Hepatitis classification

INT-247

(MACHINE LEARNING FOUNDATION)

A project report

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelors' of Technology

(Computer Science & Technology)

Submitted by:

ROBIN

11814491

KM007

A31

Github link: https://github.com/Machine-learning-2021-KM007-LPU/int-247-km007-ca-2-robinkamal

Abstract

Artificial intelligence and machine learning have promising applications in several medical fields of diagnosis, imaging, and laboratory testing procedures. However, the use of this technology in the veterinary medicine field is lagging behind, and there are many areas where it could be used with potentially successful outcomes and results. In this study, two critical predictions were explored in horses presented with acute abdomen (colic) using this technology. Those were the need for surgical intervention and survivability likelihood of affected horses based on clinical data (history, clinical examination findings, and diagnostic procedures). The two prediction parameters were explored using the application of Support Vector Machines, Logistics Regression, Random Forest. The machine learning algorithms were able to predict the need for surgery and survivability likelihood of horses presented with Accuracy 95% ,91% and 97% accuracy, respectively.

Description about dataset

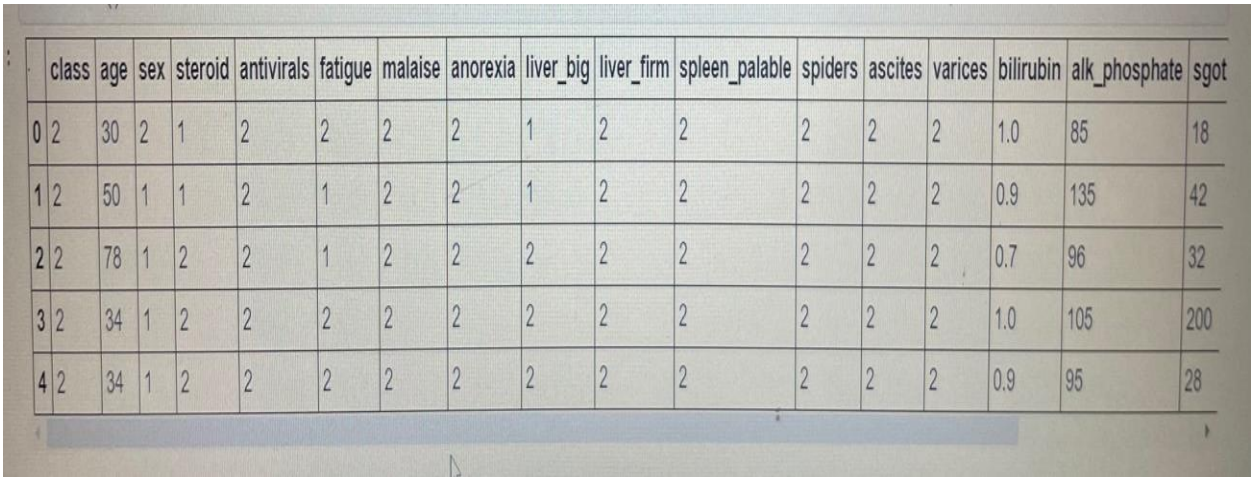
Hepatitis classification dataset consists of the data arranged in tabular form containing fixed number of rows and columns.

The data is about the health conditions of a person suffering from hepatitis and based on these health conditions we have to predict whether that person will survive or not.

There are independent variables also known as features and target class that tells whether a person survives or not.

There are 20 columns in total 19 are features and one is the class variable that determines the status of its survival. Target variable either predicts 2 that means a person survives and 1 that means person does not survive.

So, using this given dataset we have to build a ML model that will aim at correct prediction of the class.



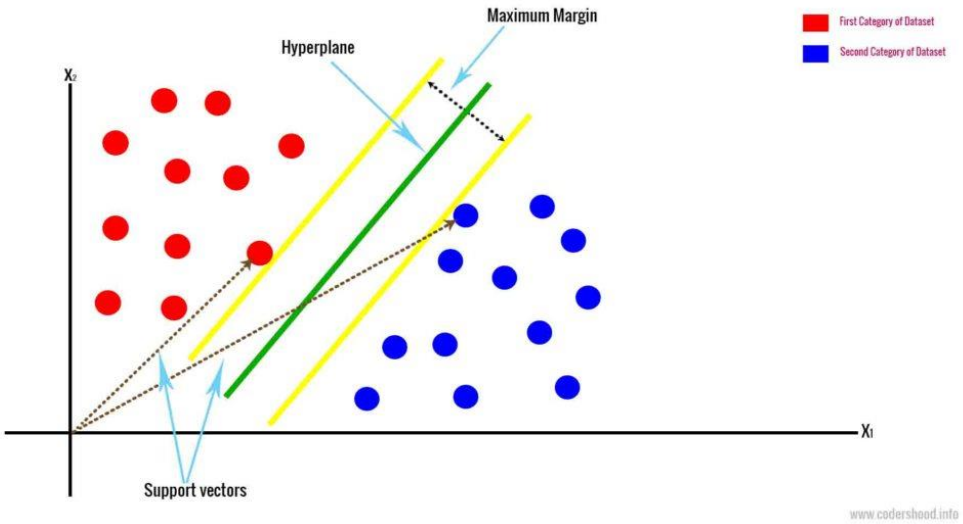
	class	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm	spleen_palable	spiders	ascites	varices	bilirubin	alk_phosphate	sgot
0	2	30	2	1	2	2	2	2	1	2	2	2	2	2	1.0	85	18
1	2	50	1	1	2	1	2	2	1	2	2	2	2	2	0.9	135	42
2	2	78	1	2	2	1	2	2	2	2	2	2	2	2	0.7	96	32
3	2	34	1	2	2	2	2	2	2	2	2	2	2	2	1.0	105	200
4	2	34	1	2	2	2	2	2	2	2	2	2	2	2	0.9	95	28

Machine Learning

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to ”learn” with data, without being explicitly programmed.

1. Support Vector Machine:

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.



The followings are important concepts in SVM –

Support Vectors − Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.

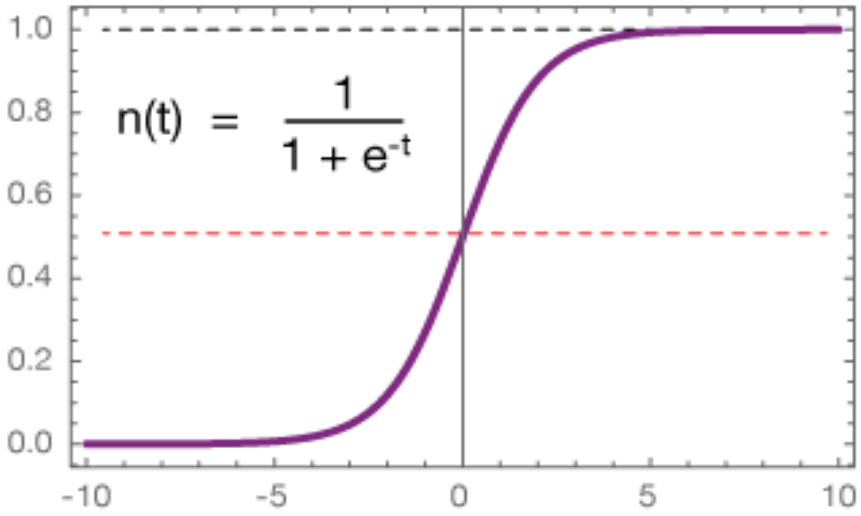
Hyperplane − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

Margin − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

2. Logistic Regression

Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. A solution for classification is logistic regression. Instead of fitting a straight line or hyperplane, the logistic

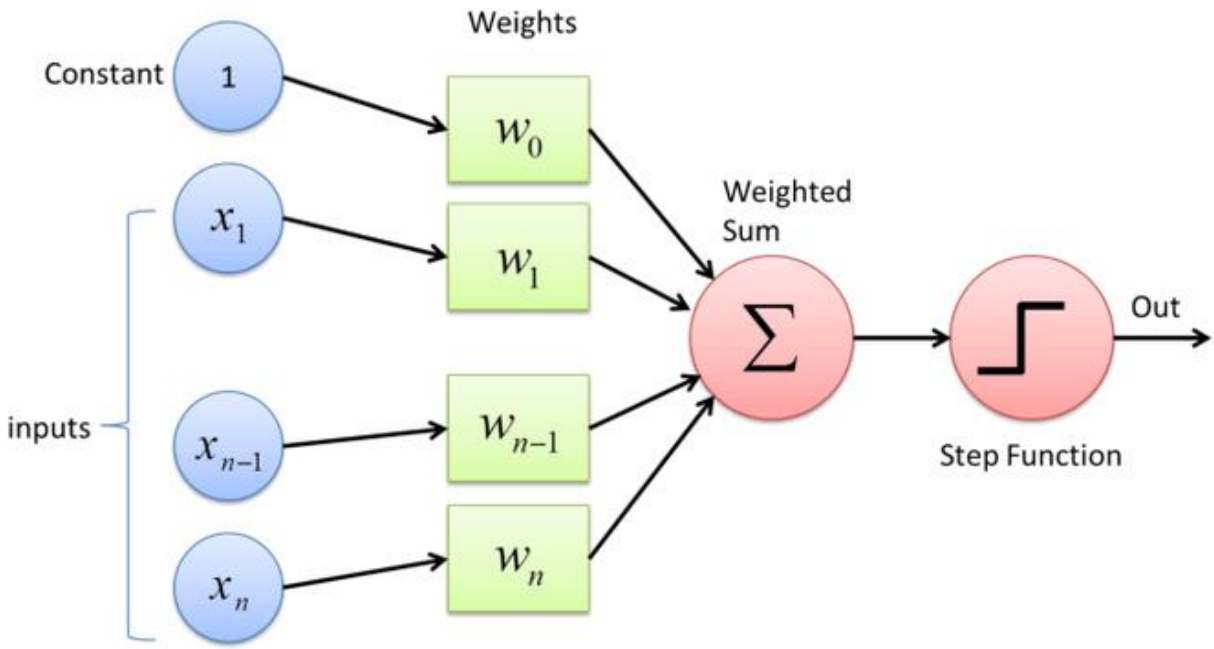
regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as:



3.Preceptron classifier

The **Perceptron** is a linear machine learning algorithm for binary classification tasks.
It may be considered one of the first and one of the simplest types of artificial neural networks. It is definitely not “deep” learning but is an important building block.

Like logistic regression, it can quickly learn a linear separation in feature space for two-class classification tasks, although unlike logistic regression, it learns using the stochastic gradient descent optimization algorithm and does not predict calibrated probabilities.



Data Exploration and Analysis

Data exploration is the initial step in data analysis, where users explore a large data set in an unstructured way to uncover initial patterns, characteristics, and points of interest. This process isn't meant to reveal every bit of information a dataset holds, but rather to help create a broad picture of important trends and major points to study in greater detail.

Data exploration can use a combination of manual methods and automated tools such as data visualizations, charts, and initial reports.

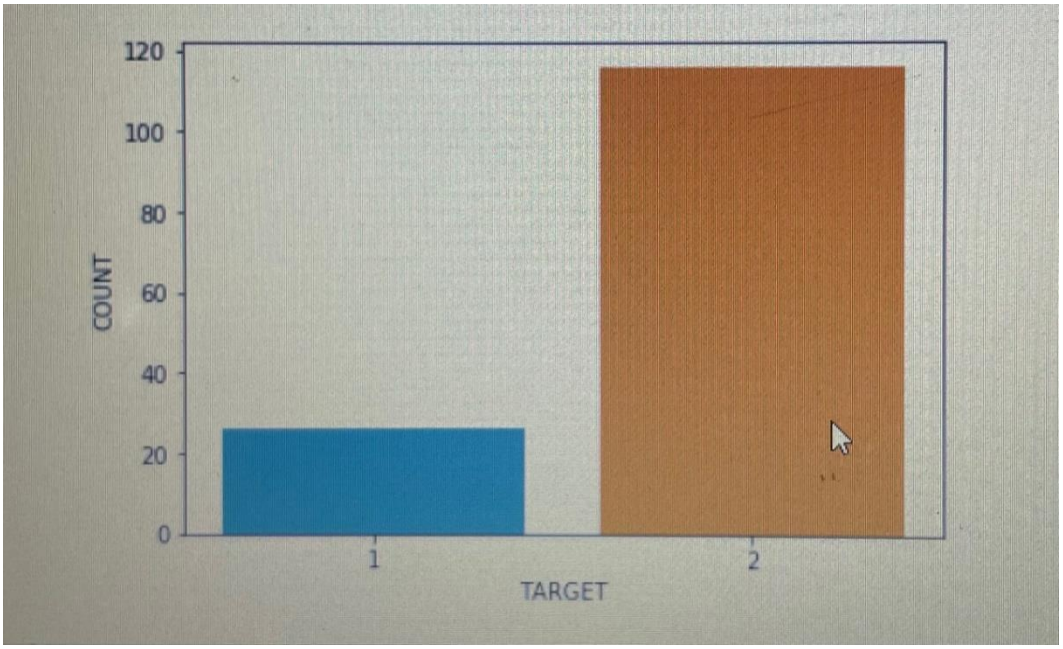
This process makes deeper analysis easier because it can help target future searches and begin the process of excluding irrelevant data points and search paths that may turn up no results. More importantly, it helps build a familiarity with the existing information that makes finding better answers much simpler.

*First, we need to check about the basic details about the data like the shape of dataset, no of tuples, no of columns, target data that the dependent variable in our data.

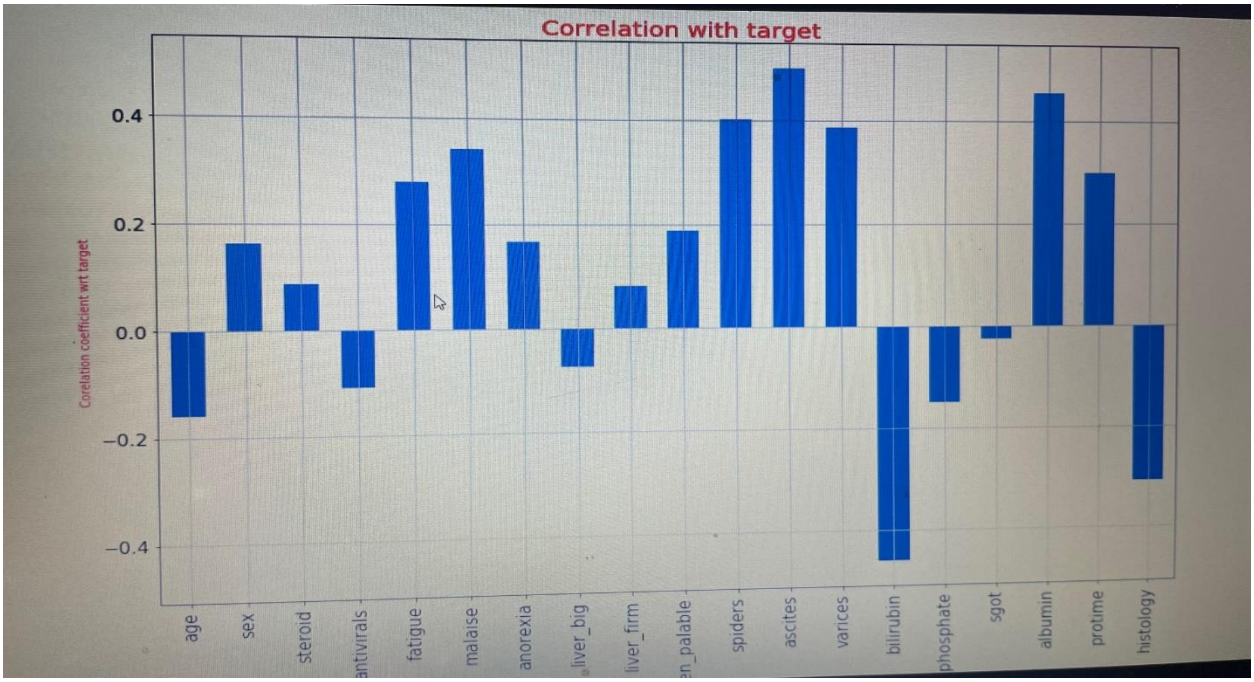
*Then we need to perform the univariate analysis about our features and collect as much information as we can.

*Then we need to check the missing values in our dataset and deal with it accordingly. For integer type features we can replace the missing values with mean or median.

*Then we check about the target data.



*Co-relation of our data features with target.



Experiment and results

1.SVM

TRAINING ACCURACY: 0.9595959595959596

TESTING DATA's ACCURACY: 0.8837209302325582

Confusion Matrix:
```
[[ 8  4]
 [ 1 30]]
```

Classification Report:

	precision	recall	f1-score	support
1	0.67	0.44	0.53	9
2	0.86	0.94	0.90	34
accuracy			0.84	43
macro avg	0.77	0.69	0.72	4
wt avg	0.82	0.84	0.82	43

2. Perceptron

TRAINING ACCURACY: 0.9191919191919192

TESTING DATA's ACCURACY: 0.7674418604651163

Confusion Matrix:
```
[[ 1  8]
 [ 2 32]]
```

Classification Report:

	precision	recall	f1-score	support
1	0.33	0.11	0.17	9
2	0.80	0.94	0.86	34
accuracy			0.77	43
macro avg	0.57	0.53	0.52	43
wt avg	0.70	0.77	0.72	43

3. Logistic Regression

TRAINING ACCURACY: 0.9292929292929293

TESTING DATA's ACCURACY: 0.8604651162790697

Confusion Matrix: [[4 5]

```
[ 1 33]]
```

Classification Report:

	precision	recall	f1-score	support
1	0.33	0.11	0.17	9
2	0.80	0.94	0.86	34
accuracy			0.77	43
macro avg	0.57	0.53	0.52	43
wt avg	0.70	0.77	0.72	43

conclusion

Since Support Vector machine has highest accuracy for the testing data of the hepatitis dataset, thus the best model for this dataset is SVM that gives more correct predictions than perceptron model and logistic regression model.

Removing of those features that has negative correlation with target variable reduces the complexity and easiness to work with dataset increases. Also, it improves the efficiency of our model.

Also, the standardization of data further improves the accuracy of the model.

Thank You