Ejercicio 1

Tutorial “Modeling”: Se utilizaron tres modelos diferentes para el data set Titanic: Naive Bayes, Decision Tree, y Rule Induction. En los tres casos, RapidMiner nos permite ver fácilmente los patrones que detecta cada modelo, y entender mejor cómo clasifican los ejemplos. Por ejemplo, se detectaba claramente en todos los modelos que era más probable que sobrevivieran las mujeres y los niños que los hombres adultos.

Tutorial “Scoring”: Se utilizó el método de Naive Bayes para modelar el data set, y se aplicó el modelo al mismo data set sin etiquetas, para obtener las predicciones sobre si cada pasajero sobrevivió o no.

Tutorial “Test Splits and Validation”: Se utilizó el 70% de los ejemplos del data set para entrenar el modelo, y el restante 30% para testearlo. Como resultado se obtiene una matriz de confusión que nos indica que tan preciso fue el modelo.

Tutorial “Cross Validation”: El operador Cross Validation nos dividió el data set en partes iguales y fue rotando entre todas, utilizando siempre una para testing y el resto para entrenamiento. Utilizar todo el data set para el testing permitió tener una mejor aproximación acerca de la precisión del modelo. Sin embargo, hay que tener en cuenta que este método utiliza mucho más poder de computación dado que se necesita entrenar tantos modelos como las particiones que se realizan en el operador.

Tutorial “Visual Model Comparison”: La curva ROC muestra qué tan bien funciona un modelo binario. Muestra la tasa de verdaderos positivos frente a la tasa de falsos positivos. El resultado es una línea que es una diagonal recta si el modelo simplemente está adivinando, y una curva que se mueve cada vez más hacia la esquina superior izquierda cuanto mejor se vuelve el modelo. En este caso se comparó la curva ROC de tres modelos diferentes, Naive Bayes, Decision Tree, y Rule Induction. En los tres casos la curva se mueve hacia la esquina superior izquierda, indicando que los tres modelos son más efectivos que adivinar al azar. El de Naive Bayes es el más alejado de la esquina, indicando que es el modelo menos efectivo para este data set en particular.