Tree Species Classification with Tree Census in Santiago de Cali

(October 2020)

Orejuela Bolaños Juan David

***Abstract*— The tree census is a dataset that contains information related to all the trees in the city of Santiago de Cali, it is for that reason that their correct feeding is important. Implementing a system which automatically classifies tree species is the overall goal. In this paper the implementation of a model for solving this classification problem is executed and evaluated with different methods of Supervised Machine Learning. The objective is creating a system which works well on unseen data to guarantee good accuracy results. The tree species classification takes the tree measurements as input. Various models are trained with different classification methods such as K-Neighbors Neighbors, One Vs Rest, and Random Forest Classifier. The results with the models are evaluated to obtain different precision according to the pre-processing of the data.**

One of the models with 96% accuracy is implemented in a web application developed with Python, Flask, Pickle and Gunicorn in Heroku. With a POST query, the information of the tree measurements is sent to the web app with API Rest and it will return a response with the tree species.

***Index Terms*—Supervised Machine Learning, Tree Census, Tree Species Classification.**

I. Introduction

The census of urban trees in the municipality of Santiago de Cali, arose, in response to the need of the city, identified by DAGMA, to have a basic tool for the management of urban trees and was reflected in Agreement 0353 of October 31, 2013, by which the Urban Forestry Statute was adopted for the municipality and other provisions related to the regulation, regulation and promotion of comprehensive planned activities in the green areas of the city are issued.

The tree census project was formulated in two phases; the first was executed through agreement No. 095 of 2013 between the CVC and the Universidad Autónoma de Occidente, through which 9 communes were registered. The second phase corresponding to the execution of the 13 remaining communes, was executed through agreement N0 049 of 2014 between the CVC and the Autonomous University of the West. In 2015 CVC made an adjustment to the tree census, tagged with metallic plates 52,604 tree individuals, identified in phase 1 the need to be reviewed. This represents a large amount of data about the trees of the city that is currently stored in physical (paper) and virtual (Excel files) documents.

The digital version of the information is published on the open data website of the Colombian government.

II. Description of the Phenomenon, Process Modeled and Problem to be Addressed

Most public employees are through a limited-time contract, so the workers in these jobs are frequently changed (Usually the normal period is 4 years, which is the duration of a mayor's work period). The technician who visits and analyzes the trees is a job that suffers this problem. So most of the time, the technicians are not very specialized in the taxonomy of a tree, which generates a tree labeling problem, since the technician knows how to take measurements of a tree but does not know how to correctly identify its species.

The previous situation generates waste of time because after collecting the information, this information must be analyzed by employees who have experience in taxonomy and make the respective corrections of the trees. For example, this is a fragment (in Spanish originally) of the description of the tree census on the open data website of the Colombian government:

“1.224 changes were made due to misidentification of the census specimen, equivalent to 2.32% of the total number of individuals to identify. 63% of the changes were made on comunas 3, 21, 22 and 9 corresponding to misidentification, mostly, Acacias, Palmas, Chiminangos, Samanes, Guayacanes, among others”

In the previous paragraph you can see that a large number of species have to be corrected again.

III. Process of Obtaining or Generating the Dataset

The measurements are taken by technicians who visit the trees and check them. Below is a list with the data that are taken by the technicians:

Id_Arbol, Barrio, Comuna, Nombre Común, Nombre

Cientifico, Familia, Vegetación, Edad, Perimetro, Cobertura, Confinamiento, Emplazamiento, PB, PAP, DAP, PAPDEL, PAPGRUESO, Altura_Arbol, Diferencia, Diametro de Copa, Inclinación, Orientación, Fuste, Tallos, Raiz, Copa, Densidad_Copa, Vitalidad, Norte y Este.

Technicians also describe tree needs or problems such as pruning or special treatments for diseases, but for the classification of species these are unnecessary data, according to an expert asked. According to the project development time, 11 continuous variables and a discrete variable (Vegetation) were taken as part of the original dataset. Below is a list of the dataset variables:

Nombre_Comun, pb, pap, dap, dap2, papdel, papgrueso, altura_fuste, altura_arbol, diferencia, diametro_copa, tallos, vegetacion.

The tree census is published on the open data website of the Colombian government or on the website of the mayor of Santiago de Cali.

IV. DESCRIPTION OF THE MACHINE LEARNING PROBLEM

According to the situation described above, the problem presented is an incorrect CLASSIFICATION of the tree species. Therefore, it is a problem that will be addressed with some of the existing forms of data classification in Machine learning.

The model developed in this project must be able to classify trees between three selected species: Palma areca, Palma real de Cuba y Limon Swinglea. Only three species were chosen because they are the most frequent in the tree census and because it has around 632 species, which would imply the development of a very robust machine learning model in a short time of development.

V. DATA PRE-PROCESSING

A. *Experts Asked*

Two experts and DAGMA employees from the Ecosystem Conservation group were asked about the problem and how to address it in this project. The experts said that the variables related to the location, treatments or diseases of a tree would not be very important variables to classify it, because they are very generic data that could describe a type of tree A or a tree B as well. But experts said the tree's vitality could be an important characteristic for the dataset, because measurements such as the height of a young Areca palm could be confused with the height of a mature Guacimo tree. For this reason, the dataset was limited to healthy and mature trees.

B. *Data Cleaning*

The data from the open data website is in CSV format, these files were processed with Microsoft Excel. Typing errors such as: Blank spaces, misspelled species names, commas instead of dots for decimal numbers, and the correction of values in Vegetation variable for example Areca Palm is Palma and Limon Swinglea is a tree.

VI. TRAINING AND EVALUATION

For this project, the classification problem will be addressed with three Machine Learning models: K-Neighbors Classifier, One versus Rest and Random Forest Classifier.

The dataset also has a pre-processing such as Scaling and Dimensionality Reduction that were not explained in the previous section since they were not applied in all cases to have different results.

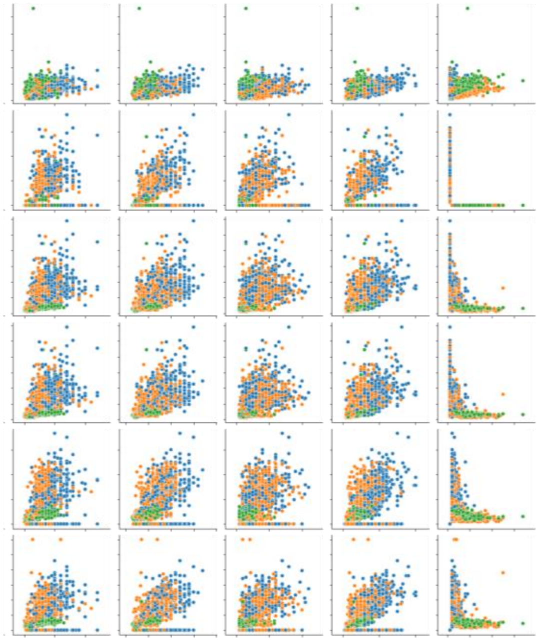


Figure 1. Dataset Pairplot

The previous image shows a pair plot between some of the characteristics of the dataset. It can be deduced that models like SVM could have more difficulties in learning and classifying information in this case.

In each case, a complexity evaluation was carried out, in Random Forest Classifier the model was trained with different values of estimators, and K-Neighbors was trained with different values of K. The dataset was divided into two parts, 70% of the data is for training and 30% of the data is for testing.

The results are shown below in the different cases:

A. *Case 1: Random Forest Classifier. Dataset only with Continuous Variables (Without "Vegetación" variable)*

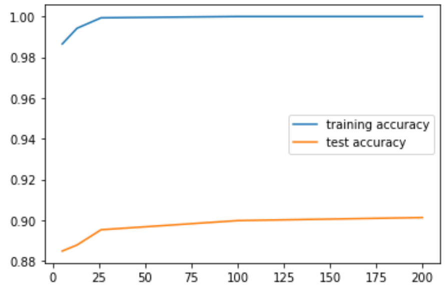


Figure2. RFC Model Accuracy Evaluation

In this case, the result with the best accuracy 90,13% was achieved with 26 estimators but obtaining an overfitting with the training dataset, which probably generates problems to generalize with this model. The dataset only contains continuous variables without the vegetation variable which is a discrete variable. The different estimators for this test were 13, 26 and 100 estimators.

B. *Case 2: Random Forest Classifier. Dataset with Continuous and Discrete Variables with One-Hot Encoding*

For this case, the discrete variable "Vegetación" was coded with the One-Hot method so that it could be used in the training of the Random Forest Classifier model. The values of the Vegetation variable are "Tree" or "Palm". With the coding, the dataset obtained two new columns which are: "Veg_Palma" with values 0 when the species is a tree and 1 when it is a palm; and "Veg_Arbol" with values 0 when the species is a palm and 1 when it is a tree. The different estimators for this test were 13, 26 and 100 estimators.

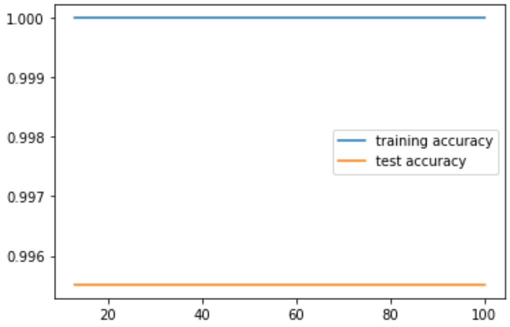


Figure 3. RFC Model Accuracy Evaluation with Discrete Variable

By including the discrete variable and by printing the importance of the characteristics in the tree according to the impurity, it can be seen that the discrete variables are more important than the continuous ones.

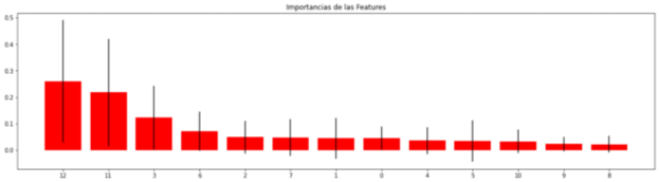


Figure 4. Feature Importance

C. *Case 3: Random Forest Classifier. Dataset with Continuous and Discrete Variables with Dimensionality Reduction PCA and Standard Scaler*

In this exercise, before using the dataset to train the model, it was scaled with the StandardScaler and dimensionality reduced with a PCA model, to capture most of the information in just two components.

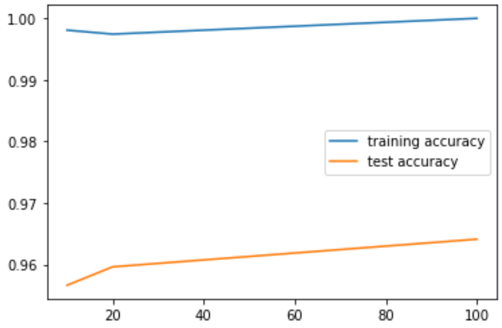


Figure 5. RFC Model Accuracy Evaluation with PCA reduction y standard scaling

With these results, it can be seen that the overtraining offered by the Random Forest Classifiers is reduced a little bit, and the accuracy of the model with the training dataset tends to increase.

D. *Case 4: K-Neighbors Classifier. Dataset with Continuous and Discrete Variables with Dimensionality Reduction PCA and Standard Scaler*

When noticing an improvement in the performance of the model with scaling and dimensionality reduction, it was decided to change the machine learning model for that of K-Nearest Neighbors. This Machine Learning method is used without thinking about the computational cost that could represent working with this model to classify the real number of tree species (625 species), its precision is the only metric to be analyzed for the classification of the three species that are being studied.

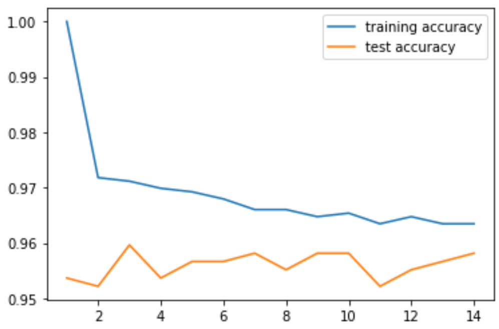
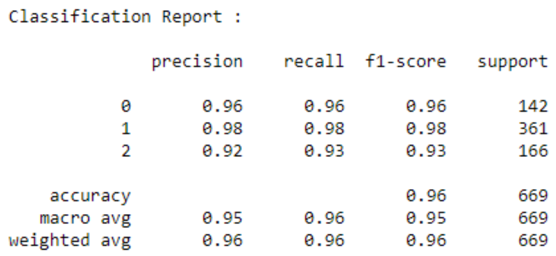


Figure 6. KNN Model Accuracy Evaluation with PCA reduction y standard scaling

In the previous graph, we can see the precision of the model trained under the KNN method, with a sweetspot in the 10 neighbors and with a 96% of precision with the test dataset and 96,7% with the training dataset. With 3 neighbors the model has greater precision but has greater flexibility (96,2%). The decrease in overfitting is an important point in this results as well.

E. Case 5: One Vs Rest. Dataset with Continuous and Discrete Variables with Dimensionality Reduction PCA and Standard Scaler

It was decided to work with the One Vs Rest method because despite the closeness between the data sets of the classes and that makes their separation by a plane or hyperplane difficult, it is very important to know the performance of a model with a reduced data (Its dimensionality). The results of the model's accuracy are shown below:



```
Classification Report :

              precision    recall  f1-score   support

           0       0.96      0.96      0.96       142
           1       0.98      0.98      0.98       361
           2       0.92      0.93      0.93       166

    accuracy                           0.96       669
   macro avg       0.95      0.96      0.95       669
weighted avg       0.96      0.96      0.96       669
```

Figure 7. One Vs Rest Model Accuracy Evaluation with PCA reduction y standard scaling

Obtaining 96% precision with this model, the reduction of dimensionality has facilitated the separation of classes with the use of hyperplanes.

VII. IMPLEMENTATION

A web application is built in Python with the Gunicorn library and a Rest API with Flask that allow the reception of the data corresponding to a new tree. The web application contains the model trained with the KNN method and that was packaged with the Pickle Python library for use in the web application without the need for training. The data can be sent with a POST Query to the web application in Heroku with the trained model. The Query has this information: pb, pap, dap, dap2, papdel, papgrueso, altura_fuste, altura_arbol, diferencia, diametro_copa, tallos, vegetación.

Below is a flow diagram of the training, implementation and operation of the machine learning model for the prediction of tree species according to their measurements:

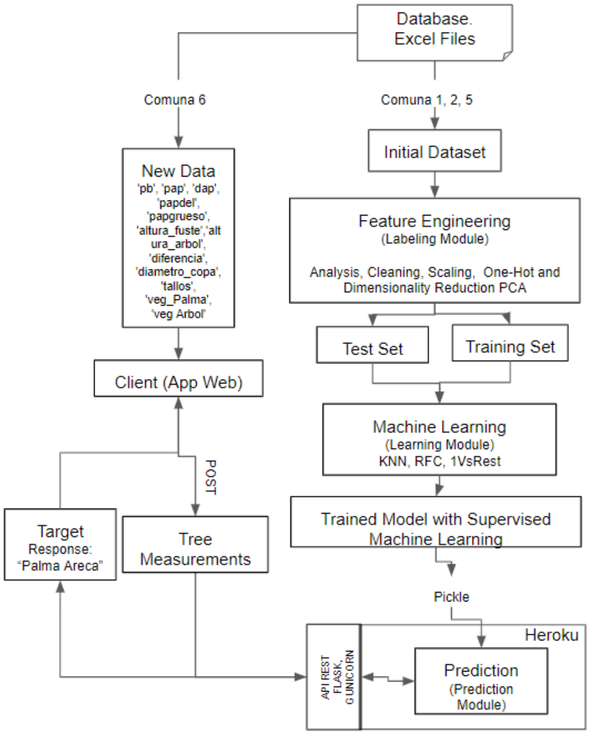


Figure 8. Diagram of Learning, Implementation and Operation of model.

VIII. CONCLUSION

A visual analysis of the datasets is important to be able to determine the models with which it is possible to work. Features of a discreet nature are very important when training models with Random Forest Classifier. Dimensionality reduction can aid in the separation and classification of data.

REFERENCES

[1] Alcaldía de Cali. (2019, March). Censo arbóreo de Santiago de Cali 2015. http://datos.cali.gov.co/dataset/censo-arboreo-de-santiago-de-cali

[2] Kaggle. (2020, June). Visualizing NYC Tree Census 2015 using Tableau. https://www.kaggle.com/vivekhn/visualizing-nyc-tree-census-2015-using-tableau

[3] Yiu T. (2019, June). Understanding Random Forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[4] Huneycutt J. (2018, May). Implementing a Random Forest Classification Model in Python. https://medium.com/@hjhuney/implementing-a-random-forest-classification-model-in-python-583891c99652

[5] Patel S. (2017, May). Chapter 5: Random Forest Classifier. https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1