

**Öğrencinin;**

**ADI**: Üveys

**SOYADI**: SARAÇ

**NO**: 1621221025

**BÖLÜM**: Bilgisayar Mühendisliği

**Projenin;**

**KONUSU**: Belirlenen veri seti ile 2 makine öğrenmesi algoritması kullanarak sınıflandırma

**Dersin**;

**ADI**: Introdiction to Data Science

**EĞİTMEN:** Dr. Öğr. Üyesi Gönül ULUDAĞ, Zeki KUŞ

Dataset :

Attribute Information:

1. cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2. cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
3. cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. bruises: bruises=t, no=f
5. odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6. gill-attachment: attached=a, descending=d, free=f, notched=n
7. gill-spacing: close=c, crowded=w, distant=d
8. gill-size: broad=b, narrow=n
9. gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
10. stalk-shape: enlarging=e, tapering=t
11. stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12. stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
13. stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
14. stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15. stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16. veil-type: partial=p, universal=u
17. veil-color: brown=n, orange=o, white=w, yellow=y
18. ring-number: none=n, one=o, two=t
19. ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20. spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21. population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22. habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

Project Aim :

Comparing two very different machine learning models on the [Mushroom Classification Dataset](https://www.kaggle.com/uciml/mushroom-classification) for the task of predicting whether a given mushroom is **poisonous** or **edible**.

The first model is [Logistic Regression](https://en.wikipedia.org/wiki/Logistic_regression) with some parameters tuning, and the second model is KNN with neighbor number tuning.

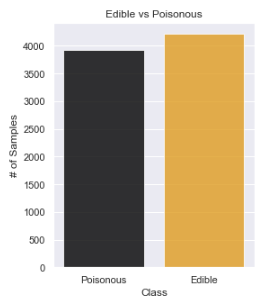
As usual, for the Supervised Learning Algorithm (Logistic Regression), we simply train the model on **80%** of the mushroom data, and then test it's performance on the remaining **20%.**

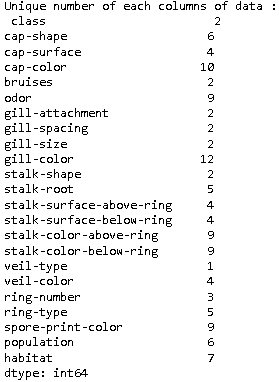
The KNN algorithm is applied in the same way.

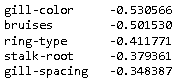
Finally, we will see which algorithm is better for this dataset.

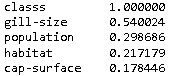
What was done :

Eda steps:

* It was seen that there are 8124 instances and 23 features.
* It was seen that all features are objects.
* No features of any instance have null values.
* In the graph drawn, the numbers of poisonous and edible are very close to each other. So we have a balanced data set.



* There was a remarkable value when looking at the unique values. This is the unique value of 1 in a feature. If there is such a thing, it means that the feature is the same for all instances. And that means that feature doesn't matter to us. Also, if a unique value is 1, that tells us that its variance is 1.
* This is the feature "veil-type" with a unique number of 1 and we drop it from our dataset.
* As mentioned before, all of our features are objects, ie strings, and we need to enumerate them all.
* There are 2 methods to encode them, label encoder and one hot encoder. We used label encoder here. If we used one hot, there would be a lot of features because all the data we have is categorical. That's why we preferred label encoder.
* From the correlation chart, we saw the featues that had the most impact on the class for us.

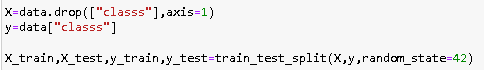


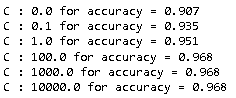
* As seen above, correlation rates are given according to classes. In this case, the highest correlation belongs to gill-size. The lowest correlation belongs to bruises and gill-color properties. In other words, these 3 features are the ones that affect the class value the most and are the most important for us.

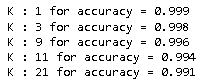
Feature Selection :

* A feature selection was made according to the Variance treshold. Threshold value is given 0.05. This means that if the feature is 95% the same for all instances, it is dropped.
* In addition, the selecfrommodel method has been applied and the number of our features has been reduced to 18.

Application of machine learning algorithms :



* Class property and other properties are reserved for the implementation of algorithms. Then it was separated as test and train. One of the important points here is that we have given the random state manually. The reason for this is to use the same test and train sets, using the same random state number, if we separate them as test trains.
* Logistic regression was applied with different C parameters. And the results were best seen as 0.968 at C=100.



* KNN was applied with different K parameters. And the results were best seen as 0.999 at C=1.

These are the results on our data without feature selection applied. And it seems that the KNN algorithm does a good classification with an acuuracy value of 0.99.

The results in the logistic regression with the dataset in which we removed 4 features with Feature Selection were better than the results of the data without Feature Selection.



We achieved accuracy = 1 with feature selection and parameter tuning in KNN.



In the last classification with KNN, the accuracy score was 1.0, which means that we can classify any mushroom given to us with 100% accuracy whether it is edible or poisonous by looking at its characteristics.