# PD2:

## 1.” Modelling”

* La creación de modelos es sumamente sencilla una vez se cuenta con los datos apropiadamente sanitizados
* El modelo Naive-Bayes usa curvas para representar las diferencias entre atributos numéricos, basándose en la densidad de fallecidos o no para cada valor del atributo. Utiliza gráficos de barras para atributos polinomiales

## 2. “Scoring”

* Significa utilizar modelos para predecir campos para datos nuevos
* Se utilizan datos “unlabeled” que significa que el campo de sobrevivió está en blanco/ no lo tiene. Son los datos a los que queremos predecir.
* El modelo retorna 3 columnas, una de la confianza de que la persona sobrevivió, una de que no y una de la predicción final.
* 9 de los 10 más probables a sobrevivir son mujeres.
* El decisión tree presenta valores de confianza mayores que el naive bayes.

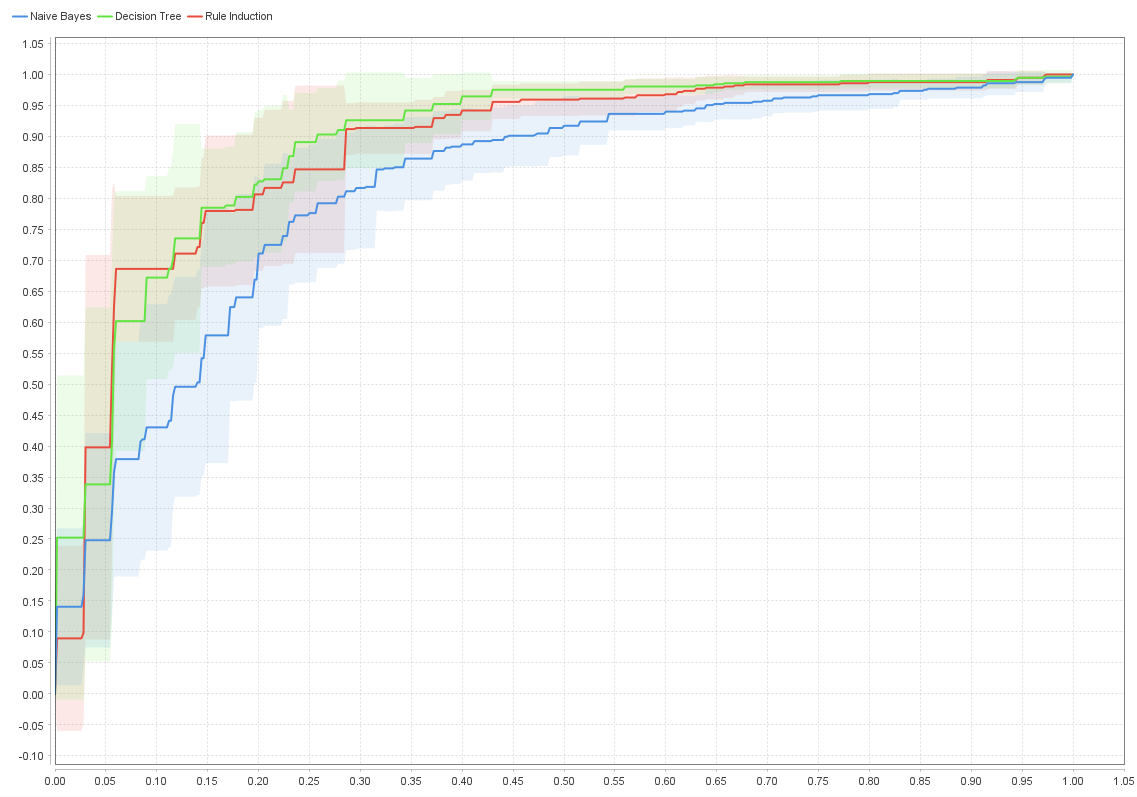
## 3. “Test Splits and Validation”

* Se dividen los sets de entrenamiento y se reservan una parte para probar el modelo luego.
* Un buen radio de separación es 70/30
* Performance retorna tanto un conjunto de datos sobre el performance, como matrices de confusión y porcentajes de precisión.
* Los datos resultados se pueden filtrar para ver predicciones correctas e incorrectas
* El modelo fallo 270 predicciones, esto también se puede contar sumando los valores de las matrices de confusión.

## 4. “Cross Validation”

* Es un método de validación donde se dividen los datos y se van usando cada uno para entrenar y el resto para testear iterativamente. Es una forma mas correcta de verificar y entrenar el modelo.
* Se puede elegir la cantidad de particiones que usa para validar.
* Aparte de los promedios de precisión, nos da desviaciones o error, dando una imagen más completa de desempeño del modelo.
* Existe “wisdom of the crowds” para parámetros puntuales de módulos.

## 5. “Visual Model Comparison”

* Para medir performance se pueden usar ROCs (Receiver Operating Characteristics). Son curvas que muestran valores de falsos y verdaderos positivos en el éxito de la predicción y la confianza
* Se pueden testear varios modelos a la vez
* Como cross validation, podemos decidir cuantas iteraciones usar, 3-10 es un rango aceptable, ya que afecta performance.
* 
* Mientras la curva se acerca más rápido hacia arriba a la izquierda, más preciso es el modelo. Esto depende enteramente de los datos utilizados