COMPARING THE PERFORMANCE OF ARTIFICIAL NEURAL NETWORKS WITH FEATURE SELECTION VERSUS SUPPORT VECTOR MACHINE WITH FEATURE SELECTION

Team Members	PeopleSoft ID
Abhishek Vanamala	1884291
Jahnavi Valisetty	1879688

ARTIFICIAL INTELLIGENCE

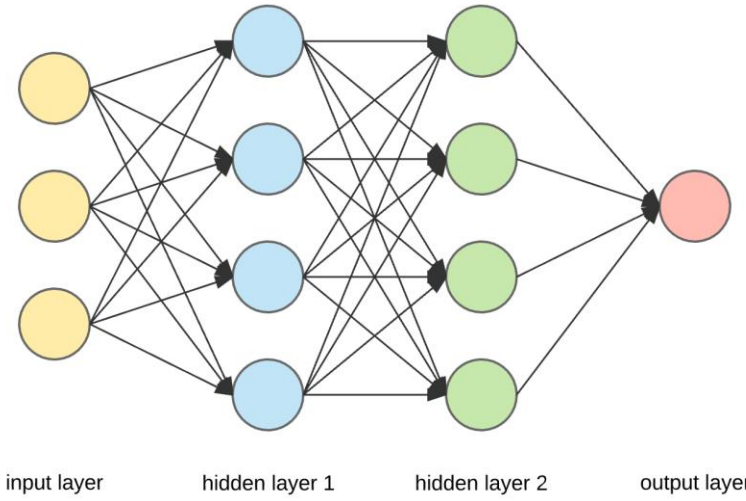
Artificial Intelligence is the wide-ranging branch of computer science that is concerned in building smarter machines which are capable to performing tasks that requires human intelligence. AI makes it possible for the machines to learn from experience, adjust to new inputs and to perform human-like tasks. Most of the AI examples that we hear today from chess playing computers to self-driving cars, they are relying heavily on deep learning and natural language processing. Using technologies like these, systems can be trained to achieve certain tasks by processing huge amounts of data and identifying patterns in it. In this huge interesting subject, we would like to explore more into neural networks.

Domain: NEURAL NETWORKS

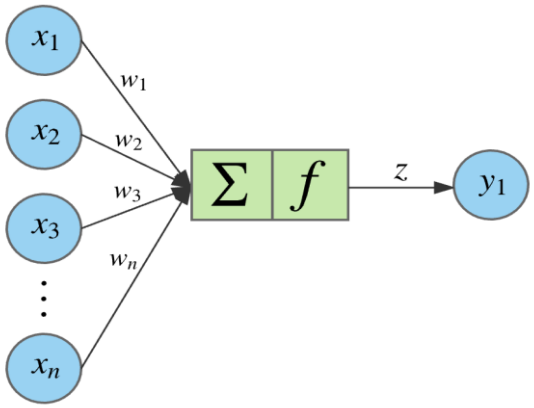
A set of algorithms which are modelled loosely that are designed to identify patterns after the human brain are called neural networks. By perception, labelling and clustering raw input data, they elucidate sensory data. The type of patterns that they identify are numerical which are contained in vectors into which all the real-world data must be interpreted. They help us cluster and classify the data as clustering and classifying can be imagined as two layers on top of the data that we store and want to process. They classify data when they have a labelled dataset and they group the unlabeled data into clusters according to the similarities in the inputs. For clustering and classification, neural networks can also extract features that are given to other algorithms.

ARTIFICIAL NEURAL NETWORKS (ANNs)

Artificial Neural Networks are the multiple layered fully connected neural networks that consists of input layer, hidden layer and output layer. Every node in one layer is connected to every other node in the next layer as shown in the below figure where node is just a place where computation happens. By increasing the number of hidden layers, we can make the network deeper.



If we zoom in into one of the output or hidden nodes, what we will come across is shown in the figure below:



The node takes the weighted sum of all the inputs and is passed through an **activation function** which is non-linear. This is the output which we get from the node which becomes the input of another node in the next layer and this goes on and the signal flows from left to right. The final output is calculated by carrying out the procedure for all the nodes. When we are training this deep neural network, it means that we are learning all the weights associated with all the edges.

Essentially what each layer of the artificial neural network does is a non-linear transformation of the input from one vector to another. And the main advantage of using deeper model is to be able to do non-linear transformations of the input and extracting more complex decision boundary.

Deep Neural networks are artificial neural networks with multiple layered structure which is incorporated within the input-output layers. They can model complex non-linear representations and can produce computational models where the object is demonstrated in terms of the layered composition of primitives.

FEATURE SELECTION:

In today's generation of IOT and Big Data, we are easily getting loaded with datasets with extremely high dimensions. In order to get any insights from such high dimensional data or to perform any machine learning task, feature selection becomes very important. Since there are some features which may be less significant or insignificant to the dependent variable. This unnecessary inclusion to the model leads to increase in time complexity for a model to get trained, makes a model harder to interpret by increasing the complexity of the model and it results in a less reliable or inaccurate predictions.

Hence performing the feature selection becomes the crucial need as it is very essential, and a must needed component in machine learning specially when dealing with high dimensional datasets.

For a dataset which contains n features, if we try with all possible combinations - we get (2^n – 1) model which needs to be evaluated for certain significant set of features. This approach takes lots of time and hence we use feature selection techniques in order to efficiently find smallest set of features.

There are three types of feature selection techniques:

1. Filter method
2. Wrapper method
3. Embedded methods:

Embedded methods combine the qualities of filter and wrapper methods and it is implemented by algorithms that have their own built-in feature selection methods.

DIFFERENCES BETWEEN FILTER AND WRAPPER METHODS:

The main differences between filter and wrapper methods are as follows:

- Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
- Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it.
- Filter methods might fail to find the best subset of features in many occasions, but wrapper methods can always provide the best subset of features.
- Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation.
- Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods.

MACHINE LEARNING:

Machine learning being the subtopic of artificial intelligence is very widely used and applied in various areas of the developing world today. These machine learning algorithms are responsible for the vast majority of the artificial intelligence applications and advancements that we hear about. These algorithms use various statistics in finding the patterns from massive amounts of data. And the data here consists of lots of things – images, numbers, words, number of clicks and so on. It can be fed into the machine learning algorithm if we can be digitally stored.

It comes in three flavors – supervised, unsupervised and reinforcement. In supervised machine learning, which is the most prevalent and widely used, it tells the machine exactly what patterns it should look for as it is given the data which is labelled. In unsupervised learning, the data is not labelled, and the machine looks for whatever patterns it can find. A reinforcement learning will try lots of different things and is rewarded or penalized depending on its behavior whether it helps or obstructs it from reaching its objective.

This process of machine learning **powers many of the services we have today** – recommendation systems like those on YouTube, Netflix, Spotify; search engines like Google and Safari; voice assistants like Alexa and Siri; Social media feeds like Twitter and Facebook.

In all these examples, every platform is collecting as much data possible about you – what categories you like watching, to which statuses you are reacting , what links you are clicking and then using a machine learning algorithm to guess what you might want next.

Frankly this process is quite basic: to find a pattern and to apply the pattern but it pretty much runs the world. The promise and the contribution which machine learning did today are fascinating. There are many real-life applications we are using today which are offered by machine learning and seems that it will rule the world in the coming days.

Hence, we came out with an idea of using supervised machine learning in classifying the data. We used both SVM and ANN learning algorithm to find the best classifier with higher success rate and accuracy. The below experimental section is organized as follows: the system model and design overview are explained in first section; the experimental analysis is explained in second section.

EXPERIMENTAL SECTION:

SECTION I:

SYSTEM MODEL

The proposed system is composed of feature selection and learning algorithm which is shown in the below figure. Feature selection component is responsible in extracting the most relevant features to recognize the instance of a certain class or group. The learning algorithm component builds the necessary knowledge or intelligence using the result which is found from the feature selection component. The model gets trained using the training dataset and builds its intelligence. Then to measure the accuracy of how much the model is correctly classified on the unseen data, the learning intelligences are applied on the testing dataset.

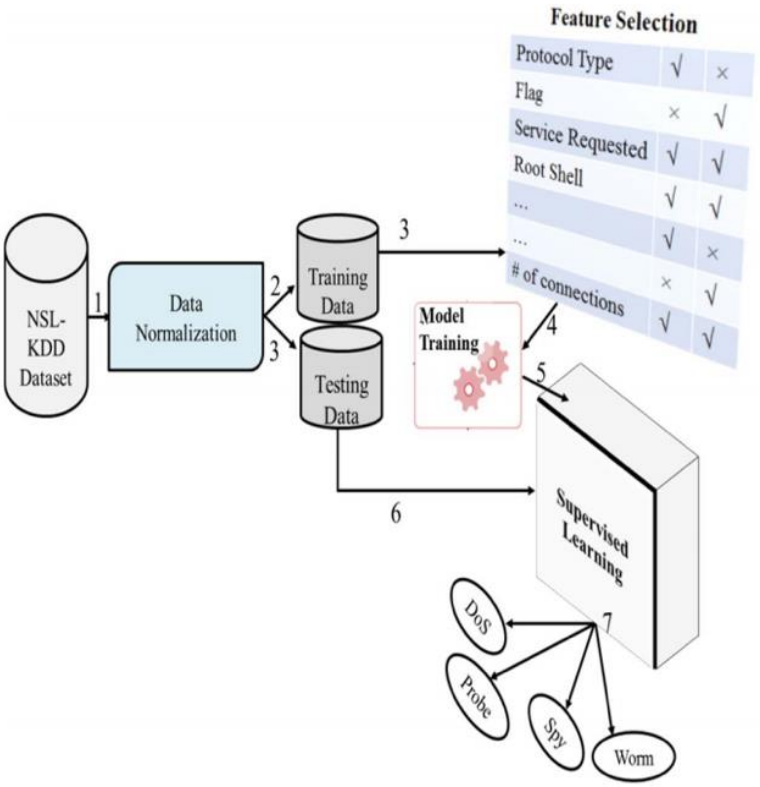


Fig: This is the proposed supervised machine learning classifier

A. FEATURE SELECTION

Feature Selection is the most important part of the machine learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection, **filter method** and **wrapper method** have been used. In feature selection, on the basis of their scores – the features are selected in various statistical tests that measure the relevance of features by their correlation with dependent variable or resultant variable. Wrapper method finds a subset of features by measuring the effectiveness and usefulness of a subset of features with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best selected feature subset depends on the machine learning algorithm which is used to train the model. In wrapper method a subset evaluator uses all possible subsets and then uses a classification algorithm to convince the classifiers from the features in each subset. The classifier considers the subset of features with which the classification algorithms performs the best. In order to find the subset, the evaluator uses different search techniques like breadth first search, random forest, depth first search or

hybrid search. Along with a ranker to rank all the features in a dataset, the filter methods use an attribute evaluator. One feature is removed here at a time which has a smaller number of ranks and then sees the predictive accuracy of the classification algorithm. Weights or ranks given by the ranker algorithm and the classification algorithm are different. Filter method is used for data mining test whereas wrapper method is used for machine learning test because data mining has thousands of millions of features.

B. BUILDING MACHINE INTELLIGENCE

Learning models are developed based on the best features selected in the feature selection process. Machine learning algorithm is used in order to develop a learning model. With the selected features, the training dataset is used to train the algorithm. In supervised machine learning, each instance in the training dataset has a class that it belongs to. The algorithm builds the learning model based on which machine learning algorithm is being used.

C. SUPPORT VECTOR MACHINE:

In SVM, depending on the type of problem and available datasets, a separating hyperplane defines the classifier. When dataset is one-dimensional, the hyperplane is a point and for two-dimensional data it is a separating line as shown in the below figure. For a three-dimensional, it is a plane and it is a hyperplane when the data dimension is higher.

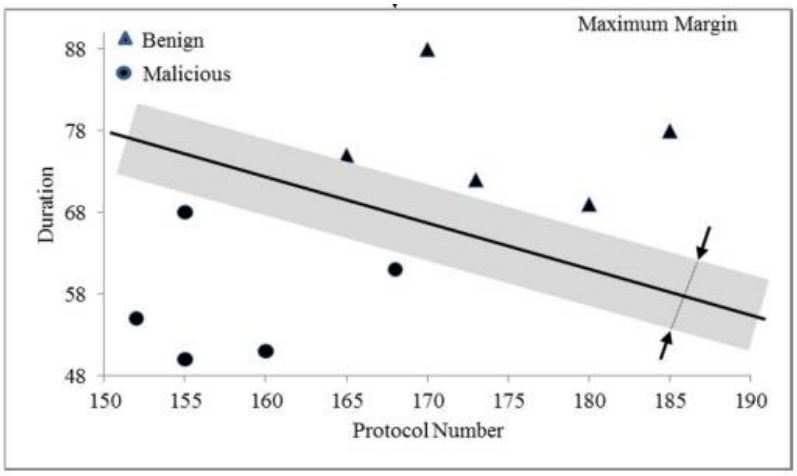


Fig. SVM classifier in two-dimensional problem spaces

For a linearly separable dataset, the classifier or the decision function will be in the form –

ax + by + c = 0 (1)

For a given data points (x, y), if ax + by >= c - the above decision function will classify the point in one class, or it will categorize if ax +by < c. The equation of a line y = ax + b can be rewritten as y-ax -b=0 which can be represented using two vectors as below-

$$W \begin{bmatrix} -b \\ -a \\ 1 \end{bmatrix} \quad \text{and } X \begin{bmatrix} 1 \\ x \\ y \end{bmatrix} \qquad (2)$$

Which says we can write the linear equation of the line using two vectors as below-

$w^T x = (-b) * 1 + (-a) * x + 1 * y$, or

$w^T x = y - ax - b$ (3)

The reason behind using the hyperplane equation instead of y=ax+b because working in two dimensions with this notation is easier and the vector w will always be normal to the hyper plane. The hyper plane can be used to make predictions once the hyper plane with maximum margin has been found. The hypothesis function will be –

$$h(x_i) = \begin{cases} +1; \text{ if } (w.x+b >= 0) \\ -1; \text{ if } (w.x+b < 0) \end{cases} \qquad (4)$$

D. ARTIFICIAL NEURAL NETWORK (ANN)

Artificial neural network is another machine learning tool. As the name infers, it is a system that is inspired from human brain and replicates the learning system of human brain. It consists of input and output layers as shown in the given figure with one or more hidden layers and uses backpropagation to adjust the outcome with the expected result or class.

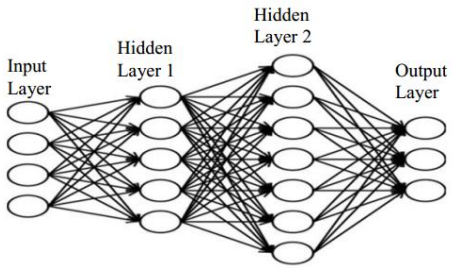


Fig. ANN showing input, output and hidden layers

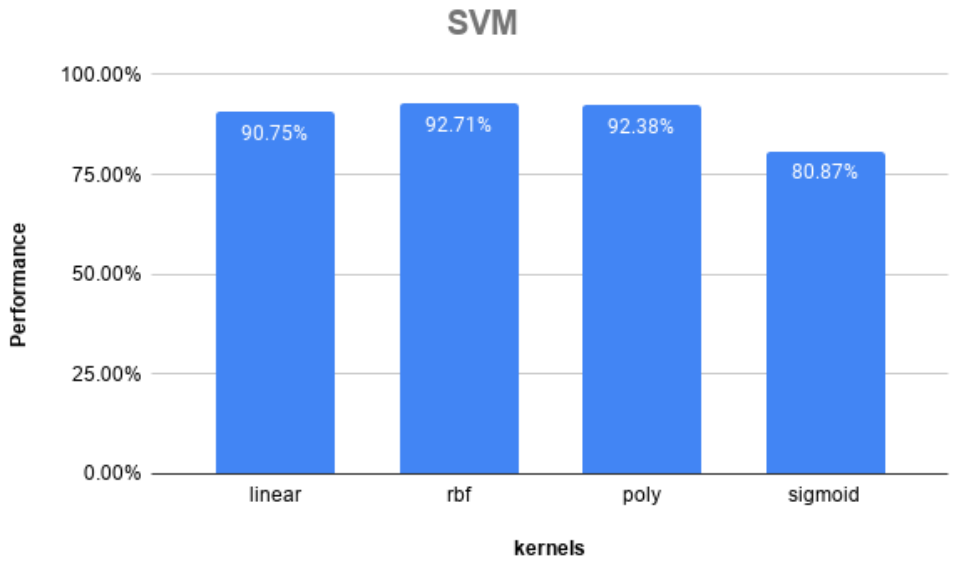
SECTION II

EXPERIMENTAL ANALYSIS

SVM Implementation using different kernels with feature selection:

SVM algorithms use a set of mathematical functions that are defined as a kernel. The kernel functions to take data as input and transform it into the required form. Different types of kernel functions are used by different SVM algorithms. These functions can be of different types which are linear, radial basis function (RBF), polynomial and sigmoid. We introduce kernel functions for sequence data, text, images, graphs, as well as vectors.

We implemented SVM with each of the different kernel functions with feature selection on the dataset (which is **NSL-KDD** dataset) and evaluated its performance and the results for the same are displayed in the below graph.



As shown in the above picture, the performance of the RBF kernel is more compared to others. And it is known to be the most used type of kernel function as it has localized and finite response along the entire x-axis. Hence, **we are going forward by using SVM with RBF kernel function with feature selection to use in our experimental analysis further.**

A. FEATURE SELECTION

The experiment is carried out using Keras tensor flow backend open source software suite which is popular for machine learning and data mining and it consists of two parts.

In the first part, we extracted the most relevant features using different feature selection methods. We used SVM classification algorithm with cross validation in wrapper method to avoid overfitting and underfitting problem. In the filter method to find the best result suitable for our proposed classifier, we used a ranker algorithm. The training data we used from NSL-KDD dataset contains 25,191 labelled examples and the results for the feature selection experiment are shown in the below **Table I**.

TABLE I: RESULT OF FEATURE SELECTION

FS Technique	FS Type	Input Features	Output Features
Correlation based	Wrapper	41	17
Chi-Square based	Filter	41	35

Correlation based feature selection found total **17** features most relevant from **41** features present in the training dataset whereas chi-square algorithm retained 35 features which are more relevant to the resultant class. These 17 and 35 retained features were used to train the model using seen or training dataset as well as to test the model using unseen or testing dataset.

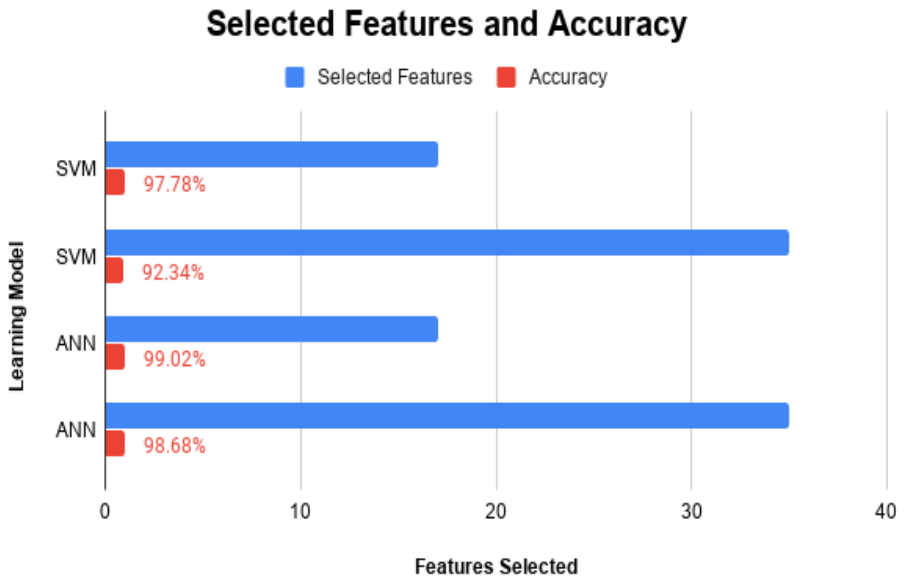
B. CLASSIFICATION

Total four models are built with Jupyter Notebook using the training dataset with the features found in feature selection part. For doing classification using supervised machine learning – it first requires training the model using training dataset. As training data, we used 20 percent of the NSL-KDD dataset which has 25,191 labelled data examples. To train the model for each type of feature selection, we used SVM and ANN learning algorithm. Hence, **we built four learning models, two using ANN and other two using SVM**. Among the two models built for each learning algorithm, one is built using 17 features and other using 35 features from the feature selection

part. Next these four models which are trained are evaluated with testing data picked from NSL-KDD dataset which has 22,542 examples. The findings are summarized in the following **Table II** –

TABLE II: RESULT OF CLASSIFICATION

Learning Model	Selected Features	Accuracy
SVM	17	92.71%
SVM	35	90.24%
ANN	17	99.20%
ANN	35	97.68%



In below table we listed the results with recently presented results in the literature. While comparing the performance of other works with the proposed model, we picked works having hypothesis of comparable aspects that are related to benchmarking datasets and to learning algorithms. **But there are certain other aspects like number of layers, number of instances, attribute reduction and learning rates used**. The detection success rate of the proposed model is compared with the other existing models in the following **Table III** -

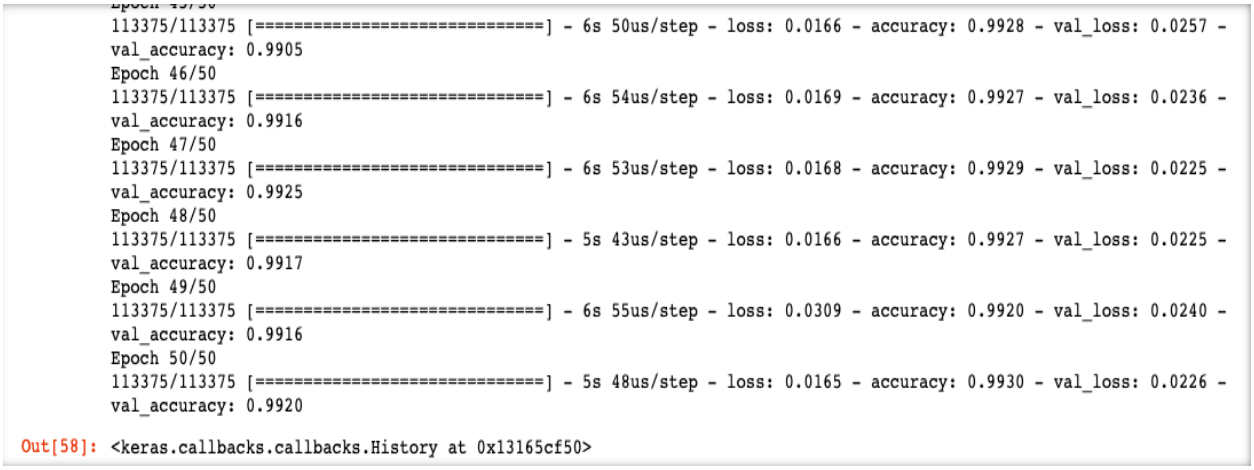


Fig: The final output of our proposed model

[Learning rate = 0.1, number of layers = 3]

TABLE III: PERFORMANCE COMPARISION WITH EXISTING MODELS

Learning Type	Our Model Accuracy	Existing Model	Existing Model
SVM	92.71%	92.84% [16]	89.52% [17]
ANN	99.20%	91.2% [18]	87.23% [19]

CONCLUSION:

We have presented different machine learning models using different machine learning algorithms and different feature selection methods to find a best model. **The analysis of the result shows that the model built using ANN and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 99.02%**. In the ANN model, we experimented with different number of hidden layer and found that the detection success rate varies with the number of hidden layers. **After several trial and error methods, we found best detection rate with 3 hidden layers and 0.1 learning rate. In the wrapper feature selection method, we also used SVM algorithm as classifier**. We believe that these findings and model will certainly contribute to research further in various domains like for building a detection system that can detect known attacks as well as novel network attacks, for gearbox fault diagnosis, for coal boiler plants using real time plant data and in various fields of medical diagnosis.

REFERENCES:

[1] https://www.extremetech.com/extreme/215170-artificial-neural-networks-are-changing-the-world-what-are-they
[2] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
[3] https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
[4] https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
[5] https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/
[6] https://www.kdnuggets.com/2020/02/deep-neural-networks.html
[7] https://towardsdatascience.com/kernel-function-6f1d2be6091
[8] https://stackabuse.com/applying-wrapper-methods-in-python-for-feature-selection/