A Project Report on

MUSHROOM CLASSIFICATION

Submitted in the partial fulfillment of the requirements for Machine Learning project of

BACHELOR OF TECHNOLOGY
In
INFORMATION TECHNOLOGY

Submitted by

V.BHARATH	18071A12B9
K.SAI SOWMITH	18071A1285
V.SAKETH	18071A12B8

DEPARTMENT OF INFORMATION TECHNOLOGY

VNR Vignana Jyothi Institute of Engineering & Technology
(Autonomous Institute, Accredited by NAAC with ‘A++’ grade and NBA)
Bachupally, Nizampet (S.O.) Hyderabad- 500 090,

April 2021

A Project Report on

ML BASED SURVEILLANCE

Submitted in the partial fulfillment of the requirements for Machine Learning project of

BACHELOR OF TECHNOLOGY
In
INFORMATION TECHNOLOGY

Submitted by

V.BHARATH	18071A12B9
K.SOWMITH	18071A1285
V.SAKETH	18071A12B8

Under the esteemed guidance of

PROJECT GUIDE
Mr. Sudhakar Yadav,
Assistant Professor,
Dept. of Information Technology, VNRVJIET

DEPARTMENT OF INFORMATION TECHNOLOGY
VNR Vignana Jyothi Institute of Engineering & Technology (Autonomous Institute, Accredited by NAAC with ‘A++’ grade and NBA) Bachupally, Nizampet (S.O.) Hyderabad- 500 090,

April 2021

VNR Vignana Jyothi Institute of Engineering & Technology

Autonomous Institute, Accredited by NAAC with ‘A++’ grade and NBA) Bachupally, Nizampet (S.O.)
Hyderabad- 500 090

Department of Information Technology

Date: July 2020

CERTIFICATE

This is to certify that the project work entitled **“MUSHROOM CLASSIFICATION”** is being submitted by **V.BHARATH SRI VARDHAN (18071A12B9), K.SAI SOWMITH REDDY (18071A1285), V.SAKETH (18071A12B8)** in partial fulfilment for the award of Degree of **BACHELOR OF TECHNOLOGY** in **INFORMATION TECHNOLOGY** to the Jawaharlal Nehru Technological University, Hyderabad during the academic year 2019-20 is a record of bona-fide work carried out by her under our guidance and supervision.

The results embodied in this report have not been submitted by the students to any other University or Institution for the award of any degree or diploma.

Project Guide	**Head of Department**
Mr. Sudhakar Yadav,	**Dr. G. SURESHREDDY**
Assistant Professor,	**Head of Department,**
Dept of IT,	**Dept of IT,**
VNRVJIET,	**VNRVJIET,**
Hyderabad.	**Hyderabad.**

VNR Vignana Jyothi Institute of Engineering & Technology

Autonomous Institute, Accredited by NAAC with 'A++' grade and
NBA) Bachupally, Nizampet (S.O.) Hyderabad- 500090.

Department of Information Technology

Date: April 2021

DECLARATION

I hereby declare that the project entitled "MUSHROOM CLASSIFICATION" submitted for the B.Tech Degree is my original work and the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles.

Signature of the Student:

V. BHARATH SRI VARDHAN	
V. SAKETH	
K. SAI SOWMITH REDDY	

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our beloved **President, Dr.D.N.Rao, VNR Vignana Jyothi Institute of Engineering &Technology** for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we record our deep sense of gratitude to our beloved **Principal, Dr.C.D.Naidu** for permitting us to carry out this project.

We express our deep sense of gratitude to our beloved professor **G. Suresh Reddy, Associate Professor and Head, Department of Information Technology, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad-90** for the valuable guidance and suggestions, keen interest and through encouragement extended throughout period of project work.

We take immense pleasure to express our deep sense of gratitude to our beloved Guide **Mr Sudhakar Yadav, Assistant Professor in Information Technology, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad,** for his valuable suggestions and rare insights, for constant source of encouragement and inspiration throughout my project work.

INDEX

S NO	CONTENTS	Page No
	ABSTRACT	i
	LIST OF IMAGES	ii
CHAPTER 1	: INTRODUCTION	9
CHAPTER 2	: LITERATURE REVIEW	10-12
	2.1 LITERATURE SURVEY	
CHAPTER 3	: ALGORITHMS	13-21
	3.1 CLASSIFICATION WORKFLOW	13
	3.2 KNN CLASSIFICATION	14
	3.3 RANDOM FOREST	16
	3.4 NAÏVE BAYES	18
	3.5 SUPPORT VECTOR MACHINES	20
CHAPTER 4	: ANALYSING ALGORITHMS & FUTURE SCOPE	22
CHAPTER 5	: CONCLUSION	23
CHAPTER 6	: REFERENCES	23

ABSTRACT

REASONABLENESS OF THE UNDERTAKING :

Mushrooms, as a sort of food, are extremely unique because of their edibility. A few nations treat mushrooms as a sort of high nourishment food. Be that as it may, just little segments of them are consumable. It is truly hazardous to eat a toxic mushroom. Hence, I need to utilize some grouping calculations to build up a best model to anticipate whether new arising mushrooms are eatable in light of the distinguished information of the mushrooms. Besides, it is a chance to look at the classifiers and furthermore see how they work.

POINT OF THE TASK :

The venture utilizes the information from Kaggle Machine Learning Repository. We mean to execute two characterization calculations to fabricate models for expectation. Simultaneously, the task mean to increment the correctness of them. And furthermore, I plan to contrast the 2 classifiers with known their benefits and inconveniences. Mushroom is one of the parasites types' food that has the most powerful supplements on the plant. Mushrooms have significant clinical benefits like slaughtering disease cells. This examination expects to track down the most fitting strategy for mushroom arrangement, and mushroom will be ordered into two classifications, harmful and nonpoisonous. The proposed approach will execute an alternate strategies and calculations like neural organization (NN), Support Vector Machines (SVM), Decision Tree, and k Nearest Neighbors (KNN), on dataset of mushroom pictures, where the dataset contains pictures with foundation and without foundation.

LIST OF IMAGES

3.1.1 CLASSIFICATION WORKFLOW

3.2.1 CLUSTERS WITH DIFFERENT K VALUES

3.2.1 COUNT PLOT

3.2.2 NO OF INSTANCES OF POISONOUS AND EDIBLE

3.3.1 PAIR PLOT

3.4.1 NAÏVE BAYES FORMULATED

3.5.1 VARIABLE IMPORTANCE IN RANDOM FOREST CLASSIFIER

INTODUCTION

Unlike plants, fungi do not get energy from sunlight, but from decomposing matter, and tend to grow well in moist environments. A shady environment is not a requirement, but it does help them retain their moisture. When the conditions are right (generally in autumn), the network of mycelium will produce fruiting bodies, which first look like pins, consisting of thin stalk and tiny cap. Although they start out small, the fruiting bodies quickly "mushroom." Once the cap, which looks like an umbrella, grows large enough, the veil (a thin membrane underneath the cap) ruptures, allowing the gills to drop spores. If the spores find their way to an appropriate growth substrate, they will germinate, and fungal filaments will appear. Some fungi require a certain amount of light before fruiting, while others can grow in dark caves. Mushrooms, the fruiting collection of growths, have been eaten by people for millennia.

All mushrooms contain protein, fiber, and the incredible cancer prevention agent selenium, however explicit types are pursued for explicit medical advantages. Shiitake mushrooms, for example, contain every one of the 8 fundamental amino acids, just as eritadenine, a compound that diminishes cholesterol. Reishi mushrooms are esteemed for their invulnerable boosting impacts, maitake for their settling sway on glucose, and porcini for their calming properties. At first, remembering mushrooms for the eating regimen implied scrounging, and accompanied a danger of ingesting noxious mushrooms. Be that as it may, starting during the 1600s, numerous assortments of mushrooms have been effectively developed. Agaricus bisporus is perhaps the most devoured mushrooms on the planet, and is developed in more than 70 nations. The top mushroom maker on the planet is China (5 million tons), trailed by Italy (762K tons), and the US (391 tons). Inside the United States, most of mushrooms are filled in Pennsylvania.

LITERATURE REVIEW

2.1 LITERATURE SURVEY :

There are various explores utilizing of various procedures that are utilized for mushrooms order. a Mushroom Diagnosis Assistance System (MDAS) was proposed by [3], which includes three segments of web application (worker), bound together information base and cell phone application (customer) which is utilized on cell phone gadgets. The Naive Bays and Decision Tree classifiers are utilized to decide the mushroom types. First and foremost, the recommended framework picks the most realized mushroom credits. Also, indicate the mushroom type. The investigation results show that Decision Tree classifier is superior to Naïve Bays classifier in right and wrong ordered occurrences, and blunder estimations. Kumar and others in [9] looked at changed arrangement strategies that are utilized in information digging for choice frameworks. An examination happen among three choice trees calculations addressed by one factual, one counterfeit neural organization, one help vector machines and one bunching calculation. The recommended approach utilizes four datasets from a few spaces to test the prescient precision, mistake rate, understandability, order record and preparing time. The test results showed that Genetic Algorithm (GA) and backing vector machines calculations are better contrasted and the others in the prescient precision metric. In choice tree-based calculations, QUEST calculation creates trees with more modest expansiveness and profundity.

Taking everything into account, the GA based calculation is the best calculation that can be utilized for their choice emotionally supportive networks. Babu and others in [10] proposed another application area that is utilized for SVM. The proposed approach utilizes the Support Vector Machine and Naïve Bayes calculations for grouping of mushrooms. The investigations results showed that SVM is better contrasted with Naïve Bayer's calculation in term of precision. Taking everything into account, the SVM is a productive strategy that can be utilized for application area. [2] utilized Multi-Layer Perception for Dataset preparing to make a model which is utilized to expectation of arranging. In the examination, just 8124 of dataset are utilized for preparing. The trial result showed that the best-covered up unit is 2, the best learning rates 0.6, the best actuation work is sigmoid, the best second rate is 0.2 and the best consequence of age is 300. Onudu in [11] recommended altered K-implies method dependent on the conventional k-mean calculation to improve the bunching absolute dataset and tackling the intrinsic issue in the customary grouping calculation. The proposed technique is relying upon Euclidean distance measure. In the recommended calculation, the informational index changed over into numeric qualities. At that point, the calculation read the info information with standardizes the numeric ascribes to stay away from the wide scope of qualities. The trial result showed that the recommended changed K-implies procedures quicker contrasted with the current calculation.

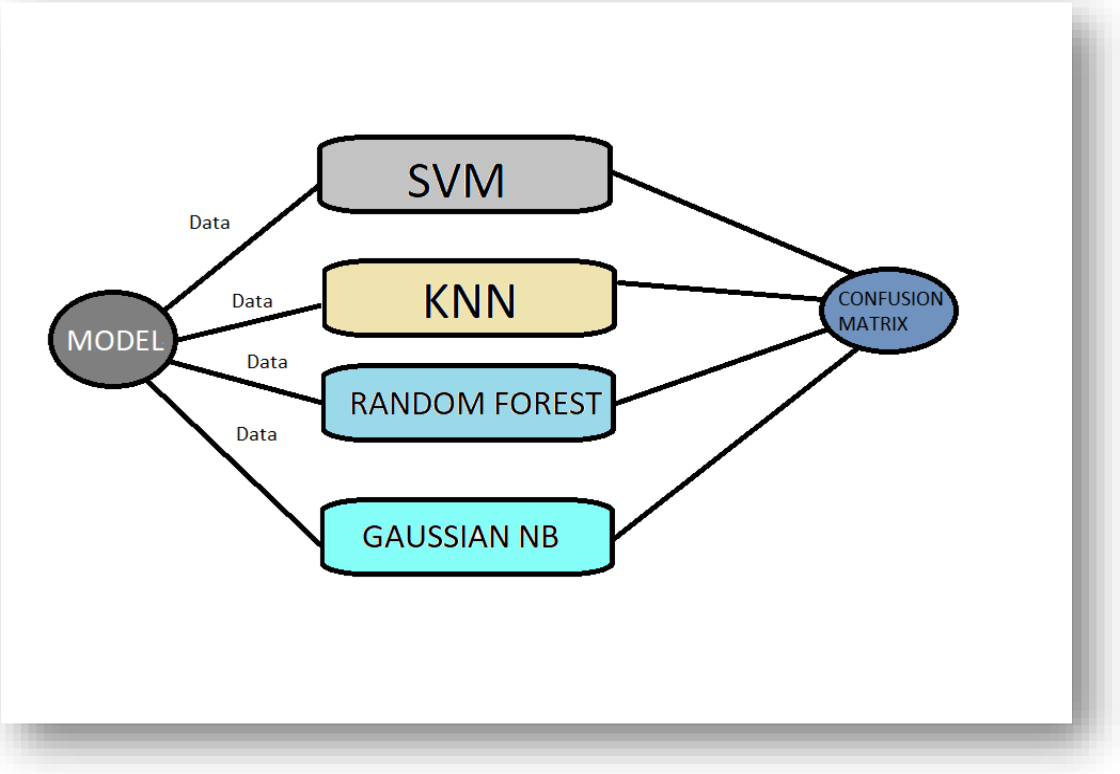
Al-mejibli and Hamad in [1] built up an application can be applied on a cell phone and web application named Mushroom Diagnosis Assistance System, the motivation behind this application is to acknowledge security when social event mushroom. They utilized choice tree and gullible bayous classifiers to bunch the mushrooms types. They relied upon the most popular mushroom credits to decide the mushroom type. This model needs to fundamental stages: preparing stage and choice stage, to allot most dynamic highlights in determination measure and find a ultimate conclusion. The test results showed that choice tree was superior to credulous sounds dependent on blunder estimations, effectively ordered examples and inaccurately grouped examples.

The creators of [12] investigated a past mushroom informational index by utilizing diverse information mining methods and Weka mining instrument. They utilized closest neighbor classifier, covering calculation to gather right standards, unpruned choice tree and a casted a ballot perceptron calculation. They came to from showing the procedures on various gatherings to investors that unpruned tree gives the best exactness result and afterward it utilized on human-machine application dependent on web to create intelligent mushroom ID. Chowdhury and S. Ojha in [13] recognized a way to recognized a few mushroom infections utilizing distinctive information mining characterization strategies. They utilized real dataset assembled from mushroom ranch by utilizing information mining like Naïve Bayes, RIDOR and SMO calculations.

They performed correlation dependent on a measurable method to identify famous manifestations for mushroom to find mushroom illness. They arrived at that gullible Bayes gives best outcome with correlations with other arrangement methods. Beniwal and Das in [14] utilized information mining grouping strategies like Zero, guileless Bayes and Bayes net to dissect mushroom dataset that contain different sorts of mushrooms, which are noxious or not harmful. They assessed characterization procedures by utilizing exactness, kappa measurement and mean absolute error. we reached that KNN classification is giving a better performance than other classification algorithms.

ALGORITHMS

3.1 CLASSIFICATION WORKFLOW :



3.1.1 CLASSIFICATION WORKFLOW

We are going to use four types of algorithms to predict the output whether the mushroom classified is poisonous or edible. First the data is of csv formatted data then we send the data to different classifiers such as KNN, Random forest, Support vector machines and Naïve bayes algorithms. These give certain accuracy, precision, recall, f-score, for the result verification and evaluate our model. Finally we use a confusion matrix to know the false negatives, false positives, true negatives and true positives i.e. correctly predicted vs wrongly predicted.

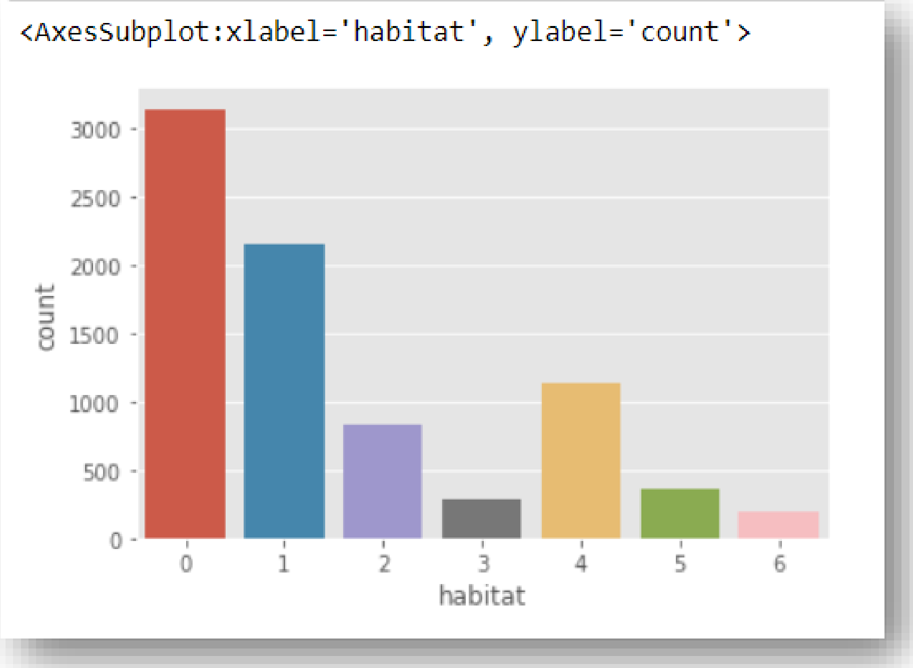
3.2 K – NEAREST NEIGBOURS :

K-Nearest Neighbors(KNN) calculation utilizes include closeness to foresee the estimations of new information focuses which further methods the new information point will be doled out to the worth dependent on how intently it coordinates with the focuses in the preparation set It utilizes euclidean distance formula,that is the distance between two focuses in the plane having organizes (x1,y1) and (x2,y2)

KNN calculation is easy to carry out and is strong to the uproarious preparing information and is more viable if the preparation information is large.There is no specific method to decide the best incentive for K ,so we need to attempt a few qualities for K ,so there is a need to track down the best out of them and euclidean distance of given information and its K closest focuses determined and it chooses the class to which the information falls in and huge estimations of K outcomes in great outcomes.

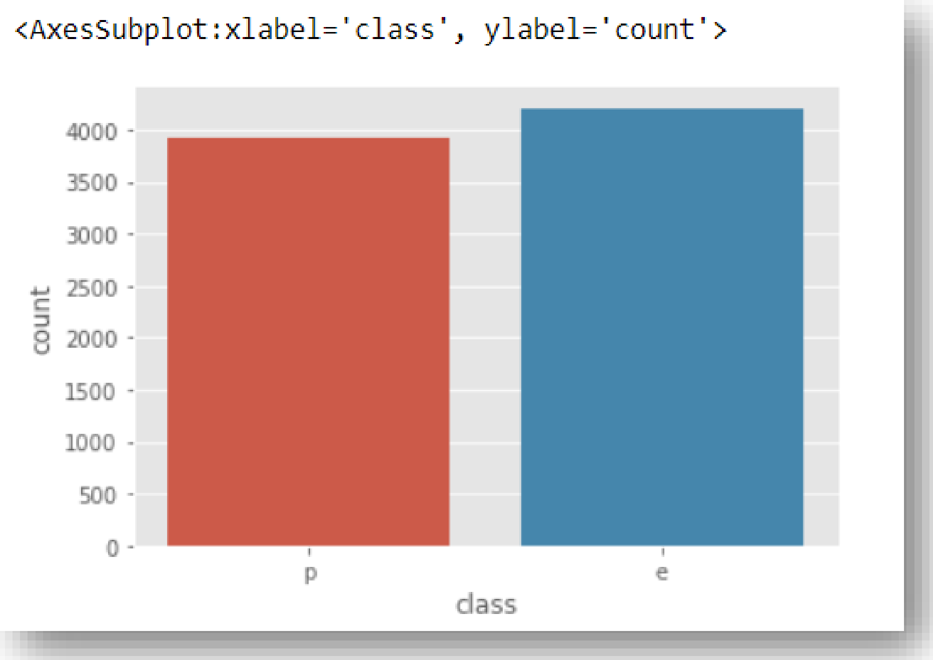
This informational index is utilized for discovering whether the mushroom is consumable or noxious from the mushroom's cap-shape,cap-surface,cap-color,bruises,odor,gill-attachment,gill-spacing,gill-size,gill-color,stalk-shape,stalk-root,stalk-surface-above-ring,stalk-surface-beneath ring,stalk-shading above-ring,stalk-shading underneath ring,veil type,veil color,ring number,ring type,spore print color,population,habitat

Pandas is an open source library that is based on top of Numpy library.It is a python bundle that offers different information design and activity controlling mathematical information and time arrangement .It is chiefly well known for bringing in and dissecting information a lot simpler



```
<AxesSubplot:xlabel='habitat', ylabel='count'>
```

3.2.1 COUNT PLOT OF HABITAT



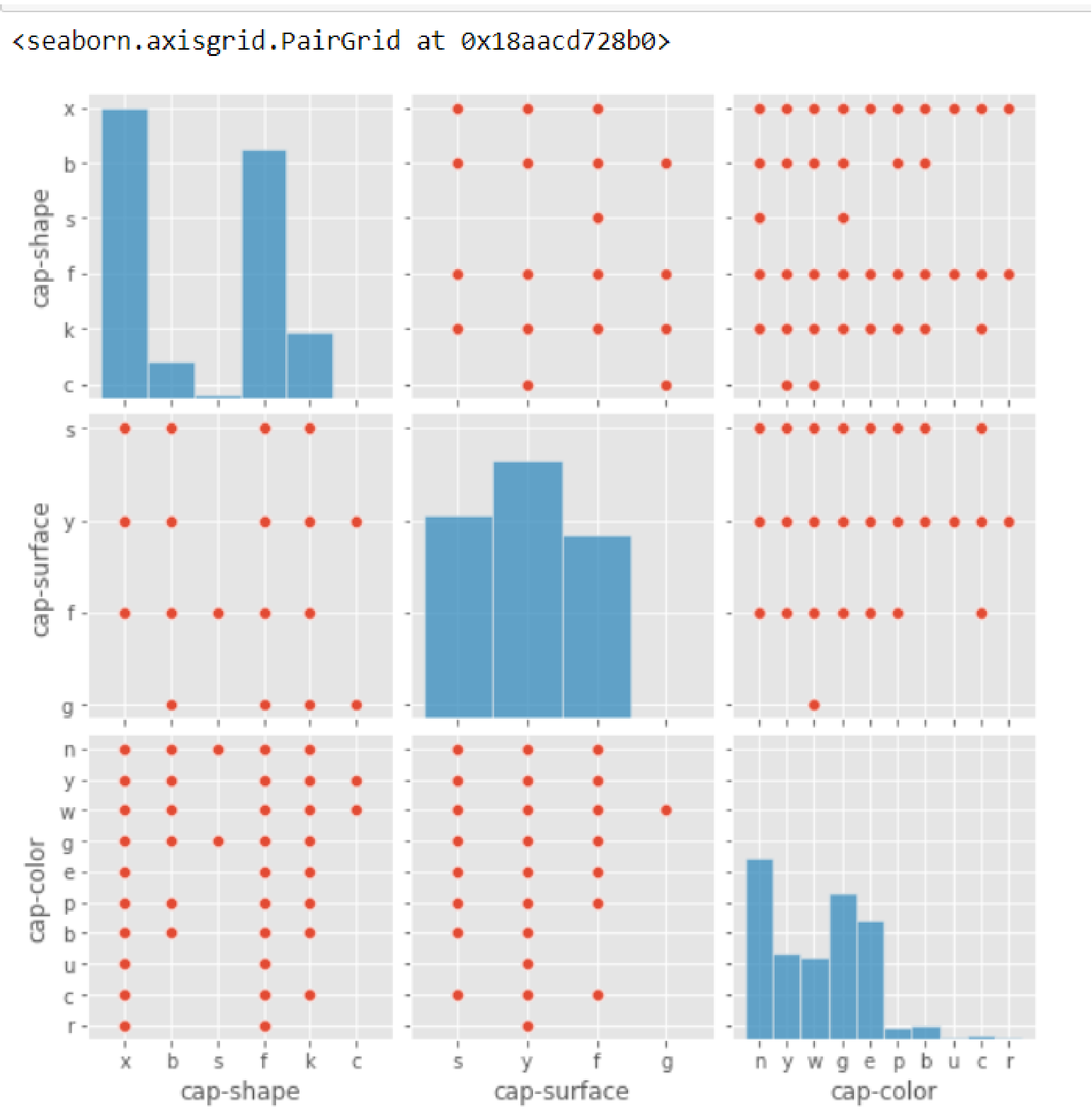
```
<AxesSubplot:xlabel='class', ylabel='count'>
```

3.2.2 NO OF INSTANCES OF POISONOUS AND EDIBLE

3.3 SUPPORT VECTOR MACHINES :

SVM [2, 5] is a classification, technique that is originated as an implementation of Vapnik's (1995) structural risk minimization principle. SVM are based on mapping input space to a high-dimensional feature space where linear separation is easier than input space. SVM have been used successfully for the solution of many problems. Consider a training set T= {Xi ,Yi } N i=1, where Xi is a real-valued n dimensional input vector (i.e. Xi Є R n) and y Є {+1,-1} is a label that determines the class of Xi . The SVM employed for two class problems are based on hyper planes to separate the data. The hyper plane is determined by an orthogonal vector w and a bias b, which defines the points that satisfy . By finding a hyper plane that maximizes the margin of separation q, it is intuitively expected that the classifier will have better generalization ability. The hyper plane with the largest margin on the training set can be completely determined by the nearest points to the hyper plane. Two such points are and and they are called support vectors (SV). Therefore, in its simplest form, SVM learn linear decision rules as F(x) = sign (Wt . X +b) so that (W, b) are determined to classify correctly the training examples and to maximize q.

The margin q can be calculated as q = So, Minimize Subject to: to get the maximum marginal classifier so we introduced Lagrange Function for the SVM quadratic problem with linear constraints as follows Where, Lagrange multiplier, ≥ 0 For L to be maximized, only training examples with = 0 (support vectors) will have ≠ 0. As practical problems are not likely to be linearly separable, the linear SVM has been extended to a nonlinear version by mapping the training data to an expanded feature space using a nonlinear transformation .Then, the maximum margin classifier of the data in the new space can be determined. With this procedure, the data that are non-separable in the original space may become separable in the expanded feature space. Since the training algorithm only depend on data through dot products. We can use a "kernel function" K such that The most commonly used function for the dot product is the RBF kernel. However, depending on the type of nonlinear mapping, the training points may not happen to be linearly separable, even in the expanded feature space. In this case, it will be impossible to find a linear classifier. Therefore, a new cost function is introduced, N non-negative slack variables are introduced to allow for training errors. C is a preselected positive penalty factor.



```
<seaborn.axisgrid.PairGrid at 0x18aacd728b0>
```

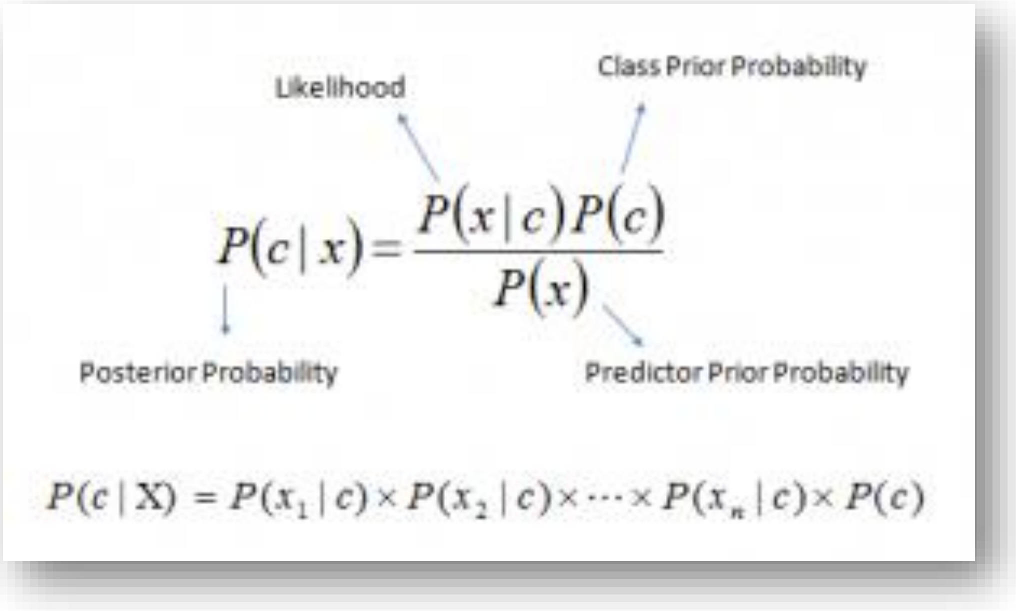
3.3.1 PAIR PLOT

3.4 : NAÏVE BAYES

Problem & Approach:

To develop a binary classifier to predict which mushroom is poisonous & which is edible. I will build a Naive Bayes classifier for prediction after basic EDA of data. Later I will also test Decision Tree & Random Forest models on this dataset.

As you might have noticed all data entities are named by initials only. Lets convert these to proper names for clarity & also convert all attributes to factors as all attributes are categorical here.



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

3.4.1 NAÏVE BAYES FORMULATED

Creating Train Test Splits

I will take 70% (5386 mushrooms) sample data for training & 30% (2438 mushrooms) for testing.

Creating Model using Naive Bayes Classifier

Naive Bayes classifier is based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Predicting Mushroom Class on Testset

Lets test our model on remaining 30% test data

3.5 : RANDOM FOREST CLASSIFIER

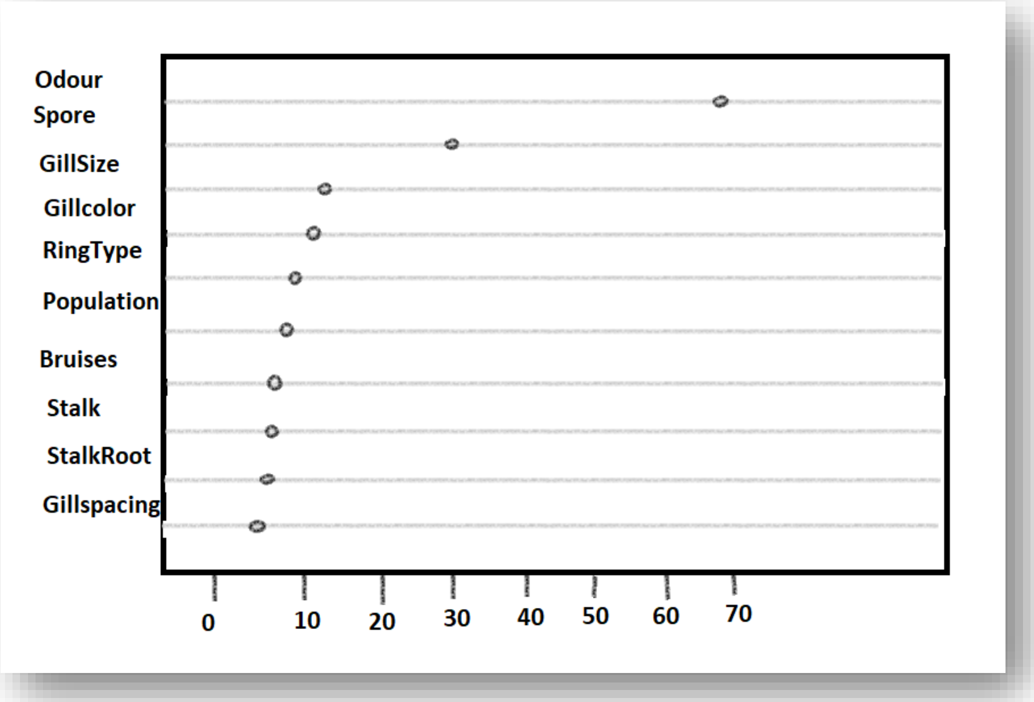
There is a plenty of arrangement calculations accessible to individuals who have a touch of coding experience and a bunch of information. A typical AI strategy is the irregular backwoods, which is a decent spot to start.This is a utilization case in R of the randomForest bundle utilized on an informational collection from UCI's Machine Learning Data Repository.

Are These Mushrooms Edible?

On the off chance that somebody gave you a large number of lines of information with many segments about mushrooms, could you distinguish which attributes make a mushroom palatable or noxious? What amount would you confide in your model? Would it be sufficient for you to settle on a choice on whether to eat a mushroom you find? (That is an awful choice generally 100% of the time).The randomForest bundle does the entirety of the substantial liftingbehind the scenes. While this "enchantment" is unbelievably pleasant for the end client, it's essential to comprehend what it is you're doing. Remember this for totally any bundle you use in R or some other language.

We need to investigate the information prior to fitting a model to find out about what's in store. I'm plotting a variable on two tomahawks and utilizing tones to consider the to be with respect to whether the mushroom is consumable or poisonous.In these plots, eatable is appeared as green and toxic is appeared as red. I'm searching for where there exists a dominant part of one color.A examination of "CapSurface" to "CapShape" shows us:

CapShape Bell is bound to be palatable
CapShape Convex or Flat have a blend of palatable and toxic and make up most of the information
CapSurface alone doesn't reveal to us a ton of data
CapSurface Fibrous + CapShape Bell, Knobbed, or Sunken are probably going to be consumable
These factors will probably build data acquire however may not be unfathomably solid



3.5.1 VARIABLE IMPORTANCE IN RANDOM FOREST CLASSIFIER

3.6 : FEED FORWARD NEURAL NETWORKS

Artificial neural networks (ANNs) are equal computational models involved thickly interconnected, versatile handling units, portrayed by an inalienable inclination for gaining as a matter of fact and furthermore finding new information Because of their incredible ability of self learning and self-adjusting, they have been widely considered and have been effectively used to handle troublesome true issues and are frequently discovered to be more productive and more exact than other arrangement strategies. Grouping with a neural organization happens in two unmistakable stages. In the first place, the organization is prepared on a bunch of combined information to decide the information yield planning. The loads of the associations between neurons are then fixed and the organization is utilized to decide the groupings of another arrangement of information.

ANALYZING ALGORITHMS

Algorithm	Precision	Recall	F1score	Accuracy
SVM	**91.5**	**90.5**	**91**	**90.85 %**
KNN	**93.5**	**93.0**	**93**	**93.07 %**
Random forest	**93.5**	**92.5**	**92.5**	**92.9 %**
Gaussian NB	**90.5**	**89.5**	**90**	**89.66 %**
Feed Forward Neural Networks	**93.5**	**93.0**	**92.5**	**93.0%**

After Analyzing all the algorithms, we came to conclusion that K – Nearest neighbors algorithm is working well for the given data since the KNN algorithm is a method to form clusters based on the distance metric we can accurately divide the clusters and classify them into edible or poisonous mushrooms. We see that KNN is performing well with 93.07% accuracy which is good. Not only KNN with distance we can also use K medoid, rock algorithms to still have better classification and reduce the number of outliers which destroys performance using KNN.

FUTURE SCOPE

After using necessary algorithms for classifying the mushrooms as poisonous or edible we wanted to use CNN algorithm to classify as poisonous or edible based on the features of the mushroom like odor, spore print color, gill size, gill color etc. This classification based on neurons and weights gives us more accuracy and we can directly send images as input into the model which will be the next level of our implementation. In future work we will attempt to remove some actual measurement from mushroom pictures like cup breadths, stem tall, shading and surface. Additionally, we will attempt to grow the dataset and utilize more pictures to improve order measure.

CONCLUSION

In the proposed approach, we utilized various calculations to get best aftereffects of mushroom grouping, we carry out every one of SVM, Random Forest, KNN and Naïve Bayes on various situations, with foundation and without foundation. We separate various highlights from mushroom pictures like Eigen highlights, histogram highlights and parametric highlights. To improve the outcomes, we eliminate pictures foundation yet shockingly this progression neglected to improve the outcome. At long last, the trial results show advantage for foundation pictures, particularly when utilized KNN calculation, and with Eigen highlights extraction and genuine components of mushroom (i.e cup breadth, stem tall and stem measurement) where exactness came to 95.86% , while the outcome in the wake of supplanting genuine measurements with virtual measurement (for example width and tallness of mushroom shape inside the pictures) is 93.07 %.

REFERENCES

- http://homepages.cae.wisc.edu/~ece539/fall13/project/Shen_rpt.pdf
- http://laurenfoltz.com/content/1-projects/1-data-mining-project/data-mining-final-report.pdf
- https://medium.com/@harinibuzu/mushroom-classification-using-knn-algorithm-dfd29507feb9#:~:text=Here%2Cthe%20first%20column%20to,the%20mushroom%20act%20as%20input.
- https://www.ijcsmc.com/docs/papers/April2014/V3I4201499b50.pdf
- https://www.stoltzmanconsulting.com/blog/random-forest-classification-of-mushrooms#:~:text=A%20common%20machine%20learning%20method,UCI's%20Machine%20Learning%20Data%20Repository.
- https://www.researchgate.net/publication/337024220_Classification_of_Mushroom_Fungi_Using_Machine_Learning_Techniques