

Actividad 2

Minería de datos

Alberto García González

77873658M



Índice

Comprensión de negocio	3
Comprensión de los datos	3
Tratamiento de los datos.....	5
Missing Values	7
Outliers.....	7
SMOTE	8
Normalizer.....	8
Feature Selection	9
Modelos predictivos	9
Árbol de decisión (Decision Tree).....	10
Boosting	11
Bootstrapping	12
Random Forest	13
Procesamiento de los resultados	14
Variables de media	15
Modelado.....	17
Año 1	19
Año 2	20
Año 3	21
Año 4	22
Año 5	23
Evaluación	24
Despliegue.....	25



Comprensión de negocio

Para esta entrega, se propone la actividad de generar predicciones de tal manera que sea capaz adivinar si una empresa entrará en quiebra del momento actual a los próximos 5 años.

Para ello, se disponen de 5 archivos con información económica de miles de empresas de Polonia, de diferentes ámbitos.

Lo que se pretende es diseñar un modelo predictivo de Knime para generar 5 predicciones (1 por archivo), tal que, cada uno de ellos, hace referencia a datos de diferentes espacios temporales, que van desde 1 año hasta los 5 años de manera que crece un año la línea temporal de cada archivo.

Esta entrega, se realiza a fin de completar la entrega de la asignatura de Minería de Datos, pero podrían considerarse como stakeholders reales los empresarios que tengan que realizar decisiones de sus empresas para, manera predictiva, tomar las que más los beneficie.

Existirá el riesgo de no cumplirse las predicciones, ya que los datos de las empresas no sean representativos o no estén completos, por lo tanto, afecte tanto al algoritmo de predicción como a la predicción como tal. La evaluación de cumplimiento de las predicciones se hará también en este documento.

Comprensión de los datos

Previamente mencionado, se disponen de 5 archivos, tal que hacen referencia a los datos de miles de empresas que contienen relaciones económicas, indicados por los primeros 64 atributos, además de indicar si en el espacio temporal al que incide el archivo si dicha empresa entró en quiebra o no (variable objetivo). Esto queda indicado de tal manera que, si se muestra un 0, hace referencia a que la empresa no entra en quiebra; y si muestra un 1, la empresa entra en quiebra en dicho espacio temporal.



Aquí la referencia de lo que muestra exactamente cada uno de los atributos numéricos:

X1 net profit / total assets	X33 operating expenses / short-term liabilities
X2 total liabilities / total assets	X34 operating expenses / total liabilities
X3 working capital / total assets	X35 profit on sales / total assets
X4 current assets / short-term liabilities	X36 total sales / total assets
X5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	X37 (current assets - inventories) / long-term liabilities
X6 retained earnings / total assets	X38 constant capital / total assets
X7 EBIT / total assets	X39 profit on sales / sales
X8 book value of equity / total liabilities	X40 (current assets - inventory - receivables) / short-term liabilities
X9 sales / total assets	X41 total liabilities / ((profit on operating activities + depreciation) * (12/365))
X10 equity / total assets	X42 profit on operating activities / sales
X11 (gross profit + extraordinary items + financial expenses) / total assets	X43 rotation receivables + inventory turnover in days
X12 gross profit / short-term liabilities	X44 (receivables * 365) / sales
X13 (gross profit + depreciation) / sales	X45 net profit / inventory
X14 (gross profit + interest) / total assets	X46 (current assets - inventory) / short-term liabilities
X15 (total liabilities * 365) / (gross profit + depreciation)	X47 (inventory * 365) / cost of products sold
X16 (gross profit + depreciation) / total liabilities	X48 EBITDA (profit on operating activities - depreciation) / total assets
X17 total assets / total liabilities	X49 EBITDA (profit on operating activities - depreciation) / sales
X18 gross profit / total assets	X50 current assets / total liabilities
X19 gross profit / sales	X51 short-term liabilities / total assets
X20 (inventory * 365) / sales	X52 (short-term liabilities * 365) / cost of products sold
X21 sales (n) / sales (n-1)	X53 equity / fixed assets
X22 profit on operating activities / total assets	X54 constant capital / fixed assets
X23 net profit / sales	X55 working capital
X24 gross profit (in 3 years) / total assets	X56 (sales - cost of products sold) / sales
X25 (equity - share capital) / total assets	X57 (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X26 (net profit + depreciation) / total liabilities	X58 total costs / total sales
X27 profit on operating activities / financial expenses	X59 long-term liabilities / equity
X28 working capital / fixed assets	X60 sales / inventory
X29 logarithm of total assets	X61 sales / receivables
X30 (total liabilities - cash) / sales	X62 (short-term liabilities * 365) / sales
X31 (gross profit + interest) / sales	X63 sales / short-term liabilities
X32 (current liabilities * 365) / cost of products sold	X64 sales / fixed assets

Además, se incluye en la siguiente imagen, la cantidad de empresas escrutadas, con referencia a los resultados totales clasificados por quiebra de cada archivo y su línea temporal, además de los datos referentes al momento y lugar de recolección de los datos de dichas empresas:

Data Set Information:
The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS, www.emis.com), which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.
Based on the collected data five classification cases were distinguished, that depends on the forecasting period:
- 1stYear: the data contains financial rates from 1st year of the forecasting period and corresponding class label that indicates bankruptcy status after 1 year. The data contains 7027 instances (financial statements); 271 represents bankrupted companies, 6756 firms that did not bankrupt in the forecasting period.
- 2ndYear: the data contains financial rates from 2nd year of the forecasting period and corresponding class label that indicates bankruptcy status after 2 years. The data contains 10173 instances (financial statements); 409 represents bankrupted companies, 9764 firms that did not bankrupt in the forecasting period.
- 3rdYear: the data contains financial rates from 3rd year of the forecasting period and corresponding class label that indicates bankruptcy status after 3 years. The data contains 15053 instances (financial statements); 495 represents bankrupted companies, 14558 firms that did not bankrupt in the forecasting period.
- 4thYear: the data contains financial rates from 4th year of the forecasting period and corresponding class label that indicates bankruptcy status after 4 years. The data contains 19752 instances (financial statements); 516 represents bankrupted companies, 19236 firms that did not bankrupt in the forecasting period.
- 5thYear: the data contains financial rates from 5th year of the forecasting period and corresponding class label that indicates bankruptcy status after 5 years. The data contains 5910 instances (financial statements); 410 represents bankrupted companies, 5500 firms that did not bankrupt in the forecasting period.

Revisados los datos, existen valores vacíos o nulos en los diferentes archivos, de tal manera que, en el modelo, dicho inconveniente tendrá que ser tratado. Otro punto para tener en cuenta es no contemplar los casos extremos, es decir, valores que, por diferentes razones, para una empresa en concreto, será extremadamente diferente al del resto de empresas; a dicha casuística se le contemplará una solución. Otro punto para considerar es la normalización, para acotar los datos a unos valores y darles una interpretación más sencilla entre las variables. Por último, para tener equidad de registros con empresas en quiebra como no, también se harán nuevos registros para tener en cuenta dicha equidad.

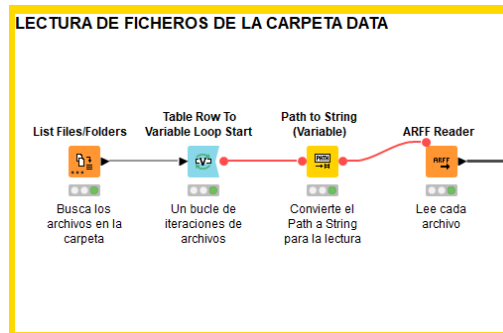
En este caso, se hace uso de selección de variables. En este caso, el método de dicha selección será de tipo Wrapper con forward, es decir, se irán viendo las relaciones entre las variables y la clase a predecir, empezando desde únicamente habiendo un atributo, añadiendo de uno en uno y viendo las relaciones entre el resto.

Por último, para generar las predicciones, se diferenciarán de manera pseudoaleatoria las empresas, de tal manera que se utilizarán parte de ellas para entrenar los algoritmos de predicción y otras para aplicar lo aprendido por estos la asignación de las predicciones a estas.



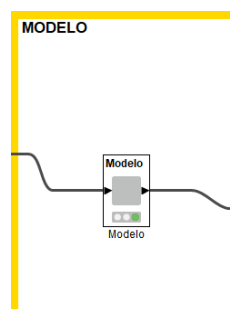
Tratamiento de los datos

Originalmente, se cuenta con una carpeta con 5 archivos, cada uno representando 5 líneas temporales de 1 año cada uno. Se solicita que la lectura de estos archivos sea a través de la carpeta, para ello se ha preparado la siguiente parte del modelo:



Se procede a la lectura de la carpeta que contiene los archivos. Estos, se leerán de forma secuencial utilizando un bucle manejado por la variable del nombre del archivo, el cual, será tratado para poder tratar dicho nombre como una cadena, y no como una ruta. Cuando se obtiene dicho nombre del archivo como cadena, se puede indicar la lectura del archivo de tipo ARFF mediante el nombre del mismo pasado como variable en el bucle.

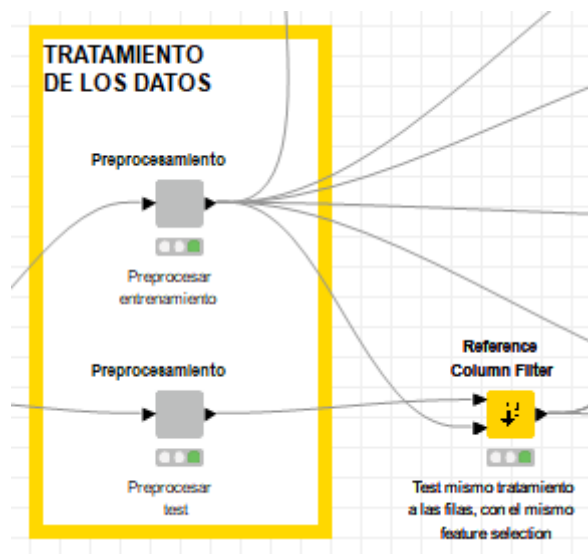
Con este tratamiento de los archivos, todo tratamiento o uso de métodos se harán a cada archivo. En este caso, todo lo realizado estará dentro de un componente llamado *Modelo*.



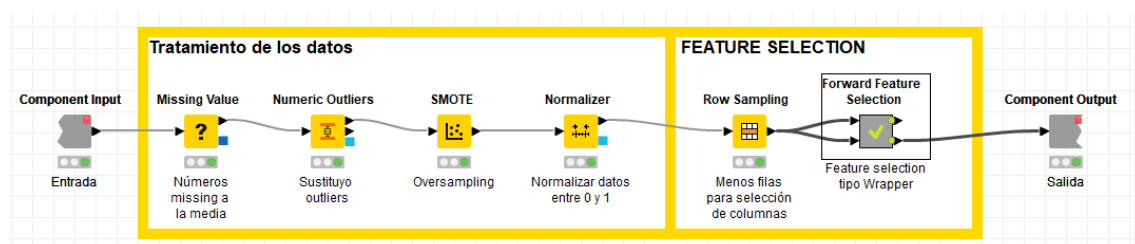


En este caso, se pretende utilizar validación cruzada, de 10 particiones, es decir, segmentar el problema en varias partes para que las predicciones sean mejores. Para ello, se hará uso del nodo X-Partitioner, que generará los segmentos de registros por cada una de las iteraciones que, terminará con X-Agregator.

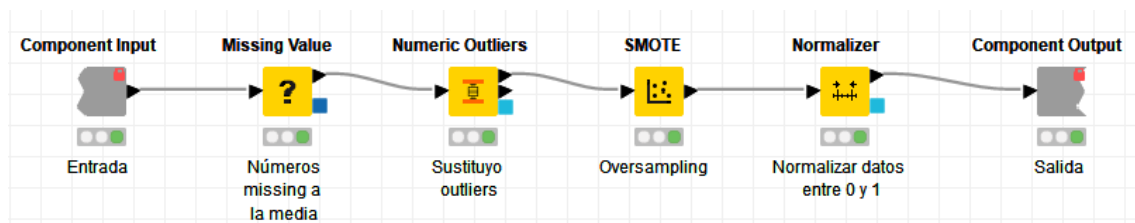
Para cada partición, se pretende hacer un tratamiento de los datos adecuado tanto para los datos de aprendizaje como para los de prueba. Dentro de este tratamiento, se encuentra la selección de variables, que únicamente se hará una vez por iteración y que se utilizarán tanto para entrenamiento como para test. En este caso, se generarán dichas columnas en el preprocesamiento del entrenamiento únicamente, y se aplicarán las columnas seleccionadas a la prueba de la misma iteración.



El preprocesamiento del entrenamiento, es decir, el que hace uso del feature selection, quedaría así:



El preprocesamiento del testing, es decir, el que no hace uso del feature selection, quedaría así:





En este caso, los tratamientos de datos comunes realizados son los siguientes:

Missing Values

Se aplica una corrección a los valores nulos de los atributos. En este caso, se aplica la misma corrección de establecer estos al valor medio de cada atributo a los valores numéricos. Los valores de tipo String nulos no tendrán aplicada ninguna corrección.

Aquí la configuración:

Number (double)	Mean
String	Do nothing

Outliers

Se aplica una corrección a los valores desajustados, tanto por encima como por debajo de los considera el rango normal. Se puede decidir qué hacer con esos valores, tal que, en este caso, se decide acercar los valores desajustados al extremo del rango decidido como normal más cercano.

Aquí se ve la configuración:



SMOTE

Se solicita para esta entrega una equidad de las clases haciendo uso de un método de oversampling, es decir, crear nuevos registros de tal manera que la clase minoritaria estuviera equilibrada en registros con la clase mayoritaria. Para ello, se hace uso del nodo SMOTE, que permite crear filas para la clase minoritaria, indicando que el método de creación será teniendo de referencia los registros que ya existen.

Settings | Flow Variables | Job Manager Selection | Memory Policy

Class column:

Nearest neighbor:

☐ Oversample by:

☒ Oversample minority classes

☒ Enable static seed

Normalizer

Se aplica un tratamiento de acotamiento común de los valores de las columnas, para que la interpretación sea más útil y legible entre los diferentes atributos. Para ello, se utiliza el nodo Normalizer, de tal manera que acote el rango de valores de cada atributo al rango entre 0 y 1, haciendo la misma escala del rango de valores, siendo el valor más bajo el 0 y el mayor el 1.

Methods | Flow Variables | Job Manager Selection | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

☒ Enforce exclusion

Include

Filter

Attr 1
Attr 2
Attr 3
Attr 4
Attr 5
Attr 6
Attr 7
Attr 8

☐ Enforce inclusion

Settings

☒ Min-Max Normalization Min: Max:

☐ Z-Score Normalization (Gaussian)

☐ Normalization by Decimal Scaling

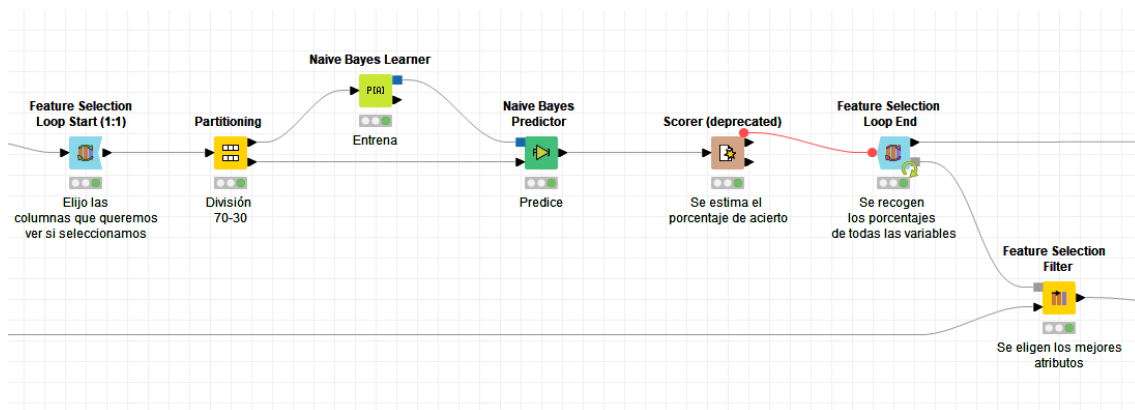


Feature Selection

Para esta entrega, se solicita un método de selección de variable, en este caso, tipo Wrapper. El tipo de Wrapper se puede diferenciar entre Forward (de menos a más columnas), y Backward (de más a menos columnas). En este caso, se aplica el contexto de Forward, ya que se enfoca la gran cantidad de atributos que es más eficiente en lo referido al costo computacional, aunque sea peor a nivel de tiempo.

Para solucionar el problema de tiempo, se opta por reducir el número de registros que se utilizarán para hacer las predicciones de las columnas. En este caso, teniendo en cuenta el tiempo de ejecución total de un archivo al completo y la cantidad de registros de cada archivo, siendo el más pequeño de casi 6.000 registros, se opta por utilizar 500 registros estratificados por la variable objetivo.

El modelo de feature selection es el siguiente:



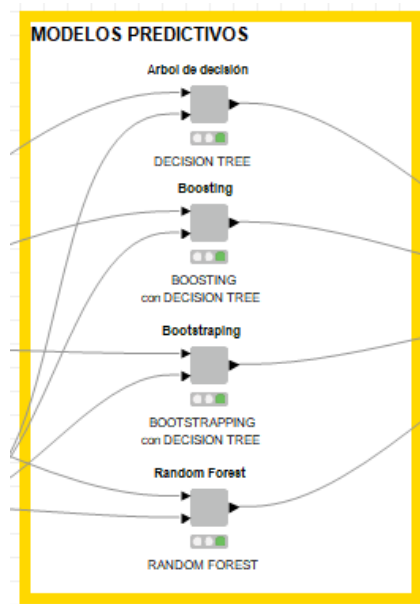
Se miran las filas empezando por una, y añadiendo una por iteración, se aplica el método de predicción Naives Bayes para ver como mejora la predicción de datos. En caso de que dicha predicción tenga un porcentaje de aciertos mayor al contemplado por el fin del bucle, se contemplará o no como aplicable al mejor conjunto de atributos. No se contempla como atributo seleccionable la variable objetivo, ya que ha de ser utilizada obligatoriamente para hacer las predicciones. Elegidas las columnas, se aplican el Feature Selection a todos los registros.

Modelos predictivos

Para esta entrega, se precisaba el uso de cuatro métodos predictivos diferentes que dieran datos concluyentes de si, usando la misma validación para todos por el método de *Cross validation* y aplicando los métodos predictivos en clase de *Decision Tree*, *Boosting*, *Bootstrapping* y *Random Forest*, capaces de determinar la quiebra o no de las diferentes empresas.



En el modelo KNIME, cada uno se encuentra embebido en un componente que lleva su nombre:



Todos los componentes además tendrán un tratamiento de los datos posterior a las predicciones para adecuar una salida útil y limpia. En puntos posteriores se detalla uso y utilidad.

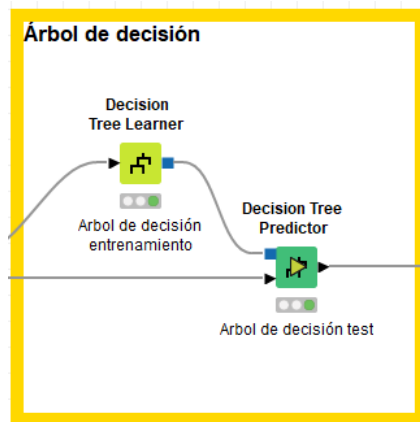
Se detallan los diferentes métodos predictivos a continuación:

Árbol de decisión (Decision Tree)

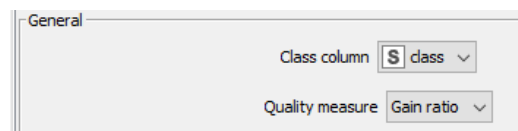
Este método genera subconjuntos a partir del conjunto original, en función de los posibles valores de los atributos. Cada subconjunto se denomina nodo, y los cuales se van desgranando hasta haber completado todas las características. Al tener el árbol completo, y tener los registros a predecir, se sigue el árbol mediante las ramas que indican la condición del nodo (posible valor de una característica). El valor del de la clase objetivo del nodo final para las condiciones de la prueba será la predicción.



Aquí se muestra la aplicación del método:



Con la comprobación de la configuración solicitada de aplicar Gain ratio como medida de calidad:



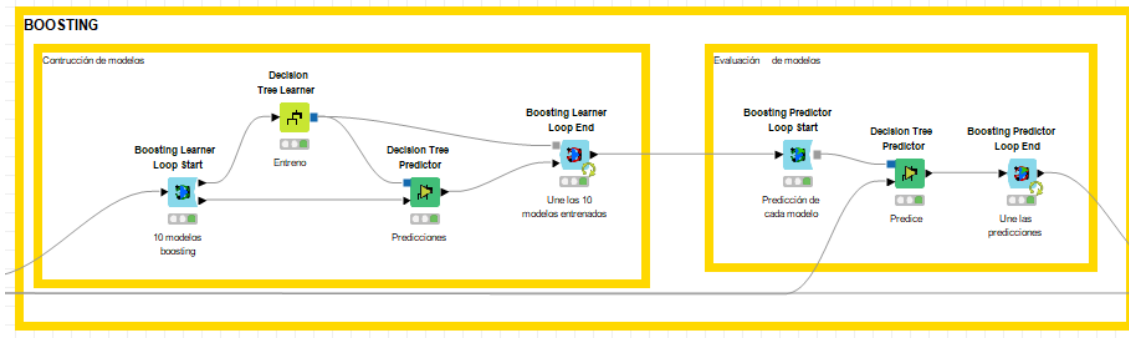
Boosting

El modelo predictivo de Boosting trata de generar varios pequeños modelos otro modelo predictivo, en este caso, aplicando Decision Tree. Con la generación de cada pequeño modelo, se le asigna la contribución al modelo, llamado peso, al principio la misma, luego de ir comprobando por iteración el error ponderado, se van actualizando los pesos. Cuando se ha terminado y se genera un modelo de predicciones en función del peso de los diferentes modelos.

En este caso, se solicita que se apliquen 10 modelos, y que la medida de calidad sea Gain ratio.



Aquí se muestra la aplicación del método:



Con la comprobación de la configuración dada:

Decision Tree Learner:

General	
Class column	<input type="text" value="S"/> class
Quality measure	Gain ratio

Boosting Learner Loop End:

Real class column	<input type="text" value="S"/> class
Predicted class column	<input type="text" value="S"/> Prediction (class)
Number of iterations	10
Use seed for random numbers	<input checked="" type="checkbox"/>
Seed	0

Bootstrapping

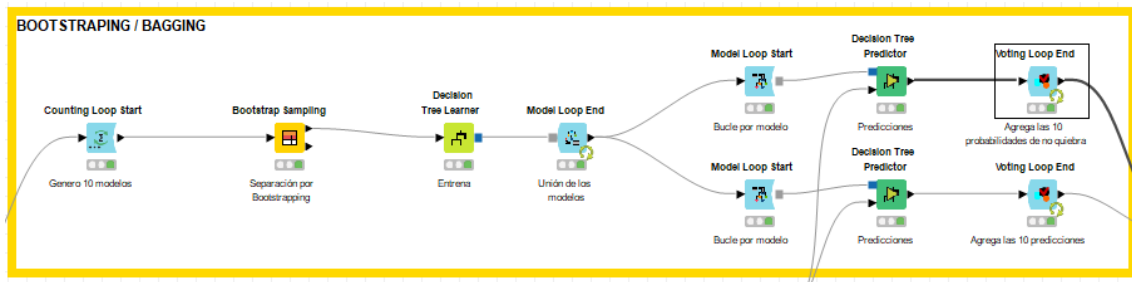
Se utilizará el modelo de predicción de Bootstrapping, que trata de aplicar una manera de separación entre registros de entrenamiento y testing en varios modelos, tal que los mismos modelos podrán repetir registros del resto de modelos.

Cuando se hace dicha separación, se aplica el método de predicción de Bagging, tal que cada modelo entrena y predice de manera independiente, de tal manera, que el resultado final serán 10 predicciones diferentes con todas las características, con una final que será la moda de los modelos parciales. En este caso, el método genera únicamente un tipo de predicción. Ya que después preciso la predicción global y el porcentaje de predicciones acertada, generaré dos tablas, una con cada una de mis predicciones, pero sobre el mismo aprendizaje aplicado a las mismas pruebas, por lo tanto, se corresponden los valores de las predicciones entre ellos.

El método de predicción usado para Bagging/Bootstrapping es Árbol de Decisión, con la configuración de medida de calidad en Gain Ratio, el Bootstrapping generará probará con el 100% del modelo inicial y creará 10 modelos segmentados de la inicial.



Aquí se muestra la aplicación del método:



Con la comprobación de la configuración dada:

Bootstrapping Sampling:

Decision Tree Learner:

Counting Loop Start:

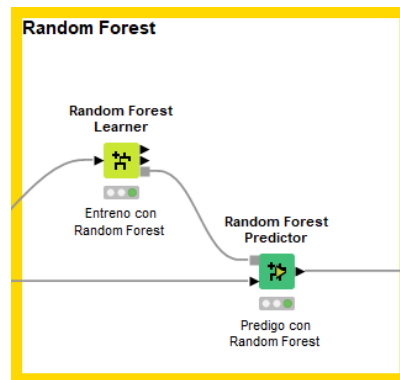
Random Forest

El método de Random Forest genera varios árboles de decisión y hace predicciones sobre estos utilizando una parte de las características, de manera aleatoria. La predicción final sobre cada registro será la moda del conjunto total de predicciones de dicho registro.

En este caso, se generarán 50 árboles y la medida de calidad será Gain ratio.



Aquí se muestra la aplicación del método:



Con la comprobación de la configuración dada:

Tree Options	
Split Criterion	Information Gain Ratio
<input type="checkbox"/> Limit number of levels (tree depth)	10
<input type="checkbox"/> Minimum node size	1
Forest Options	
Number of models	50

Procesamiento de los resultados

Cuando ya se han aplicado las correcciones de los valores incompletos o no relevantes y aplicarles los métodos de validación y predicción, se prepararán los datos para generar las gráficas pertinentes, hacer evaluaciones y decidir las conclusiones del modelo.

Para ello, se unen los resultados de los métodos, y se muestran por una curva ROC. Dicha curva, representará la relación que hay entre el porcentaje de aciertos en las predicciones de una casuística y el porcentaje de error de la misma casuística. En este caso, la casuística predicha será la no quiebra de la empresa. Para ello, se hace la comparativa de el caso real de quiebra/no quiebra de una empresa y lo que han predicho cada modelo.

Se precisan las predicciones globales, de tal manera que se hará una predicción global única generada por la moda de las 4 predicciones, que será la que X-Aggregator vea su grado de acierto en la predicción.

También se necesitarán los porcentajes de acierto sobre la no quiebra, ya que esto será lo que se muestra en la gráfica de la curva ROC.

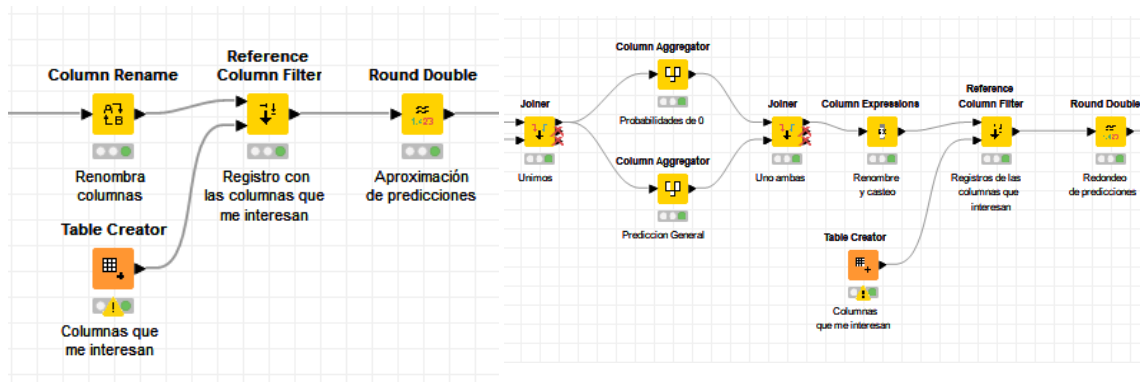
Por último, uno de los modelos llevará la clase objetivo, que será con la que se harán las comparaciones de lo real con las predicciones.



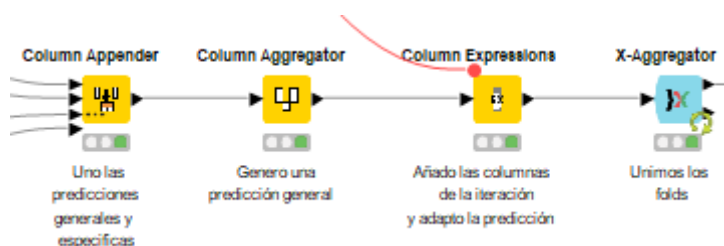
La unión, preparación y evaluación de los datos de cada método se hace de la siguiente manera:

Tratamiento datos general:

Tratamiento datos específico de Bootstrapping:



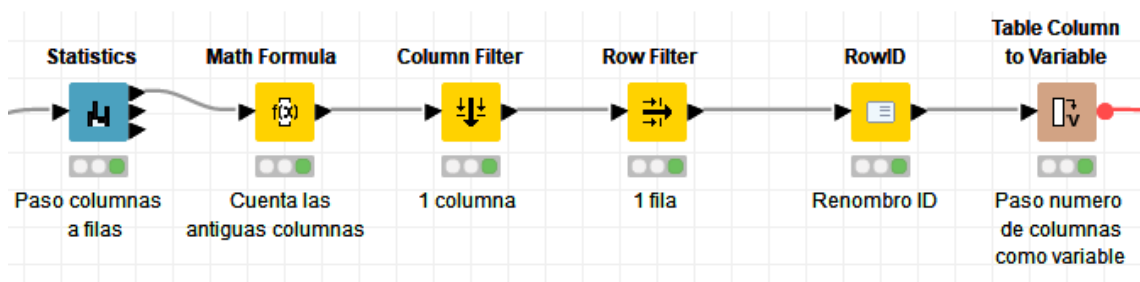
Cuando se consiguen todos los datos por separado, se unen y preparan de la siguiente manera:



Variables de media

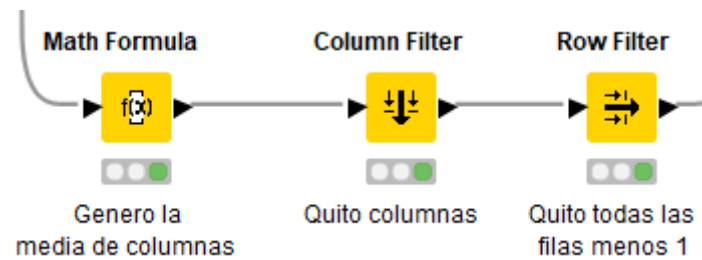
Además de la curva ROC, se solicita que, para cada archivo, se indique la media de variables que se han utilizado en cada archivo, haciendo la media de estas de las iteraciones de segmentos. Para ello, guardamos las columnas que han salido como resultado tras hacer el Feature Selection y las llevamos al punto final de la iteración, que se guardará en los registros de esta.

Para ello, se aplica el siguiente flujo, que inicia del preprocesamiento de los registros de training y finaliza como variable de flujo para añadirla como atributo al nodo Column Expressions:





Cuando se guarden en las filas de cada iteración el número de variables, se hace la media y queda registrado como una nueva característica:

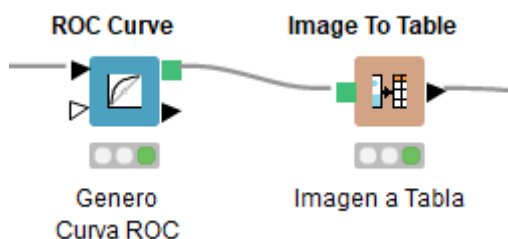


The screenshot shows the configuration for a 'Filter' tool. At the top, there are settings for 'Class column' (set to 'S class'), 'Positive class value' (set to '0'), and 'Limit data points for each curve to' (set to '2,000'). Below these, there are two selection methods: 'Manual Selection' (selected) and 'Wildcard/Regex Selection'. The 'Exclude' section (outlined in red) contains a list of items: 'P (class=1)Naives10Folds' and 'P (class=1)'. The 'Include' section (outlined in green) contains a list of items: 'P0Naives10Folds' and 'P04000NN'. At the bottom, there are two radio buttons: 'Enforce exclusion' (selected) and 'Enforce inclusion'.



Modelado

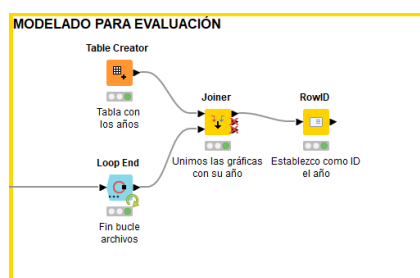
La curva ROC para la evaluación se muestra de la siguiente manera:



Con esta configuración:

Los nombres hacen referencia a los porcentajes de aciertos de 0 de cada uno de los métodos correspondientes.

Cuando ya se han obtenido los datos, estos hay que presentarlos de la manera solicitada. En este caso, se pide una tabla con dos columnas de 5 filas, de tal manera que los IDs de cada fila sea el año al que se hace referencia en cada archivo, es decir, en cada iteración. Para ello, se ha realizado la siguiente implementación:

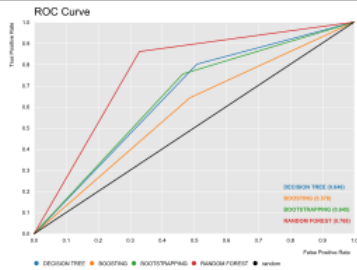
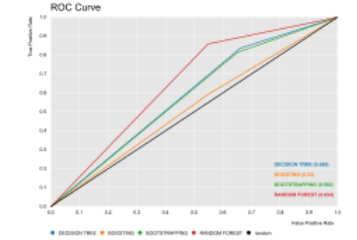
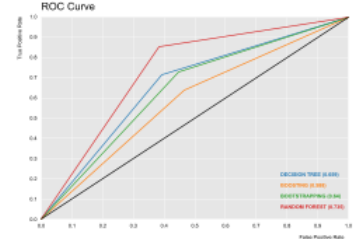
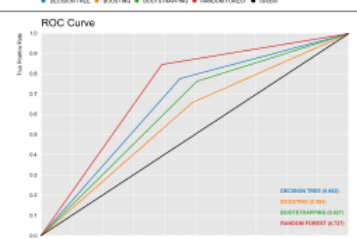
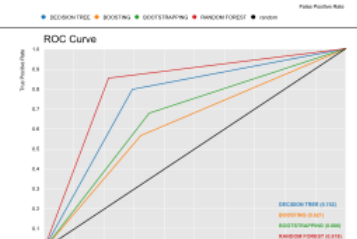




Cuando se consigue ejecutar el modelo de cada archivo, se finalizará el bucle, el cual, como el modelo tiene de salida una imagen transformada a tabla, devolverá la gráfica deseada de cada uno con la media de variables, y este, cuando termine, generará una tabla con dichos finales de iteraciones, y por tanto con las 5 gráficas y 5 medias.

Cuando tengamos las 5 gráficas de Curva ROC, se establece en el mismo orden de iteración de archivos, un atributo llamado YEAR, el cual tiene el nombre que se le querrá asignar como ID a cada tabla. Para hacer dicha asignación, después de unir las columnas, se llamará al nodo RowID para decir que los valores de la tabla se asignen ordenadamente por iteración, por tanto, por año.

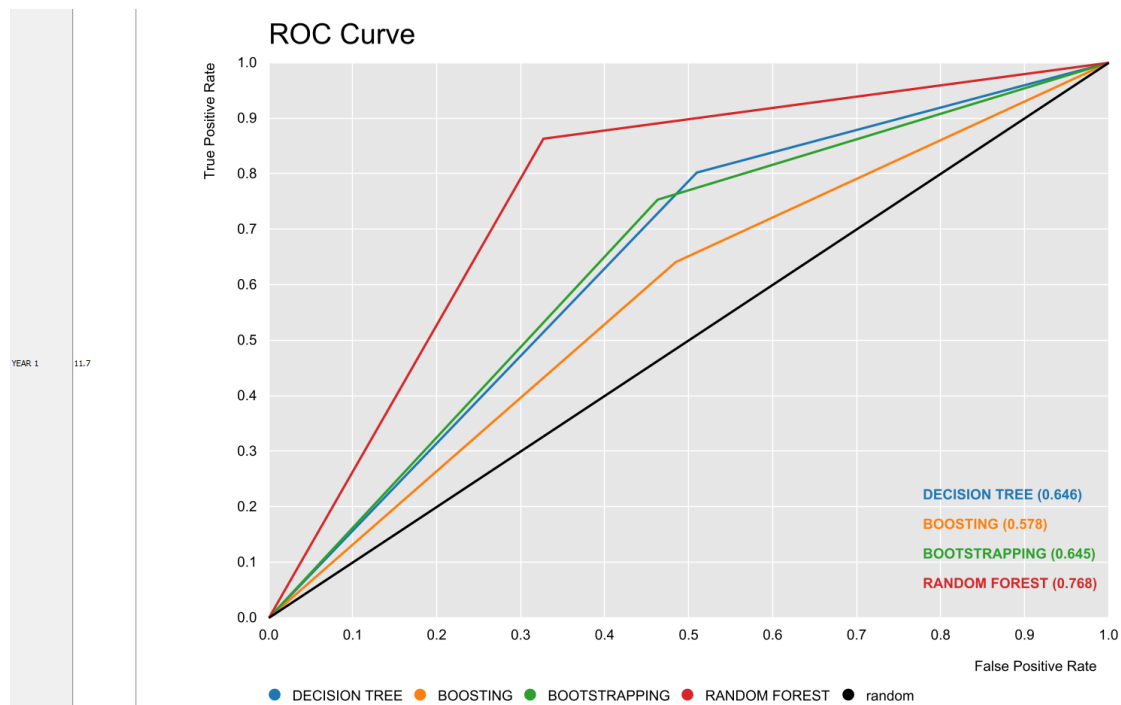
El resultado es el siguiente:

Row ID	Variables	Image
YEAR 1	11.7	
YEAR 2	9.8	
YEAR 3	14.002	
YEAR 4	11.2	
YEAR 5	11.2	



Mostradas las gráficas y sus variables de media, se procede a la parte de evaluación de los métodos para los datos proporcionadas, a nivel de efectividad. Para ello, iremos gráfica a gráfica. Hay que indicar primero que siempre serán los mismos dos métodos mencionados en este documento, usando para *Random Forest* la curva roja, para *Bootstrapping* el color verde, para *Decision Tree* el color azul y para *Boosting* el color naranja. El porcentaje de acierto es sobre las empresas que no entran en quiebra.

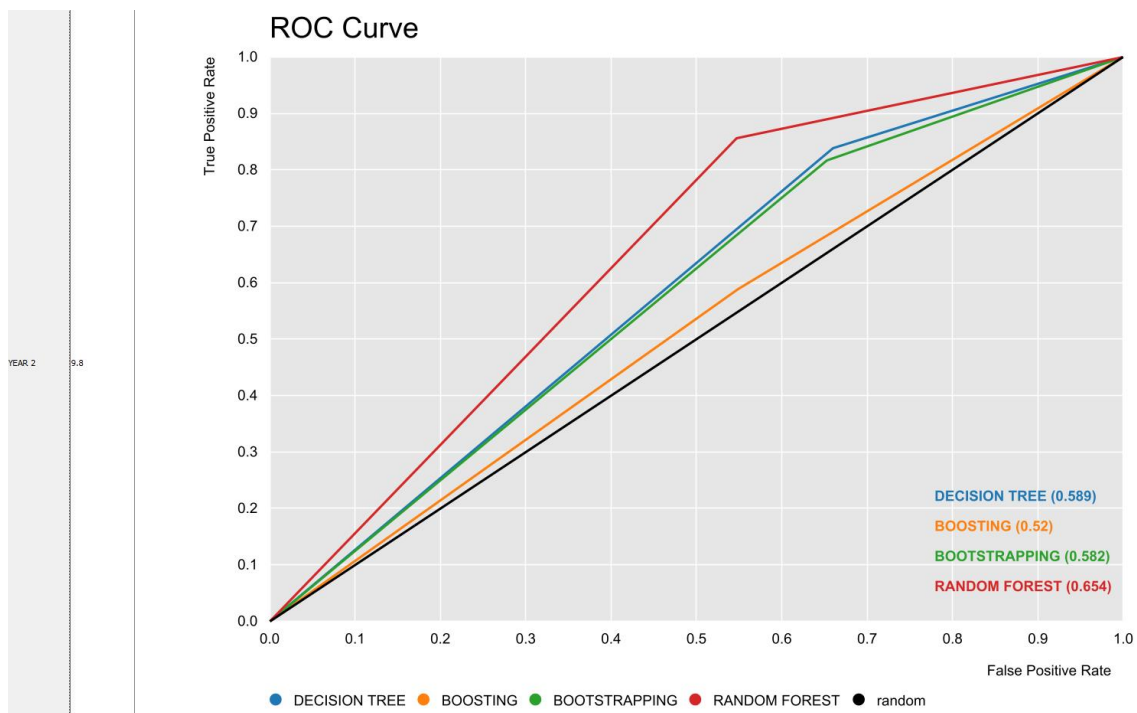
Año 1



En este caso, viendo tanto el valor numérico como las formas de la curva, para el método de predicción Random Forest la predicción acierta el 0.768 sobre 1, para Decision Tree, la predicción acierta el 0.646 sobre 1, para Bootstrapping, la predicción acierta el 0.645 sobre 1 y para Boosting, la predicción acierta el 0.578 sobre 1. En este caso, Random Forest asegura más aciertos, por tanto, es mejor método predictivo de los cuatro. En este archivo, se han usado de media 11.7 variables.



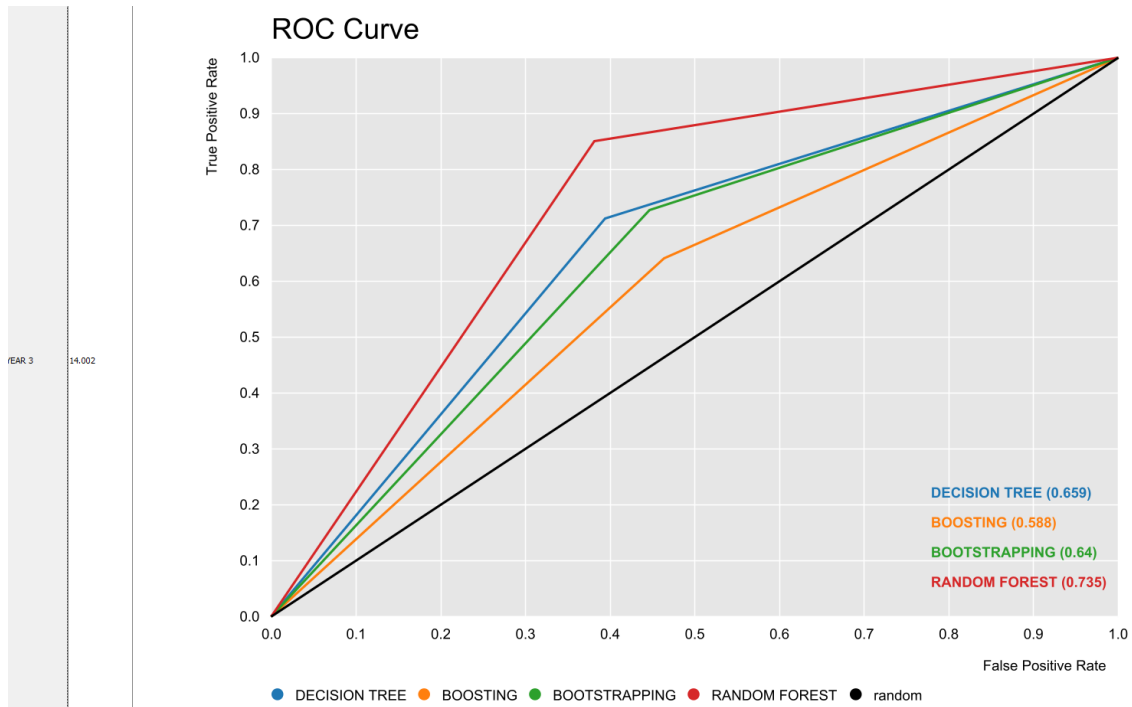
Año 2



En este caso, viendo tanto el valor numérico como las formas de la curva, para el método de predicción Random Forest la predicción acierta el 0.654 sobre 1, para Decision Tree, la predicción acierta el 0.589 sobre 1, para Bootstrapping, la predicción acierta el 0.582 sobre 1 y para Boosting, la predicción acierta el 0.52 sobre 1. En este caso, Random Forest asegura más aciertos, por tanto, es mejor método predictivo de los cuatro. En este archivo, se han usado de media 9.8 variables.



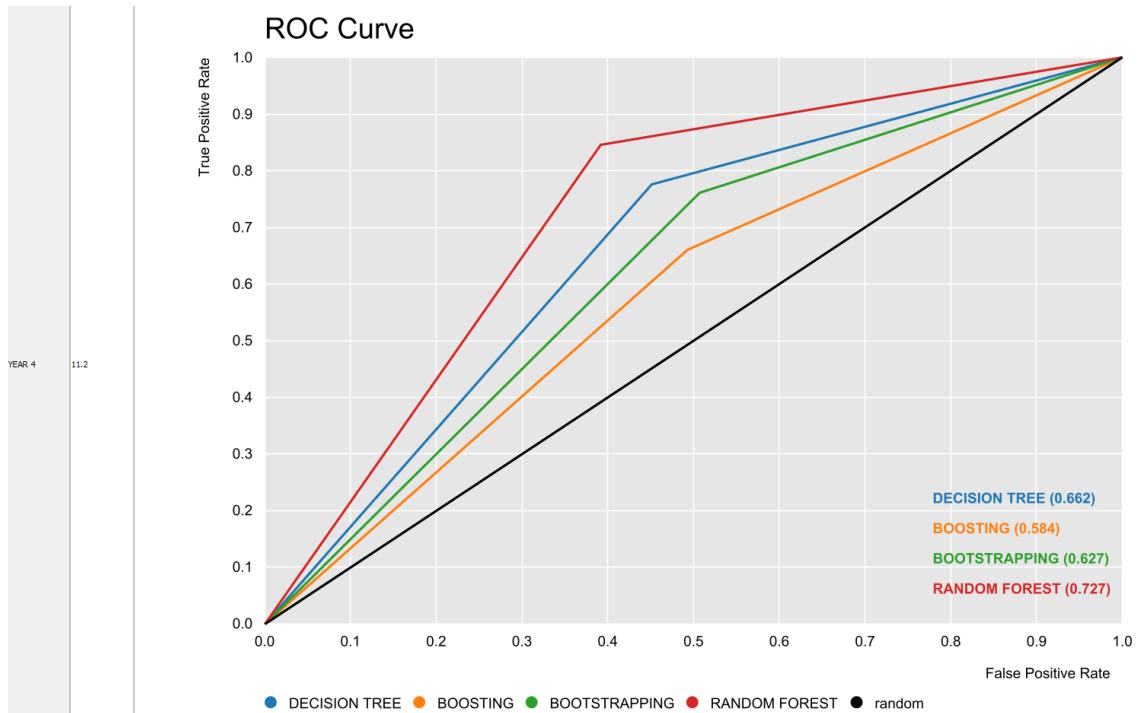
Año 3



En este caso, viendo tanto el valor numérico como las formas de la curva, para el método de predicción Random Forest la predicción acierta el 0.735 sobre 1, para Decision Tree, la predicción acierta el 0.659 sobre 1, para Bootstrapping, la predicción acierta el 0.64 sobre 1 y para Boosting, la predicción acierta el 0.588 sobre 1. En este caso, Random Forest asegura más aciertos, por tanto, es mejor método predictivo de los cuatro. En este archivo, se han usado de media 14.002 variables.



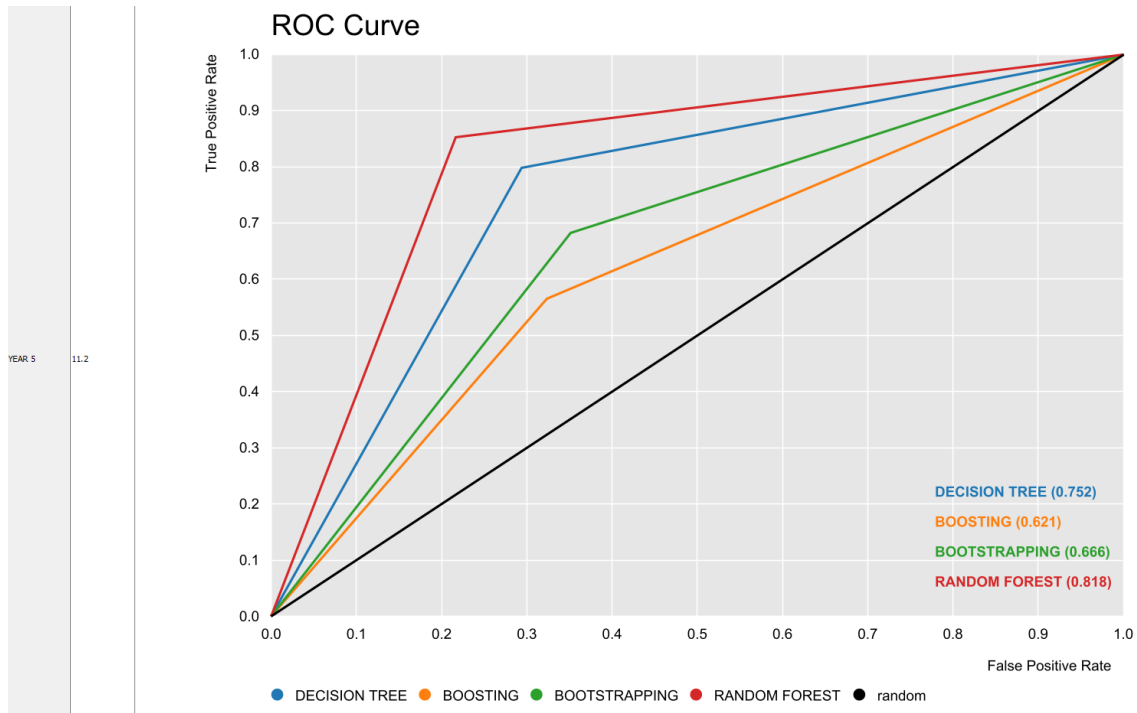
Año 4



En este caso, viendo tanto el valor numérico como las formas de la curva, para el método de predicción Random Forest la predicción acierta el 0.727 sobre 1, para Decision Tree, la predicción acierta el 0.662 sobre 1, para Bootstrapping, la predicción acierta el 0.627 sobre 1 y para Boosting, la predicción acierta el 0.584 sobre 1. En este caso, Random Forest asegura más aciertos, por tanto, es mejor método predictivo de los cuatro. En este archivo, se han usado de media 11.2 variables.



Año 5



En este caso, viendo tanto el valor numérico como las formas de la curva, para el método de predicción Random Forest la predicción acierta el 0.818 sobre 1, para Decision Tree, la predicción acierta el 0.752 sobre 1, para Bootstrapping, la predicción acierta el 0.666 sobre 1 y para Boosting, la predicción acierta el 0.621 sobre 1. En este caso, Random Forest asegura más aciertos, por tanto, es mejor método predictivo de los cuatro. En este archivo, se han usado de media 11.2 variables.

Como conclusión, viendo que, para los 5 conjuntos de datos, el método pensado y ejecutado de predicción con validación de 10 grupos y modelo de predicción de Random Forest es superior, ya sea en menor o mayor medida, por lo tanto, para casos nuevos, independientemente de la casuística temporal, aplicar la predicción de quiebra o no de empresas con los mismos atributos.



Evaluación

Para tener en cuenta también a la hora de hacer el modelo, conseguir datos nuevos, eficacia y eficiencia del de la ejecución del modelo, se tienen en consideración los siguientes aspectos.

A nivel de otras metodologías de predicción, no se han optado por diferentes opciones ya que el modelo de la entrega se limita a los especificados en el enunciado.

A nivel de los valores significativos de los atributos, se ha aplicado una selección de variables especificada en el enunciado, por tanto, no se han practicado más que la distinción entre Wrapper en contexto Forward y Backward que, por ensayo/error y la explicación de eficacia/eficiencia en un apartado anterior, se decantó por Forward.

A nivel eficacia, se pueden comprobar que no solo es capaz de predecir los modelos, si no que es capaz de acertar un gran número de veces, alcanzando más del 80% de aciertos, y bajando como mucho al 65% de aciertos. Por lo tanto, rondando el 73% de efectividad en las predicciones, a valoración de la empresa si el modelo de predicción es lo suficientemente alto como para comprobar internamente si está tomando decisiones correctas para cambiar o no los valores de los atributos.

A nivel eficiencia, se ha cronometrado el tiempo de ejecución del modelo, dadas las condiciones iniciales de 5 archivos con cada uno su cantidad de empresas. El tiempo de ejecución del modelo ha sido aproximadamente de unos 10 minutos. Contando que aproximadamente, cada archivo cuente con gran parte del tiempo de preprocesamiento. Desde el punto de vista de rendimiento, se considera, según la configuración del programa Knime, puede precisar de más memoria para agilizar el proceso, pudiendo ralentizar por momentos el sistema dependiendo de las características de este.

A consideración de las empresas el tiempo de ejecución y el rendimiento del modelo por cada empresa, pero por lo dado en dicho tiempo, se considera un trabajo eficiente.

Como conclusión dados los puntos redactados, la relación efectividad/eficiencia se declara como alta, dado el nivel de aciertos en la predicción, tiempo de ejecución y nivel de rendimiento durante las predicciones. A valorar resultados dados nuevos valores de otras empresas.



Despliegue

Para la entrega de este proyecto, se proporcionará este documento en PDF de nombre GarciaGonzalezAlberto77873658M.pdf como memoria de la actividad propuesta y el proyecto ejecutado en KNIME llamado Actividad2.knwf. Ambos archivos estarán dentro de un archivo comprimido de nombre GarciaGonzalezAlberto77873658M.zip. El proyecto cuenta con pequeñas descripciones en cada uno de los nodos indicando la razón de uso.

Para las próximas entregas, se tendrán en cuenta las explicaciones y conclusiones detalladas en este documento y se trabajará sobre el mismo proyecto en caso de tener que aplicar cambios o actualizaciones.

Dada la entrega de los archivos citados, se podrá visualizar el PDF con cualquier de este tipo de archivo. Para visualizar el proyecto, se debe hacer, desde Knime, la importación del archivo del proyecto mencionado.