Question\_4

a) Which features seem useful for classification? Which ones are correlated with the labels?

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturulduExamining the top 100 features with the highest variability:

yin\_0\_T42 2974.523703

yin\_0\_T43 2930.339281

yin\_0\_T1 2929.136716

yin\_0\_T41 2912.135665

yin\_0\_T2 2896.436401

...

bandwidth\_0\_T32 862.156972

bandwidth\_0\_T30 861.755976

bandwidth\_0\_T33 857.786545

bandwidth\_0\_T34 848.928959

centroid\_0\_T19 845.872043

Length: 100, dtype: float64

**The features with the highest variability are mostly yin\_0\_TXX features at various timestamps. High variability in a feature across our dataset implies that the feature values differ significantly from one sample to another. This can be an indication that the feature captures meaningful differences in the audio signals that could be useful for distinguishing between different words or sounds.Features with high variability might be particularly useful for classification tasks because they suggest the presence of distinct patterns or characteristics in the data that can differentiate between classes (in our case, possibly different words or sounds).**[**¶**](http://localhost:8888/notebooks/DataExploration_SML.ipynb#Features-with-high-variability-might-be-particularly-useful-for-classification-tasks-because-they-suggest-the-presence-of-distinct-patterns-or-characteristics-in-the-data-that-can-differentiate-between-classes-(in-your-case,-possibly-different-words-or-sounds).) **The fact that yin\_0\_TXX features dominate the top variability list suggests that pitch-related characteristics of the audio recordings vary widely across our dataset.**

melspect\_7\_T18 0.316700

melspect\_8\_T18 0.312075

melspect\_32\_T16 0.310866

melspect\_33\_T17 0.308488

melspect\_32\_T17 0.307993

melspect\_7\_T17 0.307864

melspect\_33\_T16 0.307111

melspect\_31\_T16 0.306584

melspect\_8\_T17 0.305117

melspect\_31\_T17 0.303191

melspect\_6\_T18 0.303009

melspect\_32\_T15 0.300192

melspect\_31\_T15 0.299863

melspect\_34\_T17 0.298479

These are the first 14 features that are highly correlated with the labels. Here is the code:

label\_encoder = LabelEncoder()

word\_labels\_encoded = label\_encoder.fit\_transform(copyy.index)

features\_array = copyy.to\_numpy()

correlations = pd.Series(np.corrcoef(features\_array.T, word\_labels\_encoded)[-1][:-1], index=copyy.columns)

sorted\_correlations = correlations.abs().sort\_values(ascending=False)

print(sorted\_correlations.head(14))

b) Do similar words (e.g., Haus - aus, Ofen - offen, or nicht - Licht) have a similar feature distribution?

diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturulduThe median value for each word is indicated by the line within each box. As an example, the feature distributions of “Haus - aus, Ofen - offen, or nicht - Licht” have been calculated for the timestamp two for the 4 features. The boxplot illustrates the fact that there is a high similarity between Haus and aus in the feature melspect\_17\_T2, however we couldn’t see that high similarity between these two in the feature contrast\_3\_T2. Therefore, it shouldn’t be generalised that there will always be a similarity if both words contain similar letters. However as a result, there does not seem to be a drastic difference in the distributions of the features.

diyagram, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturulduFor a futher understanding, same samples have been calculated but with the different feature and timestamps. It seems the median values for "Haus" and "aus" are somewhat similar across the timestamps, which suggest a degree of similarity in their feature distributions. However between the words that have no much letters in common, tend to differentiate in terms of their feature distributions, i.e aus and Ofen.