IPM Decisions Data Exploration and Modelling

# Preprocessing

## Added reorder filter (v1.2)

Class: Already\_used\_DSS

Yes 1

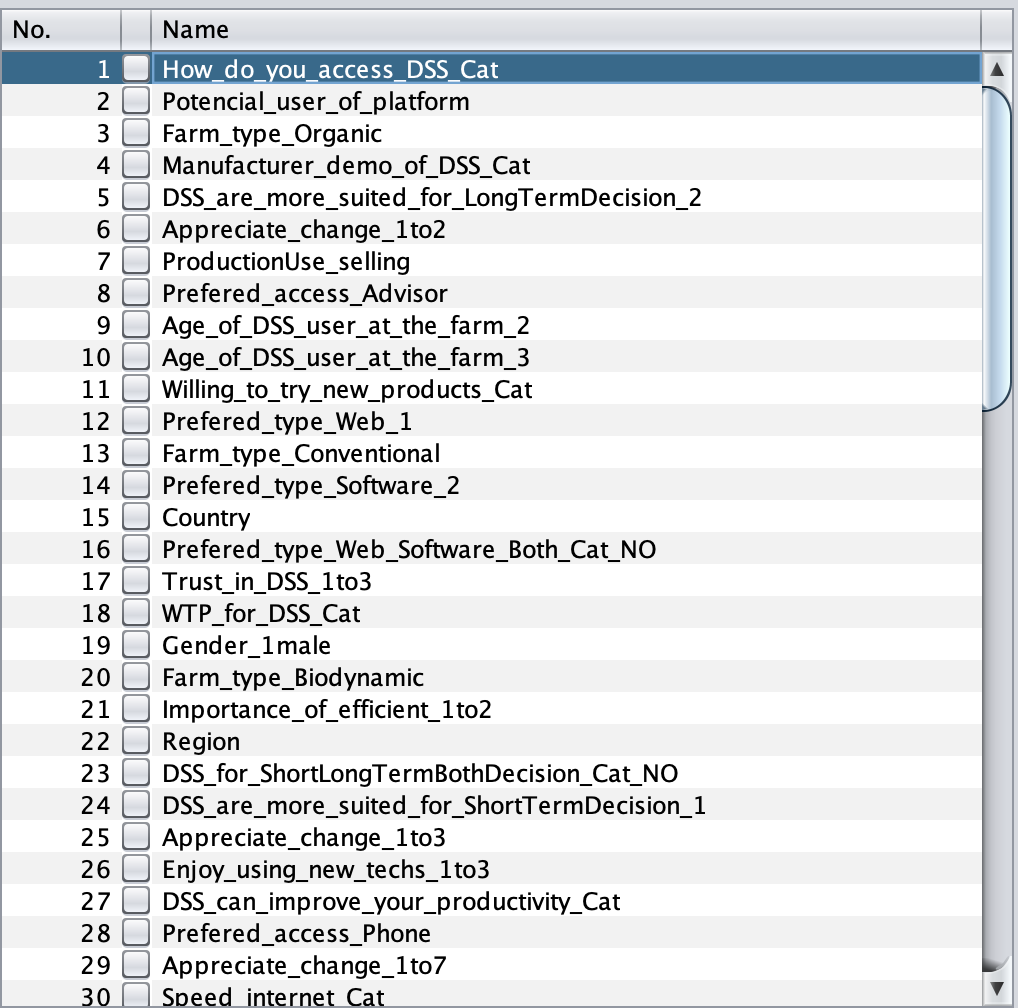
No 0

## Target (Class) missing for Greece

* Left for now.
* What to do?
  + S1: remove all Greece instances?

## Feature ranking (Filter) (v1.3)

* Applied filter: Attribute selection (CorrelationAttributeEva)



* Could not use infoGain due to missing values in class

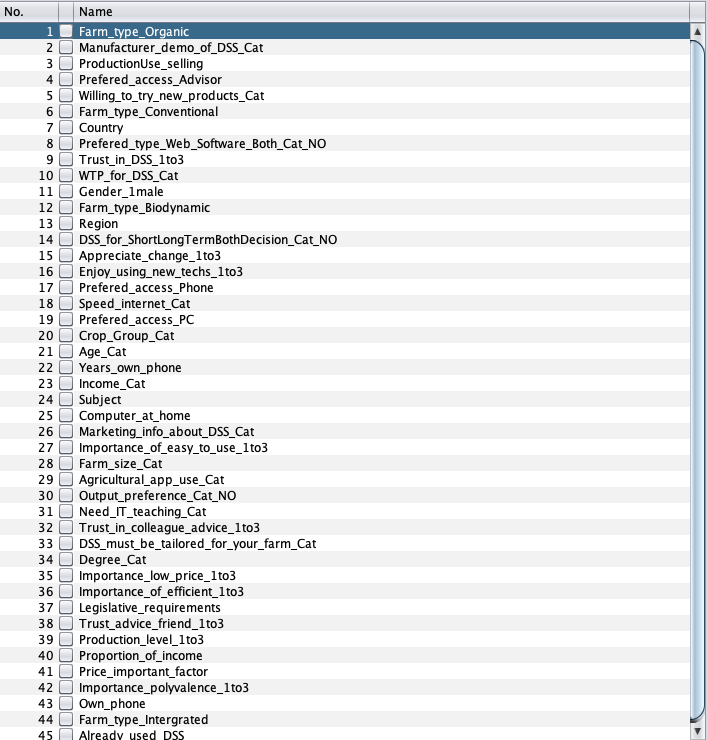
## Remove attributes (see excel)

* + Total 81
  + 10 attributes removed,  **directly related to target variable** (Total remaining 71)
  + 26 duplicate attributes removed (Total remaining 45)

*ver: farmers\_removed\_attr\_29042021.arff*

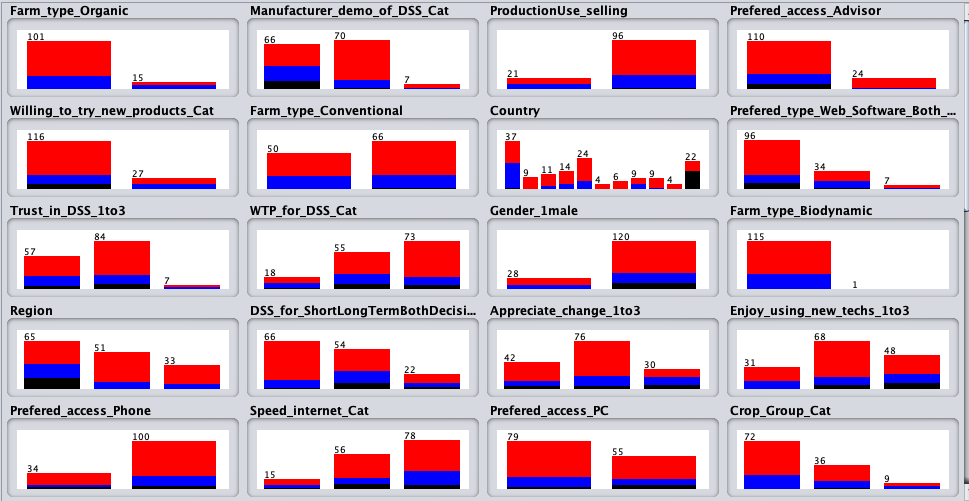
## Feature ranking (Filter) (v1.3)

* Applied filter: Attribute selection (CorrelationAttributeEva)



*ver: farmers\_removed\_attr\_29042021\_1.arff*

## Attribute visualisation



## Feature ranking (Select attributes)

* InfoGain:

Farm\_type\_Organic

Manufacturer\_demo\_of\_DSS\_Cat

ProductionUse\_selling

Prefered\_access\_Advisor

Willing\_to\_try\_new\_products\_Cat

Farm\_type\_Conventional

Country

Prefered\_type\_Web\_Software\_Both\_Cat\_NO

Trust\_in\_DSS\_1to3

WTP\_for\_DSS\_Cat

Gender\_1male

Farm\_type\_Biodynamic

Region

DSS\_for\_ShortLongTermBothDecision\_Cat\_NO

Appreciate\_change\_1to3

Enjoy\_using\_new\_techs\_1to3

Prefered\_access\_Phone

Speed\_internet\_Cat

Prefered\_access\_PC

Crop\_Group\_Cat

Age\_Cat

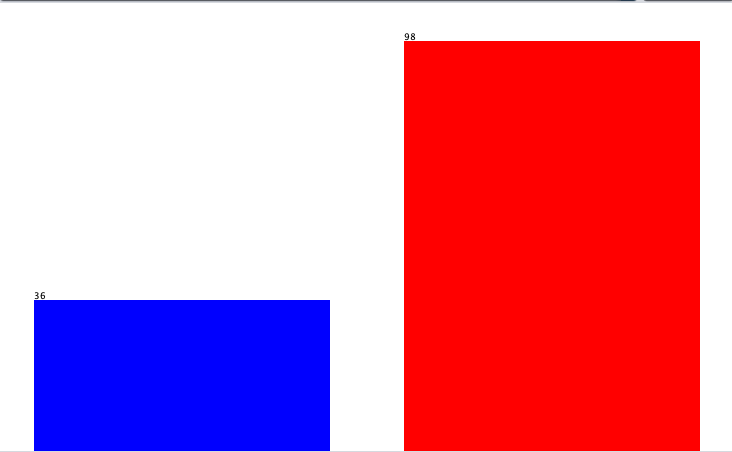
Years\_own\_phone

Income\_Cat

Old infoGain:

* + Farm\_type\_Organic
  + Manufacturer\_demo\_of\_DSS\_Cat
  + Prefered\_access\_Advisor
  + ProductionUse\_selling
  + Region
  + Willing\_to\_try\_new\_products\_Cat
  + DSS\_for\_ShortLongTermBothDecision\_Cat
  + WTP\_for\_DSS\_Cat
  + Farm\_type\_Biodynamic
  + Farm\_type\_Conventional
  + Prefered\_type\_Web\_Software\_Both\_Cat
  + Gender\_1male
  + Appreciate\_change
  + Trust\_in\_colleague\_advice
  + Prefered\_access\_Phone
  + Prefered\_access\_PC
  + Age\_Cat
  + Enjoy\_using\_new\_techs
* OneRAttributeEval:
  + 81.343 26 Marketing\_info\_about\_DSS\_Cat
  + 79.851 19 Prefered\_access\_PC
  + 79.851 4 Prefered\_access\_Advisor
  + 79.851 17 Prefered\_access\_Phone
  + 75.373 8 Prefered\_type\_Web\_Software\_Both\_Cat\_NO
  + 75.373 1 Farm\_type\_Organic
  + 74.627 5 Willing\_to\_try\_new\_products\_Cat
  + 73.881 16 Enjoy\_using\_new\_techs\_1to3
  + 73.881 28 Farm\_size\_Cat
  + 73.881 14 DSS\_for\_ShortLongTermBothDecision\_Cat\_NO
  + 73.881 31 Need\_IT\_teaching\_Cat
  + 73.881 12 Farm\_type\_Biodynamic
  + 73.881 41 Price\_important\_factor
  + 73.881 9 Trust\_in\_DSS\_1to3
  + 73.881 11 Gender\_1male
  + 73.134 6 Farm\_type\_Conventional
  + 73.134 13 Region
  + 73.134 7 Country
  + 73.134 18 Speed\_internet\_Cat
  + 73.134 23 Income\_Cat
* Old OneRAttributeEval:
  + 80.1587 29 Marketing\_info\_about\_DSS\_Cat
  + 78.5714 15 Prefered\_access\_Phone
  + 78.5714 16 Prefered\_access\_PC
  + 78.5714 3 Prefered\_access\_Advisor
  + 74.6032 39 Importance\_of\_easy\_to\_use
  + 73.8095 1 Farm\_type\_Organic
  + 73.0159 6 Willing\_to\_try\_new\_products\_Cat
  + 73.0159 34 Trust\_advice\_friend
  + 72.2222 27 Need\_IT\_teaching\_Cat
  + 72.2222 9 Farm\_type\_Biodynamic
  + 72.2222 12 Gender\_1male
  + 72.2222 11 Prefered\_type\_Web\_Software\_Both\_Cat
  + 72.2222 40 Price\_important\_factor
  + 72.2222 28 Farm\_size\_Cat
  + 71.4286 21 Crop\_Mixed
  + 71.4286 14 Trust\_in\_colleague\_advice
  + 71.4286 10 Farm\_type\_Conventional
  + 71.4286 20 Trust\_in\_DSS
* ReliefFAttributeEval:
  + 0.121476510067114 13 Region
  + 0.11543624161073808 7 Country
  + 0.08590604026845618 10 WTP\_for\_DSS\_Cat
  + 0.08523489932885887 16 Enjoy\_using\_new\_techs\_1to3
  + 0.07427293064876943 26 Marketing\_info\_about\_DSS\_Cat
  + 0.0718120805369126 15 Appreciate\_change\_1to3
  + 0.06957494407158822 2 Manufacturer\_demo\_of\_DSS\_Cat
  + 0.05302013422818784 30 Output\_preference\_Cat\_NO
  + 0.0521252796420581 14 DSS\_for\_ShortLongTermBothDecision\_Cat\_NO
  + 0.04577181208053686 28 Farm\_size\_Cat
  + 0.03993288590604023 11 Gender\_1male
  + 0.037583892617449634 19 Prefered\_access\_PC
  + 0.034228187919463075 21 Age\_Cat
  + 0.03422818791946307 8 Prefered\_type\_Web\_Software\_Both\_Cat\_NO
  + 0.032550335570469775 5 Willing\_to\_try\_new\_products\_Cat
  + 0.02975391498881431 39 Production\_level\_1to3
  + 0.028859060402684485 4 Prefered\_access\_Advisor
  + 0.02617449664429532 18 Speed\_internet\_Cat
  + 0.020357941834451908 36 Importance\_of\_efficient\_1to3
  + 0.020134228187919472 27 Importance\_of\_easy\_to\_use\_1to3
  + 0.02013422818791935 17 Prefered\_access\_Phone
  + 0.017785234899328865 41 Price\_important\_factor
  + 0.01543624161073826 9 Trust\_in\_DSS\_1to3
  + 0.013087248322147655 1 Farm\_type\_Organic
  + 0.012885906040268458 23 Income\_Cat
  + 0.012304250559284116 31 Need\_IT\_teaching\_Cat
* Old ReliefFAttributeEval:
  + 0.10469798657718093 5 Region
  + 0.0800894854586128 8 WTP\_for\_DSS\_Cat
  + 0.06017897091722587 29 Marketing\_info\_about\_DSS\_Cat
  + 0.057718120805369026 13 Appreciate\_change
  + 0.056823266219239305 2 Manufacturer\_demo\_of\_DSS\_Cat
  + 0.051006711409395895 7 DSS\_for\_ShortLongTermBothDecision\_Cat
  + 0.04295302013422817 11 Prefered\_type\_Web\_Software\_Both\_Cat
  + 0.041834451901565956 26 Output\_preference\_Cat
  + 0.03473154362416106 32 Production\_level
  + 0.034228187919463075 16 Prefered\_access\_PC
  + 0.033557046979865716 3 Prefered\_access\_Advisor
  + 0.03221476510067114 17 Age\_Cat
  + 0.03154362416107383 27 Need\_IT\_teaching\_Cat
  + 0.030872483221476496 6 Willing\_to\_try\_new\_products\_Cat
  + 0.029817833173537877 18 Enjoy\_using\_new\_techs
  + 0.02617449664429532 14 Trust\_in\_colleague\_advice
  + 0.022147651006711417 28 Farm\_size\_Cat
  + 0.02214765100671137 15 Prefered\_access\_Phone
  + 0.019798657718120814 40 Price\_important\_factor
  + 0.01946308724832216 23 Income\_Cat
  + 0.019463087248322155 1 Farm\_type\_Organic

## Problem: Class imbalance



* #Solution1: adding Weka filter: weka.filters.supervised.instance.ClassBalancer
  + It added weight 0.8 / 1.2 to instances
  + Prediction accuracy falls, but ROC is slightly higher (2%)
  + I left it out from further analysis for now
* #S2 Oversampling [option]: duplicate instances with class 0
* # Aneta: cost sensitive learning
  + Kaznuješ napačne napovedi manjšinskega razreda

## Some attributes have too many possible values

* #[option] bucketing

# Modelling

## Baseline classifier

* ZeroR: PA **73.1343 %** (uses mode to predict)
* Questionable how informative the attributes are if other classifiers perform around this mark or worse

*Ver: farmers\_removed\_attr\_29042021\_2.arff*

## Decision Rules

*Ver: farmers\_removed\_attr\_29042021\_3.arff*

* OneR (simplest classification structure - one rule)
  + When removed Attr “Subject”, classification accuracy is 77.6119 %
  + MCC: 0.341
    - \*The Matthews correlation coefficient (MCC), instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.
    - Marketing\_info\_about\_DSS\_Cat:
      * 1 [No] -> 1
      * 2 [occas] -> 1
      * 3 [Yes, subsc] -> 1
      * ? [missing v] -> 0
* PART
  + PART:
    - PA: 68.6567 % (with country attr)
    - MCC: 0.202
  + PART [23:15]: 69.8413 % (without country attr)
* JRip
  + classification accuracy is 79.1045 %
  + MCC: 0.419
  + 3 rules:
    - (Country = Italy) and (Manufacturer\_demo\_of\_DSS\_Cat = 1) => Already\_used\_DSS=0 (19.0/4.0)
    - (Farm\_type\_Organic = 1) => Already\_used\_DSS=0 (11.0/5.0)
    - => Already\_used\_DSS=1 (104.0/15.0)

## Naive Bayes

* NaiveBayes
  + Classification accuracy: 74.6269 %
  + MCC: 0.332

## K Nearest Neighbour

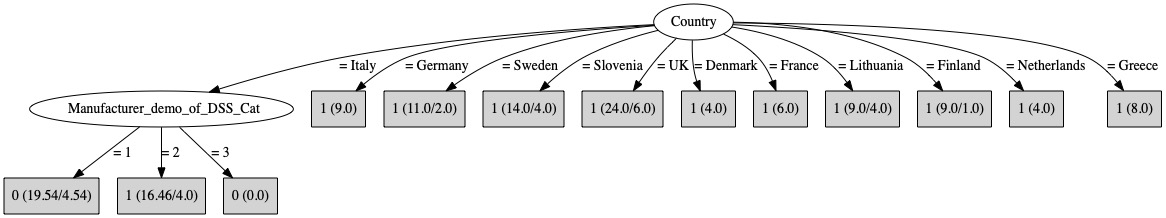
* Lazy.IBk:
  + PA: 76.8657 % (K=10)
  + MCC: 0.313

## Build Decision trees

* J48 , M2
  + PA 71.6418 %
  + MCC 0.084
* J48 , M10, binarySplits\_TRUE (Whether to use binary splits on nominal attributes when building the trees.)
  + PA: 79.1045 %
  + MCC: 0.413

## 

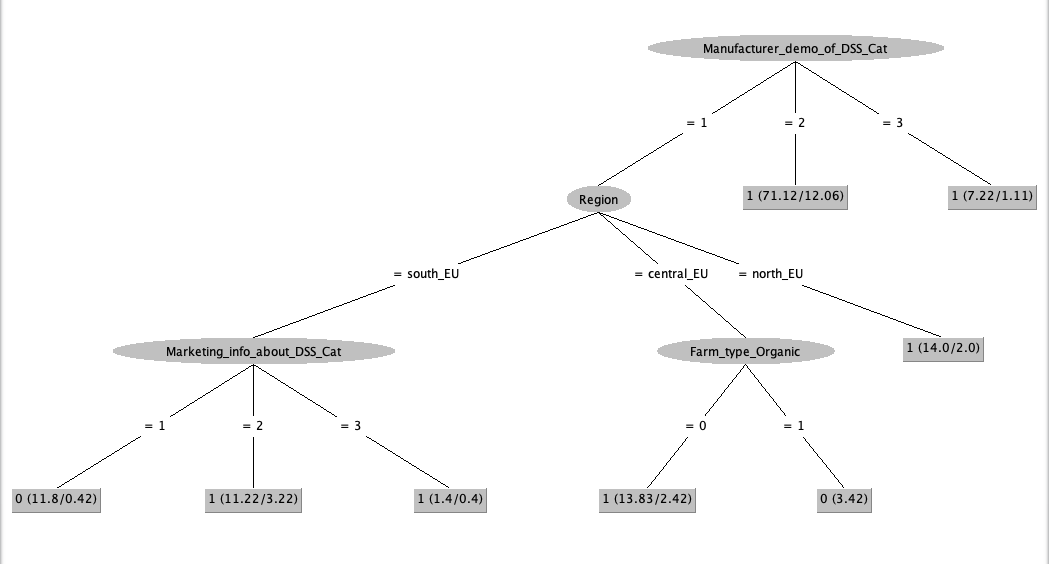
* J48 , M10, binarySplits\_FALSE, Unporuned\_TRUE
  + PA: 79.1045 %
  + MCC: 0.413



* **Removed “Country” attribute!**

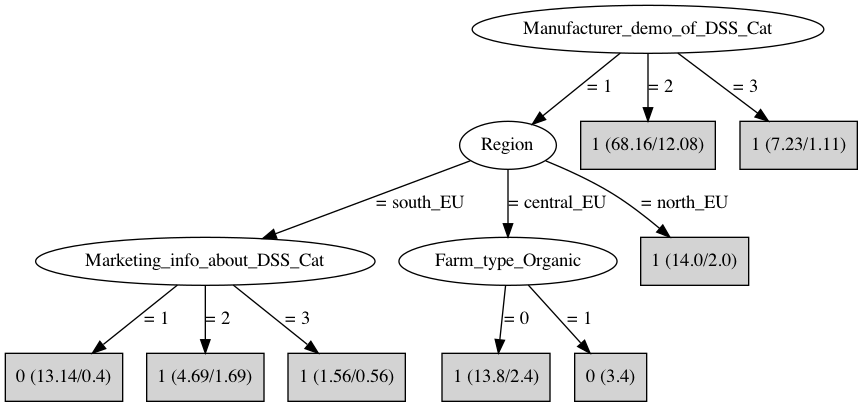
*Ver: farmers\_removed\_attr\_29042021\_4.arff*

* J48 , M2, binarySplits\_FALSE, Unporuned\_TRUE
  + PA: 71.6418 %
  + MCC: 0.140



old:

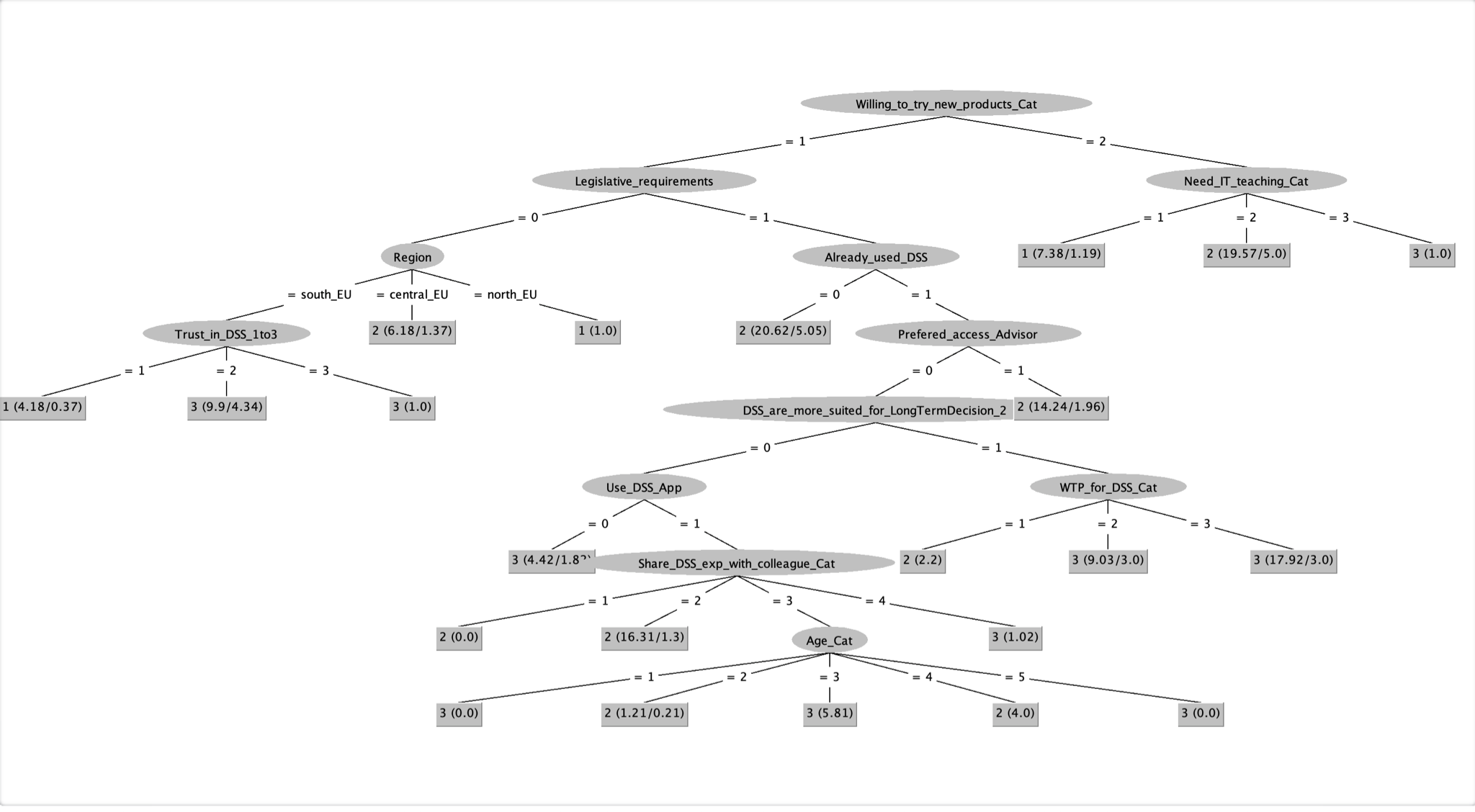
## 



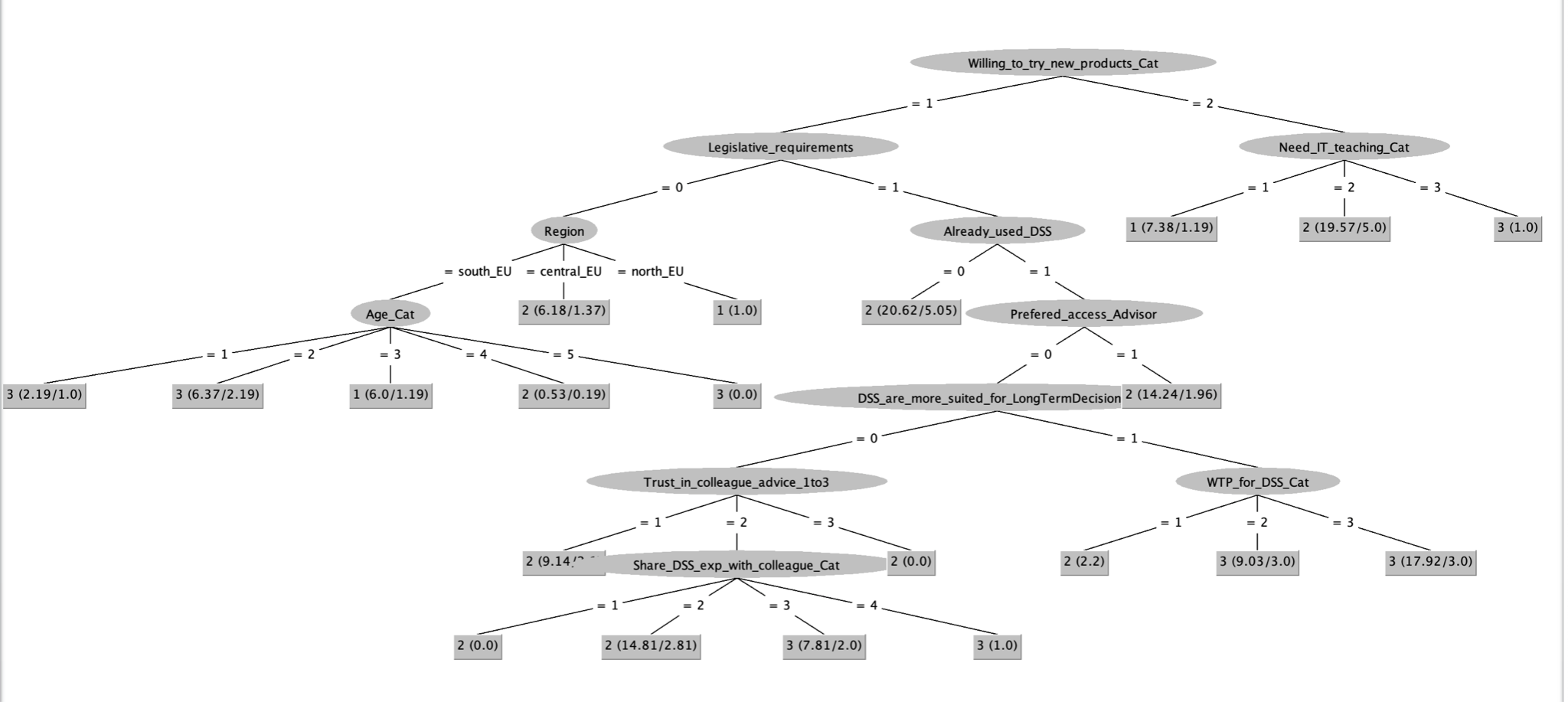
**Target Class: DSS\_can\_improve\_your\_productivity\_Cat**

\*Removed duplicated (lef 1-3)

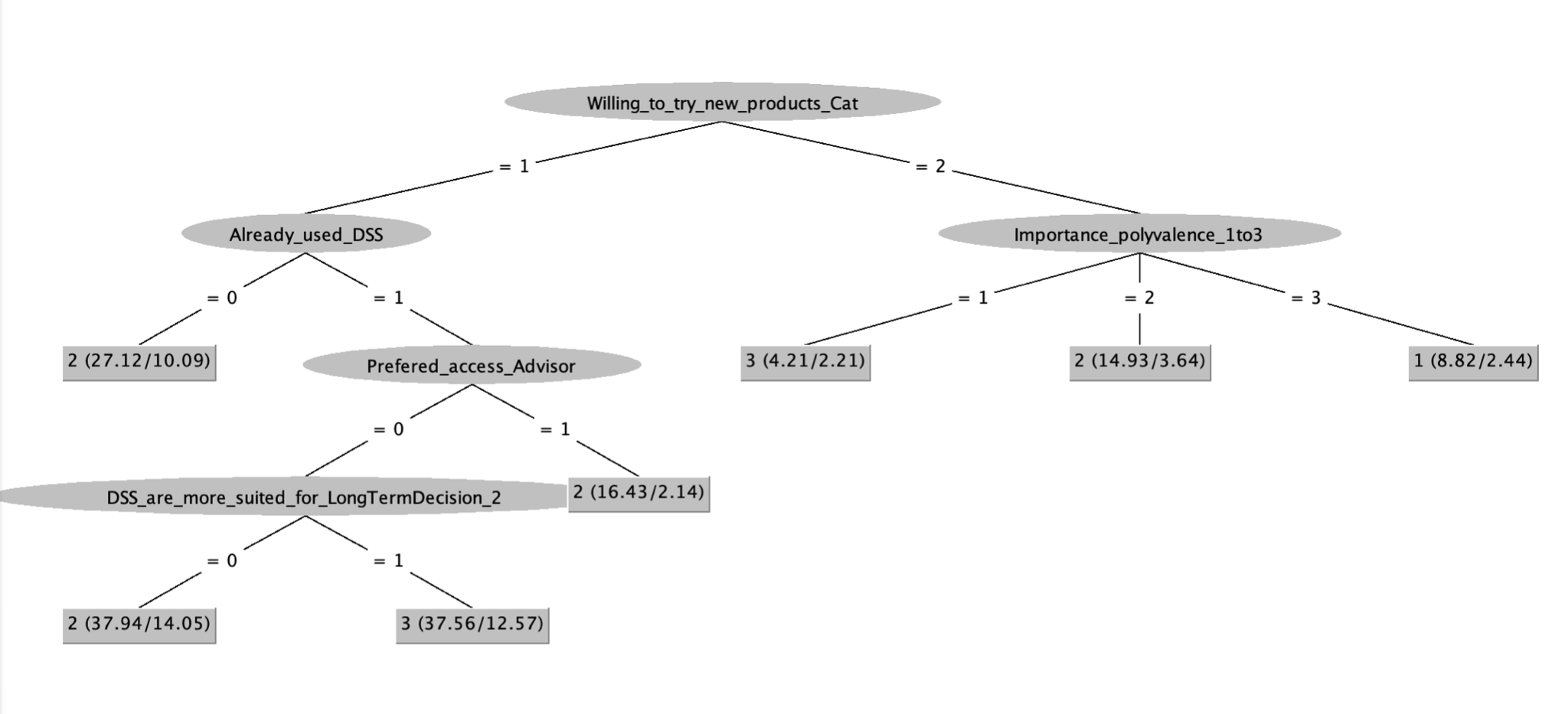
* ZeroR:
  + PA: 54.4218 %
* J48 , M4,
  + PA: 55.7823 %
  + MCC: 0.215



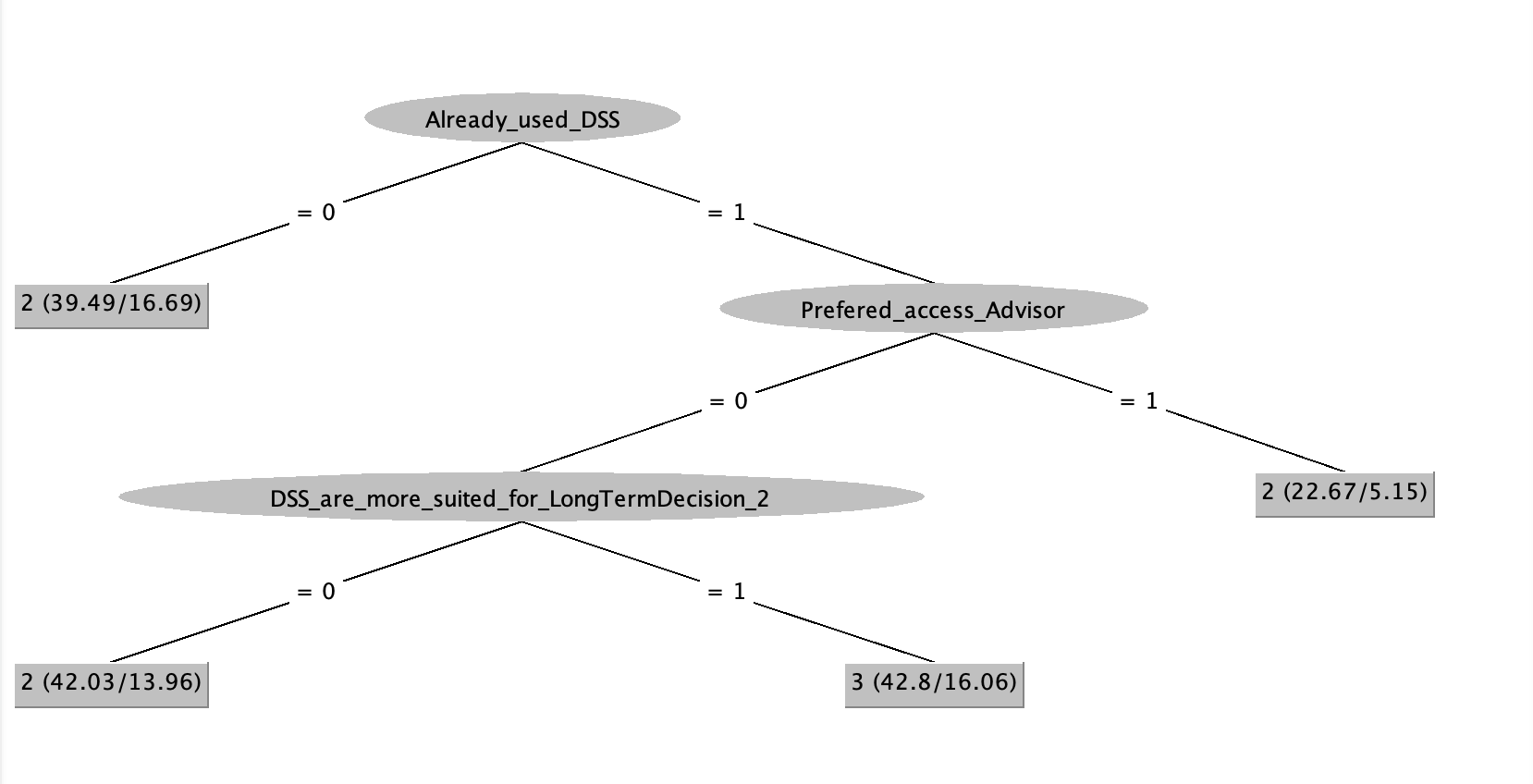
* J48 , M5,
  + PA: 56.4626 %
  + MCC: 0.205



* J48 , M8,
  + PA: 55.102 %
  + MCC: 0.146



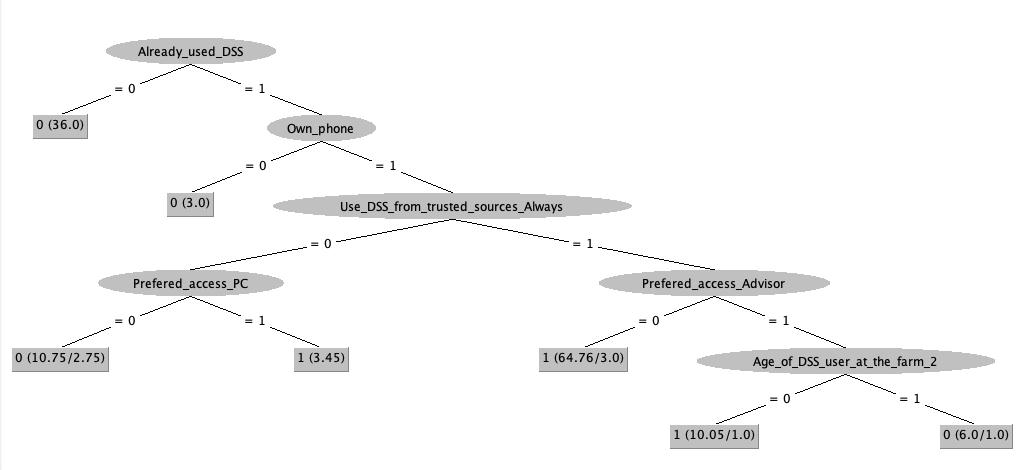
* J48 , M2,
  + PA: 59.1837 %
  + MCC: 0.217



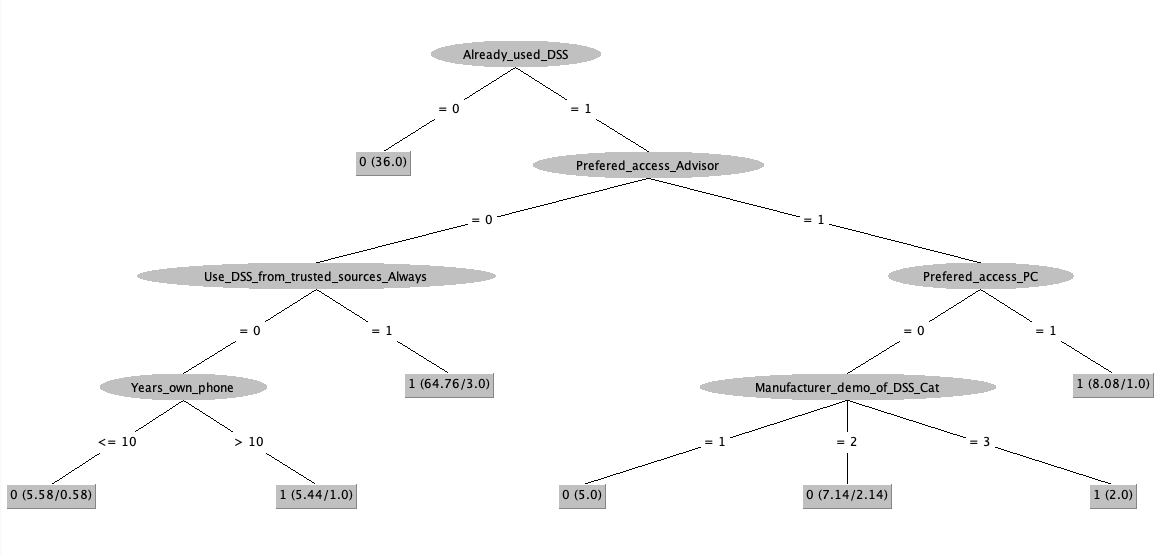
**Target Class: Potencial\_user\_of\_platform**

Removed attr:

* duplicated (lef 1-3)
* How\_do\_you\_access\_DSS\_Cat
* ZeroR:
  + PA: 58.209 %
* J48:
  + PA: 85.8209 %
  + MCC: 0.707

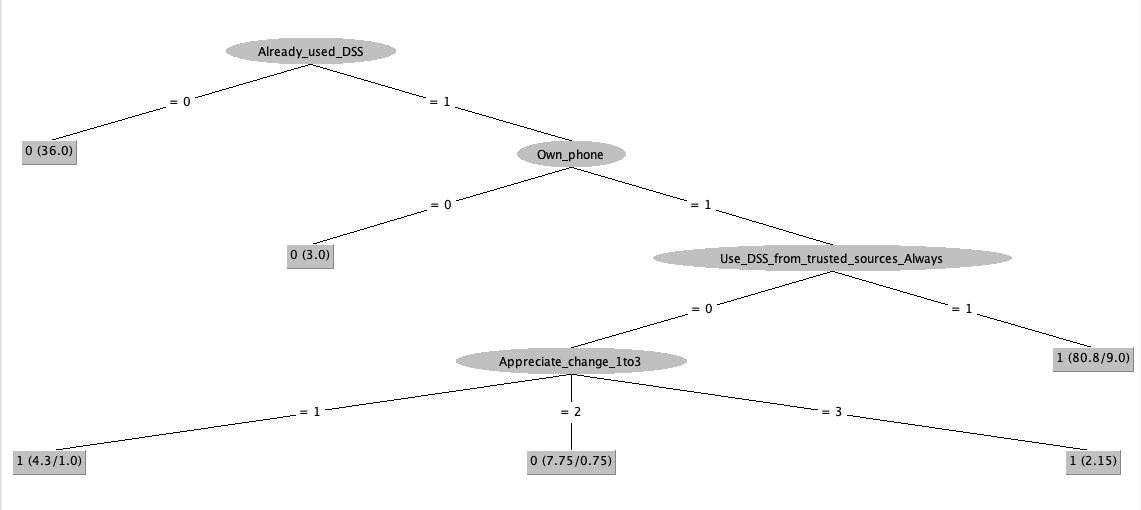


* **J48, M5:**
  + PA: 85.8209 %
  + MCC: 0.708



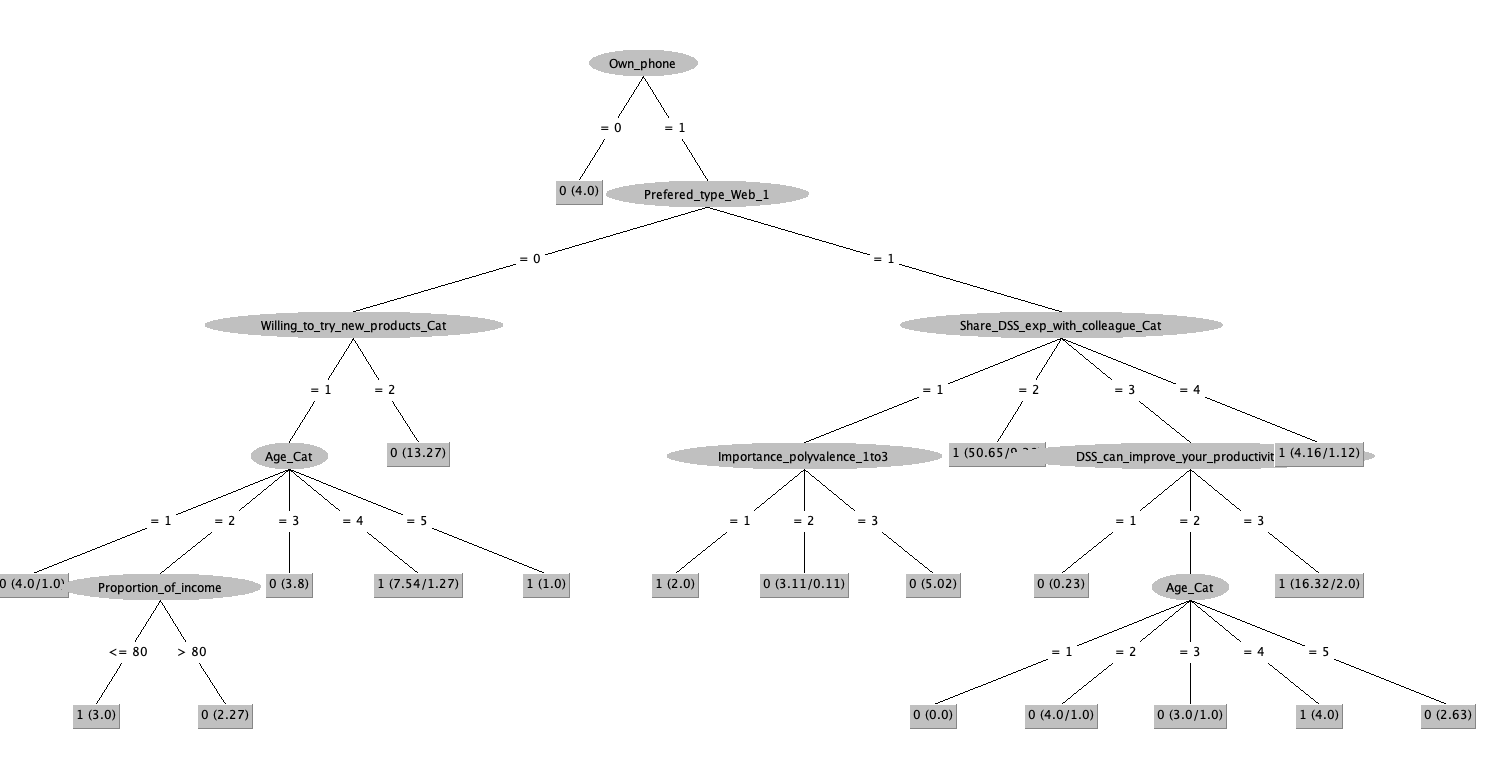
Removed 3x attr “prefered\_acceess”

* J48, M2:
  + PA: 87.3134 %
  + MCC: 0.741

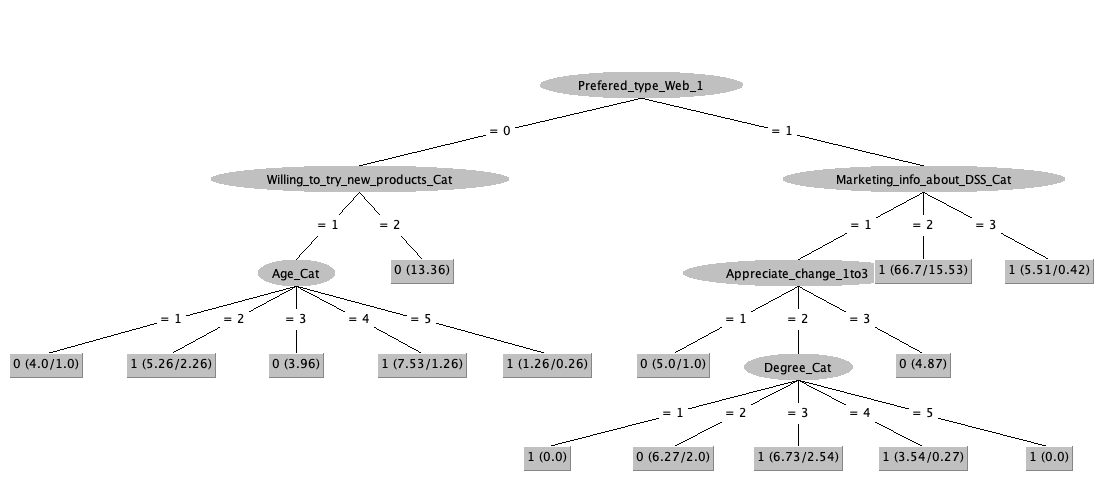


Removed attr “Already\_used”, duplicates 1-7 & 1-7,

* J48, M2:
  + PA: 64.1791 %
  + MCC: 0.245



* J48, M5:
  + PA: 65.6716 %
  + MCC: 0.277



# Clustering (Python)

## Dim Reduction (tsne, UMAP)

* Uniform Manifold Approximation and Projection (UMAP)
  + reducing the dimension in a way that preserves as much of the structure of the data as possible we can get a visualisable representation of the data allowing us to “see” the data and its structure and begin to get some intuition about the data itself;
  + Ultimately supervised UMAP will be similar in some ways to a KNN classifier.
  + # Need one-hot encoding and using metric='dice'
  + # [option] categorical and numerical data types can be handled using gower-distance metric. code for gower distance metric from [here](https://sourceforge.net/projects/gower-distance-4python/).
  + # [2019] at the moment best way to handle this data in UMAP is to split the numeric and categorical variables, then one-hot encode the categoricals (pd.get\_dummies) with a dice metric, then merge the two splits and continue.
  + # One alternative would be to check out the 0.4dev branch which has a (very!) experimental class called DataframeUMAP that would take a pandas dataframe and a tuple of (column name, metric) lists (similar to the ColumnTransformer in sklearn) and does all the required manipulations.
  + UMAP reduces down to 2D. Each row of the array is a 2-dimensional representation of the corresponding instance

## Try clustering methods for binary data (try different similarity metrics)

## 