Pattern Mining in Sentinel 2A Satellite Images Using KNIME Analytics Platform

Rudiney Soares PEREIRA, Elisiane ALBA, Juliana MARCHESAN, Mateus SCHUH, Roberta FANTINEL, Brazil

Key words: Data mining; Knime; satellite images, land use, land cover

SUMMARY

In this article we present a land use patterns and land cover mining tool designed to intelligently manage knowledge from Sentinel 2A series satellite image data. This tool uses integrated plugins in Knime Analytics Platform. The work was elaborated on Knime platform through the selection of configured and connected nodes and plugins constituting a workflow composing all the methodological phases in order to produce results of each of the process steps with the application of numerous multispectral image processing techniques such as: preprocessing activities (preparation of image data); image segmentation; application of digital filters; pattern classification; pattern mining and visualization. The input data consisted of hundreds of small multispectral images, color compositions, obtained by from 10 spectral bands with spatial resolutions of 10m and 20m from the MSI sensor aboard the Sentinel 2A satellite. This tool is expressed in the form of a workflow that contains each of the phases required for pattern mining, requires no knowledge of programming languages, and is based on the connection of plugins configurable according to the purpose of processing. Connected to each other, these plugins allow you to receive different configurations and defined the methodology workflow in the main phases: loading and viewing images; feature extraction which consisted of extracting from each image a non-redundant numerical vector that characterizes land use and land cover; the creation of attributes for each target (land use class and land cover; testing phase and predictive model evaluation. At the end of the processing, the patterns were extracted and these could be filtered using regular expressions based on the performance of the classifiers for the images. The algorithm that showed the highest performance was Random Forest when compared to Decision Tree. Thus, it is concluded that it is possible to do intelligent knowledge management.

Pattern Mining in Sentinel 2A Satellite Images Using the KNIME Analytics Platform

Rudiney Soares PEREIRA, Elisiane ALBA, Juliana MARCHESAN, Mateus SCHUH, Roberta FANTINEL, Brazil

1. INTRODUCTION

In a few decades, we have seen a significant increase in data collected by sensor instruments, opening enormous perspectives for treatment and analysis. For example in the context of climate change, monitoring of land use and land cover changes, studies of environmental impacts and disasters, agricultural activities, among others. Geographic information systems (GIS) are generally used for this. However, time-space data analysis, especially for large areas, is difficult for users to visually interpret. The number of possible interactions, particularly in data analysis, grows exponentially due to the complexity and size of the data collected (massive data, heterogeneity, inaccuracy, noise and multi-scalability).In order to contribute to the landscape studies for example, different software provide methods to analyze the landscape structure based on satellite images and remote sensing techniques [11]. If we still consider the possibility of working with time series data, we would need a very large set of data. Data mining offers solutions especially when it comes to finding patterns of phenomena and their evolution. Standard mining in TSSI (Temporal Series Satellite Images) was studied in [1] and [2]. The authors considered the images as a sequence of labeled pixels, and patterns extracted to find frequency of evolution. In [3], the authors developed the research considering a set of multispectral images. According to this author [3], a limitation of this method is due to not considering independence of the spatial dimension and the sequence of pixels in the processes. KNIME (Konstanz Information Miner) is a platform that allows integration, processing of data arranged in files arranged in the form of tables or images of different origins and formats, performing exploration, comprehensive and easy-to-use analysis. KNIME was developed using rigorous software engineering practices and has the development and support of more than 6,000 professionals worldwide, both in industry and academia. Because it is a modular data exploration platform that allows the user to create data flows visually (usually called pipelines), selectively perform some or all of the analysis steps and then investigate the results through interactive visualizations of data and models [6]. The purpose of mining image patterns is to extract valuable knowledge from image data. If we consider the supervised image classification process, what we want is to assign a label to images considering their visual content. This entire process is identical to the standard data mining process. We train a classifier from a set of previously classified images. Then, we can apply a new image to the classifier to process the categorization into classes. The peculiarity in this case is that we must extract a vector of numerical resources from the image before starting the machine learning of the classifier algorithm in the implementation phase. The theme is not new, however, easy access is recent. Two reasons are justified: first, the volume of images available on the web means that we have skilled statisticians and data scientists. In this case, the challenge is increasingly present, it is necessary to extract information from

images; the second reason is a variety of tools that are easy to use for data mining. A while ago, we needed a lot of computer programming. Today, there are efficient tools that allow complex data analysis without being an image processing expert. These tools include scikit-image, developed for high-level programming languages such as Python. The power of the data processing, analysis and exploration tools allows to achieve what is essential to extract from the data, optimizing the time necessary to explain in detail, the low-level structures of images. Although this knowledge becomes important, it is necessary to make parameter adjustments in our analyzes.

2. MATERIAL AND METHODS

2.1 Images dataset

The data set comprised a total of 1,370 files with 30 files per spectral band, sample fractions, in TIFF file format, with size 10 X 10 pixels of 10 spectral bands of images from the Sentinel 2A satellite (Figure 1). These fractions received labels differentiating samples that contained exclusively the forest category (label FN_) and fractions with other category, receiving the label NF_. All files were arranged in a folder so that they could be processed by the Knime Analytics platform with the "Knime Image Processing" module installed.

Figure 1

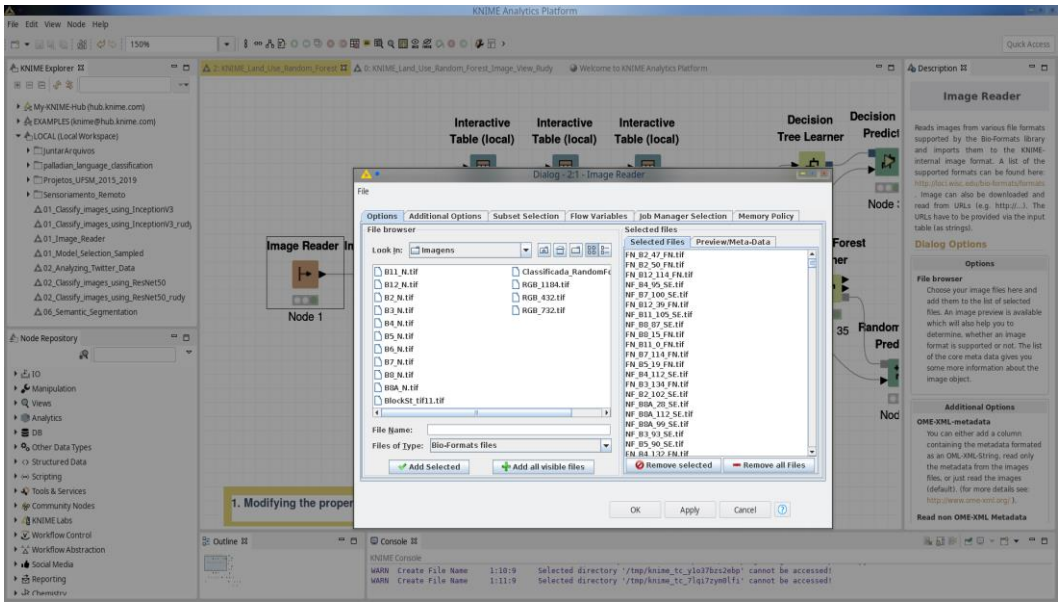

Figure 1 – Image files, the three first letters specify the class membership

2.2 Building the Workflow in Knime

The workflow is elaborated in Knime by placing nodes with their different functions and interconnected with each other to flow data processing. The construction of the flow consisted of grouping in three main steps to be mentioned: a) preparation of data such as reading, extracting characteristics and filtering data; b) data partitioning, machine learning in the decision tree and random forest algorithms and classification prediction; and c) performance analysis of the classification algorithms. In the Knime workflow editing panel, each of the nodes, interconnected, has at least and in general, two phases necessary to be fulfilled, the first deals with the configuration of the node and the second with the actual execution processing phase that allows you to advance in the different stages. Figure 2 presents a partial view of the workflow comprising the stages of preparation (reading, character extraction and filtering) and data partitioning.

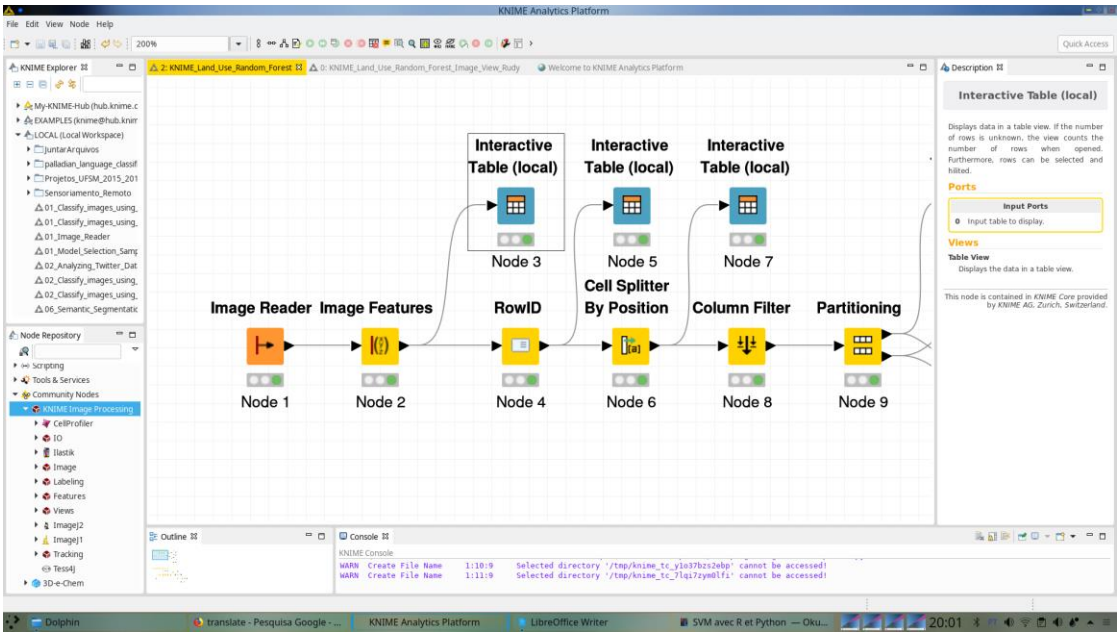

Figure 2 – Partial view of the workflow (data preparation and partitioning).

2.3 Feature extraction

The extraction of characteristics from the images was performed by the node called "Image Features" and consisted of configuring two extraction options. The first option used the parameters of the first order statistic: values of minimum, maximum, average, geometric mean, standard deviation, variance among others and, option two, comprised parameters proposed by Haralick [6] such as: contrast, correlation, variance, entropy among others, and the diagonal, antidiagonal, horizontal and vertical matrices. Figure 3 presents in table form a

partial view of the results of the processing of the extraction of the characteristics of the images according to the options chosen in the configuration of the "Image Features" node.

2.4 Creating the target attribute and learning and evaluating the predictive model

In the data grid showing the characteristics extracted from the images (Figure 3), the "Row ID" column represents the name of the data files. These are the sample images used in the extraction of the characteristics and these are categorized using the first three characters of the name of these files. The nodes of these procedures are called RowID and Cell Splitter By Position, respectively. In the procedure of learning and predictive evaluation of the model, unnecessary columns such as ID and Other were removed. The node used for this procedure is called Column Filter and precedes the Partitioning node. Figure 4 shows a partial view of the connections between these nodes.

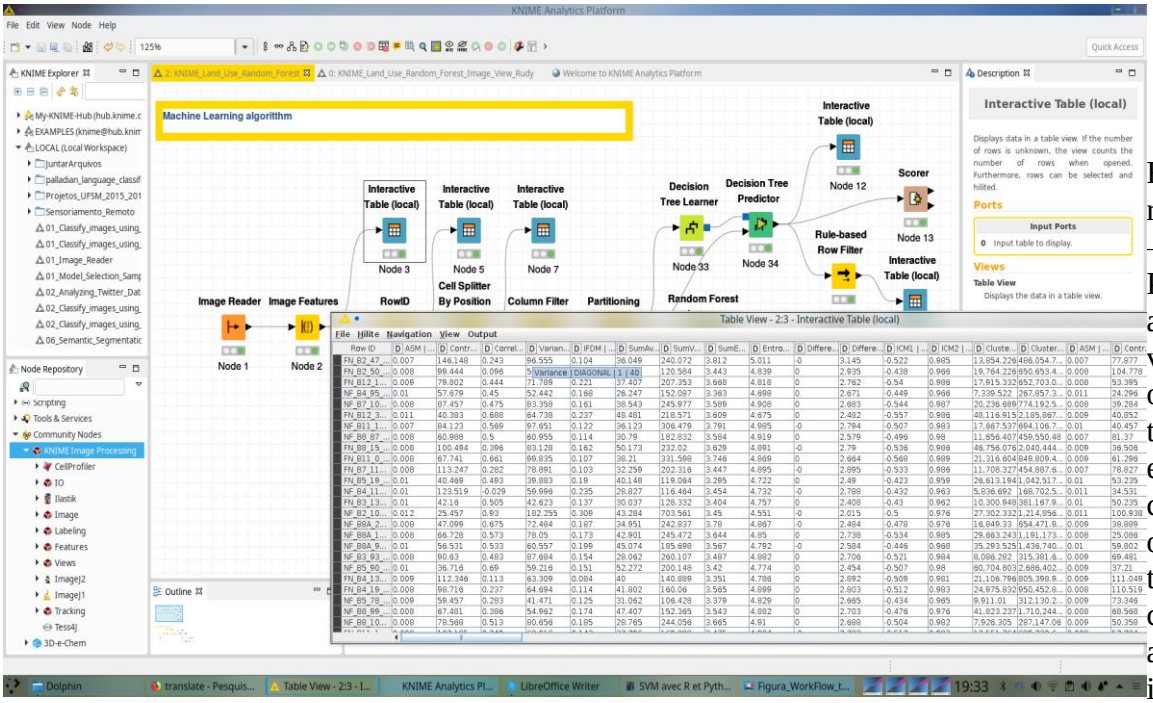

Figure 3 – Partial view of the extraction of the characteristics of the images.

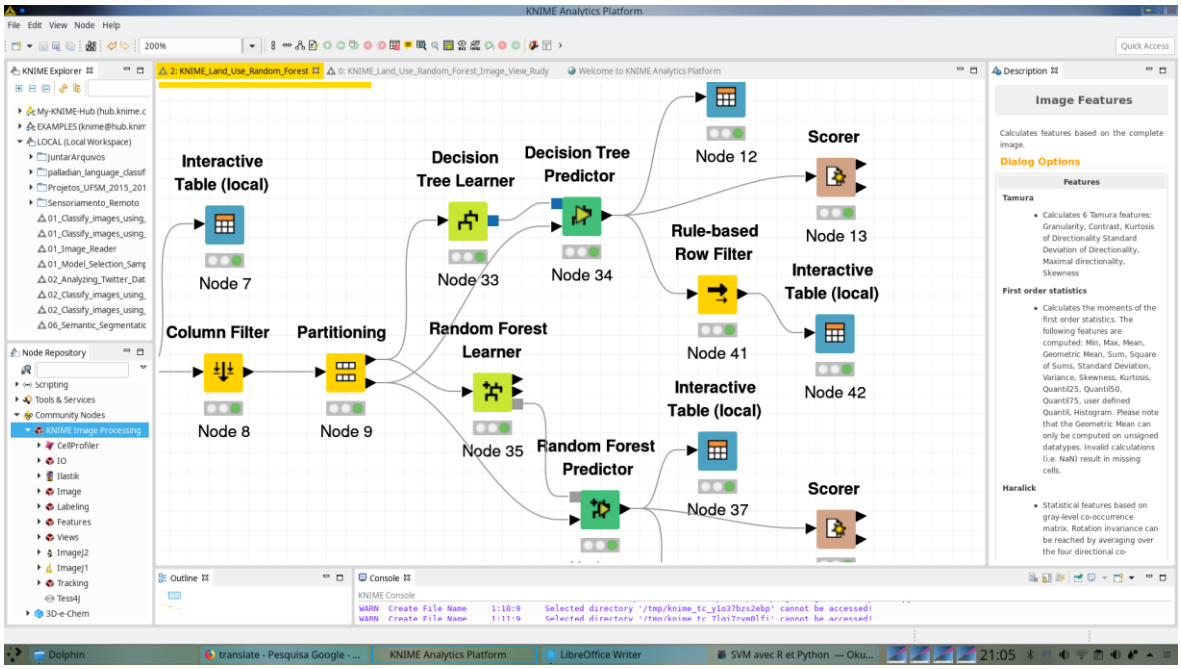

Figure 4 – Partial view of the connections between Column Filter and Partitioning node.

The image data set was partitioned into two other sets, training and testing. In this step, the Partitioning node was configured with 70% of training data with stratified sampling. The adjustment of the model in the algorithm, in the Decision Tree Learner node, followed the definition of the Forest Class and the quality of the measurement, used the Gini Index with a mean cutoff point. In the Random Forest algorithm, the Random Forest Learner node was configured for the cutoff point using the Information Gain Ratio parameter.

3. RESULTS AND DISCUSSION

The Figure 5 presents a partial view of the connections of the two classification algorithms (Decision Tree and Random Forest) both, connected to the antecedent Partitioning node and in sequence, respectively to the Decision Tree Predictor and Random Forest Predictor nodes.

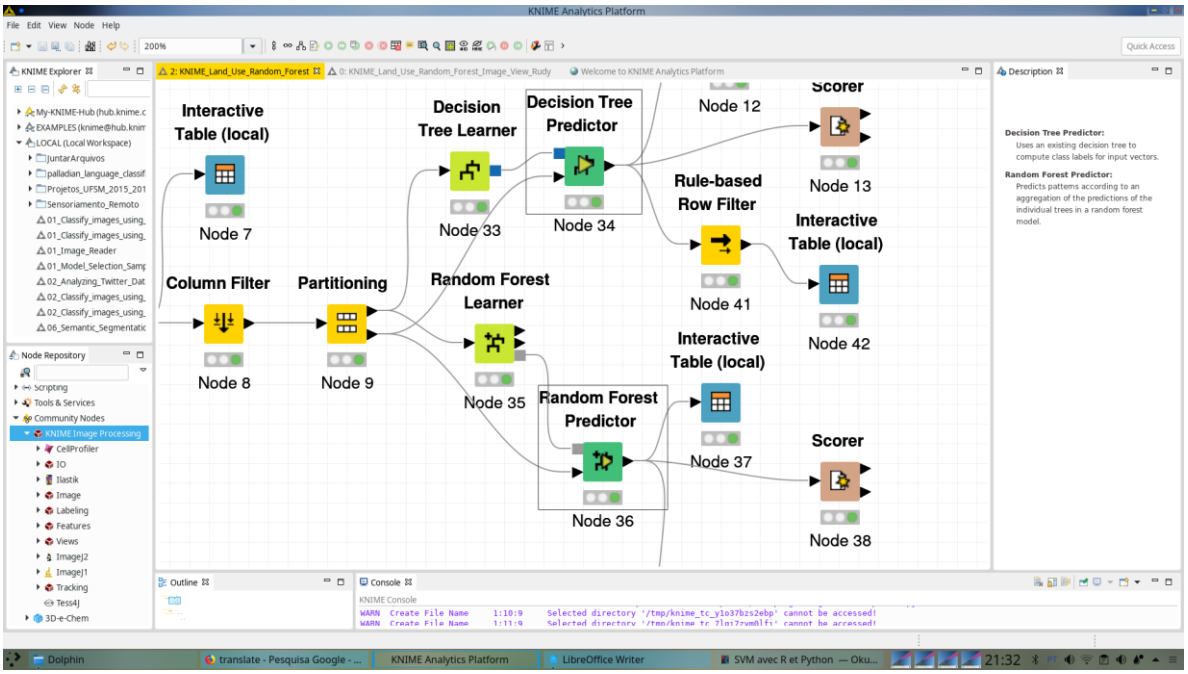

Figure 5 – Partial view of the connections Decision Tree and Random Forest

The workflow area for comprising nodes 34, 36, 12, 41, 37, 44, 13, 42, 38 and 43 makes up the production area of the final classification results by the Decision Tree and Random Forest algorithms with the production of performance analysis of these classifiers. In the Decision Tree algorithm, the node produces two information to analyze the classifier's performance, the first consists of a graphical information in the form of a tree in which it presents the weights and percentages of each of the parameters, used in the characterization process of data image. The Figure 6 shows this condition mentioned. The second possible information is represented by Figure 7 which shows the performance values of the classifier such as: samples classified correctly, classification errors and accuracy. The Decision Tree classification algorithm had an accuracy of 87.778% accuracy, an error (incorrectly classified samples) of 12.22%, with the Cohen's Kappa k = 0.756 index, considered satisfactory.

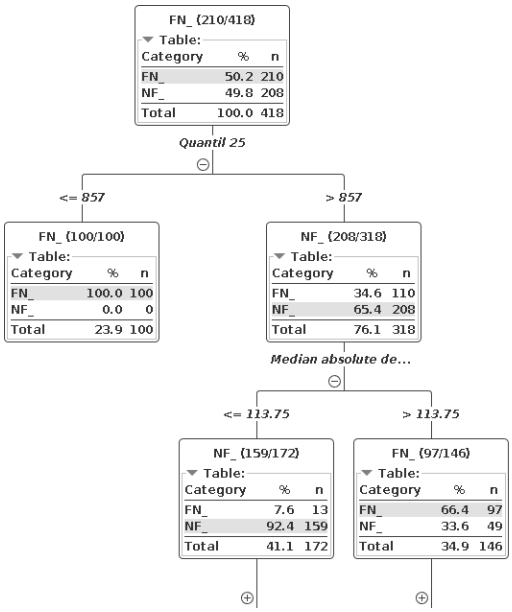

Figure 6 – Partial view of the decision tree produced by the Decision Tree algorithm

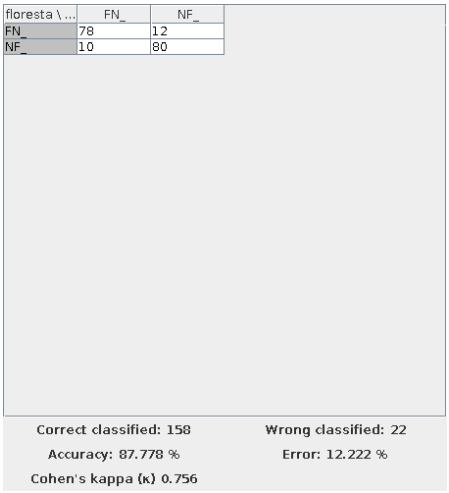

Figure 7 - Performance of the Decision Tree Predictor algorithm - Confusion matrix

The Random Forest classification algorithm produced the performance results as shown in Figure 8 which shows the performance values of the classifier such as: samples classified correctly, classification errors and accuracy. The Random Forest classification algorithm had an accuracy of 93.333% accuracy, an error (incorrectly classified samples) of 6.667%, with the Cohen's Kappa k = 0.867 index, considered satisfactory.

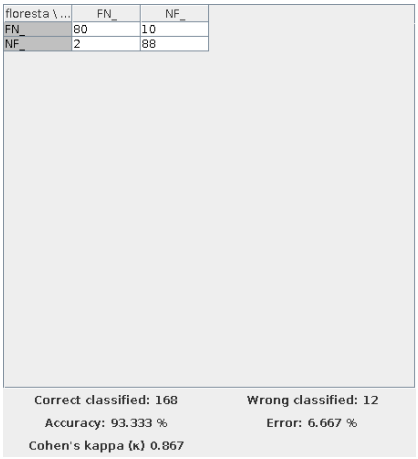

Figure 8 - Performance of the Random Forest - Confusion matrix

The Random Forest algorithm showed the highest accuracy when compared to the performance of the Decision Tree algorithm. In this study, then, the choice for Random Forest produced the most precision in the classification process of forest samples. The Figure 9 shows the general workflow for conducting the work. It is possible to observe that all the processes involved can be simplified using the logic of this solution and the necessary configurations in each of the nodes connected to each other. These nodes composed the workflow allowing the performance comparison of the two algorithms used Decision Tree Predictor and Random Forest. This workflow can be exported in "knwf" format and, later, be used by any user who intends to apply it with other image data.

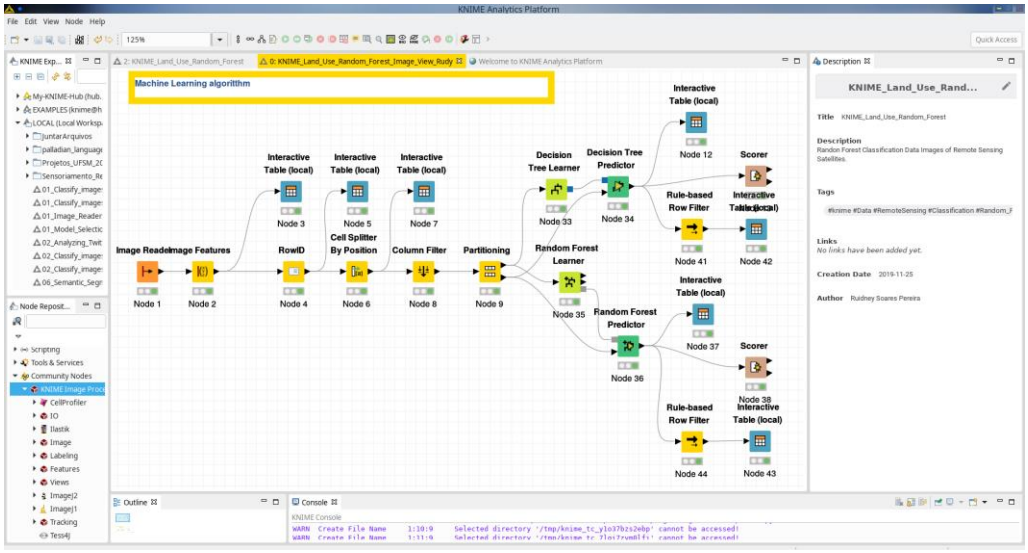

Figure 9 - Overview of Workflow Knime for image classification

4. CONCLUSION

The Knime Analytics platform appears as a high-performance tool for complex analyzes without requiring a single line of code with a programming language, although it is possible to introduce specialized scripts in the workflow to meet the purpose of the analysis. The flow model used allows it to be improved since it is possible to export it, edit it and adapt it to the interests of each user with this, characterized by its flexibility of use. The analysis of the performance of the Decision Tree and Random Forest algorithms allowed us to conclude that it is possible to have images classified with the necessary accuracy. Random Forest was the one that presented the best performance in the classification of images with the target of interest Forests. Thus, it is concluded that it is possible to do intelligent knowledge management.

REFERENCES

1. Julea N.A. Méger E. Trouvé P. Bolon "On extracting evolutions from satellite image time series" IGARSS 2008 pp. 228-231 July 8–11 2008.

2. Julea N.A. Méger C. Rigotti E. Trouvé P. Bolon V. Lazarescu "Mining pixel evolutions in satellite image time series for agricultural monitoring" 11th Industrial Conference ICDM 2011 pp. 189-203 August 30-September 3 2011.

3. E. Christophe J. Inglada "Open source remote sensing: Increasing the usability of cutting-edge algorithms" IEEE Geoscience and Remote Sensing Newsletter vol. 35 no. 5 pp. 9-15 2009.

4. F. Petitjean P. Gançarski F. Masseglia G. Forestier "Analysing Satellite Image Time Series by Means of Pattern Mining" in Springer Berlin Heidelberg pp. 45-52 2010.

5. Sanhes J. F. Flouvat C. Pasquier N. Selmaoui-Folcher J. Boulicaut "Weighted path as a condensed pattern in a single attributed DAG" in IJCAI 2013 Beijing China pp. 1642-1648 August 2013.

6. KNIME (Konstanz Information Miner), Available at: http://www.knime.org/, (Accessed 12 December 2019).

7. Cerf J. L. Besson C. Robardet J.-F. Boulicaut "Data-Peeler: Constraint-Based Closed Pattern Mining in n-ary Relations" SIAM vol. Proc. SIAM pp. 37-48 2008.

8. Collin M., F. Flouvat and N. Selmaoui-Folcher, "PaTSI: Pattern Mining of Time Series of Satellite Images in Knime," *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, Barcelona, 2016, pp. 1292-1295. doi: 10.1109/ICDMW.2016.0187

9. Berthold M. R., N. Cebron F. Dill T. R. Gabriel T. Kötter T. Meinl P. Ohl C. Sieb K. Thiel B. Wiswedel "KNIME: The Konstanz Information Miner" Studies in Classification Data Analysis and Knowledge Organization (GfKL'07) 2007.

10. Selmaoui-Folcher N. F. Flouvat C. Mu J. Sanhes J. Boulicaut "Extraction complète efficace de chemins pondérés dans un a-dag" in EGC 2015 27–30 Janvier 2015 Luxembourg pp. 179-190 2015.

11. Zaragozí, B.; Belda, A.; Linares, J.; Martínez-Pérez, J.E.; Navarro, J.T.; Esparza, J. 2012. A free and open source programming library for landscape metrics calculations. Environmental Modelling & Software. v. 31, p. 131 – 141.

CONTACTS

Dr. Pereira, Rudiney
Federal University of Santa Maria, Brazil
Professor
Av. Roraima, 1000 Campus Universitário, UFSM-CCR
Santa Maria
97105-900
Rio Grande do Sul State
Brazil
Tel. +55 55 3220-9468
Fax +55 55 3220-8261
Email rudiney.s.pereira@ufsm.br
Web site: http://www.ufsm.br/labsere