

Introducción a Data Mining

Trabajo Práctico (grupal) Nº 3 – IDT , NB, IBL

Un importador obtuvo datos sobre los embarques que pasaron por diferentes aduanas de la Argentina en el período ene-jul de 2017. Cuenta con varios atributos descriptivos generales y considera como variable objetivo el CANAL que la Aduana le preasignó a cada embarque. El importador desea utilizar esta evidencia para predecir el CANAL que se le asignaría a los embarques que pretende importar en agosto de 2017. Tenga en cuenta que los valores que puede tomar la variable CANAL son:

- R (rojo): el embarque será fiscalizado de forma intrusiva y exhaustiva.
- N (naranja): dependiendo de la disponibilidad de recursos y la decisión del inspector el embarque será fiscalizado con mayor o menor detenimiento.
- V (verde): el embarque se dejará pasar libremente sin fiscalización.

Por lo tanto para el importador, la clase más riesgosa es R, en menor medida N y nula V; si es que pretende cometer algún ilícito o simplemente mitigar los tiempos del embarque en tránsito.

Utilizando los siguientes datasets:

- “embarques aduana - 201701_201706.xls” como TRAIN SET.
- “embarques aduana - 201707_201707.xls” como TEST SET.

Se pide:

1. Utilizando un método de inducción de árboles de decisión debe desarrollar un modelo para predecir la variable CANAL. Elaborar y validar diferentes modelos utilizando el mismo algoritmo pero variando la complejidad del mismo para obtener el que considere mejor según una relación complejidad/performance. Documente las pruebas y la decisión.
2. Utilizando un método de predicción basado en distancias debe desarrollar un modelo para predecir la variable CANAL. Elaborar y validar diferentes modelos utilizando el mismo algoritmo pero variando los parámetros correspondientes para obtener el que considere mejor. Documente las pruebas y la decisión.
3. Utilice los mismos dataset para obtener un modelo basado en Bayes Ingenuo.
4. Compare la performance de los métodos seleccionados y cierre con conclusiones sobre los resultados obtenidos.

Considere que:

5. Debe tomar y justificar todas las decisiones que considere necesarias para el desarrollo del trabajo.
6. Puede utilizar cualquier software que implemente estos algoritmos. Debe especificar que entorno utiliza.
7. Debe entregar un informe de los resultados obtenidos de no más de 3 páginas.
8. Sea sintético, no incluya texto que esté en la bibliografía, use referencias.
9. Se evaluará la compleción del trabajo realizado, la aplicación de los pasos metodológicos, la razonabilidad de las decisiones tomadas y la calidad del trabajo presentado.



UNIVERSIDAD
AUSTRAL

| **INGENIERÍA**