CLASSIFICATION BETWEEN Musk AND Non-Musk COMPOUNDS USING CONVOLUTIONAL NEURAL NETWORKS

By:

Pranav viswanathan

[Vpranav009@gmail.com](mailto:Vpranav009@gmail.com)

INDEX

Problem statement 3

Variable identification-3

Hypothesis-4

EDA and Preprocessing-4

Model creation-6

Performance Metrics-7

Conclusion-9

Problem statement:

The given dataset contains details about organic chemical compounds including their chemical features, isomeric conformation, names and the classes in which they are classified. The compounds are classified as either ‘Musk’ or ‘Non-Musk’ compounds.

A model is created by using CNN (convolutional neural network) to classify the dataset into ‘Musk’ or ‘Non-Musk’ compounds.

Here I have used IBM Watson studio for the purpose of creating the model.

Variable identification:

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Variable name | Description | Type |
| 1 | molecule\_name | It tells the molecule name of compound | Numeric |
| 2 | conformation\_name | Name | Numeric |
| 3 | F1-F166 | Features of the compounds | Numeric |
| 4 | class | Class :whether musk or not | Categorical |

Hypothesis:

Ho: There are no features to classify between compounds

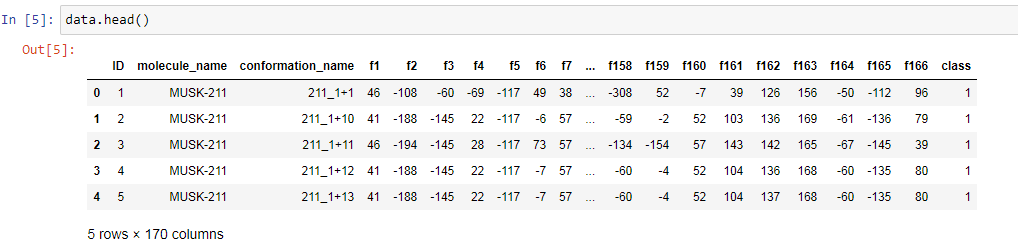
Ha: There is at least one feature to classify between

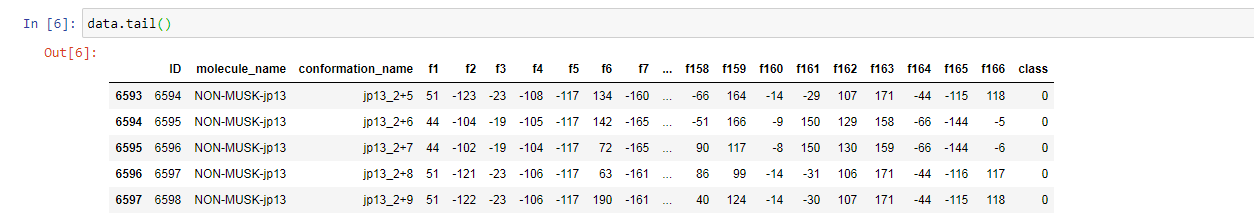
EDA and Preprocessing:

**The objectives of EDA are to:**

1. Suggest hypotheses about the [causes](https://en.wikipedia.org/wiki/Causality) of observed [phenomena](https://en.wikipedia.org/wiki/Phenomenon).
2. Assess assumptions on which [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference) will be based.
3. Support the selection of appropriate statistical tools and techniques.
4. Provide a basis for further data collection through [surveys](https://en.wikipedia.org/wiki/Survey_sampling) or [experiments](https://en.wikipedia.org/wiki/Design_of_experiments).

EXPLORING THE DATASET:

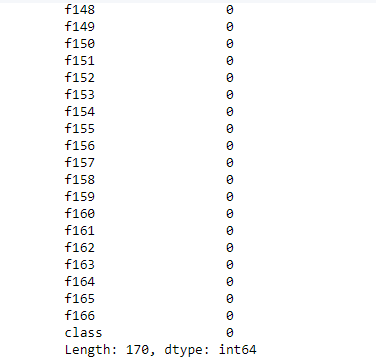
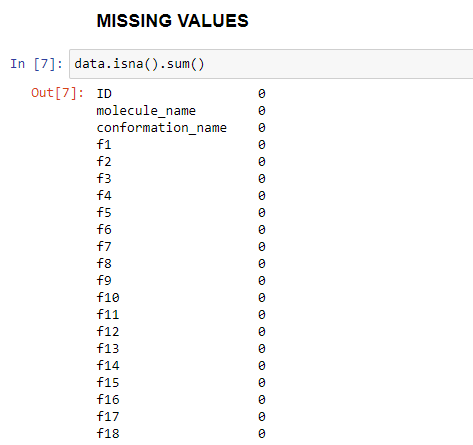




The above pics depict the top and bottom 5 rows of the dataset.

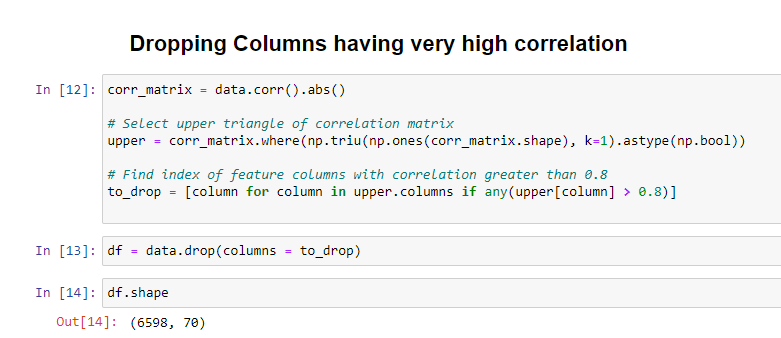
We see that there are 170 columns .

CHECKINF FOR MISSING VALUES IN THE DATASET:



From, the above pics we see that the dataset has no missing values.

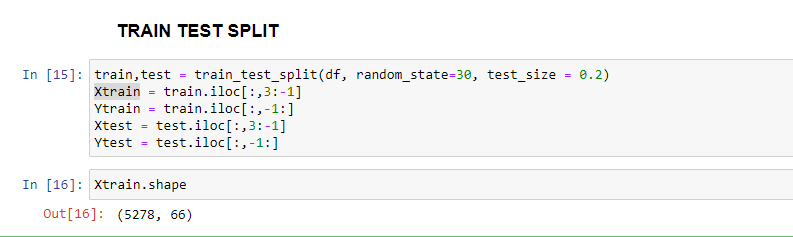
DATASET REDCUTION:



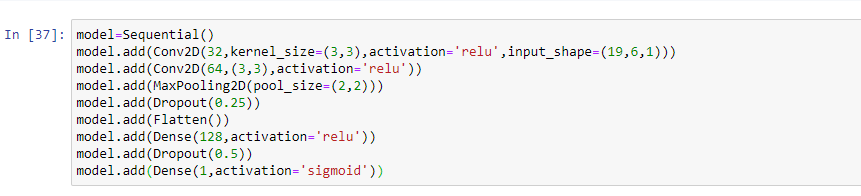
The dataset has reduced as most of them with correlation greater than 0.92 has been deleted .

Model creation:

1)TRAIN AND TEST SPLIT



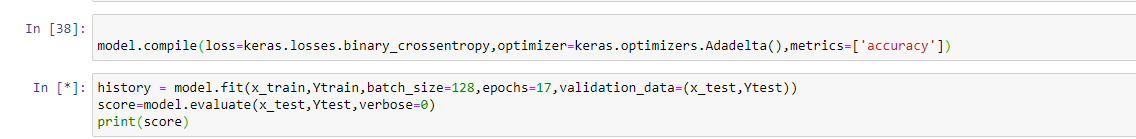
2)CNN MODEL CREATION:

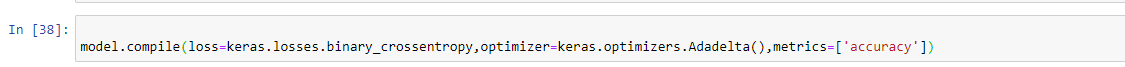


Here a multiple layers are created for better classification.

First layer has a filter of dimension 32 and activation as relu.

Similarly we define for different layers.

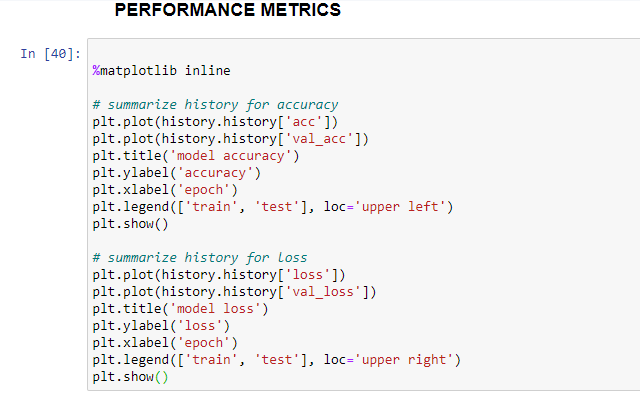


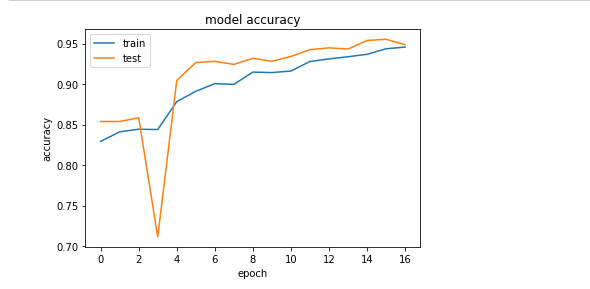


Finally the model is validated.

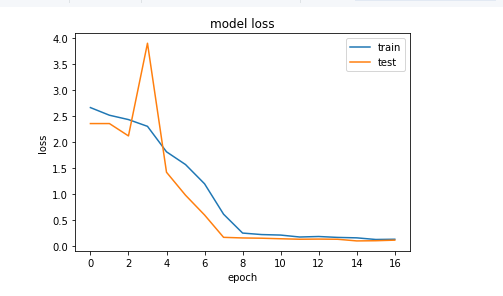
Performance Metrics:

We plot the accuracy and model loss plot and carry out different parametrs for validation of the model created.



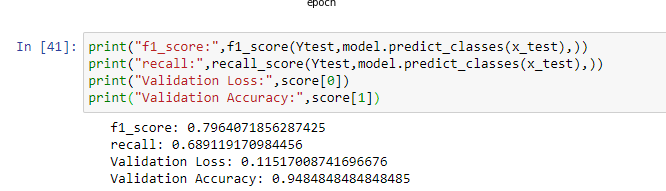


The above plot is model accuracy of test and train.



The above plot is model loss of test and train.

VARIOUS OTHER PARAMETRS



|  |  |
| --- | --- |
| PARAMETER | VALUE |
| F1 score | 0.79 |
| Recall | 0.68 |
| Validation loss | 0.11 |
| Validation Accuracy | 0.94 |

Conclusion:

A CNN model is created of classifying between musk and non-musk compounds and validated using various performance metrics.