# **Task Description**

Hello,

As CTO and head of Blackwell's eCommerce Team, I'd like to welcome you aboard. I'm excited to get started on this project, but I'd first like to give you a bit of background to get you up to speed. Blackwell has been a successful electronics retailer for over 40 years, with over 30 stores in the Southeast. A little over a year ago we launched our eCommerce website. We are starting to build up customer transaction data from the site and we want to leverage this data to inform our decisions about site-related activities, like online marketing, enhancements to the site and so on, in order to continue to maximize the amount of revenue we generate from eCommerce sales.

For example, our VP of Sales, Martin Goodrich, thinks that customers who shop in the store are older than customers who shop online and that older people spend more money on electronics than younger people. He is considering some marketing activities and potentially some design changes to the website to attract older buyers. I, on the other hand, believe that the differences in transactions and customer demographics may be regional. Before we even consider any additional activities related to the website, we want to gain insight into any factors that can explain how our customers shop and how much they spend.

To that end, I would like you to explore the customer transaction data we have collected from recent online and in-store sales and see if you can infer any insights about customer purchasing behavior. Specifically, I am interested in the following:

1. Do customers in different regions spend more per transaction? Which regions spend the most/least?
2. Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?
3. We need to investigate Martin’s hypothesis: Is there any correlation between age of a customer and if the transaction was made online or in the store? Do any other factors predict if a customer will buy online or in our stores?
4. Finally, is there a relationship between number of items purchased and amount spent?

To investigate this, I’d like you to use data mining methods to explore the data, look for patterns in the data and draw conclusions. I have attached a data file of customer transactions; it includes some information about the customer who made the transaction, as well as the amount of the transaction, and how many items were purchased. Once you have completed your analysis, please create a brief report of your findings and conclusions and an explanation of how you arrived at those conclusions so I can discuss them with Martin.

# **Task Solution**

A data set (Blackwell\_Hist\_Sample.csv) provided by the CTO, a quick review to ensure data is tidy and clean was conducted, no missing values were found in the data set. Based on this outcome, the data can be used for analysis.

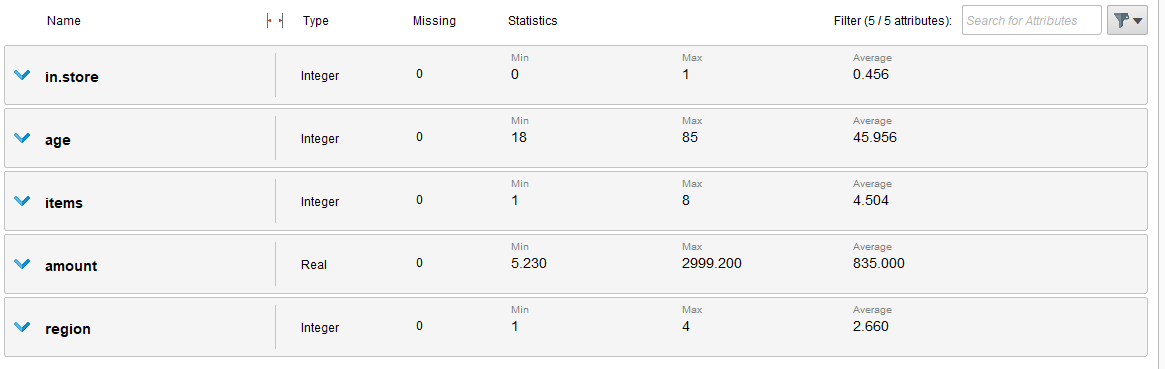


Figure 1 - Data Set Statistics

**Q1 - Do customers in different regions spend more per transaction? Which regions spend the most/least?**

Yes, the region 4 (Central) has the higher spend and region 2 (West) showed the lowest spend, based on the results listed in figure 2 and 3.



Figure 2 – Relationship between regions and the amount per transaction

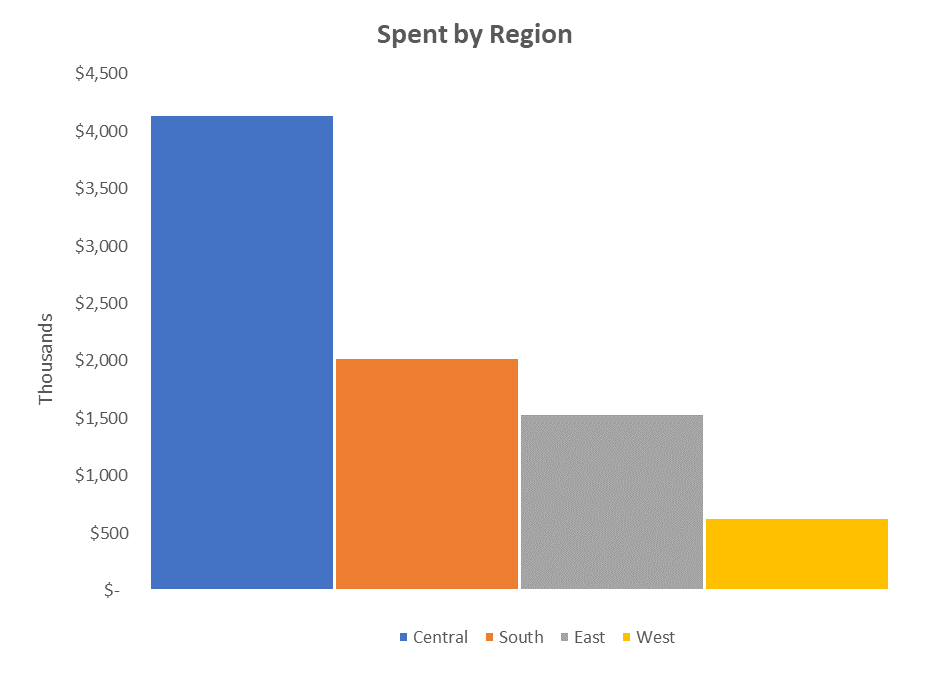


Figure 3 – The region with the highest spend is Region 4 (Central), with a spent over $4 MM

**Q2 - Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?**

Yes, there are differences in the age of customers between regions, as is show in figure 4. There are four ranges:

|  |
| --- |
| **Age Range (years)** |
| 18 - 35 |
| 35 - 52 |
| 52 - 68 |
| 68 - 85 |

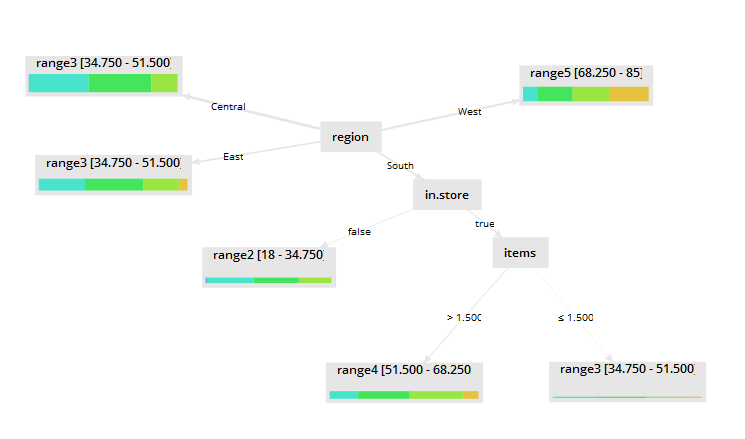
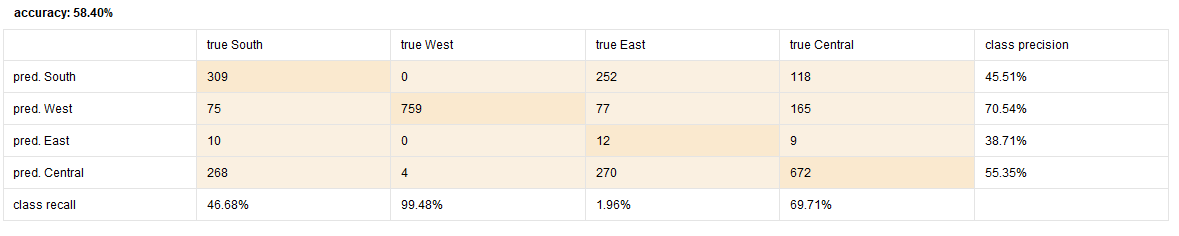


Figure 4 – Difference in the age of customers

This model is not a reliable model to predict the age of customers, since 1 out of 4 regions a low error margin (Accuracy 58,40%), the other 3 regions has accuracy values lower than 30% in one instance is below 2%, as is show in figure 5.



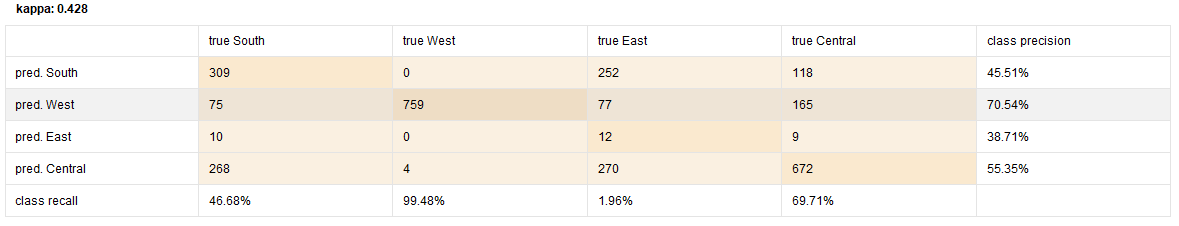
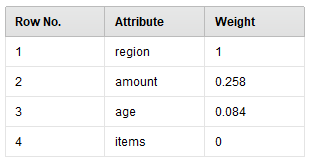


Figure 5 – Accuracy and Kappa values for a trained model

**Q3 - We need to investigate Martin’s hypothesis: Is there any correlation between age of a customer and if the transaction was made online or in the store? Do any other factors predict if a customer will buy online or in our stores?**

There is no correlation between customer’s age and in.store, since the correlation matrix ran for this model, indicates a value of -0.176 which is considered low, based on rapidminer references (Figure 6).



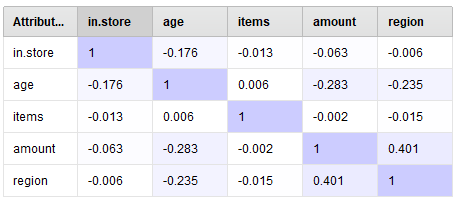
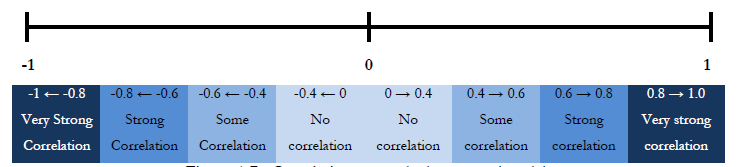


Figure 6 – Correlation matrix and Weight by information (Label = in.store)

There are other two factors with some degree of correlation with the in.store attribute, region and amount. In the correlation matrix, the bond between Amount – Region is the strongest relationship and using the weight by information gain operator, the weight of attributes with respect to the class attribute by using the information gain, shows that Region and Amount have the highest weights.

This can be easily interpreted in the decision tree listed below (Figure 7). Using the in.store attribute as Label, the Regions and the amounts are probability nodes.

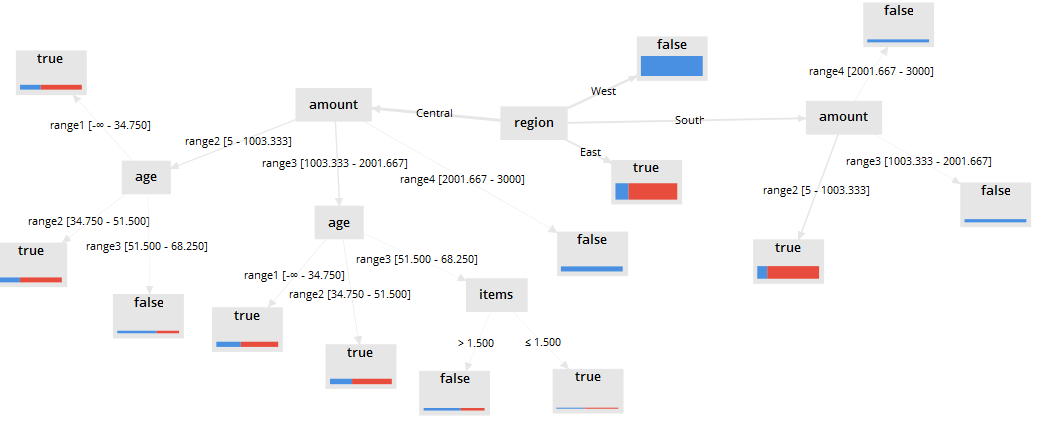


Figure 7 – Decision tree using in.store as label.

**Q4 - Finally, is there a relationship between number of items purchased and amount spent?**

The relationship between items purchased and amount spent is not clear or at least not obvious in this model, since correlation matrix shows a low relationship -0.002 and also the weight of attributes with respect to the class attribute by using the information gain, shows a weight of zero for the items attribute, regardless the label attribute.

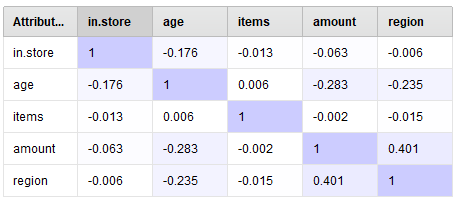
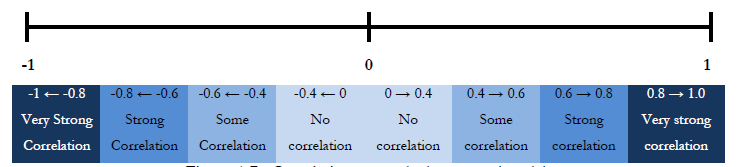




Figure 8 – Correlation matrix and Weight by information (Label = Region)

**Results Discussion**

The model developed for this task, helps to identify a difference in the spent within regions, age ranges and purchases made online or in stores.

The model has some limitation in terms of predictability for the customer’s age between regions and purchased items vs. spent. With the current data set or these correlations are not obvious, a different data set shall be tested, since different algorithms and model optimization were conducted.

The model accuracy was not significant between the different algorithms (decision Tree = 59,37%, Random Forrest = 59,37% and Gradient Boosted Trees = 59,41), an optimized version of the decision tree algorithm was ran with results of 59,57%. This give us some clues about the data set, since the model optimization and algorithm variation didn’t provide higher improvements.

**Models**

The figures listed below are proof of the models used to determine the case responses.

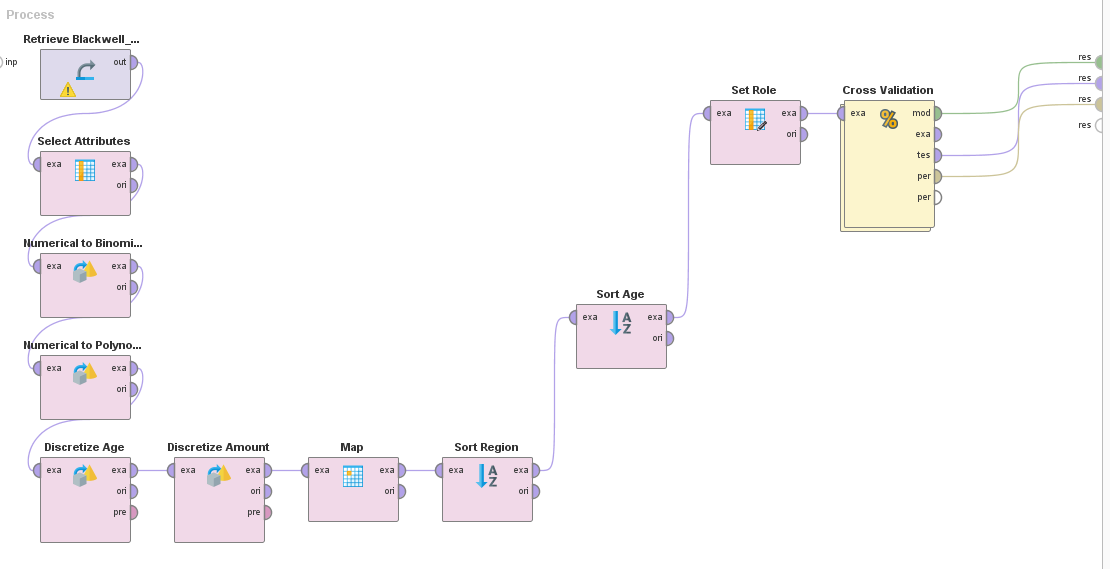


Figure 9 – Decision Tree Algorithm

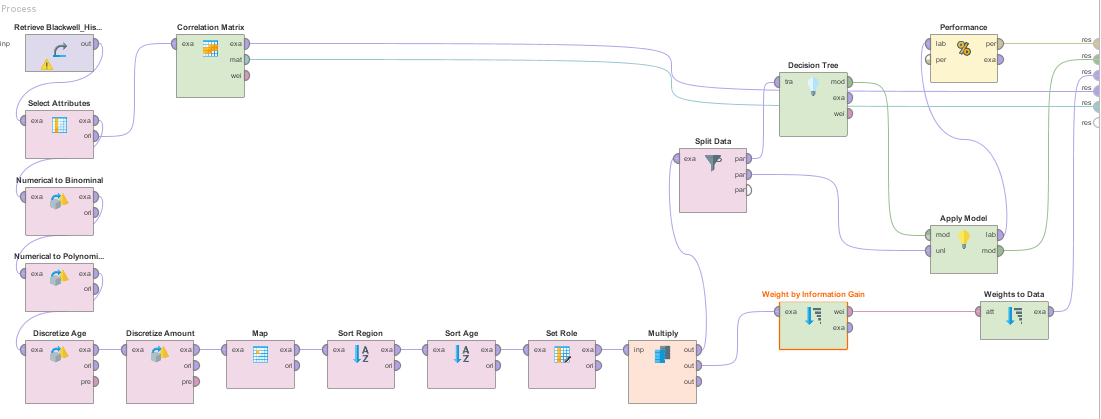


Figure 10 – Decision Tree Algorithm Trained model (70% training data / 30% unlabeled data)

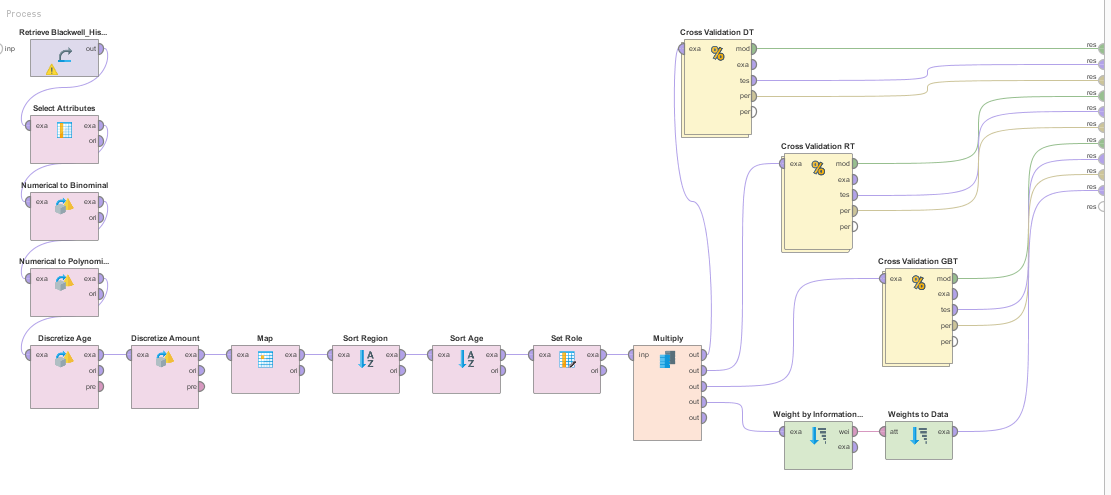


Figure 11 – Decision Tree, Random Forest, Gradient Boosted Trees algorithms

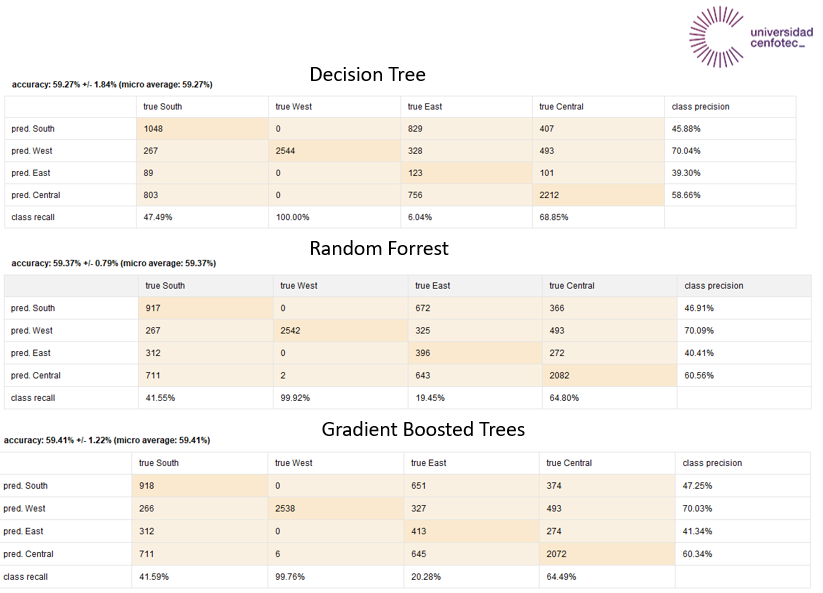
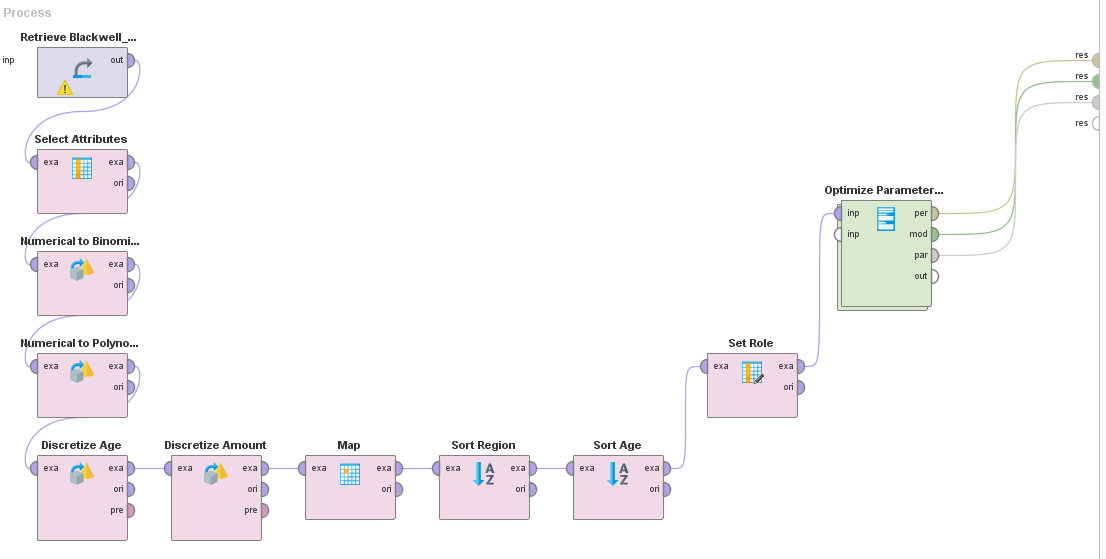


Figure 12 – Accuracy comparison between 3 different algorithms

**Optimized model – Using Label = Region**



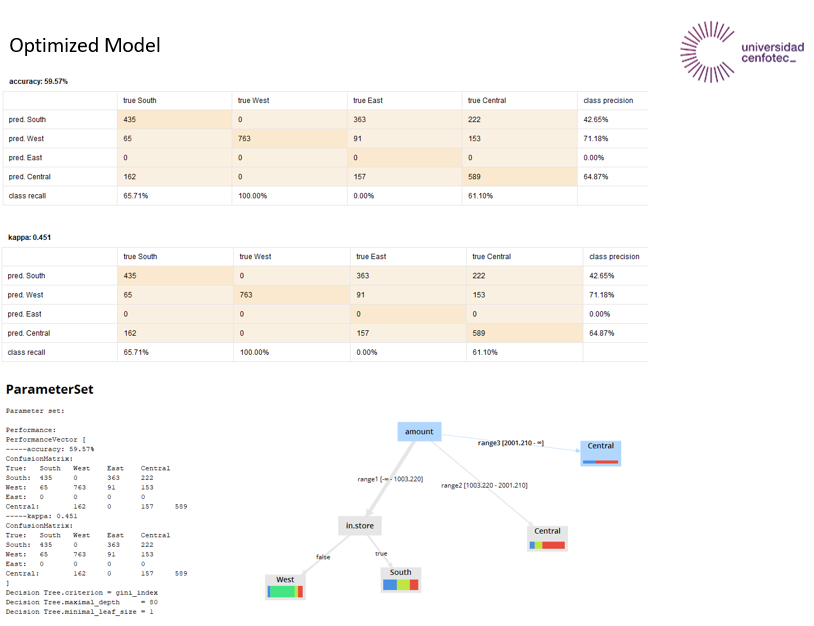


Figure 13 – Decision tree optimized model