SIV895
Special Module On Intelligent Info. Processing
Project Report

Submitted By -
Om Prakash - 2019MCS2567
Kishore Yadav - 2019MCS2656

Indian Institute of Technology Delhi, India
Department of Computer Science
Prof, P.K Kalara
June 24, 2020

1 Problem Statement

We have landed on mars and we are trying to colonize it. But the main problem we are facing is which plants we can grow at different geographical regions so that maximum number of plants would survive. Fortunately we have the data from earth to give us insights. Please help us in predicting which kind of plant we can grow at different locations depending on the environment.

2 Objective

1. Exploratory Data Analysis
2. Data Preprocessing
3. Data Engineering
4. Data Preparation for Predictive Modeling
5. Multiple Classification Model Predictions with hyperparameter tuning
6. Comparison of model using performance KPIs, Training and Testing Time
7. Final predictive model recommendation

3 Dataset

Following is the link to download the training and test dataset.
`https://drive.google.com/drive/folders/1bCp1zhz-qodXDQiMJRHUvodtHrf6VgG1?usp=sharing`

1. Data size (test and train) : (116203 * 13 and 464809 *13)
2. Target Type : Plant Type

4 Details of Dataset

It's a multiclass classification problem. Following are the classes to be predicted. Sevral different Machine learning models are trained with different paramters.

4.1 The Target Classes

- Assassin vine
- Ascomoid
- Basidirond
- Kelpie
- Myconid
- Hangman tree
- Dark tree

4.2 Frequency of Each Class

- Ascomoid : 169487
- Assassin vine : 226705
- Basidirond : 28488
- Dark tree : 2184
- Hangman tree : 7598
- Kelpie : 13931
- Myconid : 16416

4.3 Missing Values in Data

- In the **training Data**, 93033 values are missed for the attribute 'Shadow In Midday'
- In the **test Data**, 23079 values are missed for the attribute Shadow In Midday

5 Machine Learning Models

Following are the details of the multiple different Machine Learning models are trained. At the end the Model which performs best will be recommended.

5.1 Forest Classifier (with default parameters)

- Time taken to train the Model in minutes : 2.19
- Total samples : 116203
- Misclassified samples: 4546
- Accuracy on test data: 0.96

Note:- An acceptable accuracy on test data. The model performed very well,and took nearly 2 minutes to get trained.

5.2 Random Forest Classifier (with default parameters on scaled data)

- Time taken to train the Model in minutes : 3.76
- Total samples : 116203
- Misclassified samples: 112109
- Accuracy on test data: 0.04

Note:- The model performed **very badly**. It performed worst among all The models.

5.3 Logistic Regression (with default parameters and max iterations = 1000)

- Time taken to train the Model in minutes : 5.69
- Total samples : 116203
- Misclassified samples: 36647
- Accuracy on test data: 0.68

Note:- The classifier was run for maximum 1000 iterations, but it didnt converge for these many iterations.

5.4 Logistic Regression (with default parameters,normalized data and max iterations = 1000)

- Time taken to train the Model in minutes : 1.25
- Total samples : 116203
- Misclassified samples: 43399
- Accuracy on test data: 0.63

Note:- The classifier was run for maximum 1000 iterations, but it didnt converge for these many iterations. This model performed a little worse on normalized data.

5.5 Logistic Regression (with default parameters,scaled data and max iterations = 1000)

- Time taken to train the Model in minutes : 0.58
- Total samples : 116203
- Misclassified samples: 33792
- Accuracy on test data: 0.71

Note:- The classifier was run for maximum 1000 iterations, but it didnt converge for these many iterations. This model performed a little better on scaled data.

5.6 Logistic Regression (with default parameters,robust scaled data and max iterations = 1000)

- Time taken to train the Model in minutes : 3.01
- Total samples : 116203
- Misclassified samples: 33774
- Accuracy on test data: 0.71

Note:- The classifier was run for maximum 1000 iterations, but it didnt converge for these many iterations. This model performed a little better on robust scaled data, but no improvement over the model trained on the scaled data.

5.7 Support vector machines (C=0.01, penalty="l1", dual=False)

- Time taken to train the Model in minutes : 5.01
- Total samples : 116203
- Misclassified samples: 38921
- Accuracy on test data: 0.67

Note:- The classifier was run for default 1000 iterations, but it didnt converge for these many iterations.

5.8 Support vector machines (C=0.001, penalty="l1", dual=False)

- Time taken to train the Model in minutes : 4.68
- Total samples : 116203
- Misclassified samples: 39073
- Accuracy on test data: 0.66

Note:- The classifier was run for default 1000 iterations on best paramter C=0.001, but it didnt converge for these many iterations. The performance is not improved.

5.9 One v/s Rest Linear SVM (with default parameters)

- Time taken to train the Model in minutes : 5.01
- Total samples : 116203
- Misclassified samples: 74415
- Accuracy on test data: 0.36

Note:- The classifier was run for default 1000 iterations, but it didnt converge for these many iterations.

5.10 K-Nearest Neighbors

- Time taken to train the Model in minutes : 0.04
- Total samples : 116203
- Misclassified samples: 5141
- Accuracy on test data: 0.96

Note:- With best value of K=3, the Model KNN performed well with accuracy of 0.96.

5.11 Conclusion

The performances of different models may vary on a dataset. A machine learning model may perform well on a dataset while may perform badly on the other dataset. For this problem **K-Nearest Neighbor**s and **Random Forest Classifier** performed the best with 0.96 accuracy on test dataset, therefore, these are the recommended models for this problem.