Applying data visualization techniques in Mushroom dataset (edible or poisonous) based on R

Most. Lailun Nahar Aurthy  
 BSc. In CSE  
 American International University-Bangladesh  
 Dhaka, Bangladesh  
 [20-43997-2@student.aiub.edu](mailto:20-43997-2@student.aiub.edu)

Md. Iftear Hossain Ratul

BSc. In CSE  
American International University-Bangladesh  
 Dhaka, Bangladesh  
[iftear.ratul@gmail.com](mailto:iftear.ratul@gmail.com)

Md. Faisal   
 BSc. In CSE  
American International University-Bangladesh  
 Dhaka, Bangladesh  
[faisalahmed7595@gmail.com](mailto:faisalahmed7595@gmail.com)

Abir Hasan  
BSc. In CSE  
American International University-Bangladesh  
 Dhaka, Bangladesh  
 [21-45672-3@student.aiub.edu](mailto:21-45672-3@student.aiub.edu)

ABSTRACT

This study investigates the efficacy of data visualization methodologies in discerning the edibility of mushrooms, drawing from a Kaggle-acquired dataset featuring diverse attributes including cap diameter and gill attachment. Employing histograms, bar graphs, and box plots, the research elucidates intricate data patterns, crucial for discriminating between edible and poisonous fungal species. Methodologically, the study underscores the utility of visualization in comprehending mushroom attributes and underscores the role of robust preprocessing techniques in ensuring dataset integrity. Insights derived from skewness analysis, correlation assessments, and distribution examinations inform the development of sophisticated classification algorithms such as K-nearest neighbors (KNN) and Multilayer Perceptron (MLP). This research accentuates the transformative potential of visualization-driven approaches in mushroom identification, paving the way for practical implementations such as mobile applications tailored to enhance food safety protocols.

**INTRODUCTION**

The data visualization process of extracting information from large set database. With the proposal on this data set apply some reasonable technique of the data visualization to identify for the mushroom is edible or not. This highly satisfying food source contains amino acids, carbohydrates, fiber, important vitamins, and minerals. Mushroom are also used into the pharmaceutical industry for making necessary medicine. There have many types of edible mushroom, as like button mushroom, portobello mushroom, shiitake mushroom, oyster mushroom, porcelain mushroom etc. Those mushrooms are use most for edible sector. Apply the data visualization, on the mushroom data set to ensure and improve the safety in daily life of the people who use the mushroom. Edible mushrooms are mushrooms that can be consumed as fresh or dried fruit parts. With the visualization process clean the data by logistic with theoretical acknowledgment to maintain or cover all the points to show the clearance. This data was divided into two parts: 70% training and 30% testing.

In the mushroom data set there have two types of data. One is numerical, on the other hand other is categorical. In the data visualization techniques those work combinedly for the nine individual columns and show the relation for the target attributes (edible mushroom). In manually, reach to all the attributes are very time consuming. For that reason, the data visualization is more preferable for this data set.

With the advancements in visualization process and pattern recognition process, there is an opportunity to develop automated systems for identifying edible mushrooms. The data visualization process, provides a promising avenue by using labeled training data to construct predictive models.

**LITERATURE REVIEW**

The classification of mushroom particularly distinguishing between edible and poisonous varieties, which is crucial given the commonness of mushroom species in Indonesia. The study shows the comparison between three popular classification algorithms in data mining. They are- Decision Tree (C4.5), Naïve Bayes and Support Vector Machine (SVM). Using the WEKA tool to conduct this experiment. Testing between two families of mushroom Agaricus and Lepiota, it is found that both C4.5 and SVM have 100% accuracy where C4.5 is faster process. The decision tree by C4.5 reduces the attribute numbers to 22 to 5 for identification any mushroom. The proper algorithm selection is significance of enhancing accuracy and efficiency in mushroom classification. Also, the author suggested for implementing practical application such as mobile-based application and image processing techniques to identify the edible or poisonous mushrooms.

Data mining process is used to extract information from a large dataset. Here in-depth exploration of data mining techniques applied of mushroom classification algorithm such as ID3, CART and Hoeffding Tree (HT). The study evaluates the accuracy, efficiency and error rates through these algorithms which conducted in the R studio software environment. Performing and comparing the algorithms, the ultimate finding that Hoeffding Tree performs better than both ID3 and CART algorithms, exhibiting higher accuracy (100%) and lower error rates (0%). The time required to construct a decision tree is lower (0.3 sec) than other two methods. Also, the finding of the research helps to understand the impact of data mining techniques in real-world scenarios, specially health-related risks associated with food consumption.

Edible mushroom identification has been a topic of interest and research in several domains, including mycology, food science, and computer science. This an overview of significant studies and approaches that have contributed to the progress of edible mushroom identification, notably the use of supervised learning techniques. Supervised learning techniques, a subset of machine learning algorithms, have been increasingly applied to edible mushroom identification. These techniques make accurate classification possible by using labeled training data to estimate the class of unseen instances. Studies have explored various supervised learning algorithms, including decision trees, support vector machines, neural networks and features indicative of edible mushroom species. Intelligence visualization techniques play a crucial role in conveying information and knowledge derived from mushroom-related data. Real-time processing, analytics, and batch processing are essential considerations, requiring solutions that prioritize speed, reliability, and scalability. The classification of mushroom data into structured and unstructured formats further complicates processing requirements, necessitating tailored strategies for each data type. Data processing and display provide substantial obstacles in edible mushroom identification, particularly for big data and data science research. Visualization techniques, such as tree maps, dendrograms, and hierarchical clustering methods, have been instrumental in representing the evolutionary relationships among mushrooms based on genetic data.

This paper investigates the visual analysis of different classification algorithms applied to mushroom datasets. It compares the performance of algorithms such as (Decision trees, support vector machines, and neural networks) in classifying mushrooms based on visual features. Focuses on exploring various interactive visualization techniques for mushroom datasets. It discusses the importance of visualization in understanding complex datasets and presents several interactive visualization tools specifically designed for mushroom datasets. The utilization of data visualization techniques, coupled with machine learning algorithms like K-Nearest Neighbors (KNN), has garnered attention for its efficacy in analyzing mushroom datasets. KNN, being a simple yet effective algorithm, is often used in conjunction with dimensionality reduction techniques for visualization purposes. Recent research has highlighted the relationship between data visualization techniques and the k-Nearest Neighbors (KNN) algorithm in the context of mushroom datasets. The Mushroom dataset, a common benchmark in this domain, presents categorical attributes describing various ecological features of mushrooms alongside their edibility status. Visualization tools such as scatter plots, pair plots, and decision boundary visualizations have been instrumental in elucidating relationships between attributes and interpreting the behavior of KNN models. The researchers have leveraged the interpretability of the KNN algorithm to gain insights into its decision-making process. Additionally, efforts have been directed towards evaluating and optimizing KNN models using visualization-based approaches, aiming to enhance model performance and accuracy. Despite the progress made, challenges persist in effectively integrating these techniques, warranting further exploration and development of novel visualization methodologies tailored to mushroom classification tasks.

The study described in the passage focuses on the classification of mushrooms into edible or poisonous categories using machine learning techniques and data mining algorithms. The experiment compared classification algorithms such as Decision Table, JRip, Logistic Regression, SGD, and SMO, utilizing a data mining tool to analyze the mushroom dataset, specifically testing the families Agaricus and Lepiota. While the accuracy rates of the algorithms were not explicitly provided, the emphasis on proper algorithm selection highlights the importance of accurate and efficient mushroom classification to prevent potential health risks associated with consuming poisonous varieties. The author suggests practical applications like developing a mobile application with image processing capabilities for mushroom classification, underscoring the significance of accurately categorizing mushrooms for safety reasons.

The study described in the passage focuses on comparing different classification algorithms for mushroom classification. The classification algorithms compared in the study include Support Vector Machine (SVM), Random Forest (RF), and k-Nearest Neighbors (k-NN). The experiment was conducted using the Weka tool. The experiment tested mushroom families such as Agaricaceae, Boletaceae, and Russulaceae. The accuracy rate of the algorithms ranged from 92.5% to 98.5%. Proper algorithm selection is emphasized in the study to ensure accurate and efficient classification of mushrooms, which is crucial for various applications such as food safety, ecological studies, and medicinal purposes. The author suggests practical applications for mushroom classification, including automated identification systems for mycologists and mushroom enthusiasts. The classification of mushrooms is particularly important due to the potential risks associated with misidentification, such as poisoning from toxic mushrooms. The key findings of the study indicate that SVM and RF algorithms outperformed k-NN in terms of accuracy and efficiency, with SVM achieving the highest accuracy rate of 98.5%.

In this paper discuss the data visualization on the understanding and complex data also show the diversity domain of the data like business, sports, social science, humanities, environmental science and etc. The main aim or goal of this paper is data visualization tools and techniques and applications. The three categories of the data visualization tools (spread sheet, data visualization software, programing libraries). Emphasize the data visualization technique which have various domain and making decision, improve outcomes, advancing knowledge. The visualization exactly used in each domain as like marker analysis, customer segmentation, survey analysis, policy evaluation and etc. The paper serves as a comprehensive guide to data visualization tools and techniques offering instant into their applications across different fields.

This paper available for the classification of the mushroom into poisonous and aid able category using machine learning algorithm. The historically significant of the mushroom as a food source and there nutritional and medicinal properties. The data set consist for the 22 attributes both for poisonous and edible mushrooms. The study employees tree classification algorithms and random forest algorithm including their structure and the decision-making process. The method is used for the parameter optimization. Evaluation matrix and the accuracy, recall, precision is explained. The random forest algorithm outperforms of the other two algorithms with the performance. The paper cities relevant literature that contributes to understand the mushrooms and the machine learning algorithms. The study aims distinguished between addible and poisonous mushroom providing the safer.

**DATA COLLECTION AND PREPOSESSING**

In the data analysis work, data collection is often the most time-consuming part of the research. In this section, from the Kaggle website the dataset is collected and use the R language for data prepossessing and applying data visualization techniques.

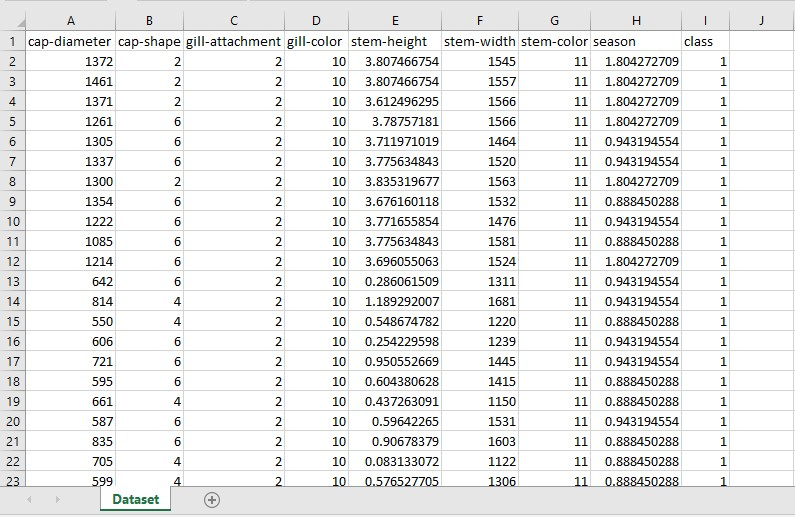


Figure 1: Mushroom Dataset

1. **DATA COLLECTION**

There is some popular website which gives us some collected dataset and Kaggle is one of them. A comprehensive mushroom dataset is gathered which includes information on their edibility or not and some relevant characteristics. The dataset contains 9 attributes and 54035 instances. The 9 attributes are- cap-diameter, cap-shape, gill-attachment, gill-color, stem-height, stem-width, stem-color, season and class. And the class is our target attribute where ‘0’ refers to edible and ‘1’ refers to poisonous. There are 24360 instances belonging to class 0 and 29675 belonging to class 1.

1. **PREPOSESSING**

Kaggle is a widely one website for collecting dataset. Those datasets are clean. Having no outlies and missing values in the dataset. So, we don’t need to prepossess the dataset for applying data visualization techniques.

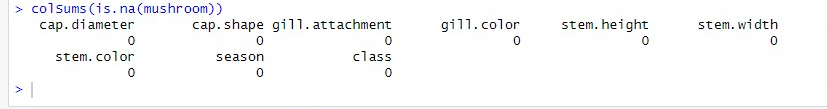


Figure 2: Prepossessed Dataset (No missing value)

**DATA VISUALIZATION TECHNIQUES**

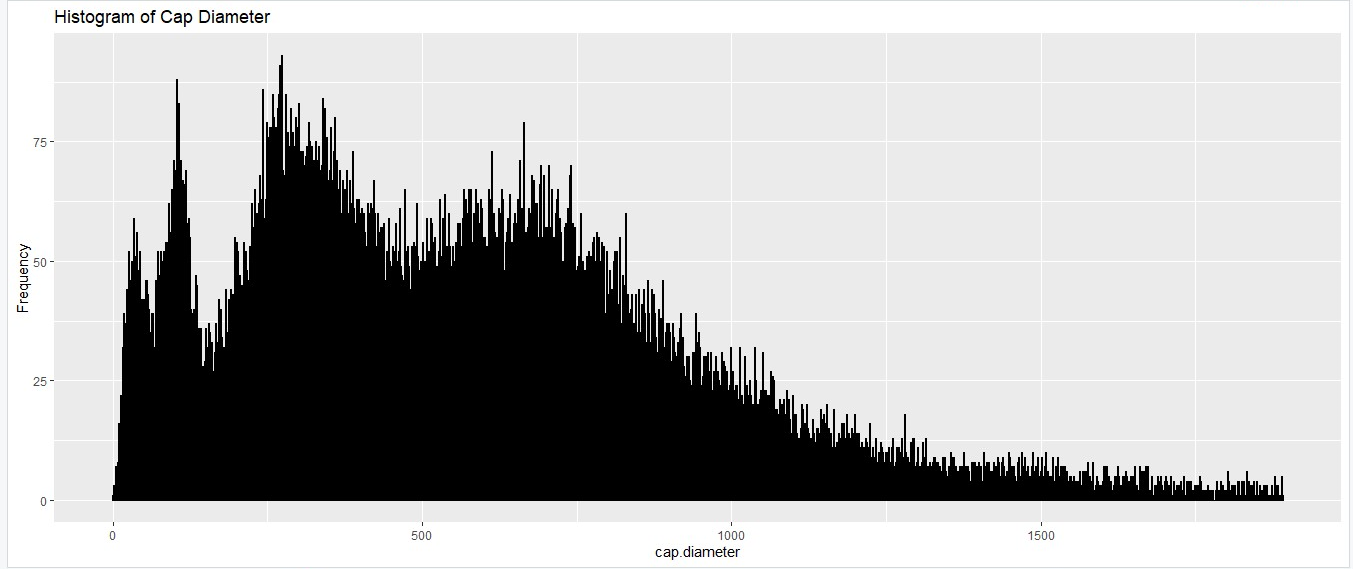
Data visualization is the technique if interpreting data and representing it in a graphical format. It helps the viewer to properly understand the significance of data by summarizing method and also describe the large dataset in an accessible, understandable and straightforward way. There are two primary uses for data visualization such as- to explore the data and to communicate the data. Some commonly used data visualization techniques are-

1. Histogram
2. Bar graph
3. Boxplot
4. Line histogram
5. Skewness
6. Line graph
7. Scatterplot
8. Scatter Matrix

**1. Histogram**

Histogram is one of the most basic visualization techniques and the graphical representation of quantitative data. It shows the distribution of the data by plotting the frequency of occurrence in a range. Histograms normally are used for finding the central location, range, and shape of distribution. Understanding the distribution can help in dealing with the missing values if the dataset contains some. Here, the attribute called cap-diameter is located in the horizontal axis and the frequency in on the vertical axis. The values with longer plots indicate that more values are concentrated there.

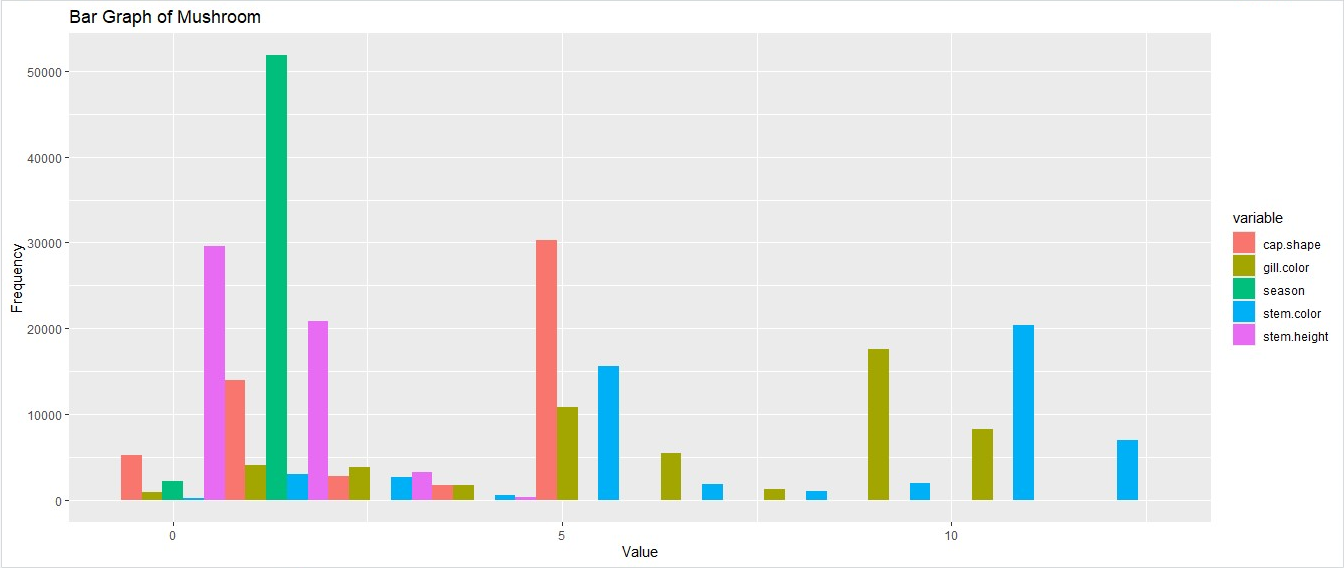
A. Wibowo, Y. Rahayu, A. Riyanto, and T. Hidayatulloh, "Classification Algorithm for Edible Mushroom Identification," in 2018 International Conference on Information and Communications Technology (ICOIACT), 2018.

****

**Figure 3: Histogram of cap-diameter**

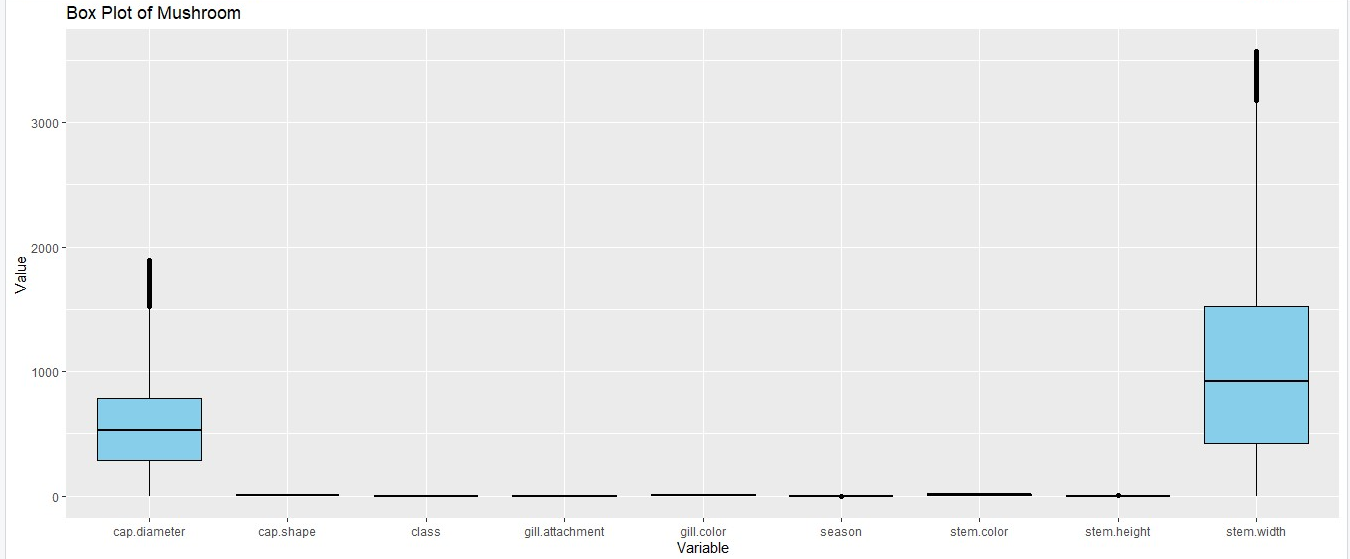
**2. Bar Graph**

It is best suited for the visualization of categorical data with rectangular bars because it allows to easily identify the difference between features values by measuring the length of the bars. The bar graph of mushroom represents the relation between frequency in y-axis and stem-height in x-axis.

****

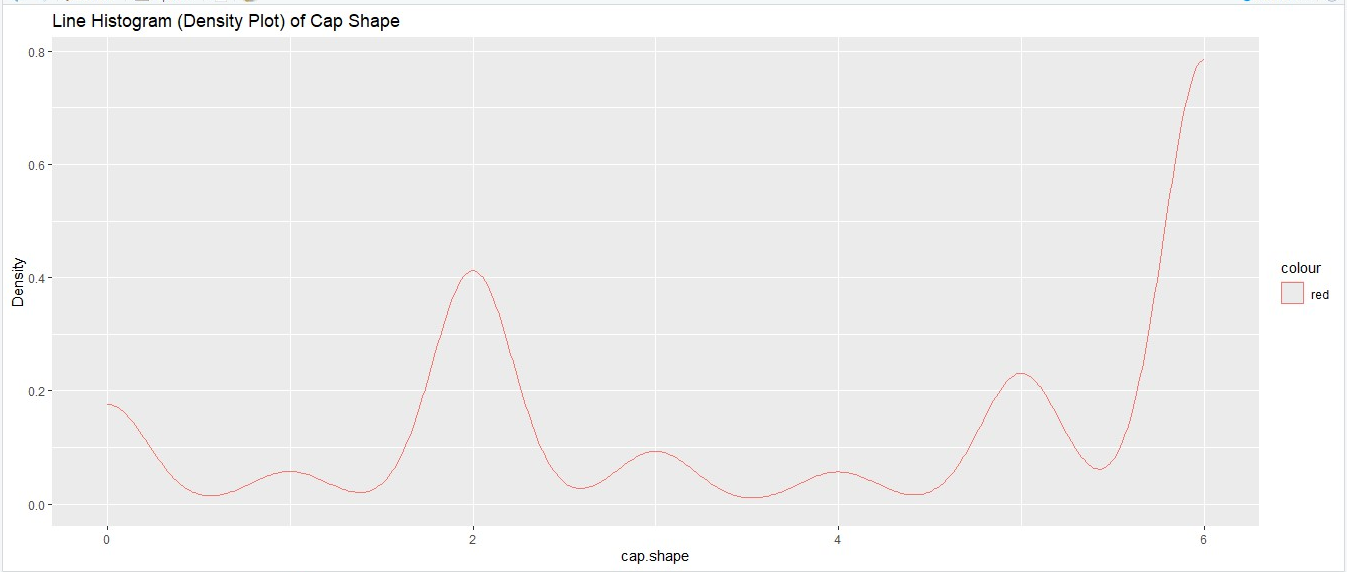
**Figure 4: Bar Graph of Mushroom**

**3. Boxplot**

Boxplot, the simplest way to detect outliers and an easy way to observe the distribution of the data. In the graph all the attributes are plotted in x-axis. The central 50% of data, with a middle line shows the median value. Lines extending from the box capture the range of the remaining data and if any data falls outside of the lines indicates an outlies. The dataset is clean so there is not outlies. The cap-diameter and the stem-width have quite large values so it is seen clearly. But rest if the other attributes like cap-shape, gill-attachment, gill-color, season, stem-color and stem-height value is too same so it is unclear. **Figure 5: Boxplot of Mushroom**

**4. Line Histogram**

Line histogram evaluates how closely the heights of the bars follow the shape of the line. If the bars follow the fitted distribution line closely, then the data fits the distribution well. Here, cap-shape attribute is in the x-axis and density is in the y-axis and the bars fitted the line closely, so the data fits the distribution well.

****

**Figure 6: Line Histogram (Density Ploy) of cap-shape**

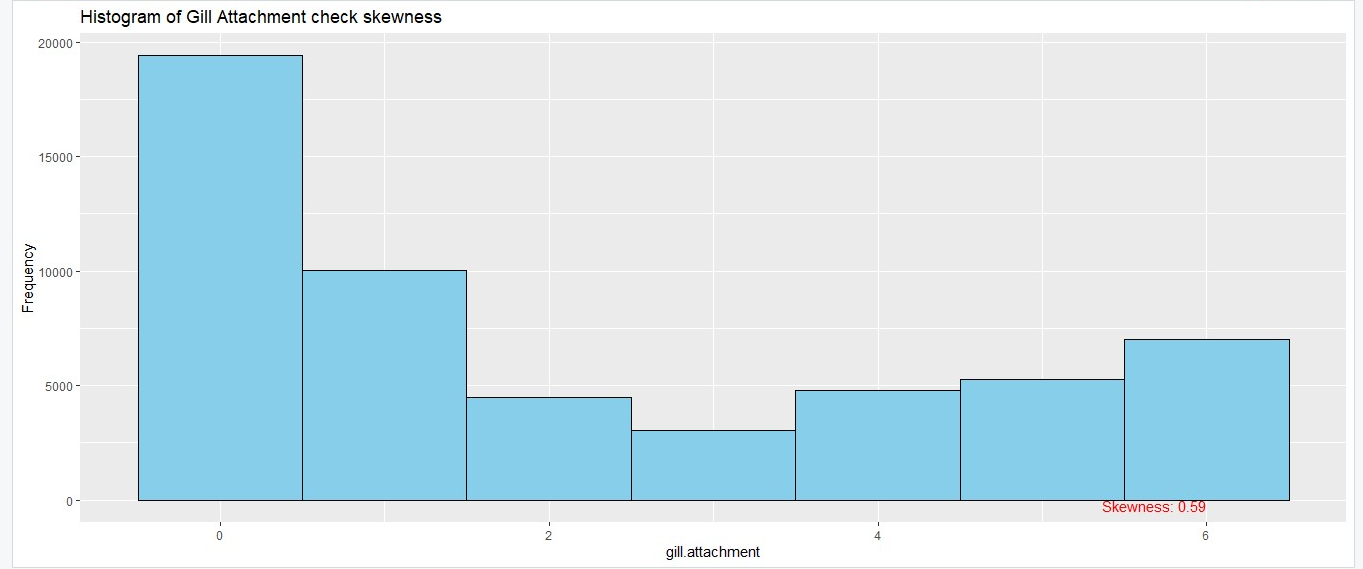
**5. Skewness**

Skewness is the degree of asymmetry observed in a probability distribution. When the mean, median and mode values are same, so there is no skewness. It is also called symmetrical skewness.

There are two types of skewness- Positive and Negative skewness. Here, in the x-axis gill-attachment attribute is located and frequency is located in y-axis.

In histogram, if most of the data falls in the right of the graph peak is call positive skewness or right-skewed histogram. In a histogram skewed to the right, mode value is to the left of the value of median and at most right mean is situated. In the positive skewness the relation between mean, median and mode is given as Mode < Median < Mean.

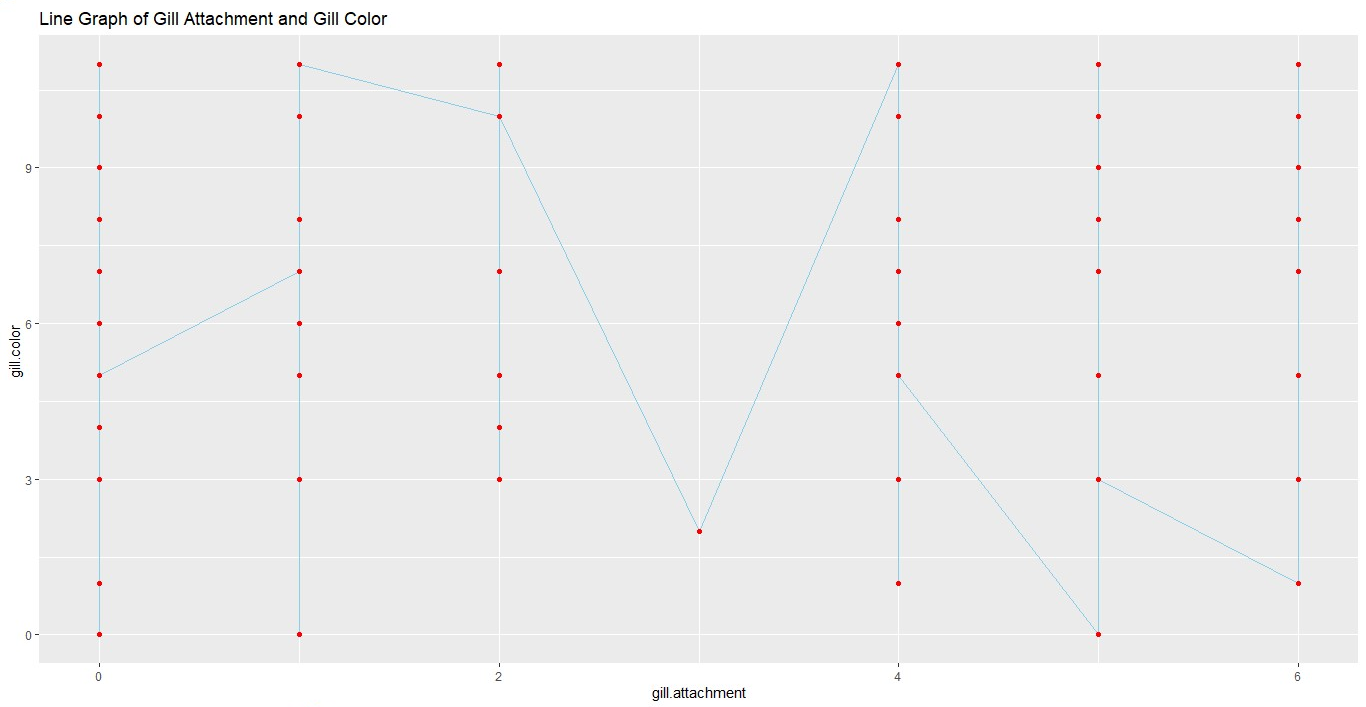
If most of the data falls in the left of the graph peak is called negative skewness or left-skewed histogram. Here, the mean value is in the left most side and mode value is in right side. Median is located in the middle as usual. The relationship can be showed as Mean < Median < Mode.



**Figure 7: Histogram of gill-attachment check skewness**

**6. Line Graph**

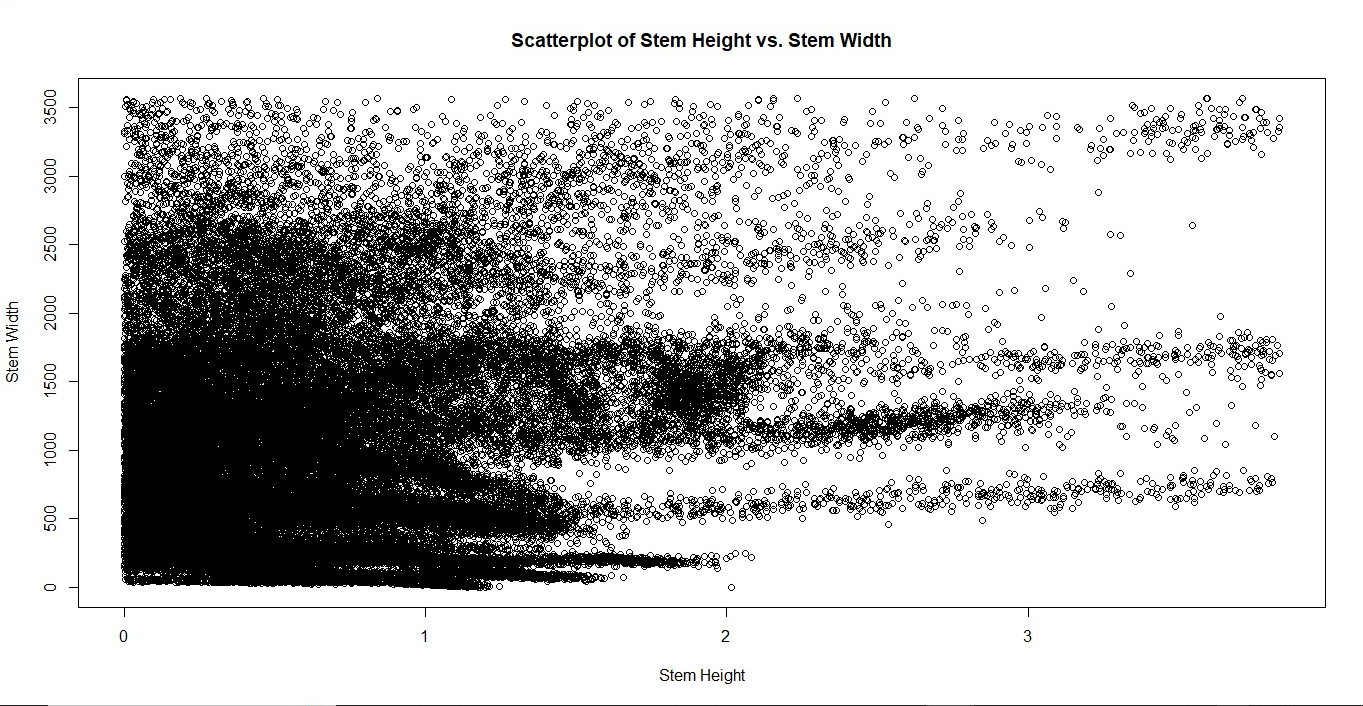
Connection of the data points by straight line between one or several numerical attributes is line graph or line chart. Line graph are also used to represent quantitative data collected over a specific subject and a specific time interval. It gives a graphical representation of the changes that had occurred over a given period of time. In the line graph of mushroom, displays the connection of gill-color and gill-attachment attribute. Where gill-color is located in the vertical axis and gill-attachment is in horizontal axis.

****

**Figure 8: Line Graph of gill-attachment and gill-color**

**7. Scatterplot**

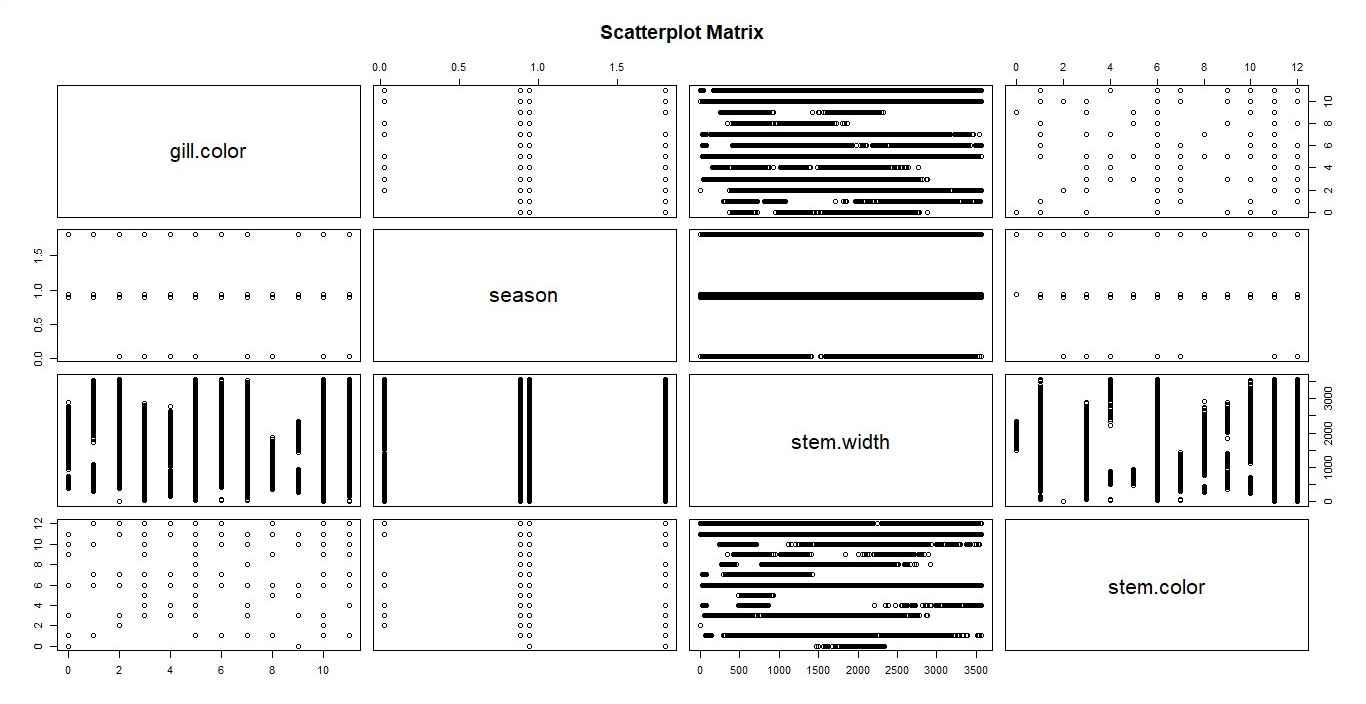
A scatterplot uses dots to represent the values for two different numeric attributes. The dots are not Scatterplot or scatter chart/graph are mainly used to identify the relationship between attributes. In the graph, where the dots are plotted on the horizontal and vertical axis that indicates values for individual data point. Identification of correlational relationships are common with scatterplot. Plotting particular horizontal value, what a good prediction would be for vertical value. The graph indicates that in the x-axis stem-height and in the y-axis stem-width is plotted. There is a large number of values are plotted so the display is quite crowded.

****

**Figure 9: Scatterplot of stem-height vs. stem-width**

8. Scatter Matrix

It is a grid of scatterplot used to visualize bivariate relationships between combinations of attributes. The matrix visualizes the relationship between pair of attributes also allow many relationships to be explored in one chart.



**Figure 10: Scatterplot Matrix**

**CONCLUSION**

The developed models can aid in preventing mushroom poisoning and promoting safe consumption. This study demonstrates the effectiveness of data visualization process that the way for future research in intelligent systems for mushroom identification and food safety. The not available value is also considered for every system and stablish the suitable and further the standard way to identify the edible. Mushroom for the general or user who have to easily find out the raw and solicit object by implemented the visualization process in the studio. The implementation shows a significant and important role into the market place, industry, medical places and other where mushroom have to use (identify the mushroom useable or not).

The proper algorithm (as like KNN, MLP, SLA etc.) for the data visualization in terms of purify the data or the deterministic value to identify the accuracy, precision, recall and the score. So, it is very clear that size of training set as well as selection of classification technique depending on the data to be analyzed is very important for data mining of patterns efficiently. It is also seen that performance of all the techniques is low when dataset size is small and the performance improves with increase in size of training set up to when training set is 70% of the whole dataset. In the future research, researcher can make the implementation convert into mobile application for easier to the user for apply the visualization. The researcher also converts to the image processing for the user to crystal clear visualization showing to the user with more effective information and further compare to the other algorithms and the techniques. With the classification determine the raw measurement by applied algorithm which means determine the main point of the errors. The showing bar diagram tell about the relation with columns data to data verification.

REFERENCES

[1] Ismail, S., A.R. Zainal, and A. Mustapha. Behavioral  
features for mushroom classification. in 2018 IEEE  
Symposium on Computer Applications & Industrial  
Electronics (ISCAIE). 2018. IEEE

[2] Ottoman, M.A., N.A. Alawad, and K. Nahar,  
Classification of mushroom fungi using machine  
learning techniques. International Journal of  
Advanced Trends in Computer Science and  
Engineering, 2019. 8(5): p. 2378-2385

[3] Hamonangan, R., M.B. Saputro, and C.B.S.D.K.  
Atmaja, Accuracy of classification of poisonous or  
edible mushrooms using naïve bayes and k-nearest  
neighbors. Journal of Soft Computing Exploration,  
2021. 2(1): p. 53- 60

[4] Maurya, P. and N.P. Singh. Mushroom classification  
using feature-based machine learning approach. in  
Proceedings of 3rd International Conference on  
Computer Vision and Image Processing. 2020.  
Springer.Conference Name:ACM Woodstock conference

[5] B.Lavanya and G.R.Preethi, “Performance Analysis of Decision Tree Algorithms on Mushroom Dataset”, *International Journal for Research in Applied Science & Engineering Technology (IJRASET****),*** Nov 2017, ISSN: 2321-9653, volume 5 Issue 11.

[6] Wibowo A., Riyanto A., Rahayu Y. and Hidayatulloh T., “Classification Algorithm for Edible Mushroom Identification”, *2018 International Conference on Information and Communications Technology (ICOIACT). Conference Short Name:WOODSTOCK’18*