Naman Gupta
Naman1102@gmail.com

Table of Contents

Abstract 2
Aim and Scope 2
Introduction 2
Experimental setup 2
Experimental procedure 2
The Dataset: 3
Methodology 3
MODLES: 4
Evaluation metrics 4
Logistic Regression: 5
Naïve Bayes 6
Support Vector Machine 7
K-Nearest Neighbour 7
Random Forest Classifier 8
Conclusion: 8
References: 9

Abstract

This paper helps to determine the best model for a unique dataset which comprises of attributes which are the same with slightly different readings. It has tried to explain the readings from a chemistry related experiment and tried to use this data to develop a model for a machine which wants to differentiate between which kind of liquid it is in. In this paper have, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Random Forest Classifier models have been used. It was found that logistic regression performs the best with this kind of dataset.

Keywords: Classification, Digital sensors, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Random Forest Classifier, Binary Classification

Aim and Scope

Testing multiple classification models (especially for binary classification) on digital sensors datasets.

Introduction

The dataset is acquired from a capacitive sensor array composed of a set of sensor electrodes immersed in three different phases: air, oil, and water. It is composed of digital signals obtained from one electrode while it was immersed in the oil and water phases at different times.

Experimental setup

[1] The experimental setup is composed of a capacitive sensor array that holds a set of sensing cells (electrodes) distributed vertically along the sensor body (PCB). The electrodes are excited sequentially and the voltage (digital) of each electrode is measured and recorded. The voltages of each electrode are converted to intensity values by the following equation:

Intensity = (|Measured Voltage - Base Voltage| / Base Voltage) x 100

Where the Base Voltage is the voltage of the electrode recorded while the electrode is immersed in air. The intensity values are stored in the dataset instead of the raw voltage values.

Experimental procedure

The aim of the experiments is to get fixed-size intensity signals from one electrode (target electrode) when being immersed in water and oil; labelled as +1 (water) or -1 (oil). For this purpose, the following procedure was applied:

- The linear actuator was programmed to move the sensor up and down at a constant speed (20 mm / second).

- The actuator stops when reaching the upper and bottom positions for a fixed duration of time (60 seconds).

- At the upper position, the target electrode is immersed in oil; intensity signals are labelled -1 and sent to the PC.

- At the bottom position, the target electrode is immersed in water; intensity signals are labelled +1 and sent to the PC.

- The sampling rate is 100 msec; since each intensity signal contains 10 values, it takes 1 second to record one intensity signal

Environmental conditions

The experiments were performed under indoors laboratory conditions with room temperature of around 23 degree Celsius.

The Dataset:

The signals included in the dataset are composed of intensity signals each with 10 consecutive values and a label in the last column. The label is +1 for a water-immersed electrode and -1 for an oil-immersed electrode.

The data has been pre-processed to remove the duplicate values, normalization was not required as all attributes are the same kind of electrodes which result in similar range of values.

After removing duplicate readings, we have 1337 unique readings in our data set.

Methodology

Testing existing models on the dataset and evaluating them on performance metrics and then finding the best classification model for the dataset. Here the dataset was split in ratio of 1:3 to testing : training datasets.

Here is a flowchart of the entire procedure of applying machine learning models to the dataset.

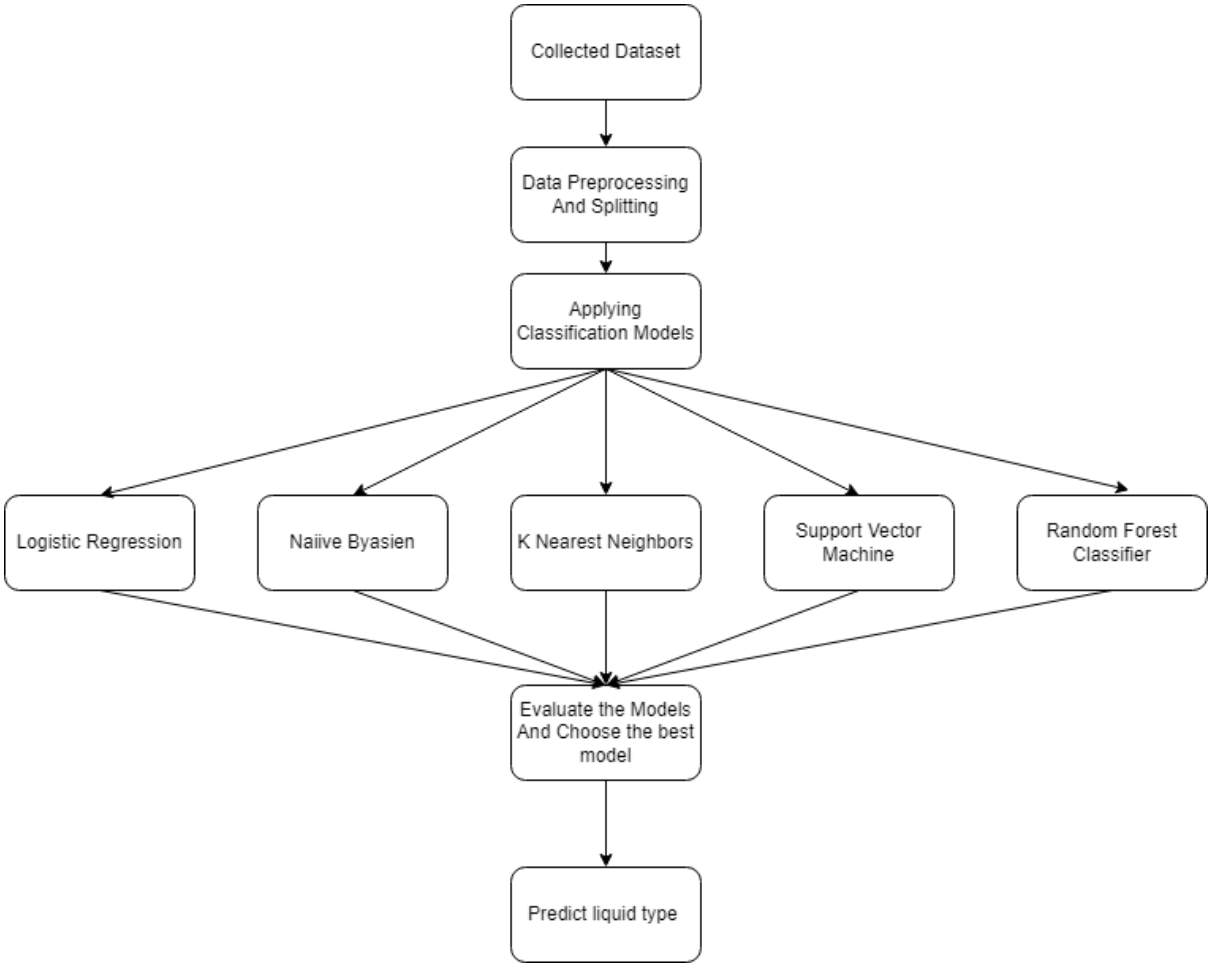

Figure 1

MODLES:

I have used 5 different classification models and evaluated them based on their accuracy score on the same data set.

Evaluation metrics

Note: All models have been used using their accuracy score. Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

These values can viewed in the confusion matrix as:

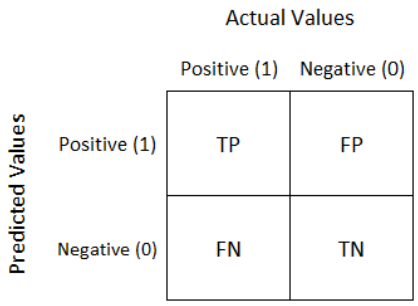

Figure 2

Logistic Regression:

[2]Here I have used binary logistic regression. Despites its name, its an algorithm vastly used for classification. It is an extremely fast algorithm to train and to predict.

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function or the sigmoid function is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Where :

- ‘e’ is the base of natural logarithms
- ‘value’ is the actual numerical value that you want to transform

When applied to my dataset, the best scores are received amongst any other classifier.

```
[[147   0]
 [  3 185]]

0.991044776119403
```

Figure 3

Naïve Bayes

[3]Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 4

Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets, and can even be used for multiclass classification although we only need binary class classification.

The results are as follows:

```
[[136  11]
 [ 58 130]]

0.7940298507462686
```

Figure 5

As we can see the accuracy score is only 0.79 or 79%. This shows that Naïve Bayes is not a good model for the dataset. This is because all the attributes in the dataset are not completely independent which is a crucial pre requisite of Naïve Bayes.

Support Vector Machine

[4] SVM seeks the best decision boundary which separates two classes with the highest generalization ability. Unlike logistic regression, which defines optimality by overall probability, SVM wants the smallest distance between data points and the decision boundary to be as large as possible.

The results of the model are as follows:

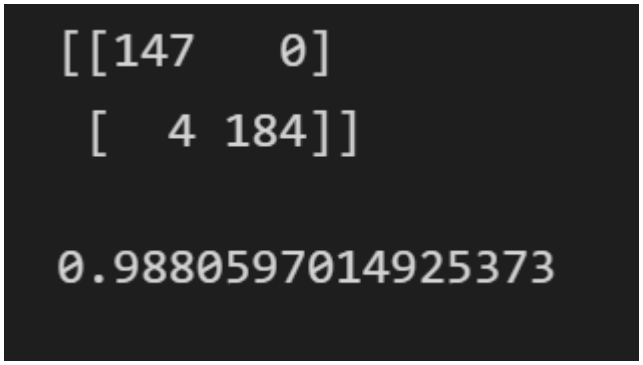

Figure 6

Here the accuracy is 98% which is very good. SVM was definitely successful at binary classification.

K-Nearest Neighbour

[5] K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

I applied KNN with 5 nearest neighbours to predict the class and the results are as follows:-

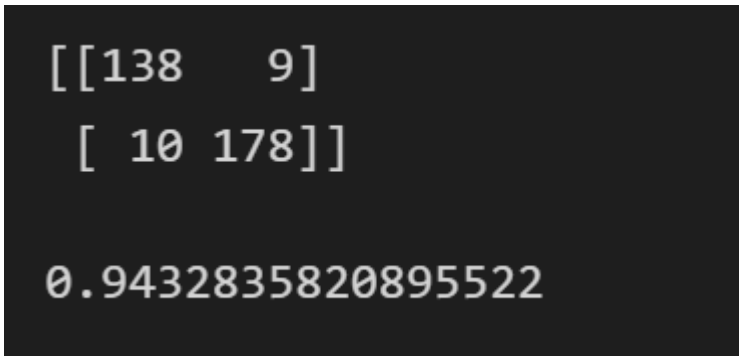

Figure 7

With 94% accuracy KNN provides a fast and reliable model that can be used.

Random Forest Classifier

[7] Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Ideally, the greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The results with 100 trees is as follows:-

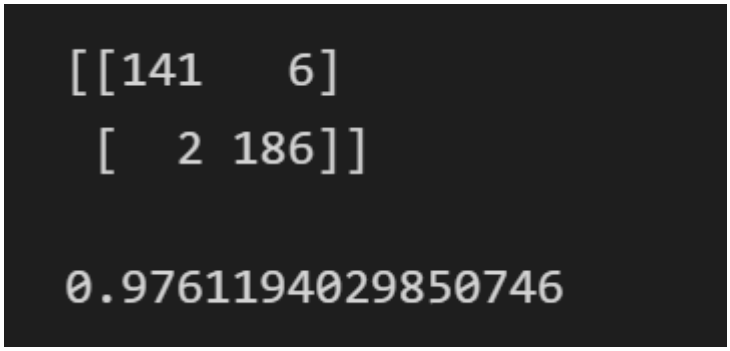

Figure 8

Random forest was able to achieve 97% accuracy which is very high. Although it ensures a high accuracy, the training time is high as well.

Conclusion:

The results can be tabulated as:-

Table 1

Machine Learning Model	Accuracy Score
Logistic Regression	99%
Naïve Bayes	79%
Support Vector Machine	98%
K-Nearest Neighbour	94%
Random Forest Classifier	97%

As we can see logistic regression gives the best results for this dataset. Because the data is taken from 10 similar electrodes, the readings are similar which makes it easier for models to predict classes except for Naïve Bayes which works mostly on probability.

Since Logistic Regression has the fastest training speed and still has the best accuracy, I choose it as the best model for Binary classification of Digital Sensor Signals.

The model can be used by a device which has to work in dual liquid areas to identify the type of liquid it is working under automatically using the same sensors used in the experiment. In

this case knowing the difference between oil and water can help a moving device set its trajectory according to the density differences between the two liquids.

References:

[1] Mahdi Saleh, Imad H. Elhajj, Daniel Asmar, October 5, 2020, "Dataset for binary classification of digital sensor signals", IEEE Dataport, doi: https://dx.doi.org/10.21227/6a44-0880.

[2] Ramosaco, Miftar & Hasani, Vjollca & Dumi, Alba. (2015). Application of Logistic Regression in the Study of Students' Performance Level (Case Study of Vlora University). Journal of Educational and Social Research. 10.5901/jesr.2015.v5n3p239.

[3] Nurul Iman Saiful Bahari *et al* 2014 *IOP Conf. Ser.: Earth Environ. Sci.* **20** 012038

[4] Sindhy Genjang Setyorini and Mustakim 2021 J. Phys.: Conf. Ser. 2049 012026

[5] F. -J. Yang, "An Implementation of Naive Bayes Classifier," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 301-306, doi: 10.1109/CSCI46756.2018.00065.

[6] Jihao You, Sasha A.S. van der Klein, Edmond Lou, Martin J. Zuidhof,

[7] Application of random forest classification to predict daily oviposition events in broiler breeders fed by precision feeding system, Computers and Electronics in Agriculture, Volume 175,2020