Project Report

Preprocessing Steps-

1) First the data was checked for any missing values. It was found that there were no missing values.
2) Then the data was checked for any categorical variables.
3) Then the data was split into training and testing data in the ratio 80:20. All the musk samples and the non_musk samples were split into 80:20 ratio in both the training and testing data.

Standardization Of the data –

Data preparation involves using techniques such as the normalization and standardization to rescale input and output variables prior to training a neural network model. As such, the scale and distribution of the data drawn from the domain may be different for each variable.

Input variables may have different units (e.g. feet, kilometers, and hours) that, in turn, may mean the variables have different scales.

Differences in the scales across input variables may increase the difficulty of the problem being modeled. An example of this is that large input values (e.g. a spread of hundreds or thousands of units) can result in a model that learns large weight values. A model with large weight values is often unstable, meaning that it may suffer from poor performance during learning and sensitivity to input values resulting in higher generalization error.

To rescale the data standard scaler was used. It is a standardization approach which transforms the data such that it has a mean 0 and standard deviation_equal to 1. Standardization is useful for data with negative values and when classification of the data is required. Some attributes of our data had negative and varying values which needed to be standardized.

Dimensionality Reduction –

We will use Principal Component Analysis for the dimensionality reduction of the data as the data had 166 attributes.This technique is widely used to reduce the number of dimensions in a data set, in order to use only the components that most contribute for tasks such as classification or regression.

All the attributes may not be independent of each other so dimensionality reduction was required to have only those components that matter the most.

It was seen from the variance vs no_components curve that around 99% of the total variance was preserved at 75 components. So we will reduce the data to a set of 75 components.

We could have also used Linear Discriminant Analysis, but PCA works better for binary data classification because the number of samples per class is less, and when the dataset is not very large.

Postprocessing Steps –

1) First an artificial neural network made up of 2 hidden layer was applied to the data. It produced an accuracy of 75% at max.
2) Different hyperparameters were used but the maximum accuracy we could get was 78%
3) Then we went for a 1 dimensional convolutional neural network.
4) We first set the number of filters to 100 and used only one layer of cnn and 1 dense layer.
5) Overfitting was occurring
6) So the number of dense layers was increased, and then the validation accuracy increased a little but still wasn't very high and was converging to 85%
7) Then the kernel size was increased from 10 to 50 and then a decrease in overfitting in the graphs was seen
8) Kernel size at 75 gave the minimum overfitting and the maximum accuracy was achieved.

The data had patterns in it that only a 1d convolutional neural network was able to identify. It was not possible using an artificial neural network.