# EVALUATIONS OF MODELS - SUPERVISED MACHINE LEARNING

## Abstract

In this paper, several supervised machine learning algorithms [five learning algorithms] are used to solve classification problems [four data sets]. The problems varied in size and in topic, while the algorithms varied in complicity. From this, comparisons were made between all the learning models accuracy scores to evaluate their performance when working with different datasets. While only looking at accuracy to evaluate each model, it is also important to note that precision is an important evaluation metric as well, which was not used for evaluations.

## 1. Introduction

For this project, four datasets were explored to show how supervised machine learning algorithms solve classification problems, comparing their performance based on their accuracies. The datasets were collected from UCI Machine Learning Repository (https://archive.ics.uci.edu/) and Kaggle (https://www.kaggle.com/). The datasets that were explored include: Stroke data which uses 14 features and 1 target to predict Strokes, Heart data which uses 13 features and 1 target to predict causes for heart conditions, Adult Income data which uses 14 features and 1 target to predict the income of Adults, and the Breast Cancer Data which uses 30 features and 1 target to predict if a mass is cancerous. Each dataset was processed to be able to run as a classification problem for each model. The models that were used include: Decision Tree (DT), SVM, Logistic Regression (LOG), K-Nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP) with 5 hidden layers.

## 2. Methodology

### 2.1 Decision Tree

The C parameters for the decision tree were 1,2,3,4, and 5. Where the optimal parameter was 1.

### 2.2 SVM

For the SVM model the regularization parameters included 0.00001, 0.0001, 0.001, 0.01, 0.1, and 1. Where the optimal parameter was 0.00001.

### 2.3 LOG

For the LOG Regression model the C parameters included 0.00001, 0.0001, 0.001, 0.01, 0.1, and 1. Where the optimal parameter was 1.

### 2.4 KNN

For the KNN model the depth of K included 1, 2, 3, 4, 5, and 6. Where the optimal parameter was 6.

### 2.5 MLP

MLP is a basic neural network, for this case it has a (5,5) hidden layer and alpha of 0.00001.

### 2.1 General Approach

Each of the models had different partitions, notified as splits. The splits that were used included 20/80 split, 50/50 split, and 80/20 split. From these splits the training accuracy was obtained, the testing accuracy, and afterwards the validation accuracy was obtained. Each model was run through a 5-fold cross validation test where the average accuracy was obtained. From these 3 accuracies [test, train, and validation], the weighted average of the 3 accuracies was noted to find the mean performance for each model, for each learning model relative to each dataset. The model with the highest accuracy was also noted for each dataset, onto a grid.

### 2.1 Stroke Data

The Stroke Data set had 5110 instances if a patient got a stroke. The best classifier had a

95% accuracy. The data that was obtained came from Kraggle/FEDESORIANO where they looked at age, marriage status, bmi, and smoking status (to name a few), to predict if the patient will have a stroke.

**2.2 Heart Data**

The Heart-Disease dataset had 303 instances of diagnosis of heart diseases. The best classifier had an 84% accuracy. The data was obtained from the Cleveland Database, where it looked at sex, age, resting blood pressure, cholesterol, and maximum heart rate (to name a few), to predict if a heart disease diagnosis would appear for a patient.

**2.3 Adult income Data**

The Adult-income dataset had 48842 instances of global income, but only US data was looked at for this paper, which left 27504 instances. Where the best classifier had a 83% accuracy. The data was extracted by Barry Becker using the 1994 Census database. It looks at features of an individual such as occupation, education, sex, race, and work class (to name a few), to predict income that is more than 50k or less than or equal to 50k.

**2.4 Breast Cancer Data**

The Breast Cancer dataset had 569 instances, where the best classifier had had a 95% accuracy. This dataset got its features from a collection of breast masses, where it used characteristics to predict if the mass was benign or malignant.

**3. Experiment**

The 20/80, 50/50, and the 80/20 splits were recorded for each data set and put into a grid that showed how all the models performed based on the accuracies. The mean of the accuracies was reported for each model to the respect for each dataset. The data was processed using one-hot-encoding when looking at the features. For the targets, since they were binary, they were converted into 0 or 1, where 1 made a positive result and 0 was a negative result. The data was also stacked and shuffled to counter pattern recognition of randomness. After, from the shuffled features and target, the features were split into 20/80, 50/50, and 80/20.

| 20/80 split | Stroke Data | Heart Data | Adult Data | Breast Cancer Data |
|---|---|---|---|---|
| DT | 0.9518 | 0.7976 | 0.8272 | 0.9592 |
| SVM | 0.9518 | 0.6654 | 0.8025 | 0.9070 |
| LOG | 0.9518 | 0.8263 | 0.8261 | 0.9429 |
| KNN | 0.9517 | 0.6835 | 0.8091 | 0.9454 |
| MLP | 0.9518 | 0.5378 | 0.8244 | 0.61352 |

| 50/50 split | Stroke Data | Heart Data | Adult Data | Breast Cancer Data |
|---|---|---|---|---|
| DT | 0.9514 | 0.8407 | 0.8288 | 0.9662 |
| SVM | 0.9529 | 0.6288 | 0.7739 | 0.9121 |
| LOG | 0.9529 | 0.8066 | 0.8240 | 0.9578 |
| KNN | 0.9527 | 0.6877 | 0.8117 | 0.9386 |
| MLP | 0.9517 | 0.5347 | 0.8245 | 0.3856 |

| 80/20 split | Stroke Data | Heart Data | Adult Data | Breast Cancer Data |
|---|---|---|---|---|
| DT | 0.9567 | 0.8093 | 0.8320 | 0.9636 |
| SVM | 0.9585 | 0.5590 | 0.7511 | 0.8979 |
| LOG | 0.9585 | 0.8021 | 0.8232 | 0.9268 |
| KNN | 0.9585 | 0.6491 | 0.8160 | 0.9240 |
| MLP | 0.9585 | 0.5604 | 0.8223 | 0.5976 |

## 4. Conclusion

From these results it was shown that DT tended to outperform the rest of the models with the overall weighted average across all splits and datasets accuracy being 89% . LOG had a consistent accuracy ranging from 80% to 95% across all dataset and splits, and an overall accuracy average of 88%. MLP being the lowest performance, with an average across all dataset and splits being 71%. These findings are interesting to see since the majority of the models are outdated ML models, we can see they are performing relatively well. Adjustments to the complexity for the DT,  LOG, SVM, and KNN [improving K, C, tree-depth by finding the best parameters] can potentially improve the accuracy scores. Adjustment to MLP potentially could have been made, where the weights for the hidden layers were potentially not enough for the data that it started to output false negatives.

## References

*Adult-income*. UCI Machine Learning Repository. (n.d.-a). https://archive.ics.uci.edu/dataset/2/adult

*Breast cancer wisconsin (Diagnostic)*. UCI Machine Learning Repository. (n.d.-b). https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

An empirical comparison of supervised learning algorithms. (n.d.). https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf

Fedesoriano. (2021, January 26). *Stroke prediction dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-datase

*Heart disease*. UCI Machine Learning Repository. (n.d.-c). https://archive.ics.uci.edu/dataset/45/heart+disease