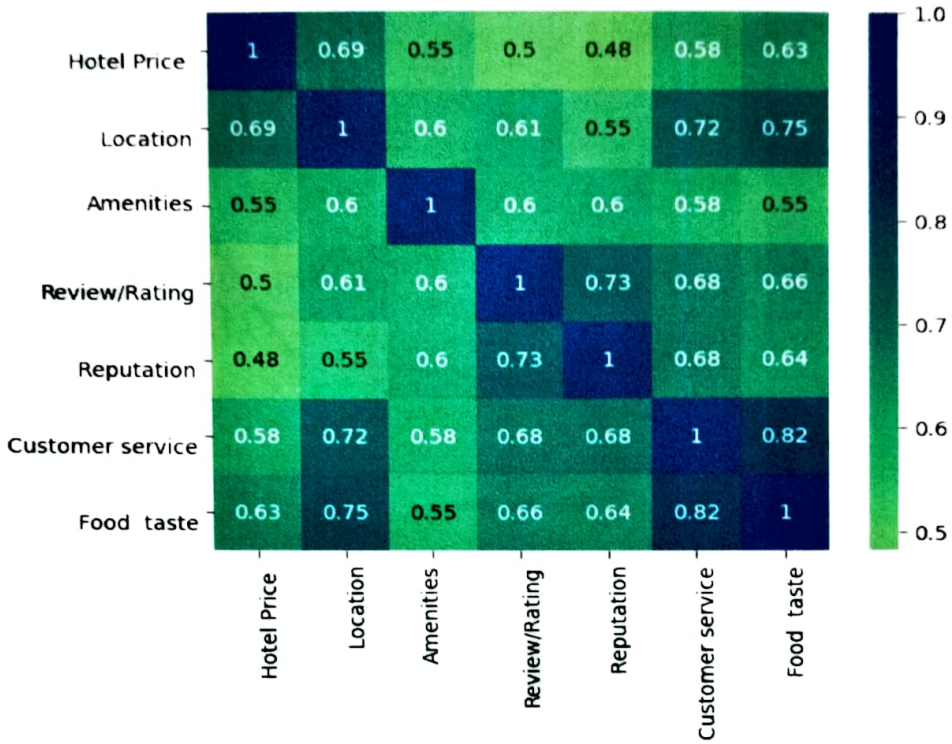


Fig 1.5: Correlation Heatmap of Factors

Conclusion: From the above heatmap, we can see that there is a correlation between Food taste and Location, Customer service and Food taste.

Scree plot of Booking Strategies:

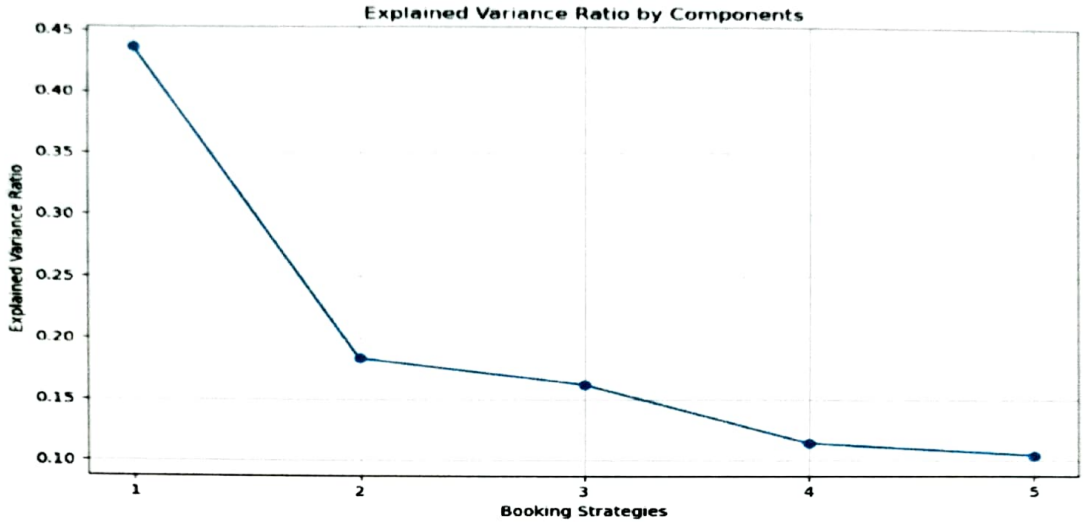


Fig 1.5: Booking Strategies

Conclusion: From the above scree plot, we conclude that most of the customers choose Early Booking Discounts strategy for booking the hotel.

Scree plot of Factors:

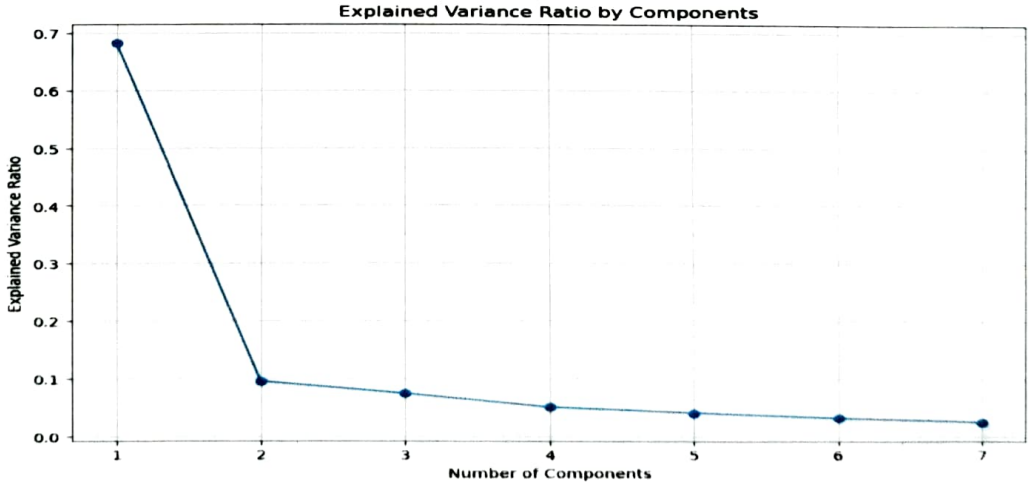


Fig 1.6: Factors

Conclusion: From the above Fig., we conclude that Food taste and customer service are the most preferred factors while choosing a hotel.

STATISTICAL ANALYSIS

Chi-square test:

➢ Testing association between Gender and Food type

Hypothesis:

H_0: There is no association between gender and food type.

V/S

H_1: There is association between gender and food type.

		Food type			Total
		Veg	Non-veg	Both (Veg &Nonveg)	
Gender	Female	62	14	143	219
	Male	31	23	98	152
Total		93	37	241	371

Table 2.3: Food type

Using SPSS:

	Value	df	p-value
Pearson Chi-Square	9.123	2	0.010
Likelihood Ratio	9.031	2	0.011

Conclusion: From the table 2.3, we conclude that there is dependency between Food type and Gender.

➢ Testing association between Gender and Service quality

Hypothesis:

H_0: There is no association between gender and service quality.

V/S

H_1: There is a significant association between gender and service quality.

Using SPSS:

	Value	d.f	p-value
Quick check-in/check-out	0.166157	1	0.683550
Friendliness and Professionalism	0.288105	1	0.591437
Responsiveness to guest requests	0.000065	1	0.993574
Room service efficiency	1.882759	1	0.170021

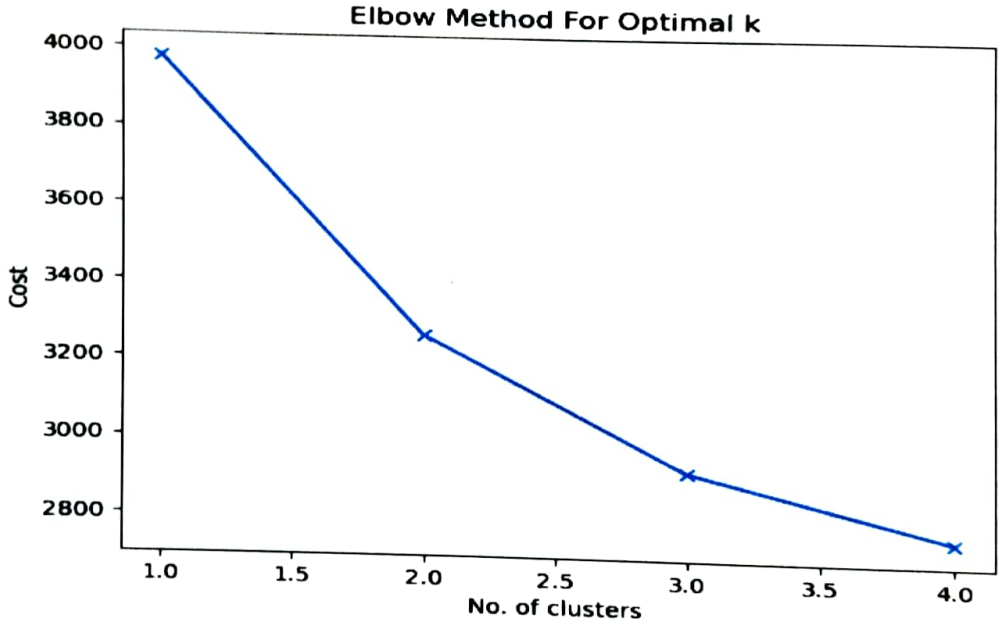
Table 2.2: Service quality

Conclusion: From the table 2.2, we conclude that there is no association between gender and service quality also we conclude that the customers give high priority to service quality.

K-Mode:

K-Modes is a clustering algorithm used in data science to group similar data points into clusters based on their categorical attributes. Unlike traditional clustering algorithms that use distance metrics, K-Modes works by identifying the modes or most frequent values within each cluster to determine its centroid. K-Modes is ideal for clustering categorical data such as customer demographics, market segments, or survey responses. It is a powerful tool for data analysts and scientists to gain insights into their data and make informed decisions.

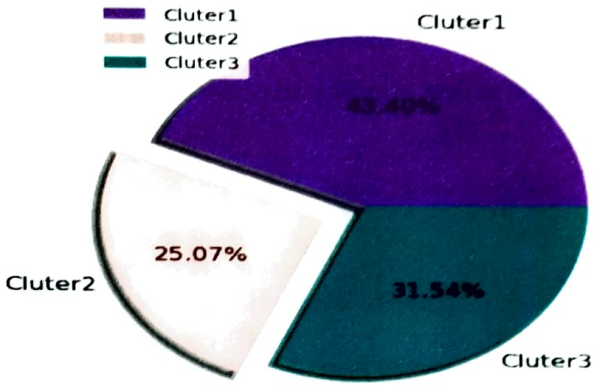
- Find the value of K:



The above graph is known as elbow graph, elbow method is a graphical representation of finding the optimal number of clusters. We will decide how many clusters to be choose using elbow point. As we can see from the graph there is an elbow-like shape at 3, suggesting that 3 clusters maybe a good choice.

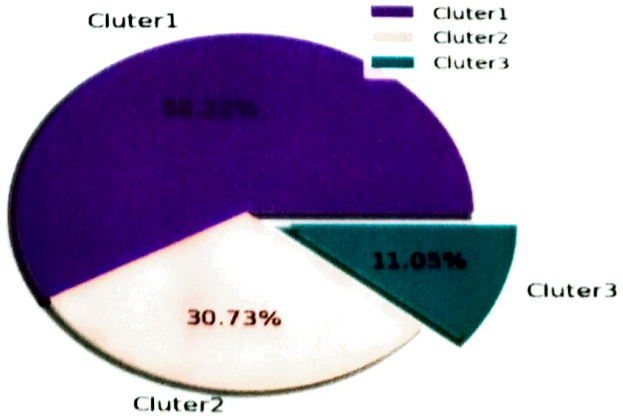
i) **Amenities:**

Cluster	Count	Centroids
Cluster1	161	[4 4 4 4 4 4 4 4 4 4 4 4 4 4 4]
Cluster2	93	[3 3 2 2 3 3 2 3 3 3 3 3 3 1 3 3]
Cluster 3	117	[5 5 5 5 5 5 5 5 5 5 5 5 5 5 5]



ii) **Factors:**

Cluster	Count	Centroids
Cluster1	216	[5 5 5 5 5 5 5]
Cluster2	114	[4 4 4 4 4 4 4]
Cluster 3	41	[3 3 3 3 3 3 3]



- While considering 16 variables related to Amenities, data is clustered into 3 clusters of sizes 161,93,117 respectively. From this, we concludes that level of likert scale (4 &5) is most important related to amenities.
- Whereas in point of view of factors, data is clustered into 3 clusters of 216 is related to those respondents who gives '5' as a response for related vriables.Similarly second cluster of 114 is related to respondents who gives '4' as a response.
- This conclude that, level of standards in Amenities as well as in factors is most important.

K-Nearest Neighbors

KNN is basically a classification algorithm that will make a prediction of a class of a target variable based on defined number of nearest neighbors. It will calculate distance from the instance you want to classify to every instance of the training dataset & then classify your instance based on the majority classes of k nearest instances.

Naïve Bayes

The Naïve Bayes classifier is simple but powerful machine learning algorithm that can be used for classification tasks. It is based on Baye's theorem which is mathematical formula that describe the probability of an event occurring given the knowledge of other event. It works by calculating the probability of each class given the input features.

Decision Tree

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. In Decision Tree the major challenge is to identification of the attribute further root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

1. Information Gain 2. Gini Index

Random Forest

The Random Forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees. Random Forests are particularly well-suited for handling large and complex datasets, dealing with high-dimensional feature spaces, and providing insights into feature importance. It is a set of decision trees (DT) from a

randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

CLASSIFICATION RESULTS

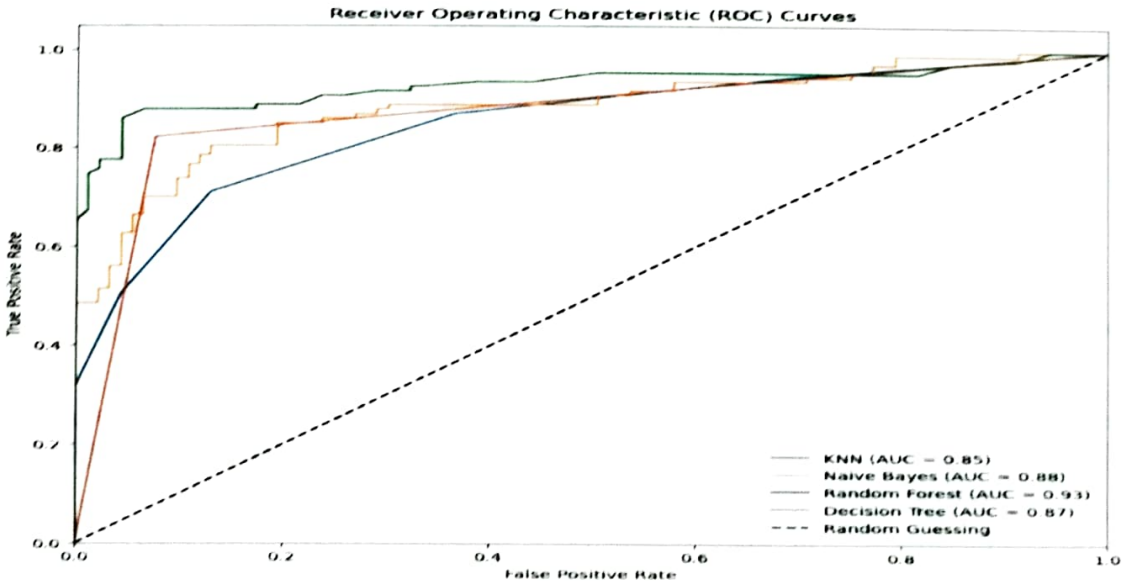
- **Amenities:**

Classifiers	Precision	Recall	F1 Score	Accuracy
KNN	0.70	0.67	0.67	0.66
Naïve Bayes	0.77	0.77	0.77	0.77
Decision Tree	0.70	0.67	0.67	0.66
Random Forest	0.85	0.83	0.83	0.83

Conclusion: Random Forest gives highest accuracy (83%) of classification.

Receiver Operating Characteristic (ROC) Curve of Amenities:

- ROC curve is a graphical representation of the performance of classification model at various classification thresholds.
- The ROC curve tells us how good our model is at telling things apart. The closer the curve is to the top-left corner, the better our model is at getting things right.
- A higher AUC value (close to 1) indicates better discrimination between classes, while a lower value suggests poorer performance.



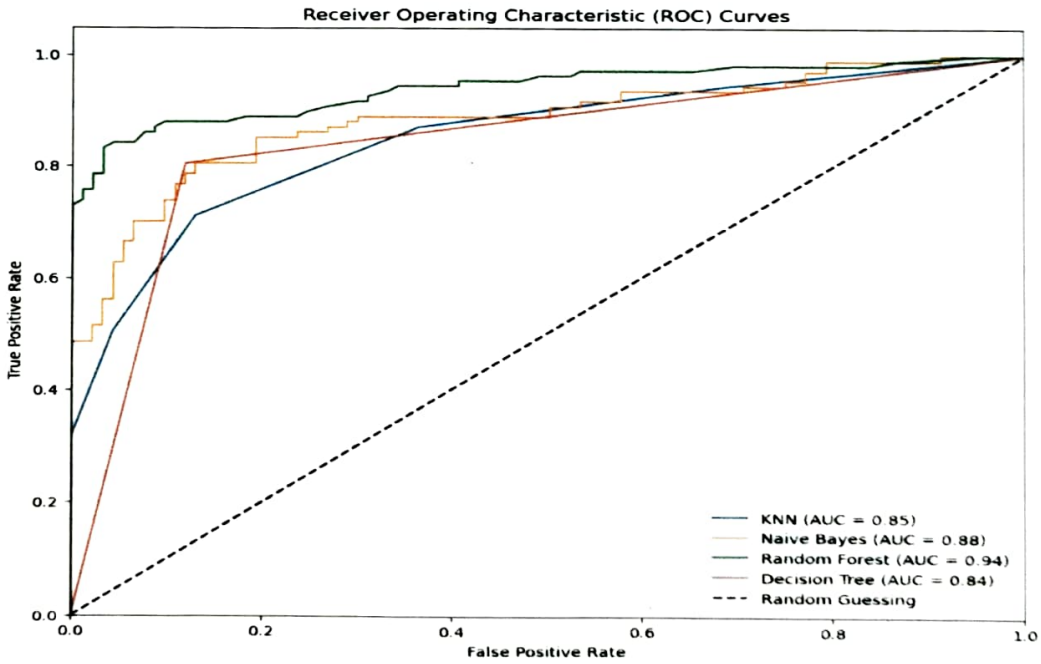
Conclusion: The area under curve is 0.93 which indicates that random forest model performed the best out of all the models we tried.

- **Factors:**

Classifiers	Precision	Recall	F1 Score	Accuracy
KNN	0.85	0.83	0.83	0.83
Naïve Bayes	0.85	0.84	0.84	0.84
Decision Tree	0.79	0.79	0.79	0.79
Random Forest	0.88	0.88	0.88	0.88

Conclusion: Random Forest gives highest accuracy (88%) of classification.

Receiver Operating Characteristic (ROC) Curve of Factors:



Conclusion: The area under curve is 0.94 which indicates that random forest model performed the best out of all the models we tried.

OVERALL CONCLUSIONS

- Customers prefers parking, complimentary breakfast, room service and free wi-fi when choosing a hotel. They also value food taste, customer service and location highly.
- Essential amenities includes parking, complimentary breakfast, gym/pool, pet-friendly options, laundry service, and spa.
- Most of the customers prefer Early Booking Discounts strategy for booking the hotel.
- There is association between Food type and Gender.
- Level of standards in Amenities as well as factors is most important.

REFERENCES

- Sim, J., Mak, B., & Jones, D.(2006). A model of Customer Satisfaction and Retention for Hotels. Journal of Quality Assurance in Hospitality & Tourism,7(3),1-23

- Tsung, M., Ma, Q., & Su, C.J. (1999). Mapping customer's service experience for operational improvement. Business Process Management, 7(1),10-18

- Rivers, M.J., Toh, R.S., & Alaoui, M.(1991). Frequent-stayer programs: the demographic, behavioral and attitudinal characteristics of hotel steady sleepers. Journal of Travel Research,30(2),41-45

- McCleary, K.W., Weaver, P. A., & Hutchinson, J.C.(1993). Hotel selection factors as they relate to business travel situations. Journal of Travel Research,32(2),42-48

- Guo, Y.; Barnes, S.J.; Jia, Q. Mining meaning from online rating and reviews: Tourist satisfaction analysis using latent Dirichlet allocation. Tour. Manag. 2017,59,467-483

SCOPE OF STUDY

- This project aims to provide valuable insights into customer preferences, helping hotels to better meet customer needs and stay competitive in the market.

RECOMMENDATION

- If we provide facilities like Parking & Complimentary breakfast, Gym/Pool & Pet-friendly environment as well as laundry service & Spa then customers will prefer to visit the hotel.

QUESTIONNAIRE

Project Title: Discovering the Hotel Selection Preference of Youth : A Statistical Approach

1. Which type of food Gender
   - ☐ Male
   - ☐ Female
   - ☐ Other

2. Occupation
   - ☐ Student
   - ☐ Job
   - ☐ Business
   - ☐ Other

3. Age Group
   - ☐ 18-25
   - ☐ 26-35
   - ☐ 36-45
   - ☐ 46-55
   - ☐ Other

4. do you prefer?
   - ☐ Veg
   - ☐ Non-veg
   - ☐ Both (Veg & Non-veg)
   - ☐ Vegan (Only plant food)

5. Which type of cuisine do you prefer at your dining at hotel?
   - ☐ Local Cuisine
   - ☐ International Cuisine
   - ☐ Chinese Food
   - ☐ Maharashtrian Food
   - ☐ Other

6. How often do you stay in hotel?
   - ☐ Never
   - ☐ Rarely
   - ☐ Occasionally
   - ☐ Frequently
   - ☐ Very Frequently

12. What factors are most important to you when choosing a hotel?
    - ☐ Price
    - ☐ Location
    - ☐ Amenities (Facility)
    - ☐ Review/Ratings
    - ☐ Reputation
    - ☐ Customer service
    - ☐ Food taste

13. Room features most important for your hotel stay?
    - ☐ Free Wi-Fi
    - ☐ Air Conditioning
    - ☐ Scenic view
    - ☐ Smart room technologies
    - ☐ Private balcony
    - ☐ Mini bar
    - ☐ Room service
    - ☐ In-room entertainment

14. Your preference for the bar area?
    - ☐ Indoor seating
    - ☐ Outdoor seating
    - ☐ No preference

15. Smoking or non-smoking area preference in the bar?
    - ☐ Smoking
    - ☐ Non-smoking
    - ☐ No preference

16. How do you prefer to book a hotel?
    - ☐ Online Travel Agencies(OTA)
    - ☐ Hotel Website
    - ☐ Phone Reservation
    - ☐ Walk-in

17. Hotel booking platforms?
    - ☐ Hotels.com
    - ☐ Trivago
    - ☐ MakeMyTrip
    - ☐ Other

18. Preferred budget range per night for accommodation?
    - ☐ Rs.500 - Rs.1,500
    - ☐ Rs.1,500 - Rs.3,000

- ☐ Rs.3,000 - Rs.7,000
- ☐ Above 7,000

19. How far in advance do you typically book a hotel?
    - ☐ Less than a week
    - ☐ 1-2 weeks
    - ☐ 1-3 months
    - ☐ Other

20. What pricing strategies do you prefer the most?
    - ☐ Early booking discounts
    - ☐ Last-minute deals
    - ☐ Package deals
    - ☐ Loyalty programme discounts
    - ☐ Promotion code & coupons

21. In off-peak season hotel booking preference?
    - ☐ Very unlikely
    - ☐ Unlikely
    - ☐ Neutral
    - ☐ Likely
    - ☐ Very likely

22. Which aspects of the service quality matter most to you?
    - ☐ Quick check-in/check-out
    - ☐ Friendliness & Professionalism staff
    - ☐ Responsiveness to guest requests
    - ☐ Room service efficiency
    - ☐ Other

23. Your preference for the hotel's check-out timing system?
    - ☐ 24 hours check-in/check-out
    - ☐ 12 hours check-in/check-out
    - ☐ Below 12 hours check-in/check-out
    - ☐ Fixed check-in/check-out

24. Your preference of location type for hotel stay?
    - ☐ City centre
    - ☐ Near nature attractions
    - ☐ Quieter areas
    - ☐ Other

APPENDIX

Correlation Heatmap

❖ Amenities

```
importpandasaspd
importnumpyasnp
importseabornassns
importwarnings
warnings.filterwarnings("ignore")
data=pd.read_csv("C:\\Users\\statv\\OneDrive\\Desktop\\Amenities1.csv")
data
data.describe()
correlation=data.corr()
data.corr()
sns.heatmap(correlation,cmap="RdYlGn",annot=True)
```

❖ Factors

```
importpandasaspd
importnumpyasnp
importseabornassns
importwarnings
warnings.filterwarnings("ignore")
data=pd.read_csv("C:\\Users\\statv\\OneDrive\\Desktop\\Project (Msc II)\\Excel files\\factors3-
1.csv")
data
data.describe()
correlation=data.corr()
data.corr()
sns.heatmap(correlation,cmap="crest",annot=True)
```

Scree plot

```
# Separate features (X) and labels (y)
X=data.iloc[:,:-1]# Exclude the last column as it represents the label
y=data.iloc[:,-1]# Last column is the label

# Standardize the features
scaler=StandardScaler()
X_scaled=scaler.fit_transform(X)

# Apply PCA
```

```
pca=PCA()
X_pca=pca.fit_transform(X_scaled)



# Plot explained variance ratio
plt.figure(figsize=(10,6))
plt.plot(range(1,len(pca.explained_variance_ratio_)+1),pca.explained_variance_ratio_,marker='o',li
nestyle='-')
plt.title('Explained Variance Ratio by Components')
plt.xlabel('Number of Components')
plt.ylabel('Explained Variance Ratio')
plt.xticks(range(1,len(pca.explained_variance_ratio_)+1))
plt.grid(True)
plt.show()

# Cumulative explained variance ratio
cumulative_variance_ratio=pca.explained_variance_ratio_.cumsum()
print("Cumulative Explained Variance Ratio:")
print(cumulative_variance_ratio)

# You can select the number of components based on the plot or cumulative explained variance
ratio
# For example, if you want to retain 95% variance, find the index where cumulative variance ratio
crosses 0.95
n_components=(cumulative_variance_ratio<0.95).sum()+1
print("Number of components to retain 95% variance:",n_components)

# Fit PCA with selected number of components
pca=PCA(n_components=n_components)
X_pca=pca.fit_transform(X_scaled)

# Visualize the transformed data (if n_components is 2 or 3)
ifn_components==2:
plt.figure(figsize=(8,6))
plt.scatter(X_pca[:,0],X_pca[:,1],c=y,cmap='viridis')
plt.title('PCA - 2 Components')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.colorbar(label='Label')
plt.show()
elifn_components==3:
frommpl_toolkits.mplot3dimportAxes3D

fig=plt.figure(figsize=(10,8))
ax=fig.add_subplot(111,projection='3d')
ax.scatter(X_pca[:,0],X_pca[:,1],X_pca[:,2],c=y,cmap='viridis')
ax.set_title('PCA - 3 Components')
ax.set_xlabel('Principal Component 1')
ax.set_ylabel('Principal Component 2')
```

```
ax.set_zlabel('Principal Component 3')
ax.legend()
plt.show
```

K-Modes Clustering

```
!pipinstallkmodes
importpandasaspd
importnumpyasnp
importmatplotlib.pyplotasplt
get_ipython().run_line_magic('matplotlib','inline')
importseabornassns
importwarnings
importtime
warnings.filterwarnings(action="ignore")

fromsklearn.metricsimportaccuracy_score
fromscipy.statsimportmode
fromkmodes.kmodesimportKModes
importmatplotlib.pyplotasplt
%matplotlib inline
```

❖ Amenities

```
data=pd.read_csv("C:\\Users\\statv\\OneDrive\\Desktop\\Amenities1_1.csv")
data.head()
data.isna().sum().sum()
kmodes_test=KModes(n_clusters=3,n_init=2)
Clusters=kmodes_test.fit_predict(data)
print(kmodes_test)
print(Clusters)

print(kmodes_test.cluster_centroids_)

df_labels=pd.DataFrame(Clusters,columns=list(["labels"]))
df_labels["labels"]=df_labels["labels"].astype("category")
df_labels=data.join(df_labels)
df_labels

pd.crosstab(index=df_labels["labels"],columns="count")

x=[161,93,117]
y=["Cluter1","Cluter2","Cluter3"]

ex=[0.0,0.2,0.0]
c=["m","pink","c"]
```

```
plt.pie(x,labels=y,explode=ex,colors=c,autopct="%0.2f%%",shadow=True,radius=1,labeldistance=
1.1,textprops={"fontsize":12},wedgeprops={"linewidth":4}
```

❖ Factors

```
data=pd.read_csv("C:\\Users\\statv\\OneDrive\\Desktop\\Project (Msc II)\Excel files\\factors2.csv")
data.head()
data.isna().sum().sum()


kmodes_test=KModes(n_clusters=3,n_init=2)
Clusters=kmodes_test.fit_predict(data)
print(kmodes_test)
print(Clusters)
print(kmodes_test.cluster_centroids_)
df_labels=pd.DataFrame(Clusters,columns=list(["labels"]))
df_labels["labels"]=df_labels["labels"].astype("category")
df_labels=data.join(df_labels)
df_labels
pd.crosstab(index=df_labels["labels"],columns="count")
x=[216,114,41]
y=["Cluter1","Cluter2","Cluter3"]
ex=[0.0,0.2,0.0]
c=["m","pink","c"]
plt.pie(x,labels=y,explode=ex,colors=c,autopct="%0.2f%%",shadow=True,radius=1,labeldistance=
1.1,textprops={"fontsize":12},wedgeprops={"linewidth":4}
```

❖ **Elbow method**

```
cost=[]
K=range(1,5)
fornum_clustersinlist(K):
kmode=KModes(n_clusters=num_clusters,init="random",n_init=5)
kmode.fit_predict(data)
cost.append(kmode.cost_)

plt.plot(K,cost,'bx-')
plt.xlabel('No. of clusters')
plt.ylabel('Cost')
plt.title('Elbow Method For Optimal k')
plt.show()
```

File Import

```
importpandasaspd
importnumpyasnp
data= pd.read_csv("C:\\Users\\statv\\OneDrive\\Desktop\\Project (Msc II)\\Excel
files\\df_labels.csv")
data
```

Data Splitting

```
fromsklearn.model_selectionimporttrain_test_split
x=data.drop("labels",axis=1)
y=data["labels"]
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

Data Balancing

```
fromimblearn.over_samplingimportSMOTE
print(x_train.shape,y_train.shape)
print("x_test.shape,y_test.shape")
print("before oversampling,count the labels'0':{}".format(sum(y_train==0)))
print("before oversampling,count the labels'1':{}".format(sum(y_train==1)))
print("before oversampling,count the labels'2':{}".format(sum(y_train==2)))
sm=SMOTE(random_state=30)
x_train_res,y_train_res=sm.fit_resample(x_train,y_train.ravel())
print("after oversampling,count the labels'0':{}".format(sum(y_train_res==0)))
print("after oversampling,count the labels'1':{}".format(sum(y_train_res==1)))
print("after oversampling,count the labels'2':{}".format(sum(y_train_res==2)))
sm=SMOTE(random_state=30)
```

Classification Techniques

❖ KNN

```
fromsklearn.neighborsimportKNeighborsClassifier
k=KNeighborsClassifier(n_neighbors=5)
k.fit(x_train_res,y_train_res)
pred=k.predict(x_test)
print("Accuracy from KNN is",accuracy_score(y_test,pred))
print(classification_report(y_test,pred))
fromsklearn.metricsimportconfusion_matrix,accuracy_score,mean_squared_error,classification_rep
ort
fromsklearnimportmetrics
print(confusion_matrix(y_test,pred))
```

❖ Naïve bayes

```
fromsklearn.naive_bayesimportGaussianNB
n=GaussianNB()
n.fit(x_train_res,y_train_res)
pred2=n.predict(x_test)
pred2
print("Accuracy from Naive Bayes is ",accuracy_score(y_test,pred2))
print(classification_report(y_test,pred2))
print(confusion_matrix(y_test,pred2))
```

❖ Decision Tree

```
fromsklearn.treeimportDecisionTreeClassifier
d=DecisionTreeClassifier(max_depth=5)
```

```
d=d.fit(x_train_res,y_train_res)
pred4=d.predict(x_test)
print("Accuracy from Decision Tree is",accuracy_score(y_test,pred4))
print(classification_report(y_test,pred4))
print(confusion_matrix(y_test,pred4))
fromsklearnimporttree
print(tree.export_text(d))
importmatplotlib.pyplotasplt
T1=tree.plot_tree(d)
```

❖ Random Forest

```
fromsklearn.ensembleimportRandomForestClassifier
r=RandomForestClassifier(n_estimators=10)
r.fit(x_train,y_train)
pred5=r.predict(x_test)
print("Accuracy from Random Forest is",accuracy_score(y_test,pred5))
print(classification_report(y_test,pred5))
print(confusion_matrix(y_test,pred5))
```

ROC Curve

```
importnumpyasnp
importmatplotlib.pyplotasplt
fromsklearn.datasetsimportmake_classification
fromsklearn.model_selectionimporttrain_test_split
fromsklearn.preprocessingimportStandardScaler
fromsklearn.neighborsimportKNeighborsClassifier
fromsklearn.naive_bayesimportGaussianNB
fromsklearn.ensembleimportRandomForestClassifier
fromsklearn.treeimportDecisionTreeClassifier
fromsklearn.metricsimportroc_curve,auc

# Generate some example data
X,y=make_classification(n_samples=1000,n_features=20,n_classes=2,random_state=42)
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=42)

# Standardize the features
scaler=StandardScaler()
X_train=scaler.fit_transform(X_train)
X_test=scaler.transform(X_test)

# Initialize all classifiers
classifiers={
"KNN":KNeighborsClassifier(),
"Naive Bayes":GaussianNB(),
"Random Forest":RandomForestClassifier(),
"Decision Tree":DecisionTreeClassifier()
}


# Train all classifiers and compute ROC curves
```

```
plt.figure(figsize=(10,8))
for name,clf in classifiers.items():
clf.fit(X_train,y_train)
y_score=clf.predict_proba(X_test)[:,1]
fpr,tpr,_=roc_curve(y_test,y_score)
roc_auc=auc(fpr,tpr)
plt.plot(fpr,tpr,label=f'{name} (AUC = {roc_auc:.2f})')
```

```
# Plot ROC curve for each classifier
plt.plot([0,1],[0,1],linestyle='--',color='k',label='Random Guessing')
plt.xlim([0.0,1.0])
plt.ylim([0.0,1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curves')
plt.legend(loc='lower right')
plt.show()
```