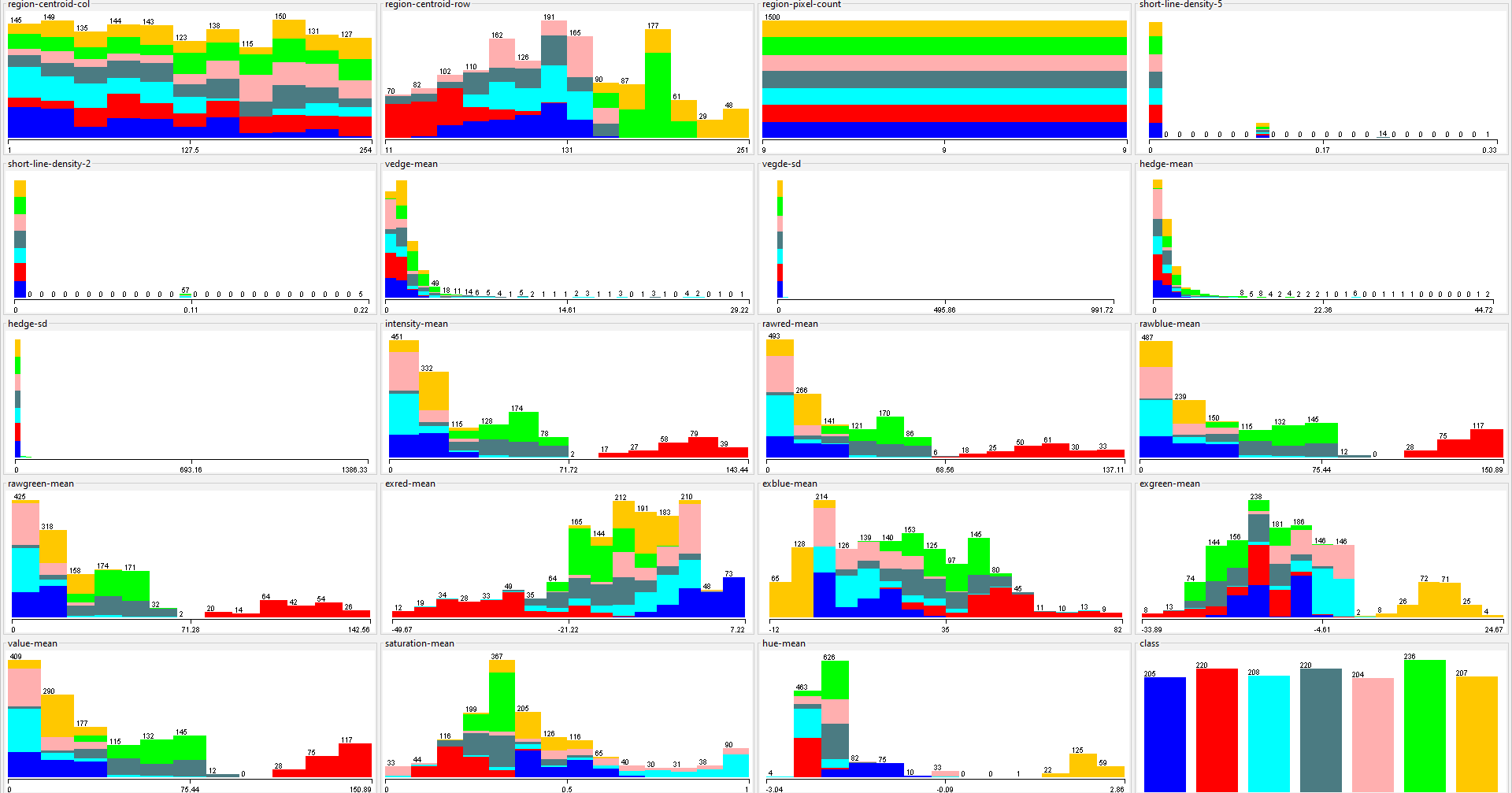
Anàlisi del data set "**segment-challenge.arff**" amb Weka:

-Després de provar 3 models, troba el millor per poder fer la predicció. Cal incloure algun gràfic que et sembli significatiu.

En l’anàlisi inicial del data set, cap dels atributs té valors nuls i per la visualització dels atributs amb la corresponent graficació (img de sota) es mostra l’atribut “region-pixel-count” com una constant amb el valor de 9 (3ªcolumna,1ªfila) i al no aportar cap informació útil per la classificació, es pot eliminar. Després tant l’atribut "short-line-density-5" com "short-line-density-2" tenen representació menor a 0.15% respecte al total de les mostres qualsevol valor diferent a 0. Per tant analitzarem el data set sense aquests 3 atributs per reduir el soroll i així simplificar el model d’interpretació.

En aquest anàlisi, faré servir el model **Logistic**, un **RandomTree** i un **RandomForest**, els tres amb un mode de test de 10-folds-cross-validation.

L’àrea sota la **curva ROC**, mostra que el model RandomForest és un millor classificador ja que generarà abans un True Positive(Y) que un False Positive(X) amb una major proporció.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic** | **RandomTree** | **RandomForest** |
| **Exactitud (%)** | 96 | 95.67 | 97.67 |
| **Mesura F1** | 0.881 a 1.000 | 0.887 a 0.995 | 0.931 a 1.000 |

Nº instàncies: 1500

|  |  |
| --- | --- |
| MATRIU DE CONFUSIÓ | MATRIU DE CONFUSIÓ |
| **Logistic** | **RandomTree** |
|  |  |

|  |
| --- |
| MATRIU DE CONFUSIÓ |
| **RandomForest** |
|  |

Els tres models classifiquen la gran majoria d’instàncies del data set per sobre d’un 95%, la mesura-F és alta per totes les classes, havent així una bona precisió predictiva i captura les instàncies de cada classe efectivament. En quant a les matrius de confusió, hi ha molt poques classificacions errònies havent-hi més en proporció en el RandomTree.

Entre el RandomForest i el model Logistic, el primer té una exactitud mínimament superior (d’un 1.67%), mostrant que pot fer prediccions més correctes respecte a l’altre i encara que els dos models mostrin una mesura-F1 en gran part de les classes, el model RandomForest té una precisió major per algunes classes comparat amb el Logistic, mostrant així que té un poder de discriminació major i pot obtenir millor els patrons i les variacions de les dades.

El millor model d’entre els 3 escollits per a la predicció és el **RandomForest**.

|  |
| --- |
| **Logistic** |
|  |
| **RandomForest** |
|  |
| **RandomTree** |
|  |

Corves ROC