Problem Statement:

To Build a deep learning model that can classify a Musk and non Musk compound from a given set of features.

Motivation:

Musk compounds traditionally belong to the most **important** substances used in the fragrance industry. It is thus imperative to develop a robust model that can correctly classify MUSK and Non MUSK compounds from a given set of features.

Dataset explanation:

The dataset contained 6598 rows 170 columns. The columns contained
ID, molecule name, confirmation name and 166 features and 1 class label. A preview of the dataset containing 15 rows 12 column is given below.

ID	molecule_name	onformation_nam	f1	f2	f3	f4	f5	f6	f7	f8	f9
6598	NON-MUSK-jp13	jp13_2+9	51	-122	-23	-106	-117	190	-161	80	-227
6597	NON-MUSK-jp13	jp13_2+8	51	-121	-23	-106	-117	63	-161	79	-224
6596	NON-MUSK-jp13	jp13_2+7	44	-102	-19	-104	-117	72	-165	65	-219
6595	NON-MUSK-jp13	jp13_2+6	44	-104	-19	-105	-117	142	-165	68	-225
6594	NON-MUSK-jp13	jp13_2+5	51	-123	-23	-108	-117	134	-160	82	-230
6593	NON-MUSK-jp13	jp13_2+4	48	-162	-101	29	-117	-86	215	19	-93
6592	NON-MUSK-jp13	jp13_2+31	37	-25	-103	34	-117	-100	220	36	-18
6591	NON-MUSK-jp13	jp13_2+30	48	-189	-37	28	-117	-85	209	10	-110
6590	NON-MUSK-jp13	jp13_2+3	37	-63	-101	34	-117	-99	214	11	-88
6589	NON-MUSK-jp13	jp13_2+29	37	-134	-28	33	-117	-97	208	3	-108
6588	NON-MUSK-jp13	jp13_2+28	44	-106	-20	-105	-117	154	-164	35	-238
6587	NON-MUSK-jp13	jp13_2+27	48	-153	-42	28	-117	-84	96	-15	-87
6586	NON-MUSK-jp13	jp13_2+26	51	-126	-23	-107	-117	148	-159	57	-245
6585	NON-MUSK-jp13	jp13_2+25	37	-53	-46	33	-117	-96	96	-22	-82
6584	NON-MUSK-jp13	jp13_2+24	44	-104	-19	56	-117	70	-167	47	-230

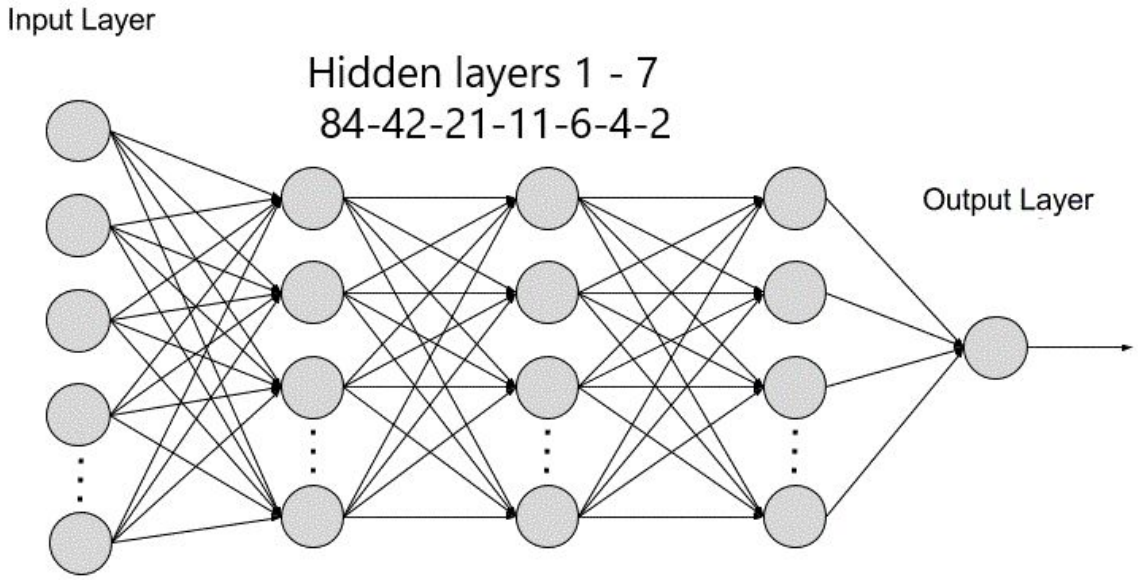
Dataset preprocessing: We begin by dropping columns corresponding to ID, molecule name, confirmation name. We are left with 166 features and the binary class label. Then we feature scale the values in feature columns so as to standardize the data values. This removes any unfair bias that might have crept in due to different scales in which features were measured in original data.
It is to be noted that the MUSK compounds are labelled as 1 and non-MUSK compounds are labelled as 0. This can be easily deduced on checking the label and molecule name columns.

Now we are ready with our pre-processed data and we will start building with the model.

Model description:

We have trained a Multi layer perceptron network consisting of 7 hidden layers. The input dimension is 166 which is the number of features for one particular row. The model consists of 7 hidden layers having 84,42,21,11,6,4,2 neurons respectively. This was done in accordance with the rule of thumb where succeeding layers have half the neurons of preceding layers as recommended by Kirill Eremenko, a prominent udemy data science educator.

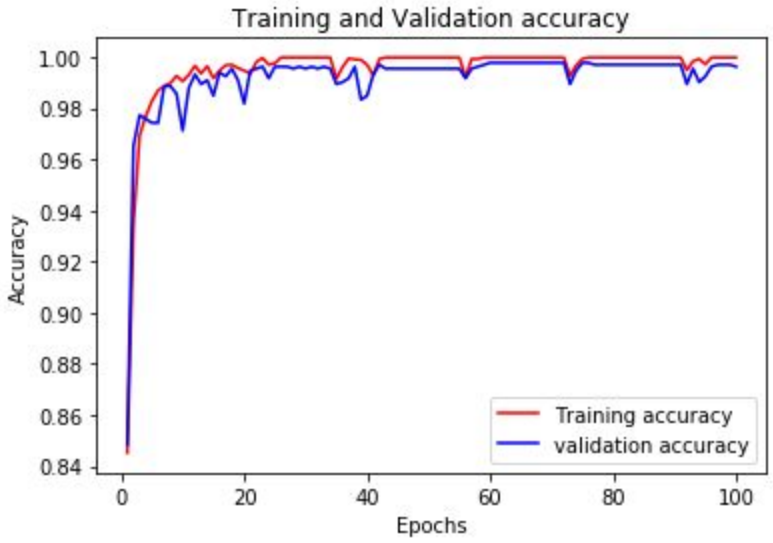
Below a diagrammatic representation of the network is shown.

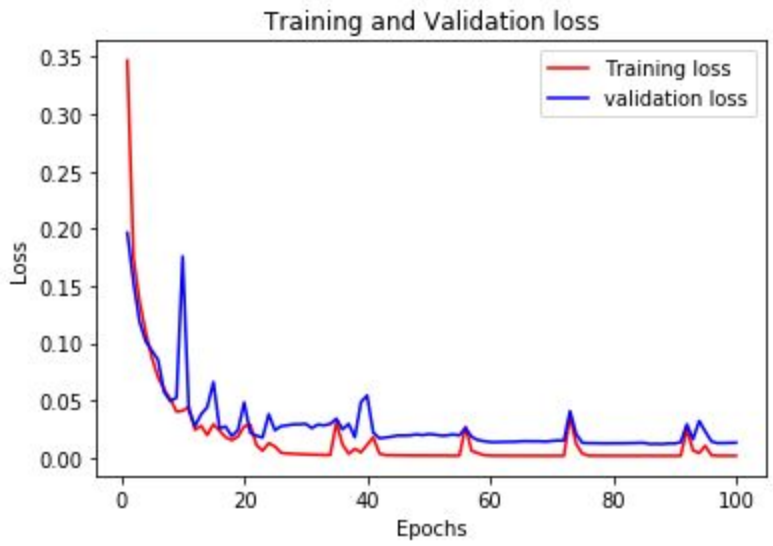


We have used relu as the activation function in the hidden layers. Since we have a binary classification task at hand so we have used binary cross entropy as the loss function and have used adam as the optimizer. Other available options for optimizers such as SGD, adagrad were also tried in combination with sigmoid as the the activation function but finally settled on relu as activation function and adam as the optimizer as the combination was giving the best accuracy on the validation data.

Results:

I trained the model over 100 epoches and a batch size of 10. The loss decreases asymptotically and starts to stabilize after 20 epoches. The train and cross validate losses and accuracy scores have been plotted against epoches.





Confusion matrix on validation data:

It can be noted that the error or misclassification in general is quite low and is almost equally distributed among the false positive and true negative classes i.e The type I and type II errors are almost equi probable.

	0	1
0	1117	3
1	2	198

Performance Metrics on Validation Data:

Accuracy	Precision	Recall	f1-score
0.9962121212121212	0.9850746268656716	0.99	0.9875311720698254