Classification of Orthopedic Patients

Mehmet Zahit ANGİ
Engineering Faculty
Eskişehir Technical University
Eskişehir, Turkey
mehmetzahitangi@eskisehir.edu.tr

***Abstract*— To classify the orthopedic patients by human orthopedic data. Three different classification algorithms are used. These are K Nearest Neighbor Classifier, Decision Trees, Support Vector Machine algorithms. In addition Confusion Matrix and Accuracy score are used for evaluation.**

Keywords-component; Classification; Machine; Learning; KNN;

I. INTRODUCTION

In this project, our aim is to classify the orthopetic patients. To provide this classification, three different machine learning algorithms has been used.

II. PROPOSED APPROACH

A. Model Selection

There are various of machine learning algorithms. Among these algorithms, three algorithms we use are as follows: K Nearest Neighbor Classifier, Decision Trees, Support Vector Machines.

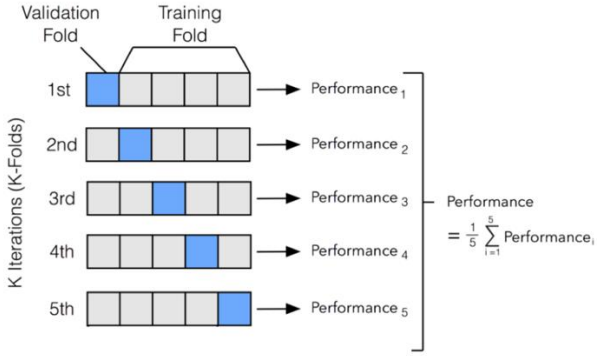
K nearest neighbor classifier is one of the algorithms used for classification and regression in supervised learning. It is considered the simplest machine learning algorithm. Unlike other supervised machine learning algorithms, it does not have a training phase. Training and testing pretty much the same thing. It is a lazy type of learning. Basically, looking for points closest to new point. K represents the amount of nearest neighbors of the unknown point. We choose k quantities of the algorithm (usually an odd number) to predict the results. In KNN classification the output indicates that it belongs to the class. An object is classified by the majority of its neighbors; the object is assigned the class most common among its nearest neighbors (k is a small positive integer). If k equals to one the object is simply assigned to the nearest neighbor's class.

Tree-based learning algorithms are considered as one of the most widely used and supervised learning methods. Tree-based methods have high accuracy, stability, and ease of interpretation. Unlike linear models, they can also map nonlinear relationships quite well. Classification or regression can be adapted to solve any given problem. Methods such as decision trees, random forest, gradient reinforcement are widely used in problems.

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression problems. However, it is mostly used in classification problems. In this algorithm, each data item is plotted as a point in the n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Next, the classification is performed by finding the hyperplane that distinguishes quite well between the classes.[1]

In addition to the methods K Fold Cross Validation Method is used. K-Fold Cross Validation is one of the methods of splitting the dataset for evaluating classification models and training the model. In K-Folds Cross Validation, we divide our data into k different subsets. We use k-1 subsets to train our data and leave the final subset as test data. The average error value obtained as a result of k experiments indicates the validity of our model.[2]

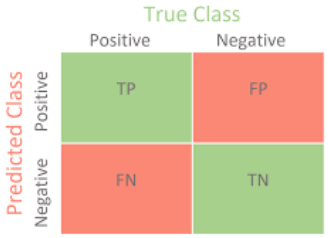
Figure 1: 5-Fold Cross Validation



B. Evaluation Metrics

In the project, two different evaluation methods are used as follows: Confusion Matrix and Accuracy. A confusion matrix is a table often used to describe the performance of a classification model on a set of test data for which the actual values are known.

Figure 2: Confusion Matrix



From this matrix we can extract the following information:

- True Positives (TP): These are instances where the true value is 1 and the predicted value is 1.

- True Negatives (TN): These are instances where the true value is 0 and our predicted value is 0.
- False Positives (FP): These are instances where the true value is 0 but our predicted value is 1.
- False Negatives (FN): These are instances where the true value is 1 but our predicted value is 0.

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right [3]. Formally, accuracy has the following definition:

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

III. Experiments and Results

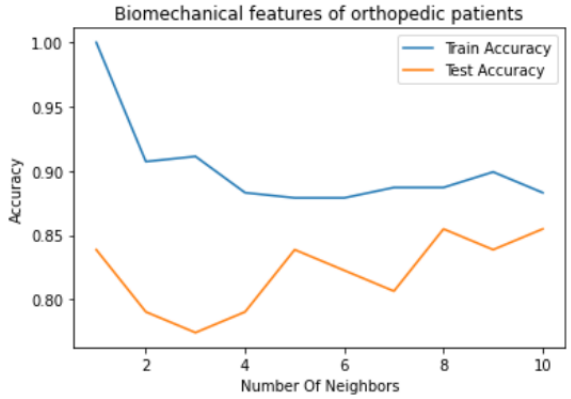
The results were compared using three different classifiers. 5 K fold cross validation method was used for all feature selections, and 80% training and 20% test data holdout method were used for algorithm training and prediction.

A. K Nearest Neighbors

For the number of neighbors, all values between 1 and 11 were tried. It has been seen that the best value for the number of neighbors is 8. This value is used for training and testing in the KNN model.

The results that have all neighbour obtained are as follows:

Figure 3: KNN All Neighbors Accuracy Results



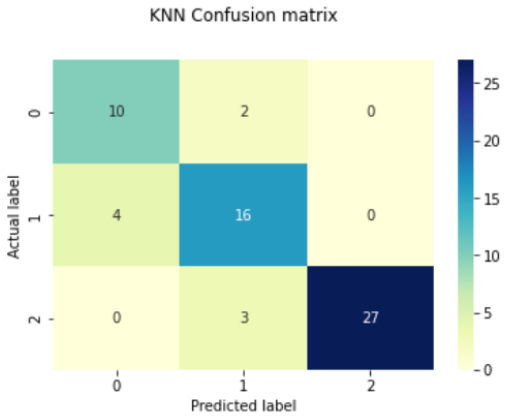
The results that have 8 neighbour obtained are as follows:

Table 1: KNN Accuracy And Classification Report Results

KNN Test Accuracy Score 0.8548387096774194
KNN Classification Report

	precision	recall	f1-score	support
Hernia	0.71	0.83	0.77	12
Normal	0.76	0.80	0.78	20
Spondylolisthesis	1.00	0.90	0.95	30
accuracy			0.85	62
macro avg	0.83	0.84	0.83	62
weighted avg	0.87	0.85	0.86	62

Figure 4: KNN Confusion Matrix Result

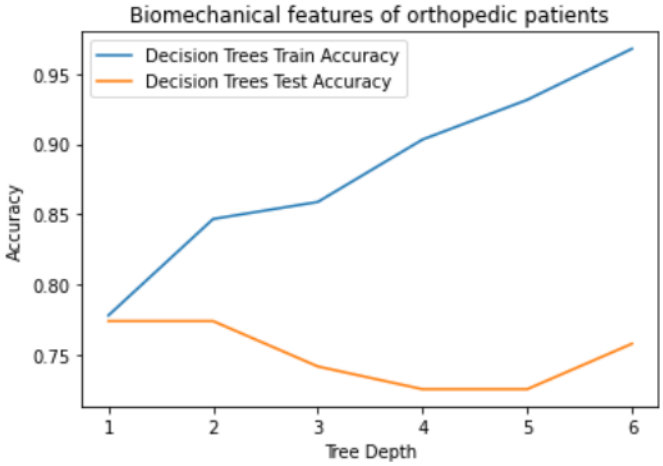


B. Decision Trees

For the tree depth, all values between 1 and 7 were tried. It has been seen that the best value for the number of tree depth is 2. This value is used for training and testing in the Decision Tree Classifier model.

The results that have all depth obtained are as follows:

Figure 5: Decision Tree All Depths Accuracy Results



The results that have 2 tree depth obtained are as follows:

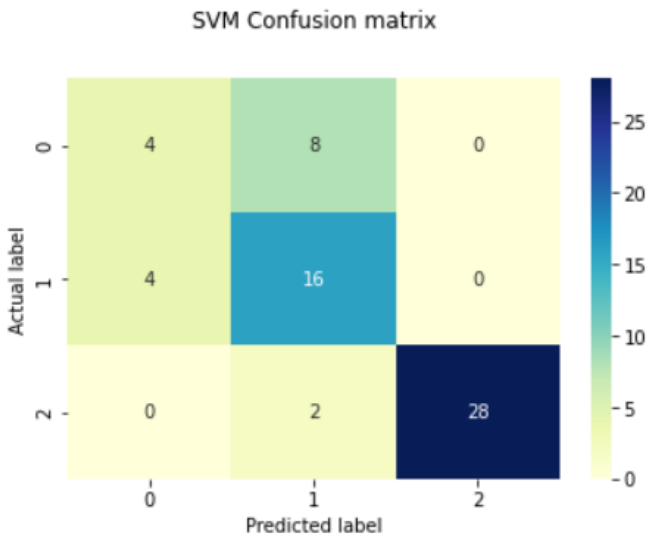
Table 2: Decision Tree Rules, Accuracy And Classification Report Results

```
Tree Rules:  |--- 5 <= 20.09
|   |--- 3 <= 28.14
|   |   |--- class: Hernia
|   |--- 3 >  28.14
|   |   |--- class: Normal
|--- 5 >  20.09
|   |--- 5 <= 31.25
|   |   |--- class: Spondylolisthesis
|   |--- 5 >  31.25
|   |   |--- class: Spondylolisthesis
```

Decision Trees Accuracy 0.7741935483870968
Decision Trees Classification Report

	precision	recall	f1-score	support
Hernia	0.50	0.33	0.40	12
Normal	0.62	0.80	0.70	20
Spondylolisthesis	1.00	0.93	0.97	30
accuracy			0.77	62
macro avg	0.71	0.69	0.69	62
weighted avg	0.78	0.77	0.77	62

Figure 6: Decision Tree Confusion Matrix Results



C. *Support Vector Machines*

Linear, radial basis function and polynomial kernels were tried. It has been seen that the best value for kernel is polynomial. This polynomial kernel is used for training and testing in the SVM model.

The results that have all kernels obtained are as follows:

Table 3: SVM All Kernels Results

```
SVM linear Kernel Accuracy: 0.8548387096774194
SVM rbf Kernel Accuracy: 0.8387096774193549
SVM poly Kernel Accuracy: 0.8709677419354839
```

The results that have polynomial kernel obtained are as follows:

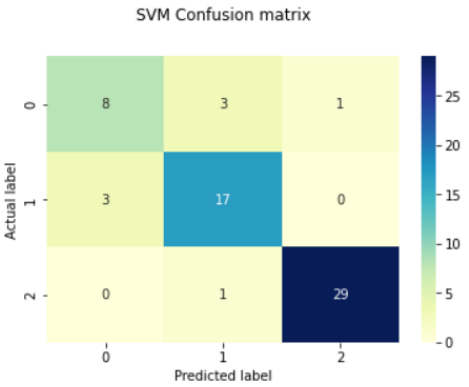
Table 4: SVM Accuracy And Classification Report Results

```
SVM Accuracy: 0.8709677419354839
SVM Classification Report
                  precision    recall  f1-score   support

          Hernia       0.73      0.67      0.70        12
          Normal       0.81      0.85      0.83        20
Spondylolisthesis      0.97      0.97      0.97        30

        accuracy                           0.87        62
       macro avg       0.83      0.83      0.83        62
    weighted avg       0.87      0.87      0.87        62
```

Figure 7: SVM Confusion Matrix Results



IV. Conclusions

In this project, the aim is to classify the patient status using orthopedic data. For this problem, 3 different classifier algorithm methods were used, and it was observed that the Support Vector Machines method gave the best results, considering the Accuracy and the confusion matrix.

References

[1] S. ALTUNAYDIN, “A Research on Machine Learning Methods and Its Applications”

[2] https://medium.com/@gulcanogundur/model-seçimi-k-fold-cross-validation-4635b61f143c

[3] https://developers.google.com/machine-learning/crash-course/classification/accuracy