Student Names: TODO
Collaboration Statement:

Turning in this assignment indicates you have abided by the course Collaboration Policy:

`www.cs.tufts.edu/comp/136/2022s/index.html#collaboration-policy`

Total hours spent: TODO

We consulted the following resources:

- TODO
- TODO
- ...

These are the official instructions for checkpoint 1. You can find instructions on how to submit at `www.cs.tufts.edu/comp/136/2022s/checkpoint1.html`

Please consult the full project description at `https://www.cs.tufts.edu/comp/136/2022s/project.html` in addition to this document when working on this checkpoint. It gives details on what we expect when choosing a model and coming up with performance hypotheses.

Exploratory Data Analysis

Describing your analysis

We followed the following data analysis steps to obtain a comprehensive understanding of our data:

1. Visualize the marginal distributions of each feature
2. Visualize the marginal distribution of the output class
3. Analyzed the correlation between features
4. Conducted principal component analysis (PCA)
5. Visualized the joint distributions of pairs of features:
   (a) highly positively correlated features
   (b) highly negatively correlated features
   (c) fairly uncorrelated features
6. Visualized the distribution of correlations between each feature and the output class

General information about the data:

This dataset contains measures of distances within different shapes (conformations) of a set of 102 molecules. The study that this data comes from used human experts to judge the smell of each molecule and determine whether it is characterized as ”musk” or ”non-musk”, which makes this a binary classification dataset.

- There are 6,598 conformations total (number of data points)
- There are 166 features (not including the molecule and conformation names)

Step 1

After importing the data into a pandas dataframe and standardizing the data using a MinMaxScaler from sklearn, the hist method was used to visualize the histograms for each of the features. We also visualized the marginal distribution of the output class which is binary.

Analyzing the results

Model and Learning Method Properties

Model

Learning method (inference)

Performance Hypotheses