Team - [Kanodiya, Hejib, Waikar, Singh, Patel]

Full Names - [Rishabh Kanodiya, Atharv Hejib,
Shreyas Waikar, Devansh Pratap Singh, Akash Patel]

EAS508 Project - 2

Conclusion and Analysis

Iris Dataset -

The Iris dataset consists of 150 samples of iris flowers, each belonging to one of the three species which are : setosa, versicolor and virginica.

There are four features for this dataset:
1) Sepal length
2) Sepal Width
3) Petal length
4) Petal Width

The dataset is organized into a 150*5 matrix, where the first four columns represent the features, and the fifth column represents the species label.

For classification purposes we have developed 5 different models that will classify these 150 observations into three different species as mentioned above.
Methods used to classify this dataset are:

1) Multinomial Logistic Regression
2) KNN (K- Nearest Neighbour) Classification
3) Decision Trees
4) Random Forests
5) SVM (Support Vector Machine)

To check error rate and prediction accuracy we have used k-fold cross validation with k set to 10.

We have also used confusion matrix and mean error rate for different models.

Below is one by one explanation for each model accuracy and error rate.

Multinomial Logistic Regression for IRIS Dataset

Introduction - We have built a multinomial logistic Regression model for the IRIS Dataset which contains 150 data points with values for 4 parameters namely - sepal length, sepal width, petal length and petal width for three classes namely - Setosa, Versicolor and Virginica.

Data Split - The total data consists of 150 observations, 50 each for each of the three classes.
We have split the data into training and testing sets, 70% data is used for training (105) observations while rest 30% is for testing (45) observations.

Model - We have built a Multinomial Logistic Regression model for the IRIS Dataset, this model will classify each of the 45 testing observations in their respective correct classes based on probability that it will be calculating for each class.
The class with highest probability for a particular data point will be assigned to it.

Model Evaluation - We have made predictions for the test data and calculated the misclassificatiion rate.
The misclassification rate is coming out to be 2.22 % for our model.

Model Evaluation via Cross Validation - We have used 10 fold cross validation to determine the error for our model. The mean accuracy is coming out to be 96.38 %, meaning the error rate coming out from 10 fold cross validation is 3.65%.

Conclusion - Our model performed exceptionally well on test data with 2.22% error rate and we have validated via cross validation also that it is working as expected as we are getting a mean cross validation error rate of 3.65 %. So, we may conclude that this multinomial logistic regression model is effective in predicting the iris dataset.

KNN Classification for IRIS Dataset

Introduction - We have built a Knn classification model for the IRIS Dataset which contains 150 data points with values for 4 parameters namely - sepal length, sepal width, petal length and petal width for three classes namely - Setosa, Versicolor and Virginica.

Data Split - The total data consists of 150 observations, 50 each for each of the three classes. We have split the data into training and testing sets, 70% data is used for training (105) observations while rest 30% is for testing (45) observations.

Model - We have built a Knn Classification model for the IRIS Dataset, this model will classify each of the 45 testing observations in their respective correct classes based on the k nearest neighbours for the particular data point. The category that the majority of the k neighbour belong to is assigned to that data point as well.

Model Evaluation - We have made predictions for the test data and calculated the misclassification rate. The misclassification rate is coming out to be 2.22 % for our model.

Model Evaluation via Cross Validation - We have used 10 fold cross validation to determine the error for our model. The mean accuracy is coming out to be 92.67 %, meaning the error rate coming out from 10 fold cross validation is 7.33%.

Conclusion - Our model performed pretty well on test data with 2.22% error rate and we have validated via cross validation. It is working as expected as we are getting a mean cross validation error rate of 7.33 %. So, we may conclude that this Knn classification model is effective in predicting the iris dataset.

Decision Tree for IRIS Dataset

Introduction - We have built a decision tree model for the IRIS Dataset which contains 150 data points with three classes namely - Setosa, Versicolor and Virginica based on petal length.

Data Split - The total data consists of 150 observations, 50 each for each of the three classes. We have split the data into training and testing sets, 70% data is used for training (105) observations while rest 30% is for testing (45) observations.

Model - We have built a decision tree model for the IRIS Dataset, this model will classify each of the 45 testing observations in their respective correct classes based on their petal length, for example if the petal length is less than 2.5 then the class will be setosa otherwise versicolor and it will be classified further.

Model Evaluation - We have made predictions for the test data and calculated the misclassificatiion rate. The misclassification rate is coming out to be 2.22 % for our model.

Model Evaluation via Cross Validation - We have used 10-fold cross validation to determine the error for our model. The error rate coming out from 10-fold cross validation is 4.44%.

Conclusion - Our model performed well on test data with 2.22% error rate and we have validated via cross validation also that it is working as expected as we are getting a mean cross validation error rate of 4.44 %. So, we may conclude that this decision tree model is effective in predicting the iris dataset.

Random Forest Classification for IRIS Dataset

Introduction - We have built a Random Forest Classification model for the IRIS Dataset which contains 150 data points with values for 4 parameters namely - sepal length, sepal width, petal length and petal width for three classes namely - Setosa, Versicolor and Virginica.

Data Split - The total data consists of 150 observations, 50 each for each of the three classes.
We have split the data into training and testing sets, 70% data is used for training (105) observations while rest 30% is for testing (45) observations.

Model - We have built a Random Forest Classification model for the IRIS Dataset, this model will classify each of the 45 testing observations in their respective correct classes. It classifies by building multiple decision trees during training and merges their predictions to improve accuracy and avoid overfitting.

Model Evaluation - We have made predictions for the test data and calculated the misclassification rate. The misclassification rate is coming out to be 2.22 % for our model.

Model Evaluation via Cross Validation - We have used 10 fold cross validation to determine the error for our model. Then we have taken the mean for the 10 folds error that we are getting.
So, the mean cross-validated misclassification error rate is 0.833 % for our model.

Conclusion - Our model performed exceptionally well on test data with 2.22% error rate and we have validated via cross validation also that it is working as expected as we are getting a mean cross validation error rate of 0.833 %. So, we may conclude that this Random Forest Classification model is effective in predicting the iris dataset.

Report on SVM Analysis of Iris Dataset

Introduction - This report outlines the implementation and evaluation of a Support Vector Machine (SVM) model using the Iris dataset. The Iris dataset is a classic dataset in the field of machine learning and statistics, consisting of 150 records of iris flowers, each with four features: sepal length, sepal width, petal length, and petal width. The dataset contains three classes of iris species.

Methodology

Libraries Used:

e1071: For SVM model implementation.

caret: For model training and evaluation.

Data Loading and Preprocessing:

The Iris dataset was loaded into the R environment.

The dataset was divided into features (X) and response (y). The features include the physical measurements, while the response is the species of the iris.

SVM Model Training:

An SVM model with a linear kernel was trained on the entire dataset.

The model was used to predict the species of the iris flowers in the dataset.

Model Evaluation:

The accuracy of the model was calculated on the training data (without cross-validation).

The misclassification rate was also determined for the training data.

Cross-Validation:

To evaluate the model's performance, 5-fold cross-validation was conducted.

The accuracy and misclassification rate from the cross-validation were calculated.

Results

Training Data Evaluation:

Estimated Training Accuracy (Without Cross-Validation): 96.67%

Estimated Training Misclassification Rate (Without Cross-Validation): 3.33%

Cross-Validation Results:

Estimated Test Accuracy from Cross Validation: 97.33%

Estimated Test Misclassification Rate from Cross Validation: 2.67%

Conclusion - The SVM model demonstrated high accuracy in classifying the species of iris flowers both on the training data and through cross-validation. The results indicate the effectiveness of the SVM algorithm with a linear kernel in handling multi-class classification problems, as seen with the Iris dataset.