Big Homework

Jexembayev Ruslan

16.12.2023

Contents

1 Introduction

Here you can introduce the context and purpose of your research.

2 List of Datasets

2.1 Brief Description of Each Dataset

2.1.1 "Iris" Dataset (Balanced)

Description: The "Iris" dataset is one of the most famous datasets in machine learning, containing data on three species of Iris flowers (setosa, versicolor, and virginica), with measurements of four features: sepal length, sepal width, petal length, and petal width. This dataset is balanced, meaning it has an equal number of samples for each class.
Reason for Selection: The "Iris" dataset was chosen due to its status as a benchmark dataset in machine learning and its balanced nature, making it ideal for assessing the classification algorithms' ability to accurately distinguish between classes. It is also a convenient dataset for practicing with the neural FCA algorithm due to its simplicity.
Link - `https://archive.ics.uci.edu/dataset/53/iris`

2.1.2 "Airline Satisfaction" Dataset (Balanced)

Description: This dataset includes airline passenger reviews, ratings of their satisfaction with various aspects of their flight experience. The dataset is balanced and includes a wide range of features, such as service quality, seat comfort, and overall airline impression.
Reason for Selection: This dataset was selected for its relevance and the diversity of the data, which allows testing the ability of algorithms to analyze and classify complex and varied data patterns. It also demonstrates the applicability of algorithms in real-world usage scenarios, such as customer review analysis. The "Airline Satisfaction" dataset is intriguing for its numerous columns that are interesting to binarize.
Link - `https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction`

2.1.3 "Stroke" Dataset (Imbalanced)

Description: The "Stroke" dataset compiles patient data, including medical and demographic characteristics, to predict stroke risk. This dataset is imbalanced, as the number of stroke cases is significantly less than non-stroke cases.
Reason for Selection: "Stroke" was chosen for its imbalanced nature, presenting a challenge for many machine learning algorithms. Studying model effectiveness on such data helps assess their ability to handle real clinical data with high levels of imbalance. This dataset is particularly interesting due to its imbalanced nature.
Link - `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`

3 Working with Dataset 1

Let's examine the Iris dataset. Below is a snapshot of the dataset, showcasing the attributes: sepal length, sepal width, petal length, petal width, and the species classification. We will then look into the distribution of species and other attributes in the dataset.

Index	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

Table 1: Snapshot of the Iris Dataset

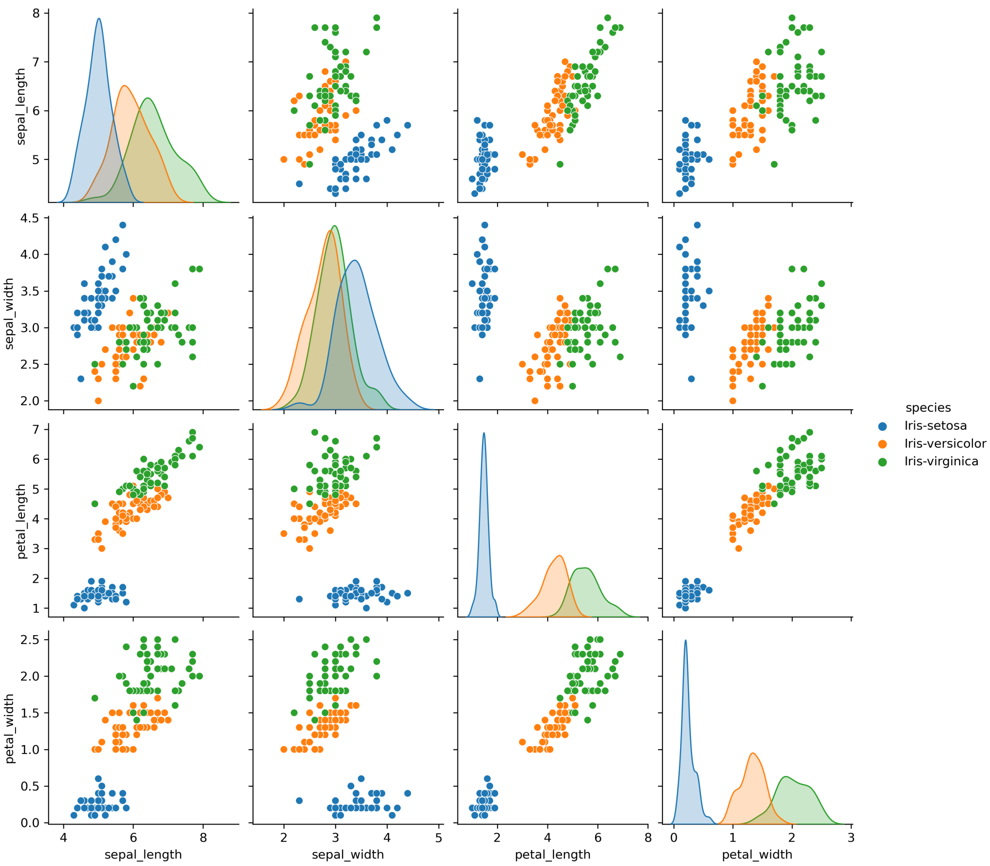

Figure 1: Your caption here

It is evident that we have three equally-sized classes in the dataset. One of them, *Iris-setosa*, can be easily classified based on the visual representation, while the other two, *Iris-versicolor* and *Iris-virginica*, pose more challenges in classification.

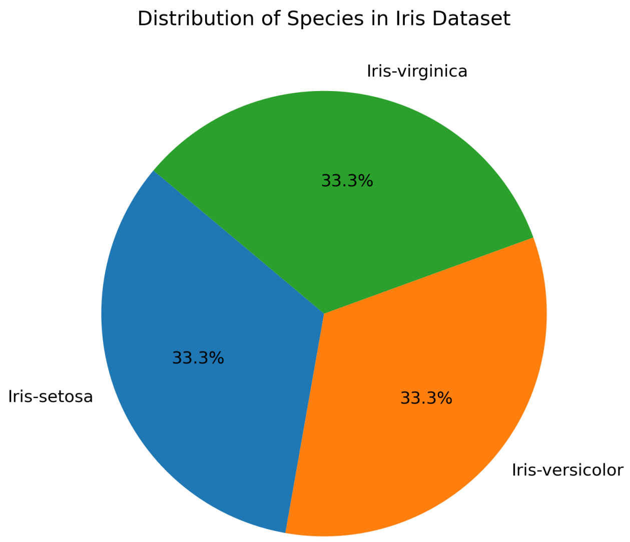

Figure 2: Your caption here

Note: We need to binarize the 'Species' attribute. Previously, I removed all but the *Iris-setosa* samples and couldn't understand why my accuracy was always 100%. Now, I realize that approach was incorrect. We will now create three separate datasets from this single dataset.

Therefore, we will evaluate each class individually, and I anticipate the accuracy to be around 100% for *Iris-setosa*, and approximately 70% for both *Iris-versicolor* and *Iris-virginica*, as the visual representation shows that the green and orange points (*versicolor* and *virginica*) are intermingled but still distinguishable.

3.1 Binarization of Attributes

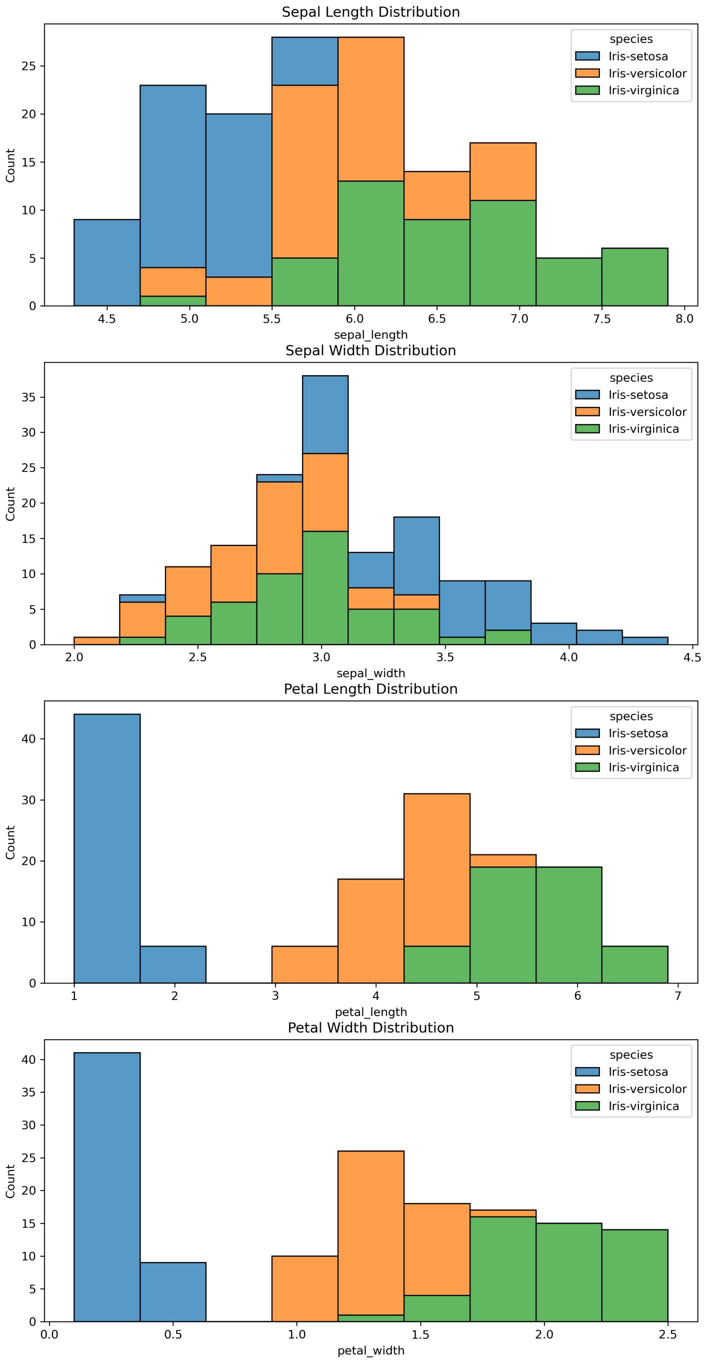

Figure 3: Your caption here

Let's explore how we can binarize the attributes. Referring to the image (figure 3), we can divide the data into intervals.

1. For **Sepal Length Distribution**, the intervals are: [4.5 - 5.5], [5.5 - 6.5], and [6.5 - 8].
2. For **Sepal Width Distribution**, the intervals are: [2.0 - 3.25] and [3.25 - 4.5].
3. **Petal Length Distribution** will be divided into: [1-3], [3-5], and [5-7].
4. Lastly, **Petal Width Distribution** will be segmented into: [0.0-1.0], [1.0-2.0], and [2.0-3.0].

Certainly, we could have calculated the optimal intervals based on the density of the given classes, but visually analyzing the image is more enjoyable than sitting and crunching numbers.

The binarization of the dataset attributes has been successfully completed as planned. The following table presents the binarized attributes with their respective ranges. Each attribute is now represented as a boolean value, indicating the presence or absence of the attribute within the specified range. This table illustrates the transformed dataset, aligning well with our binarization strategy.

#	Column	Non-Null Count
0	sepal_length(4.3, 5.3)	150 non-null bool
1	sepal_length(5.3, 6.3)	150 non-null bool
2	sepal_length(6.3, 7.3)	150 non-null bool
3	sepal_length(7.3, 8.3)	150 non-null bool
4	sepal_width(2.0, 3.25)	150 non-null bool
5	sepal_width(3.25, 4.5)	150 non-null bool
6	petal_length(1.0, 3.0)	150 non-null bool
7	petal_length(3.0, 5.0)	150 non-null bool
8	petal_length(5.0, 7.0)	150 non-null bool
9	petal_width(0.1, 1.1)	150 non-null bool
10	petal_width(1.1, 2.1)	150 non-null bool
11	petal_width(2.1, 3.1)	150 non-null bool

Table 2: Transformed Binarized Dataset

3.2 Evaluation of Machine Learning ModelsL

We've split the dataframe into three, meaning we now have three dataframes: is_setosa, is_versicolor, and is_virginica. Let's start by simply checking the accuracy for each set using standard machine learning algorithms.

3.2.1 Performance on is_setosa

height**Model**	**Accuracy**	**F1 Score**	**ROC-AUC**
DecisionTree	1.000	1.000	1.0
RandomForest	1.000	0.977	1.0
KNN	0.978	0.977	1.0
NaiveBayes	1.000	1.000	1.0
LogisticRegression	1.000	1.000	1.0

3.2.2 Performance on is_versicolor

Model	**Accuracy**	**F1 Score**	**ROC-AUC**
DecisionTree	0.889	0.887	0.893
RandomForest	0.889	0.864	0.959
KNN	0.867	0.864	0.908
NaiveBayes	0.844	0.841	0.861
LogisticRegression	0.889	0.886	0.905

3.2.3 Performance on is_virginica

Model	Accuracy	F1 Score	ROC-AUC
DecisionTree	0.844	0.643	0.864
RandomForest	0.889	0.780	0.909
KNN	0.889	0.780	0.866
NaiveBayes	0.644	0.625	0.821
LogisticRegression	0.911	0.852	0.893

3.2.4 Summary and Next Steps

As we anticipated!
The results indicate that the performance varies across different models and datasets. For the is_setosa dataset, all models achieved high accuracy, F1 Score, and ROC-AUC values.

Next, we will shift our focus to the Formal Concept Analysis (FCA) to explore how it performs on these datasets compared to the traditional machine learning models.

If you are reading this, know that the report is still being written right now....

3.3 Results

3.3.1 Neural network with fitted edge weights

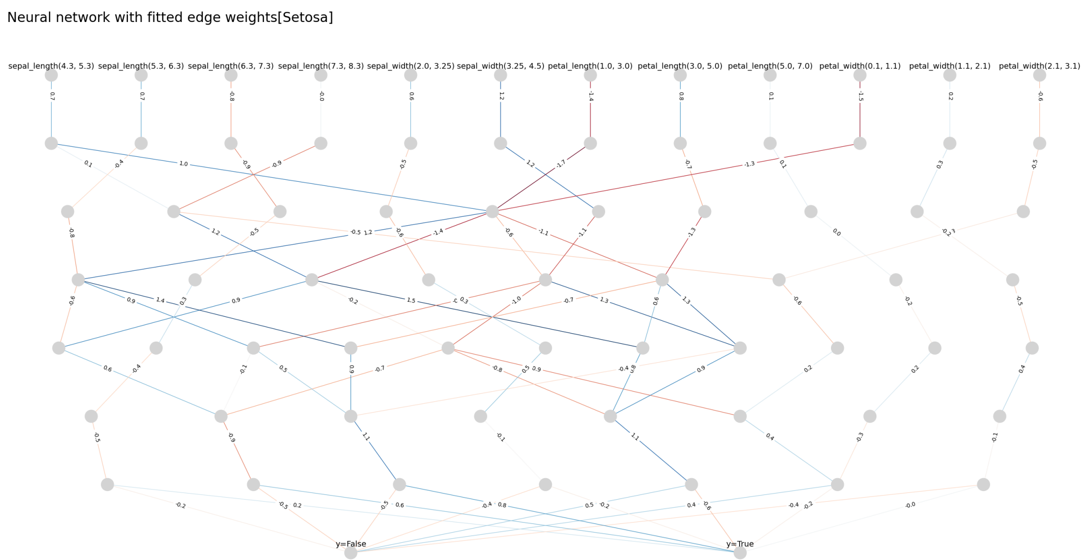

Figure 4: Neural network with fitted edge weights[Setosa]

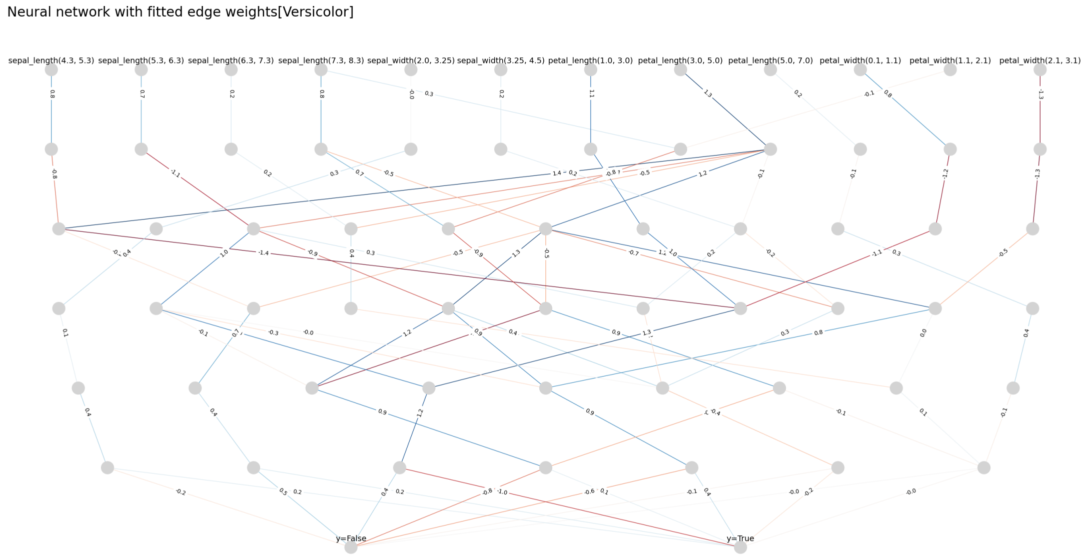

Figure 5: Neural network with fitted edge weights[Versicolor]

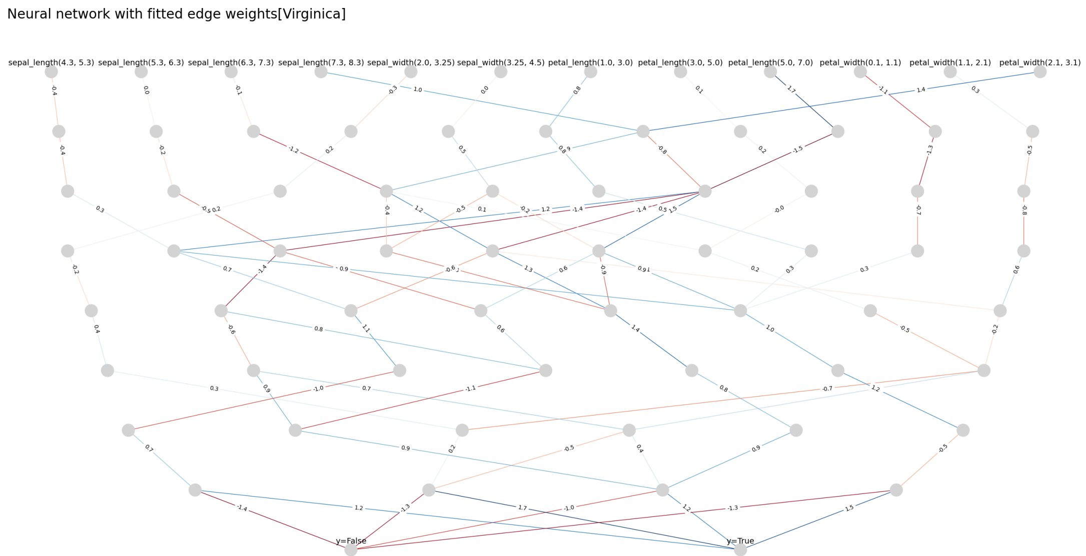

Figure 6: Neural network with fitted edge weights[Virginica]

3.3.2 Results for Versicolor Dataset

Model	Accuracy	F1 Score	ROC-AUC
FCA Model	1.0	1.0	1.0
DecisionTree	1.000	1.000	1.0
RandomForest	1.000	0.977	1.0
KNN	0.978	0.977	1.0
NaiveBayes	1.000	1.000	1.0
LogisticRegression	1.000	1.000	1.0

3.3.3 Results for Versicolor Dataset

Model	Accuracy	F1 Score	ROC-AUC
FCA Model	0.911	0.895	0.917
DecisionTree	0.889	0.887	0.893
RandomForest	0.889	0.864	0.959
KNN	0.867	0.864	0.908
NaiveBayes	0.844	0.841	0.861
LogisticRegression	0.889	0.886	0.905

3.3.4 Results for Virginica Dataset

Model	Accuracy	F1 Score	ROC-AUC
FCA Model	0.911	0.800	0.849
DecisionTree	0.844	0.643	0.864
RandomForest	0.889	0.780	0.909
KNN	0.889	0.780	0.866
NaiveBayes	0.644	0.625	0.821
LogisticRegression	0.911	0.852	0.893

3.4 Summary and Conclusion

The comparison of FCA with traditional machine learning models on the three datasets (Setosa, Versicolor, Virginica) shows that FCA performs competitively, especially in the Setosa dataset where it achieved perfect scores. In the Versicolor and Virginica datasets, FCA shows good performance, although there is some variation in results across different models.

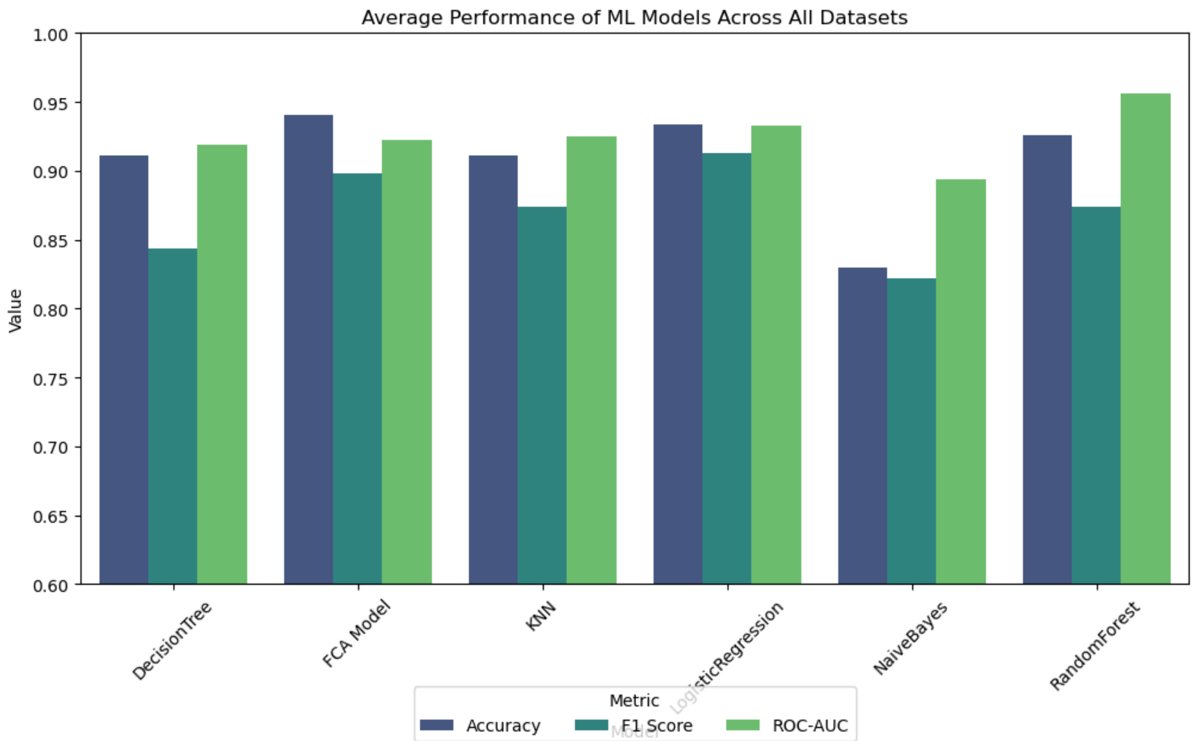

Figure 7: Average

4 Working with Dataset 2

This dataset includes a variety of attributes related to the passengers' travel experience. Key columns include 'Gender', 'Customer Type', 'Age', 'Type of Travel', 'Class', and 'Flight Distance'. It also contains ratings for various services such as 'Inflight wifi service', 'Departure/Arrival time convenient', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', and 'Cleanliness'. Additionally, the dataset tracks 'Departure Delay in Minutes' and 'Arrival Delay in Minutes'. The target variable seems to be 'satisfaction', which indicates whether a passenger was satisfied or not. This dataset provides a comprehensive view of factors that might influence passenger satisfaction during air travel, making it suitable for analyzing and predicting passenger satisfaction levels

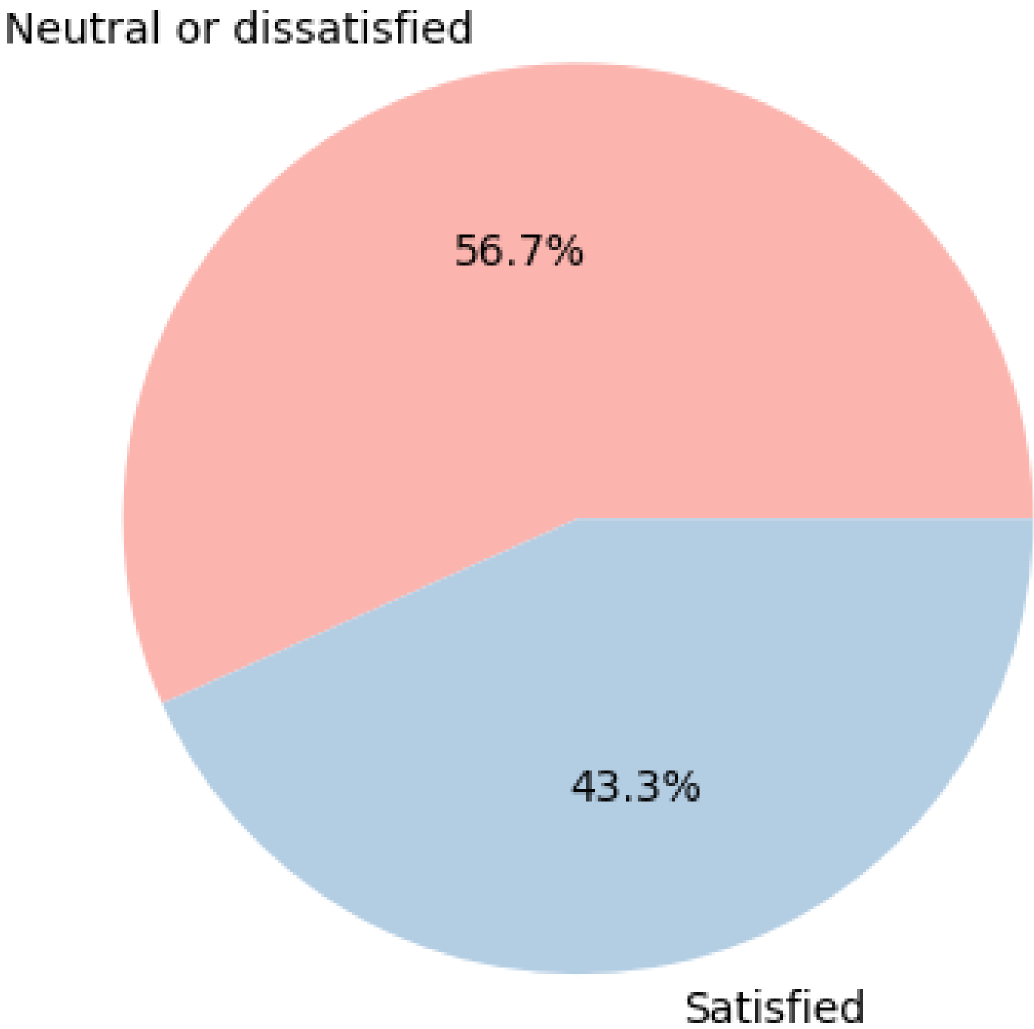

Figure 8: Your caption here

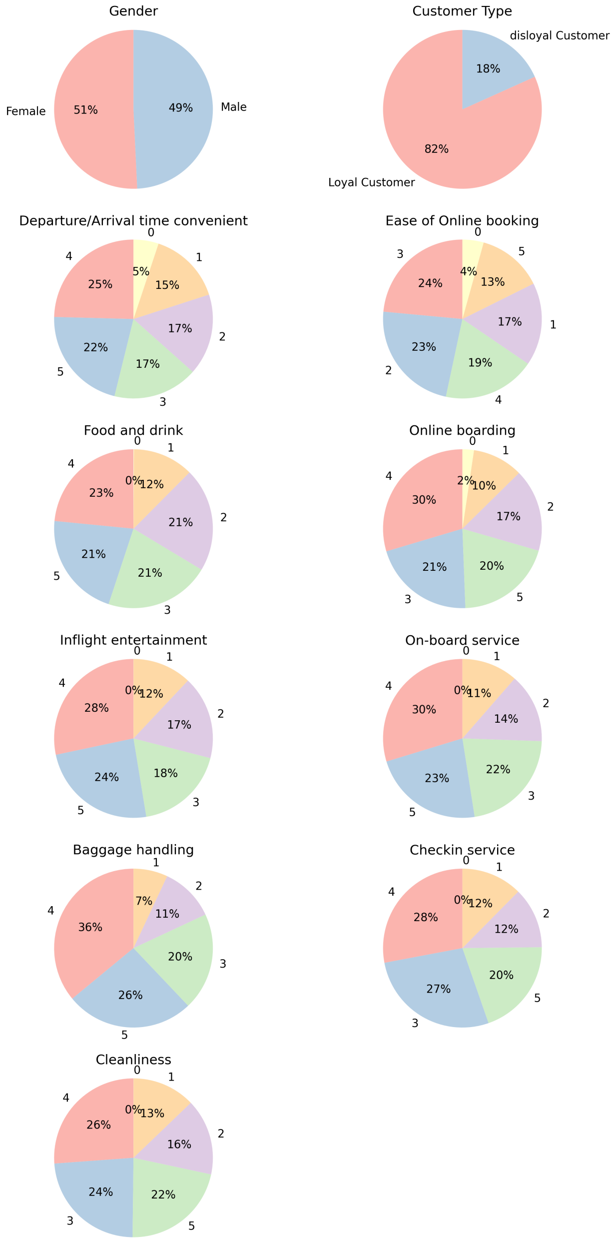

Figure 9: Your caption here

4.1 Binarization of Attributes

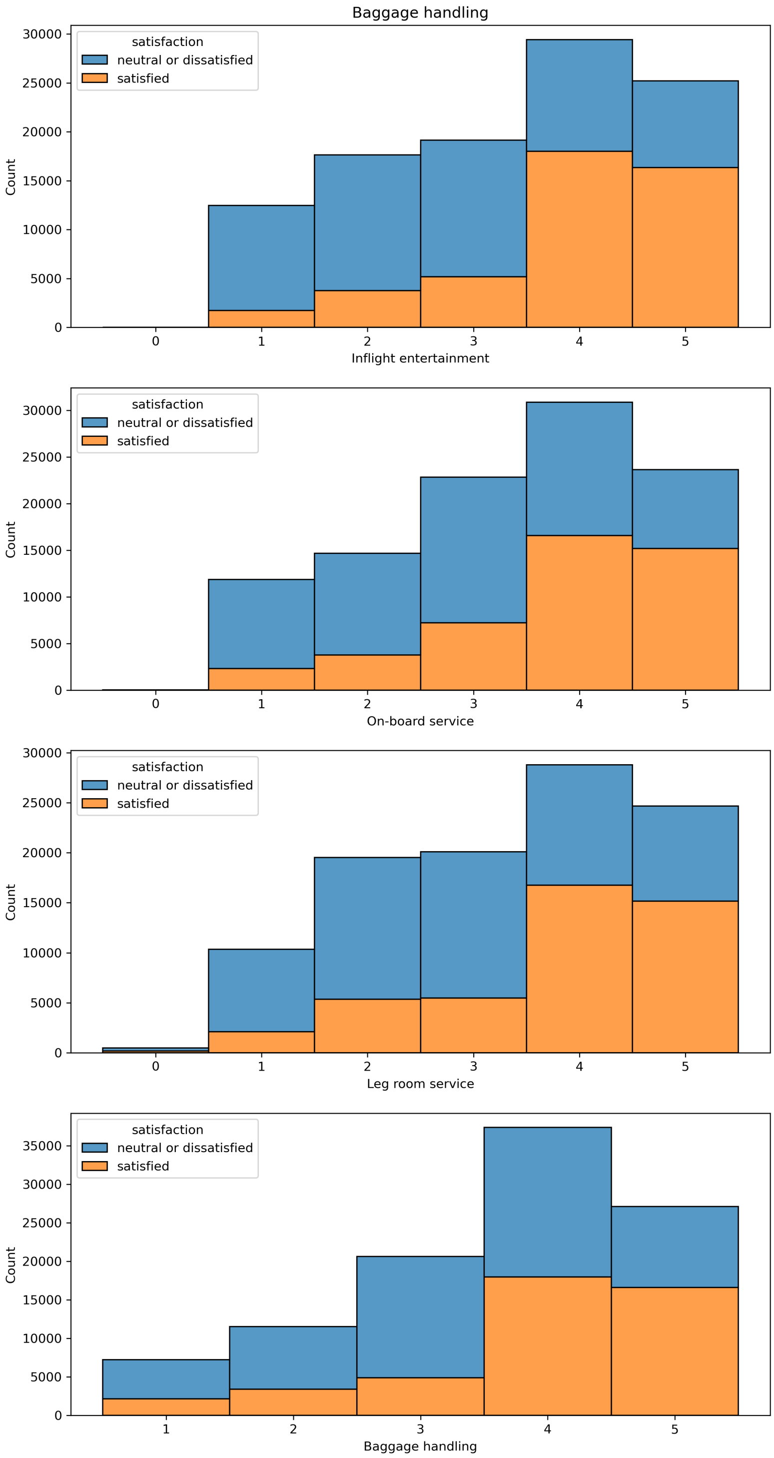

Figure 10: Your caption here

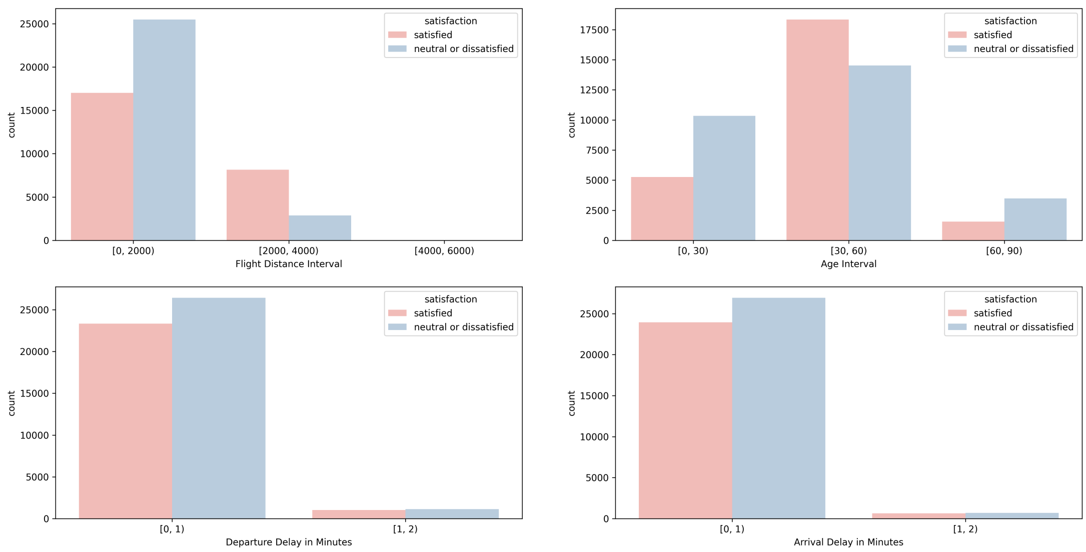

Figure 11: Your caption here

In our approach to analyzing the dataset, we simplified the ratings by categorizing scores of 0 to 3 as 'dissatisfied' and scores of 4 to 5 as 'satisfied'. This binary classification was evident across all visualizations.

For the attributes 'Age' and 'Flight Distance', optimal interval sizes were manually determined. The interval size for 'Age' was set at 30 years, and for 'Flight Distance', it was set at 2000.

The treatment of delay data was particularly interesting. After conducting several experiments, it was decided to binarize the delay data as follows: A value of 0 was categorized as 'no delay', while any value above 0 was considered a delay. This approach was based on the observation that the existence of a delay, irrespective of its length, is the critical factor, as indicated by our graphical analyses.
The dataset underwent significant transformation to facilitate the analysis. Key steps included:

- One-Hot Encoding was applied to categorical attributes such as 'Gender', 'Customer Type', 'Type of Travel', and 'Class'.
- The 'satisfaction' column was binarized, with values being classified as 'True' for 'satisfied' and 'False' otherwise.
- The classes 'Business', 'Eco', and 'Eco Plus' were simplified into two categories: 'Business' and 'Eco'.
- The 'Departure Delay in Minutes' and 'Arrival Delay in Minutes' were binarized into 'no delay' (value of 0) and 'delay' (any value above 0).

This process resulted in the following transformed dataset structure:

#	Column	Non-Null Count
0	Inflight wifi service_high	88594 non-null bool
1	Departure/Arrival time convenient_high	88594 non-null bool
2	Ease of Online booking_high	88594 non-null bool
3	Gate location_high	88594 non-null bool
4	Food and drink_high	88594 non-null bool
5	Online boarding_high	88594 non-null bool
6	Seat comfort_high	88594 non-null bool
7	Inflight entertainment_high	88594 non-null bool
8	On-board service_high	88594 non-null bool
9	Leg room service_high	88594 non-null bool
10	Baggage handling_high	88594 non-null bool
11	Checkin service_high	88594 non-null bool
12	Inflight service_high	88594 non-null bool
13	Cleanliness_high	88594 non-null bool
14	Age 7-37	88594 non-null bool
15	Age 37-67	88594 non-null bool

#	Column	Non-Null Count
16	Age 67-97	88594 non-null bool
17	Flight Distance 31-2031	88594 non-null bool
18	Flight Distance 2031-4031	88594 non-null bool
19	Flight Distance 4031-6031	88594 non-null bool
20	Satisfaction	88594 non-null bool
21	Gender Female	88594 non-null bool
22	Gender Male	88594 non-null bool
23	Customer Type Loyal Customer	88594 non-null bool
24	Customer Type Disloyal Customer	88594 non-null bool
25	Type of Travel Business Travel	88594 non-null bool
26	Type of Travel Personal Travel	88594 non-null bool
27	Class Business	88594 non-null bool
28	Class Eco	88594 non-null bool
29	Departure Is Delay	88594 non-null bool
30	Arrival Is Delay	88594 non-null bool