UNIVERSIDAD DE GRANADA

MÁSTER EN ESTADÍSTICA APLICADA

TÉCNICAS DE MINERÍA DE DATOS APLICADAS AL MARKETING ANALÍTICO PARA PREDECIR EL COMPORTAMIENTO DE LOS CONSUMIDORES

TRABAJO FIN DE MÁSTER

REALIZADO POR:

Adrián Alejandro Miranda Albarrán

TUTOR:

Juan Antonio Maldonado Jurado

2020-2021

ABSTRACT: *Marketing Analytics is a statistical and marketing-based discipline that works with data collection and analysis, to understand patterns, evaluate strategies, and make better decisions. Churn prediction means detecting which customers are likely to unsubscribe from a service. This is a critical prediction for many companies because acquiring new customers often costs more than retaining existing ones. The objective of this work is to build a classification model, based on the historical data of a telephone company, which will make it possible to predict potential customer defections. This model will allow to act proactively in the retention of clients and improve the services provided. A data set with 21 variables that influence customer attrition was used. A dependent variable (churn) was used, which is an identifier that determines whether the customer left (churn = 1) or not (churn = 1) the company's service. For the construction of this model, several algorithms were used such as decision trees, support vector machine (SVM) and neural networks with different test methods. The entire implementation was done with KNIME Analytics Platform. The results reached an accuracy of 93.79% for the decision tree classifier, which indicates that decision trees turn out to be an attractive alternative to develop prediction models of customer churn in this type of data.*
Keywords: Data mining, marketing analytics, machine learning, churn prediction, KNIME Analytics.

RESUMEN: *El marketing analítico es una disciplina basada en estadística y mercadotecnia que trabaja con la recolección y análisis de datos, para entender patrones, evaluar estrategias y tomar mejores decisiones. La predicción de abandono (churn prediction) significa detectar qué clientes es probable que cancelen una suscripción a un servicio. Esto es una predicción fundamental para muchas empresas porque a menudo adquirir nuevos clientes cuesta más que retener a los existentes. El objetivo de este trabajo es construir un modelo de clasificación, a partir de los datos históricos de una compañía telefónica, que permitirá predecir potenciales deserciones de clientes. Los resultados del modelo permitirán actuar de forma proactiva en la retención de clientes y mejorar los servicios prestados. Se utilizó un conjunto de datos con 21 variables que influyen en la deserción de los clientes. Se utilizó una variable dependiente (churn), que es un identificador que determina si el cliente abandonó (churn=1) o no (churn=1) el servicio de la empresa. Para la construcción de este modelo se utilizaron varios algoritmos como árboles de decisión, máquina de vectores de soporte (SVM) y redes neuronales con diferentes métodos de prueba. Toda la implementación se realizó con el paquete de software KNIME Analytics Platform. Los resultados alcanzaron una precisión del 93.79% para el clasificador árbol de decisión, lo que indica que los árboles de decisión resultan ser una alternativa atractiva para desarrollar modelos de predicción de deserción de clientes en este tipo de datos.*
Palabras clave: Minería de datos, marketing analítico, aprendizaje automático, predicción de abandono de clientes, KNIME Analytics.

Agradecimientos

Un agradecimiento especial a:

Mi tutor Juan Antonio Maldonado Jurado, gracias por siempre tener una pronta respuesta a pesar de mis tantos contratiempos. Enseñar es tocar una vida para siempre, directa o indirectamente.

A todas las personas que han pasado, que están, y que pasarán por mi vida. Coincidir con ustedes en esta vida es un enorme privilegio, pero también una enorme responsabilidad de testificar acerca de la Verdad. Siempre han estado en mis oraciones.

Y por sobre todos y todo, a tí, quien me da la fuerza, la vida y la esperanza de que algún día te veré cara a cara. Tú posees la verdad y la verdad está en tí, porque tuyo es el conocimiento, la sabiduría y la ciencia. Y el fin y principio de todo hombre eres tú.

Tenías razón, la vida del hombre es tan corta, casi como un suspiro.

Soli Deo Gloria

Adrián Alejandro Miranda Albarrán.

ÍNDICE GENERAL

Contenido **N° Página**

Caso práctico de aplicación de minería de datos a una compañía telefónica para predecir si un cliente renovará su contrato o no.

1. MINERÍA DE DATOS

INTRODUCCIÓN

1.1 Breve historia de la minería de datos

El proceso de hurgar en los datos para descubrir conexiones ocultas y predecir tendencias futuras tiene una larga historia. En la década de 1960, los estadísticos utilizaron los términos de *data fishing* o *data dredging* para referirse a lo que consideraban la mala práctica de analizar datos sin una hipótesis a priori. Conocido algunas veces como "descubrimiento de conocimientos en bases de datos", el término "minería de datos" no se acuño sino hasta la década de 1990. El término minería de datos apareció en la comunidad de profesionales que se dedicaba principalmente a las bases de datos, generalmente con connotaciones positivas. Otros términos utilizados incluyeron, recolección de información, descubrimiento de información, extracción de conocimiento, etc.

La extracción manual de patrones a partir de datos se ha producido durante siglos. Los primeros métodos para identificar patrones en los datos incluyen el teorema de Bayes (1700) y el análisis de regresión (1800). La proliferación y el poder cada vez mayor de la tecnología informática han aumentado drásticamente la capacidad de recopilación, almacenamiento y manipulación de datos. A medida que los conjuntos de datos han crecido en tamaño y complejidad, el análisis de datos se ha incrementado y automatizado con la ayuda de otros descubrimientos en informática, especialmente en el campo del aprendizaje automático, como las redes neuronales, análisis de conglomerados, algoritmos genéticos (década de 1950), árboles y reglas de decisión (década de 1960) y máquinas de vectores de soporte o SVM (década de 1990). La minería de datos es el proceso de aplicar estos métodos con la intención de descubrir patrones ocultos en grandes conjuntos de datos.

Gregory Piatetsky-Shapiro acuñó el término "*knowledge discovery in databases (KDD)"* y este término se hizo muy popular en la comunidad de inteligencia artificial y aprendizaje automático. Sin embargo, el término minería de datos se hizo más popular en el ámbito empresarial y de negocios. **Actualmente, los términos minería de datos y descubrimiento de conocimiento (KDD) se utilizan indistintamente.** En la comunidad académica, los principales foros de investigación comenzaron en 1995 cuando se llevó a cabo la Primera Conferencia Internacional sobre Minería de Datos y Descubrimiento del Conocimiento (KDD-95) en Montreal, Canadá bajo el patrocinio de la AAAI (Asociación para el Avance de la Inteligencia Artificial). Un año después, en 1996, Usama Fayyad lanzó la revista llamada *Data Mining and Knowledge Discovery* como su editor en jefe y fundador. Más tarde la conferencia *KDD International* se convirtió en la principal conferencia de la más alta calidad en minería de datos. Actualmente, la revista *Data Mining and Knowledge Discovery* es la principal revista de investigación del campo.

1.2 ¿Qué es la minería de datos?

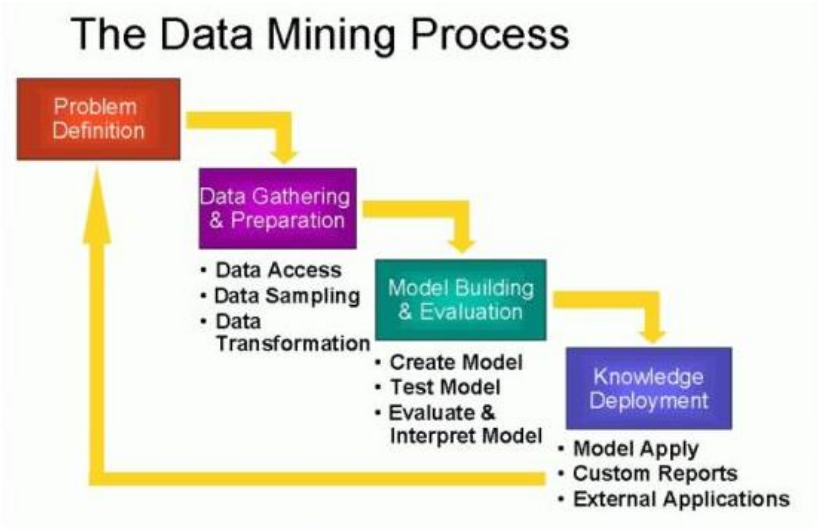
La minería de datos, también conocida como descubrimiento de conocimiento en bases de datos o *knowledge discovery in databases (KDD)*, es el proceso de descubrir patrones y otra información valiosa en grandes conjuntos de datos (*datasets*). Dada la evolución de la tecnología de almacenamiento de datos (*data warehousing*) y el crecimiento del big data, la adopción de técnicas de minería de datos se ha acelerado rápidamente durante las últimas dos décadas, ayudando a las empresas a transformar sus datos crudos en conocimiento útil. La minería de datos ha mejorado la toma de decisiones en las empresas u organizaciones a través de los análisis de datos detallados. Estos análisis se utilizan para organizar y filtrar datos aplicando una amplia variedad de técnicas para después mostrar la información más relevante. Las empresas pueden utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con los clientes, reducir riesgos, detectar fraudes, analizar el comportamiento de sus consumidores, evitar ataques cibernéticos y más.

La minería de datos es un campo interdisciplinario de la **informática** y la **estadística** con el objetivo general de extraer información (con métodos inteligentes) de un conjunto de datos y transformar la información en una estructura comprensible para su uso posterior. La minería de datos utiliza la **estadística** para estudiar numéricamente las relaciones entre los datos y las áreas de la **informática** que generalmente utiliza son los sistemas de bases de datos, la inteligencia artificial y el aprendizaje automático (*machine learning)* para desarrollar algoritmos que pueden aprender de datos para hacer predicciones. La minería de datos muchas veces se considera un subcampo de lo que hoy se conoce como ciencia de datos.

La diferencia entre el análisis de datos y la minería de datos es que el análisis de datos se utiliza para probar modelos e hipótesis en el conjunto de datos, por ejemplo, analizar la efectividad de una campaña de marketing, independientemente de la cantidad de datos; por el contrario, la minería de datos utiliza modelos estadísticos y de aprendizaje automático para descubrir patrones clandestinos u ocultos en un gran volumen de datos.

1.3 Fases de la minería de datos

El proceso de minería de datos implica una serie de pasos, desde la recopilación de datos hasta la visualización de la información relevante que fue extraída de los grandes conjuntos de datos. Las técnicas de minería de datos se utilizan para generar **descripciones** y **predicciones** sobre un conjunto de datos. Los científicos de datos y estadísticos describen los datos a través de la observación de patrones, asociaciones y correlaciones. También clasifican y agrupan datos a través de métodos de clasificación y regresión e identifican valores atípicos. La minería de datos generalmente consta de **cuatro fases principales:** (1) establecimiento de objetivos o definición del problema, (2) recopilación y preparación de datos, (3) aplicación de técnicas y algoritmos de minería de datos y (4) evaluación de resultados.



1. **Establecimiento de objetivos o definición del problema:** esta puede ser la parte más difícil del proceso de minería de datos y muchas organizaciones dedican muy poco tiempo a este importante paso. Los científicos de datos y las partes interesadas del negocio deben trabajar juntos para definir el problema empresarial, definir la estrategia y establecer los parámetros de los datos para un proyecto determinado. Es posible que los analistas también necesiten realizar una investigación adicional para comprender el contexto empresarial o del negocio de manera adecuada.

2. **Recopilación y preparación de datos:** una vez que se define el alcance del problema, es más fácil para los científicos de datos identificar qué conjunto de datos ayudará a responder las preguntas formuladas por la empresa durante la fase de definición del problema. Una vez que se recopilen los datos relevantes, los datos se limpiarán para identificar problemas de calidad en los datos como duplicados y valores atípicos. Es importante familiarizarse con los datos para identificar y retener los predictores más importantes y así garantizar una precisión óptima dentro de cualquier modelo.

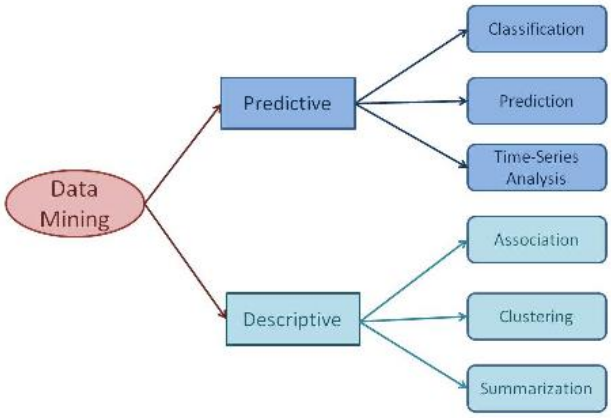
3. **Aplicación de minería de datos y construcción de modelos:** Dependiendo del tipo de análisis, los científicos de datos pueden investigar cualquier relación de datos interesante, como patrones secuenciales, reglas de asociación o correlaciones. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos sobre la forma de los datos. Por ejemplo, los algoritmos de *deep learning* se pueden aplicar para clasificar o para agrupar un conjunto de datos según los datos disponibles. Si los datos de entrada están etiquetados (es decir, aprendizaje supervisado), se puede usar un modelo de clasificación para categorizar los datos o, alternativamente, se puede aplicar una regresión para predecir la probabilidad de una asignación en particular. Si el conjunto de datos no está etiquetado (es decir, aprendizaje no supervisado), los puntos de datos individuales en el conjunto de entrenamiento se comparan entre sí para descubrir similitudes agrupándolos según ciertas características.

4. **Evaluación de resultados e implementación del conocimiento:** Una vez que las técnicas de minería de datos se han aplicado, se evalúa minuciosamente el modelo y se revisan los pasos ejecutados con los que se construyó el modelo, para asegurarse de que se cumplan adecuadamente los objetivos de la empresa. Se determina si hay algún objetivo, definido anteriormente por la empresa, que no se haya considerado y se toma una decisión sobre el uso de los resultados tras la aplicación de la minería de datos. Al finalizar, los resultados deben ser válidos, novedosos, útiles y comprensibles. Cuando se cumple este criterio, las organizaciones pueden utilizar este conocimiento para implementar nuevas estrategias, logrando sus objetivos previstos.

La minería de datos es iterativa. Un proceso de minería de datos continúa después de que se implementa una solución. Las lecciones aprendidas durante el proceso pueden generar nuevas preguntas comerciales. El cambio de datos puede requerir nuevos modelos. Los procesos de minería de datos posteriores se benefician siempre de las experiencias anteriores.

1.4 Técnicas de minería de datos

Las técnicas de minería de datos se pueden dividir en dos áreas o metas principales; pueden **describir** el conjunto de datos o pueden **predecir** resultados mediante el uso de algoritmos de aprendizaje automático.
Los métodos **descriptivos** también se conocen como métodos o técnicas de **aprendizaje no supervisado** y los métodos **predictivos** se conocen como **aprendizaje supervisado.**



Técnicas de aprendizaje no supervisadas o descriptivas

Las técnicas descriptivas o de aprendizaje no supervisado son técnicas que no involucran a una persona. Es una rama del aprendizaje automático que aprende de los datos de prueba (*test data*). Los modelos descriptivos reconocen los diseños o relaciones en los datos y descubren las propiedades de los datos estudiados. Algunos ejemplos de técnicas descriptivas son:

• **Agrupación o *Clustering*.** Esta técnica es muy similar a la clasificación (técnica predictiva) agrupando los datos en función de sus similitudes. Los grupos de *clústeres* están menos estructurados que los grupos de clasificación, lo que los convierte en una opción más sencilla para la minería de datos. En un supermercado, por ejemplo, un grupo de clúster simple podría ser los alimentos y los artículos no alimentarios podría ser otro clúster. La agrupación es diferente a la clasificación ya que no tiene clases predefinidas. Los algoritmos de agrupación en clústeres descubren colecciones de datos de modo que los objetos del mismo clúster son más idénticos entre sí que otros grupos.

• **Reglas de asociación**. La asociación en la minería de datos tiene que ver con el seguimiento de patrones, específicamente basados en variables vinculadas. En el ejemplo del supermercado, esto puede significar que muchos clientes que compran un artículo específico también pueden comprar un segundo artículo relacionado. Así es como las tiendas pueden saber cómo agrupar ciertos alimentos, o en las compras en línea pueden mostrar la sección que sugiere "La gente también compró esto".

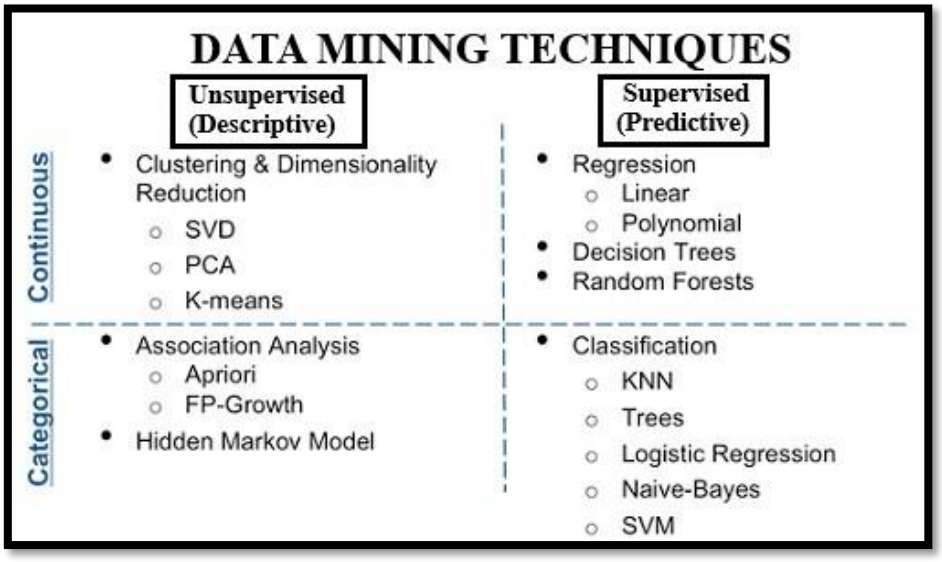
• **Detección de anomalías / valores atípicos**. Para muchos casos de minería de datos, solo con identificar el patrón general en los datos no es todo lo que se necesita, también se deben identificar y comprender los valores atípicos en sus datos. Por ejemplo, en el supermercado, si la mayoría de los compradores son mujeres, pero una semana de abril son principalmente hombres, se tiene que investigar ese valor atípico y comprender qué hay detrás.

Técnicas de aprendizaje supervisado o predictivas

El aprendizaje supervisado involucra a una persona que ayuda a aprender. El aprendizaje predice un resultado basado en ciertos criterios. El modelado predictivo es un proceso que utiliza la minería de datos y la probabilidad para pronosticar los resultados. Los modelos predictivos se utilizan para obtener resultados. Cuando se simulan los resultados, se construye un modelo estadístico. Algunos ejemplos de técnicas predictivas son:

• **Clasificación.** Esta técnica de minería de datos es compleja y utiliza atributos de datos para establecer categorías diferenciadas para posteriormente sacar conclusiones. La clasificación consta de dos pasos: entrenamiento y pruebas o *testing*. La precisión del modelo de clasificación depende del grado en que las reglas de clasificación sean verdaderas, las cuales se estiman mediante los datos de prueba (*test data*). En el ejemplo del supermercado, se puede usar la clasificación para agrupar los tipos de alimentos que los clientes están comprando, como verduras, carne, artículos de panadería, etc. Estas clasificaciones ayudan a la tienda a aprender aún más sobre los clientes, los productos, etc.

• **Análisis de regresión.** La regresión se utiliza para planificar y modelar, identificando la probabilidad de una variable específica. Por ejemplo, el supermercado puede hacer una proyección de sus precios en función de la disponibilidad, la demanda de los consumidores y su competencia. La regresión también ayuda a identificar la relación que hay entre las variables en un conjunto de datos.



SOFTWARE DISPONIBLE PARA MINERÍA DE DATOS

1.5 Software para minería de datos

El software de minería de datos se refiere al software que permite a las empresas y a otros usuarios extraer datos utilizables dentro de un conjunto de datos con el fin de encontrar correlaciones, patrones y anomalías. Los resultados del proceso de minería de datos ayudan a las empresas a tomar decisiones estratégicas. Las técnicas utilizadas por el software de minería de datos incluyen análisis estadísticos, algoritmos específicos, aprendizaje automático, estadística aplicada a bases de datos e inteligencia artificial.
En términos simples, las aplicaciones de minería de datos ayudan a las empresas a obtener conocimiento de grandes volúmenes de datos y a transformarlos en información procesable.

Hay muchos sistemas de minería de datos y algunos de ellos ofrecen funcionalidades más avanzadas o utilizan diferentes métodos para procesar información y validar resultados. Por lo tanto, la elección del software de minería de datos dependerá de las preferencias o necesidades del usuario. En esta sección presentamos algunos de los paquetes de minería de datos más utilizados tanto en la industria como en la academia.

1.6 Software de licencia comercial

1. Sisense Analytics

Sisense simplifica la analítica empresarial para datos complejos. Impulsado por tecnologías In-Chip y Single Stack, Sisense ofrece rendimiento, agilidad y valor inigualables, eliminando gran parte de la costosa preparación de datos que tradicionalmente se necesitaba con herramientas de análisis empresarial y proporcionando una herramienta única y completa para analizar y visualizar grandes conjuntos de datos. La experiencia de Sisense en datos complejos incluye tanto grandes conjuntos de datos como datos provenientes de múltiples fuentes. Sisense aprovecha la analítica en chip para mejorar drásticamente el acceso de los usuarios comerciales a la analítica avanzada en máquinas básicas de bajo costo sin la necesidad de herramientas especiales de almacenamiento de datos o personal de TI dedicado.

Principales características:

- Simplemente arrastrar y soltar para unir datos de múltiples fuentes.
- Crear tableros interactivos sin conocimientos técnicos.
- Compartir tableros interactivos en la nube.
- Brinda a los usuarios la libertad de consultar datos en tiempo real.

2. Alteryx Analytics

Alteryx Analytics es una suite de programas que incluye: Alteryx Designer, Alteryx Server y Alteryx Analytics Gallery. La suite de Alteryx permite combinar datos internos, datos de terceros y datos basados en la nube. También permite crear potentes aplicaciones de análisis espacial y predictivo basados en R sin necesidad de programación y permite compartir información detallada sobre los datos con los responsables de las tomas de decisiones empresariales. Permite también aplicar técnicas de modelado predictivo, como regresión logística o árboles de decisión, técnicas de agrupamiento, como el clúster de centroide K y análisis de componentes principales, técnicas de investigación de datos, como diagramas de dispersión y análisis de asociación. Alteryx Analytics puede implementar todas estas aplicaciones sin necesidad de programación en código.

Principales características:

- Solución en tiempo de ejecución basada en la nube.
- Se puede crear configuraciones de seguridad para cada aplicación analítica.
- Se puede personalizar las aplicaciones analíticas desde la nube.
- Prepara y combina todos los datos en un flujo de trabajo.
- Ejecuta análisis predictivos, espaciales y estadísticos sin necesidad de código.
- Puede exportar los resultados a los formatos más conocidos y utilizados en el mercado.

3. SAS Enterprise Miner

SAS Enterprise Miner es una solución para crear modelos con gran precisión, ya sean descriptivos o predictivos, sobre grandes volúmenes de datos provenientes de diferentes fuentes. SAS es el líder en software y servicios de análisis empresarial y es el mayor proveedor independiente en el mercado de inteligencia empresarial. Desde 1976 SAS, a través de soluciones innovadoras, ha ayudo a los clientes a mejorar la toma de decisiones utilizando sus datos crudos. SAS se desarrolló en la Universidad Estatal de Carolina del Norte desde 1966 hasta 1976, después de esto se fundó el ya famoso y prestigioso SAS Institute. SAS Enterprise Miner ofrece infinidad de características y funcionalidades.

Principales características:

- Interfaz gráfica de usuario fácil de usar.
- Preparación y exploración de datos complejos.
- Modelado descriptivo y predictivo avanzado.
- Integración de código abierto con R y SAS.

- Manera rápida y fácil para que los usuarios generen sus propios modelos.
- Permite la comparación entre diferentes modelos en tiempo real.
- Permite la integración de otros paquetes de SAS dentro de un flujo de proceso.
- Opción de implementación en la nube.

4. Oracle Data Mining - ODM

Se trata del sistema de análisis de datos avanzado de Oracle. Las empresas líderes en el mercado lo utilizan para maximizar el potencial de sus datos y para realizar predicciones precisas. El sistema funciona con poderosos algoritmos de datos, además, identifica tanto anomalías como oportunidades de venta cruzada y permite a los usuarios aplicar un modelo predictivo diferente en función de sus necesidades. Personaliza los perfiles de los clientes de la forma deseada. Oracle Data Mining ODM proporciona una potente funcionalidad de minería de datos y permite a los usuarios descubrir nuevos conocimientos sobre los datos ocultos. Oracle Data Mining tiene varios algoritmos de minería de datos y análisis de datos. Estos algoritmos proporcionan medios para la creación, manipulación, aplicación, prueba y despliegue de modelos. Se utilizan para clasificación, predicción, regresión, asociaciones, detección de anomalías, extracción de características y análisis especializados. Los modelos se implementan en el kernel de la base de datos de Oracle y se almacenan como objetos en la base de datos.

Principales características:

- Seguridad avanzada en datos.
- Base de datos en memoria.
- Procesamiento analítico en línea.
- Aplicaciones en tiempo real.

5. Salford Systems SPM - Minitab

El paquete de software SPM Salford Predictive Modeler es una plataforma de análisis y minería de datos de gran precisión para crear modelos predictivos, descriptivos y analíticos a partir de bases de datos de cualquier tamaño o complejidad. Este conjunto de herramientas de minería de datos incluye los productos de Salford Systems de CART, MARS, TreeNet y Random Forests. Salford Systems SPM incluye más de 70 escenarios automatizados y 4 versiones de productos diferentes entre las que los usuarios pueden elegir según lo que necesiten para la creación de sus modelos.

Principales características:

- Creación automática de registros de comandos.
- Opción para guardar conjuntos de datos procesados en varios formatos.
- Creación automática de indicadores de valor nulos (*missing values*).
- Cálculo de correlación de más de 10 tipos diferentes de correlación.
- Funciones de automatización.
- Indicadores automáticos para la estabilidad del modelo.

6. IBM SPSS Modeler / Clementine

SPSS Modeler / Clementine es un conjunto de herramientas de minería de datos que tiene como objetivo permitir que los usuarios realicen su propia minería de datos. Clementine tiene una programación visual y una interfaz de flujo de datos, lo que simplifica el proceso de extracción de datos. Las aplicaciones de Clementine incluyen segmentación, creación de perfiles de clientes para empresas de marketing, detección de fraudes, calificación crediticia, previsión de carga para empresas de servicios públicos y predicción de beneficios para minoristas. SPSS Clementine fue una de las primeras herramientas de minería de datos de propósito general. SPSS Modeler / Clementine tiene uno de los paquetes de minería de datos más populares en el mercado.

Principales características:

- Soporte para muchas fuentes de datos.
- Fácil implementación de modelos de minería de datos.
- Preparación automática de datos.
- Potente motor de gráficos.
- Flujos de análisis visual.
- Modelado automatizado.
- Amplia gama de algoritmos.
- Se puede emplear lenguajes como R y Python para ampliar las capacidades de modelado.
- Analítica de texto.
- Analítica geoespacial.

1.7 Software de licencia gratuita

1. RapidMiner Studio

RapidMiner Studio es un entorno de diseño visual que proporciona un entorno integrado para la preparación de datos, aprendizaje automático, *deep learning*, minería de texto y análisis predictivo. Es uno de los principales sistemas de código abierto para la minería de datos. El programa está escrito íntegramente en lenguaje de programación Java. Proporciona una biblioteca extensa de algoritmos de aprendizaje automático, funciones de preparación y exploración de datos y herramientas de validación de modelos. Los usuarios pueden reutilizar fácilmente el código R y Python existente y agregar nuevos algoritmos desarrollados en estos lenguajes. RapidMiner se utiliza para aplicaciones comerciales e industriales, así como para investigación y entornos educativos.

Principales características:

- Plataforma unificada.
- Diseño visual del flujo de trabajo.
- Amplia funcionalidad.
- Permite la implementación de código abierto.
- Amplia conectividad.

2. Dataiku DSS

Dataiku DSS es una plataforma colaborativa de ciencia de datos que permite a los usuarios explorar, crear prototipos, construir y entregar sus propios productos de datos de manera más eficiente. Dataiku DSS proporciona una interfaz visual interactiva donde con un solo clic los usuarios pueden usar lenguajes como SQL para extraer datos, volver a ejecutar flujos de trabajo, visualizar resultados…etc. Dataiku DSS proporciona herramientas para elaborar borradores de preparación y modelado de datos en minutos.

Principales características:

- Amplia Conectividad.
- Bibliotecas con algoritmo de aprendizaje automático.
- Bibliotecas de visualización de datos.
- Flujo de trabajo visual.
- Permite colaboración con otros sistemas informáticos de minería de datos.

3. KNIME Analytics Platform

KNIME, Konstanz Information Miner, es una plataforma de análisis, informes e integración de datos de código abierto. KNIME integra varios componentes para el aprendizaje automático y para minería de datos a través de su concepto de canalización de datos modular y proporciona una interfaz gráfica de usuario que permite el ensamblaje de nodos para el preprocesamiento de datos, modelado y análisis, y visualización de datos. KNIME Collaborative Extensions y KNIME Analytics Platform incluye KNIME TeamSpace, KNIME Server Lite, KNIME WebPortal y KNIME Server. KNIME Analytics Platform proporciona más de 1000 rutinas de análisis de datos, ya sea de forma interna o mediante extensiones con R, Python y Weka. Knime permite aplicar técnicas como estadísticas univariadas y multivariadas, minería de datos, series de tiempo, procesamiento de imágenes, análisis web, análisis de texto y mucho más.

Principales características:

- Análisis de gran alcance.
- Plataforma gratuita.
- Más de 1000 módulos y creciendo.
- Conectores para todos los formatos de archivo y para las bases de datos más importantes.
- Soporte para una gran cantidad de tipos de datos: XML, JSON, imágenes, documentos y mucho más.
- Funciones matemáticas y estadísticas.
- Algoritmos avanzados de aprendizaje automático y predictivo.
- Control visual del flujo de trabajo.
- Combinación de herramientas como Python, R, SQL, Java, Weka y muchos más.
- Informes y visualizaciones interactivas de los datos.
- Análisis de sentimiento en las redes sociales.
- Conectores para los softwares de visualización más importantes como Tableau y PowerBi

4. Orange Data mining

Orange es una herramienta de análisis y visualización de datos de código abierto. Orange se desarrolla en el Laboratorio de Bioinformática de la Facultad de Ciencias de la Información y la Computación de la Universidad de Ljubljana en Eslovenia. La minería de datos se realiza mediante programación visual o secuencias de comandos de Python. La herramienta tiene componentes para aprendizaje automático, complementos para bioinformática y minería de texto y está repleta de funciones para el análisis de datos. Los scripts de Python pueden ejecutarse en una ventana de terminal, en entornos integrados como PyCharm y PythonWin, o en shells como iPython.

Principales características:

- Código abierto.
- Visualización de datos interactiva.
- Programación visual.
- Admite integración de gráficos e ilustraciones visuales externas.
- Admite complementos que amplían la funcionalidad.
- Para todos: principiantes y profesionales.
- Ejecuta análisis de datos simples y complejos.
- Crea gráficos complejos e interesantes.

5. R Software Environment

R es un entorno de software libre para gráficos y computación estadística. Se compila y se ejecuta en una amplia variedad de plataformas como UNIX, Windows y MacOS. R es un conjunto integrado de instalaciones de software para la manipulación, el cálculo y la visualización gráfica de datos. Algunas de las funcionalidades incluyen una instalación efectiva de manejo y almacenamiento de datos, un conjunto de operadores para cálculos en arreglos, en particular matrices, una colección grande, coherente e integrada de herramientas intermedias para análisis de datos, instalaciones gráficas para análisis de datos y visualización directamente en la computadora, y un lenguaje de programación bien desarrollado, simple y efectivo.

Principales características:

- Código abierto.
- Instalación rápida.
- Conjunto de operadores para cálculos en arreglos, vectores y matrices.
- Colección amplia, coherente e integrada de herramientas intermedias para el análisis de datos.
- Instalaciones gráficas para el análisis y visualización de datos.
- Lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario e instalaciones de entrada y salida.
- Proporciona una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento) y técnicas gráficas.
- Se ejecuta en una amplia variedad de plataformas: UNIX, Windows, MacOS.
- Software estadístico ampliamente utilizado.

6. Weka Data Mining

Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Los algoritmos se pueden aplicar directamente a un conjunto de datos o aplicarlos utilizando el propio código Java del usuario. Las características de Weka incluyen aprendizaje automático, minería de datos, preprocesamiento, clasificación, regresión, agrupamiento, reglas de asociación, selección de atributos, experimentos, flujo de trabajo y visualización. Weka está escrito en Java, desarrollado en la Universidad de Waikato en Nueva Zelanda. Todas las técnicas de Weka se basan en el supuesto de que los datos están disponibles como un solo archivo plano o relación, donde cada punto de datos se describe mediante un número fijo de atributos. Weka proporciona acceso a bases de datos SQL.

Principales características:

- Código abierto.
- Interfaz gráfica.
- Interfaz de línea de comando.
- Integración con Java para crear aplicaciones propias.
- Gran variedad de documentación y manuales en la red.
- Amplia variedad de algoritmos para utilizar.

7. Pyhton

Python es el lenguaje de programación más popular que ofrece la flexibilidad y el poder para que los programadores y científicos de datos realicen análisis de datos y apliquen algoritmos de aprendizaje automático. En los últimos años, Python se ha vuelto más popular para la minería de datos debido al aumento en la cantidad de bibliotecas de análisis de datos. Las bibliotecas de Python más utilizadas para el análisis de datos son Pandas, Matplotlib y NumPy.

Principales características:

- Herramienta de manipulación y análisis de datos de código abierto.
- Rápido, potente, flexible y fácil de aprender y de usar.
- Gran popularidad.
- Liberia extensas.
- Se puede integrar en casi todas las plataformas analíticas.
- Integra herramientas para visualización.
- Gran comunidad de soporte en línea.

2. KNIME ANALYTICS PLATFORM

2.1 Introducción a Knime Analytics Platform

KNIME Analytics Platform es un software de código abierto para crear aplicaciones y servicios de ciencia de datos. KNIME es intuitivo, flexible y continuamente está integrando nuevos desarrollos. KNIME hace que la comprensión de datos y el diseño de flujos de trabajo de ciencia de datos sean accesibles para todos.

KNIME Analytics Platform puede crear flujos de trabajo visuales con una interfaz gráfica de estilo intuitivo, de arrastrar y soltar, sin necesidad de codificación.
En la página oficial de KNIME se puede encontrar extensa documentación, manuales, cursos e incluso una comunidad de usuarios que han creado un foro de asistencia en línea llamado KNIME Forum. KNIME Forum está considerado como uno de los mejores foros de asistencia y soporte en línea para softwares de ciencia de datos. Lo mejor es que todo esto está al alcance del usuario de manera gratuita.

2.2 Instalación de Knime Analytics Platform

1. Para instalar KNIME Analytics Platform se tiene que ir a la página de descarga en el sitio web www.knime.com

2. La página de descarga muestra tres pestañas que se pueden abrir individualmente:

   - ***Regístrese para* recibir *ayuda y actualizaciones*:** aquí, opcionalmente, el usuario puede proporcionar información personal e inscribirse en una lista de correo para recibir las últimas noticias de KNIME.
   - ***Descarga KNIME*:** aquí es donde se puede descargar el software.
   - ***Comenzar*:** esta pestaña brinda información y enlaces sobre lo que se puede hacer después de haber instalado KNIME Analytics Platform.

3. Ahora se tiene que abrir la pestaña de *Descargar KNIME* y se hace clic en la opción de instalación que se ajuste a su sistema operativo.

   Notas sobre las diferentes opciones para Windows:
   - El instalador de Windows extrae la carpeta de instalación comprimida, agrega un icono a su escritorio y sugiere configuraciones de memoria adecuadas.
   - El fichero autoextraíble simplemente crea una carpeta que contiene los ficheros de instalación de KNIME.
   - El fichero zip se puede descargar, guardar y extraer en la ubicación preferida del usuario.

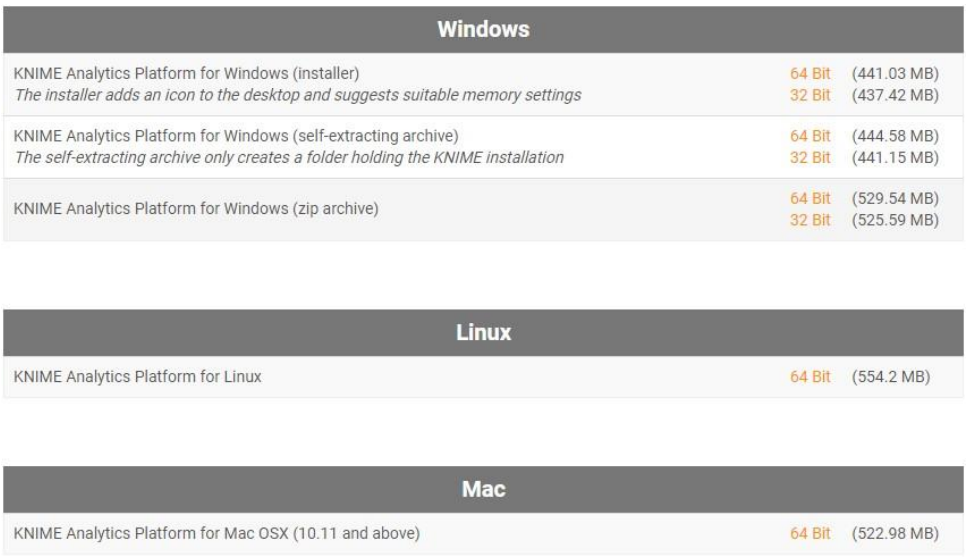


Figura 1. Versiones de KNIME Analytics Platform

4. Se debe de leer y aceptar la política de privacidad y los términos y condiciones. Luego se hace clic en *Descargar*.

5. Una vez descargado, se procede con la instalación de KNIME Analytics Platform:

   - ***Windows:*** *se* ejecuta el instalador descargado o el fichero autoextraíble. Si se ha optado por descargar el fichero zip, se debe de descomprimir en la ubicación que se eligió. Después se ejecuta knime.exe para iniciar KNIME Analytics Platform.
   - ***Linux*****:** se extrae el *tarball* descargado a la ubicación que se eligió. Se ejecuta el fichero knime ejecutable para iniciar KNIME Analytics Platform.
   - ***Mac:*** se hace doble clic en el fichero dmg descargado y se espera a que finalice la verificación. Después de esto, el ícono KNIME debe de aparecer en la sección de *Aplicaciones*. Se hace doble clic en el icono de KNIME en la lista de aplicaciones para iniciar KNIME Analytics Platform.

2.3 Instalación de extensiones e integraciones

Si se desea aumentar las capacidades de KNIME Analytics Platform, se puede instalar extensiones e integraciones. Las extensiones disponibles van desde extensiones e integraciones de código abierto gratuitas proporcionadas por KNIME hasta extensiones gratuitas aportadas por la comunidad y extensiones comerciales, incluidos nodos de tecnología novedosa proporcionados por las asociaciones con las que KNIME tiene convenio. Las extensiones e integraciones de KNIME desarrolladas y mantenidas por

KNIME contienen algoritmos de *deep learning* proporcionados por Keras, los algoritmos de aprendizaje automático de alto rendimiento son proporcionados por H2O, los procesamientos de big data son proporcionados por Apache Spark y los módulos de scripts son proporcionados por Python y R, solo por mencionar algunos módulos.

Para instalar las extensiones:

- Haga clic en *File* en la barra de menú y luego en *instalar Extensiones KNIME*. Después se abre el cuadro de diálogo que se muestra en la Figura 2.
- Se debe de seleccionar las extensiones que se desea instalar.
- Se da clic en "Siguiente" y se siguen las instrucciones
- Por último, se debe reiniciar la plataforma de análisis KNIME.

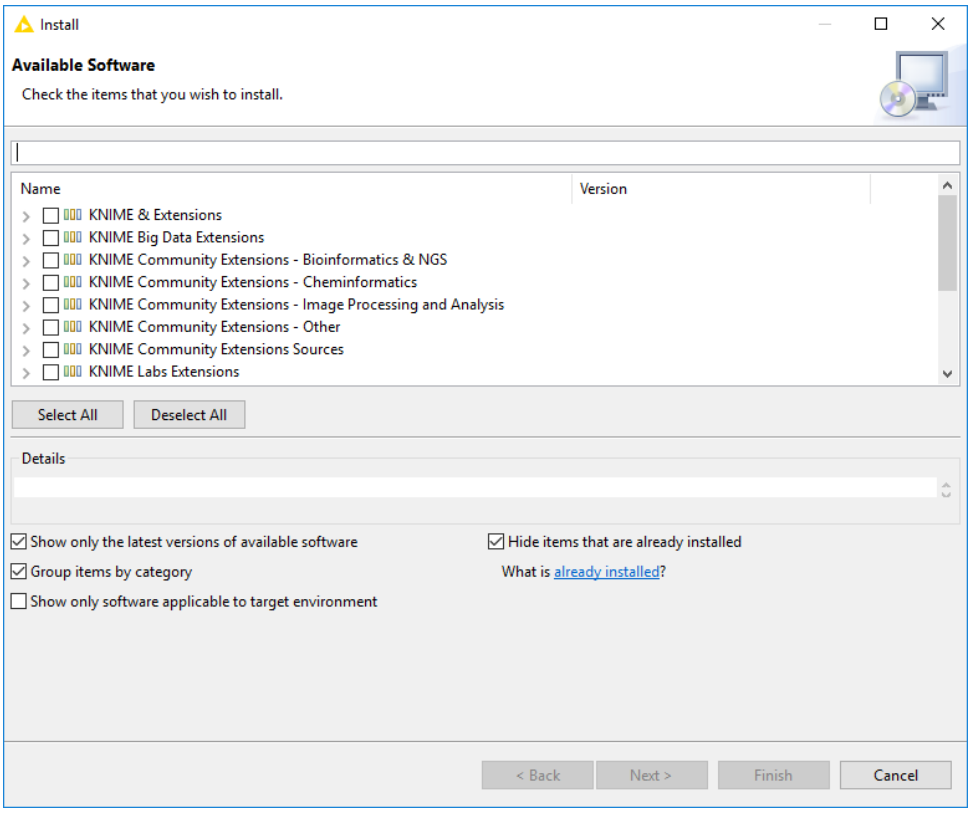


Figura 2. Instalación de Extensiones e Integraciones

2.4 Actualización de la plataforma y las extensiones de KNIME Analytics

Es bueno asegurarse de utilizar siempre la última versión de KNIME Analytics Platform y sus extensiones.
Para actualizar la plataforma y las extensiones:

1. Se da clic en *File* → *Actualización KNIME*. En el cuadro de diálogo que se abre, se selecciona las actualizaciones disponibles que se desea instalar y luego se da clic en *Siguiente*.

2. Se siguen las instrucciones hasta que KNIME Analytics Platform se reinicia de forma automática para aplicar las actualizaciones.

2.5 Introducción a la plataforma de análisis de KNIME

Una vez instalado KNIME Analytics Platform en el ordenador. Se da doble clic para iniciar KNIME Analytics Platform y cuando aparezca la ventana del KNIME Analytics Platform Launcher se tiene que definir el espacio o la carpeta de trabajo de KNIME como se muestra en la Figura 1.

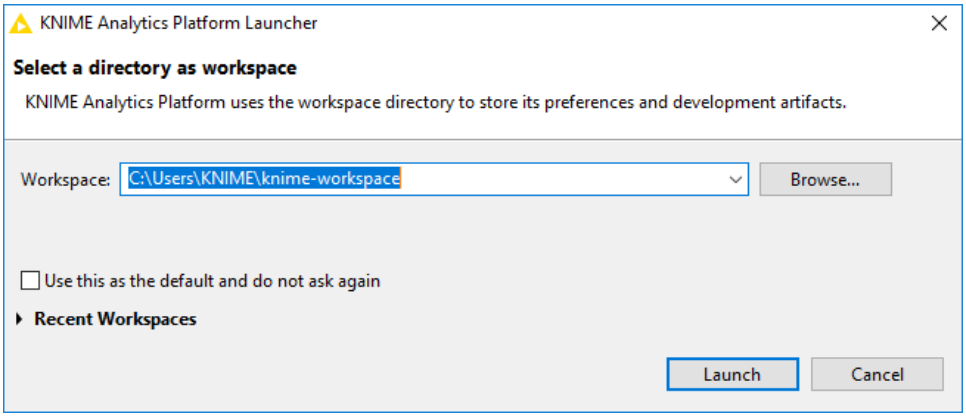


Figura 1. Lanzador de la plataforma KNIME Analytics

El espacio de trabajo KNIME es una carpeta en el ordenador local para almacenar los flujos de trabajo KNIME, las configuraciones de los nodos y los datos producidos por el flujo de trabajo. Los flujos de trabajo y los datos almacenados en el espacio de trabajo están disponibles a través de KNIME Explorer en la esquina superior izquierda de KNIME Workbench. Después de seleccionar una carpeta como el espacio de trabajo de KNIME para almacenar los proyectos, se hace clic en *Launch*.

La interfaz principal del usuario que aparece en la captura de pantalla que se muestra en la figura 2 se llama **KNIME Workbench.**

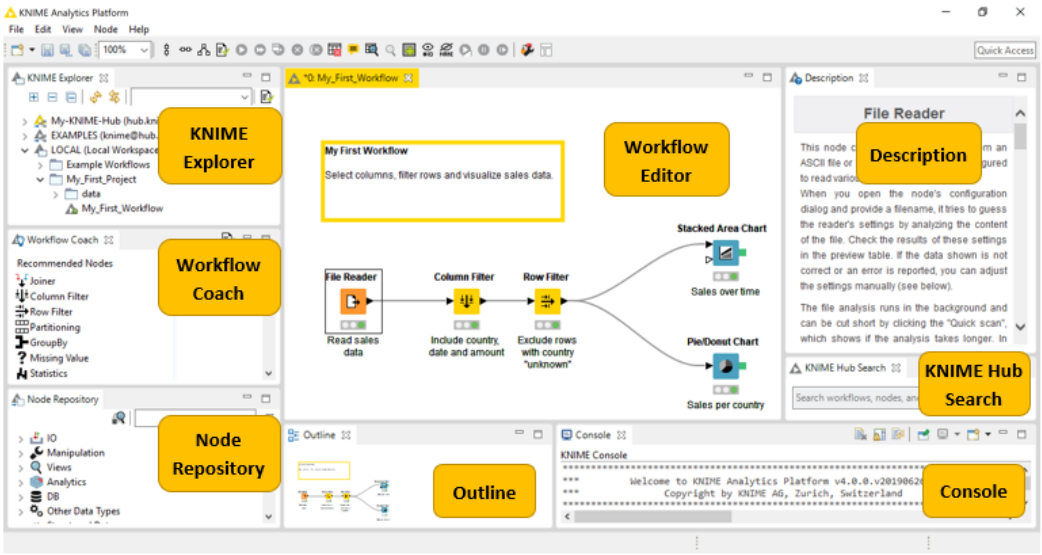


Figura 2. KNIME Workbench

La interfaz principal llamada *KNIME Workbench* se compone de los siguientes componentes:

- **KNIME Explorer:** descripción general de los flujos de trabajo disponibles en los espacios de trabajo activos, es decir, su carpeta de trabajo local y los trabajos en los servidores de KNIME.
- **Workflow Coach**: enumera las recomendaciones de nodos según los flujos de trabajo creados por la amplia comunidad de usuarios de KNIME.
- **Node Repository**: aquí se enumeran todos los nodos disponibles en el núcleo de KNIME Analytics Platform y en las extensiones que el usuario ha instalado. Los nodos están organizados por categorías, pero también se puede usar el cuadro de búsqueda en la parte superior del repositorio de nodos para buscar los nodos manualmente.
- **Workflow Editor**: espacio para editar y visualizar el flujo de trabajo activo actualmente.
- **Description**: Descripción del flujo de trabajo actualmente activo o de un nodo seleccionado (en el Editor de flujo de trabajo o en el Repositorio de nodos).
- **Outline**: se puede visualizar y enfocar una parte del flujo de trabajo activo actualmente.
- **Console**: muestra mensajes, advertencias y errores de ejecución.

2.6 Nodos y flujos de trabajo

En KNIME Analytics Platform, las tareas individuales están representadas por nodos. Cada nodo se muestra como un cuadro de color con puertos de entrada y salida, los nodos indican un estado (rojo, amarillo, verde o error) como se muestra en la Figura 3. Las entradas son los datos que procesa el nodo y las salidas son los conjuntos de datos resultantes. Cada nodo tiene una configuración específica, que podemos ajustar en un diálogo de configuración. Cuando lo hacemos, el estado del nodo cambia, mostrado por los colores de un semáforo debajo de cada nodo. Los nodos pueden realizar todo tipo de tareas, incluida la lectura y escritura de archivos, la transformación de datos, la formación de modelos, la creación de visualizaciones…etc.

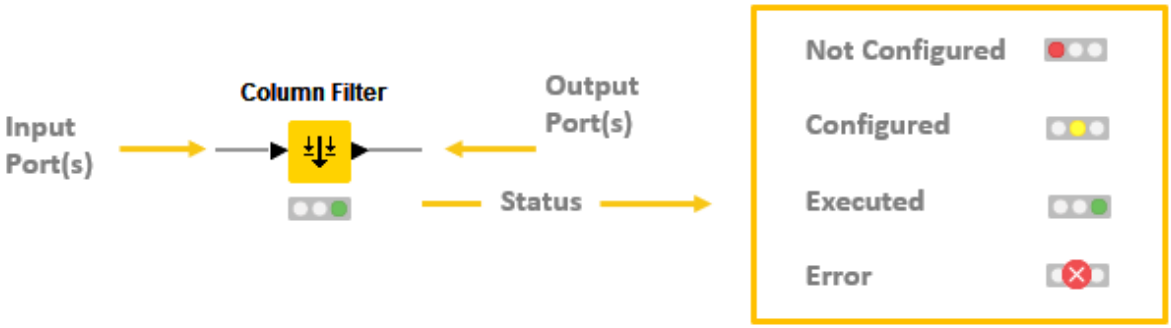


Figura 3. Puertos de nodo y estado de nodo

Una colección de nodos interconectados constituye un flujo de trabajo (workflow) y por lo general, representa una parte, o quizás la totalidad, de un proyecto de análisis de datos en particular.

2.7 Ejemplo de cómo crear un flujo de trabajo (workflow) en KNIME

El ejemplo del flujo de trabajo en la **figura 4** lee datos de un fichero CSV, filtra un subconjunto de las columnas, filtra algunas filas y visualiza los datos en dos gráficos: un gráfico de áreas apiladas (*stacked area chart*) y un gráfico circular (*pie chart*), que se pueden ver en la **figura 5.** En la figura 5 la gráfica de la izquierda muestra la evolución de las ventas a lo largo del tiempo y la gráfica de la derecha muestra la participación de diferentes países en las ventas totales.

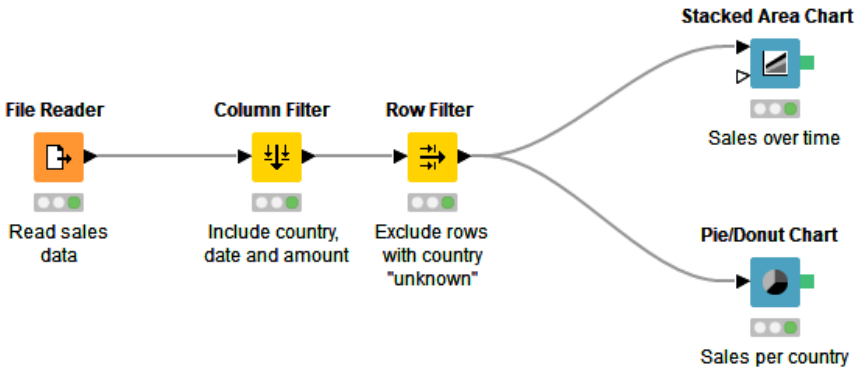


Figura 4. Ejemplo de flujo de trabajo o workflow en Knime

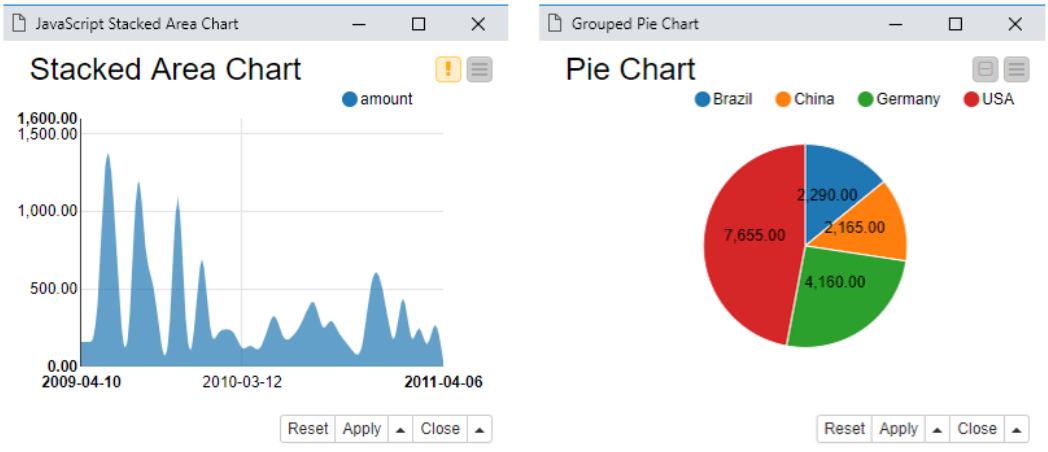


Figura 5. Ejemplo de Gráficos en un flujo de trabajo

Para crear el ejemplo de la figura 4 se deben de seguir los siguientes pasos.

Primero se debe de descargar el fichero CSV que contiene los datos que se van a usar en el flujo de trabajo. El fichero se puede encontrar en la página oficial de KNIME bajo el nombre de *First worflow example* .

A continuación, se crea un nuevo flujo de trabajo vacío de la siguiente manera:

- Se da clic en *New KNIME Workflow* en el panel de la barra de herramientas en la parte superior de KNIME Workbench

- O haciendo clic con el botón derecho en una carpeta de su espacio de trabajo local en KNIME Explorer, como se muestra en la figura 6

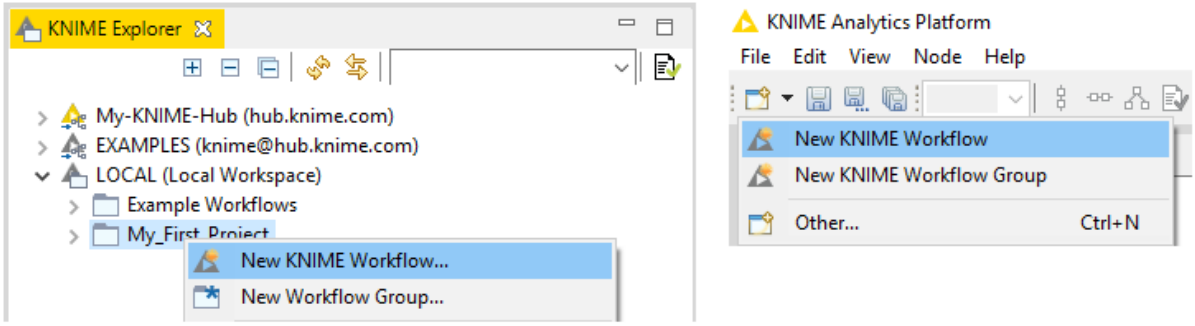


Figura 6. Creación de un nuevo flujo de trabajo vacío

El primer nodo que se necesita es el nodo *File Reader*, que se encuentra en el repositorio de nodos. Se debe de ir a *IO → Read → File Reader* o escribir una parte del nombre en el cuadro de búsqueda en el panel del repositorio de nodos.

Para usar el nodo en el flujo de trabajo se puede:

- Arrastrar y soltar desde el repositorio de nodos al editor de flujo de trabajo
- O hacer doble clic en el nodo, en el repositorio de nodos, y el nodo aparecerá automáticamente en el editor de flujo de trabajo.

Una vez que el nodo esté en el flujo de trabajo, se debe de definir ahora la configuración de ese nodo:

- Se abre el diálogo de configuración ya sea haciendo doble clic en el nodo, o haciendo clic derecho y seleccionando *Configure* como se muestra en la figura 7.

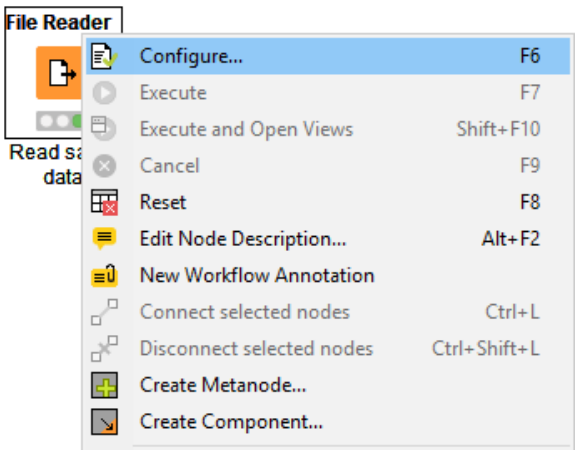


Figura 7. Configuración de un nodo

- En el cuadro de diálogo de configuración, se debe definir la ruta del archivo haciendo clic en el botón *Browse*, luego se verifica las otras configuraciones disponibles. También aquí se puede obtener una vista previa de los datos en la sección de *Preview* como se muestra en la figura 8.

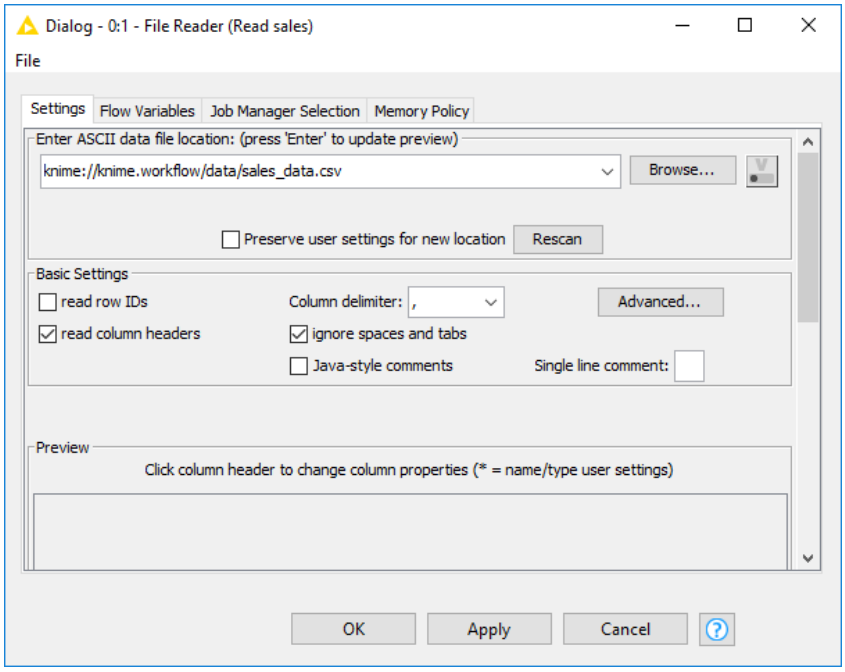


Figura 8. Diálogo de configuración del nodo File Reader

Después de verificar que el diálogo de configuraciones está correcto se da clic en *Apply* y después en *OK*.

Para ver si el fichero de datos se leyó como se pretendía se debe ir a la tabla de salida del nodo de la siguiente forma:

- Se ejecuta el nodo haciendo clic con el botón derecho y se selecciona *Execute*
- Después se abre la tabla de salida haciendo clic derecho en el nodo ya ejecutado y se selecciona la última opción en el menú: *File Table.*

Si los datos se leyeron correctamente, seguimos con nuestro flujo de trabajo y ahora se agrega el nodo *Column Filter* al editor de flujo de trabajo. Este nuevo nodo se debe de conectar a nuestro primer nodo *File Reader* de la siguiente manera:

- Se da clic en el puerto de salida del nodo *File Reader*, se mantiene presionado el botón del ratón y se suelta en el puerto de entrada del nodo *Column Filter*. De esta forma es como se conecta cualquier nodo con otro nodo en el flujo de trabajo.

Ahora se debe de configurar el segundo nodo llamado *Column filter* :

- Se mueven las columnas "country", "date" y "amount" al campo *Include* con marco verde, ya sea haciendo doble clic en ellas o usando los botones entre los campos de *Exclusion* e *Inclusion* en el cuadro de diálogo de configuración que se muestra en la figura 9

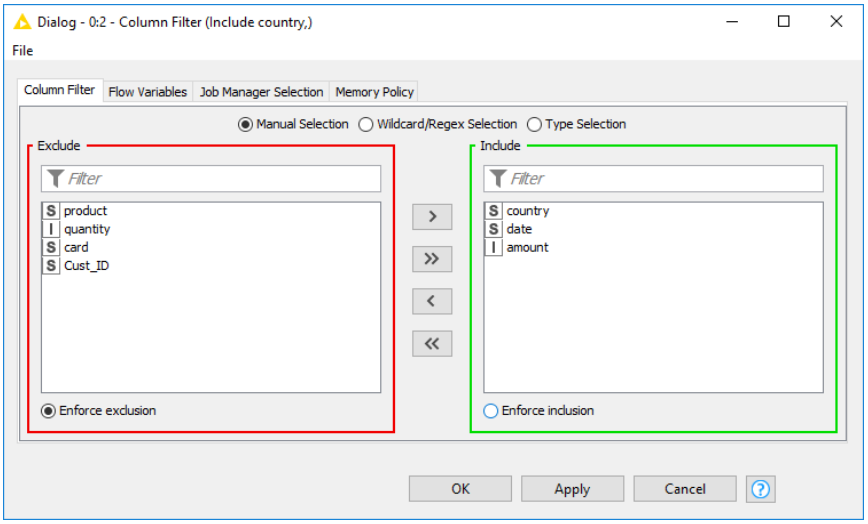


Figura 9. Configuración del nodo Column Filter

- Se finaliza la configuración haciendo clic en *OK*

Para agregar el siguiente nodo de *Row Filter* a nuestro flujo de trabajo debe de hacerse lo siguiente:

- Se agrega el nodo *Row Filter* al flujo de trabajo y se conecta al nodo *Column Filter*, siguiendo la instrucciones de conexión entre nodos que ya se mencionó anteriormente.

- Se abre el cuadro de diálogo de configuración del nodo *Row Filter* y se excluyen las filas de la tabla de entrada donde la columna "country" tiene el valor "unknown" como se muestra en la figura 10

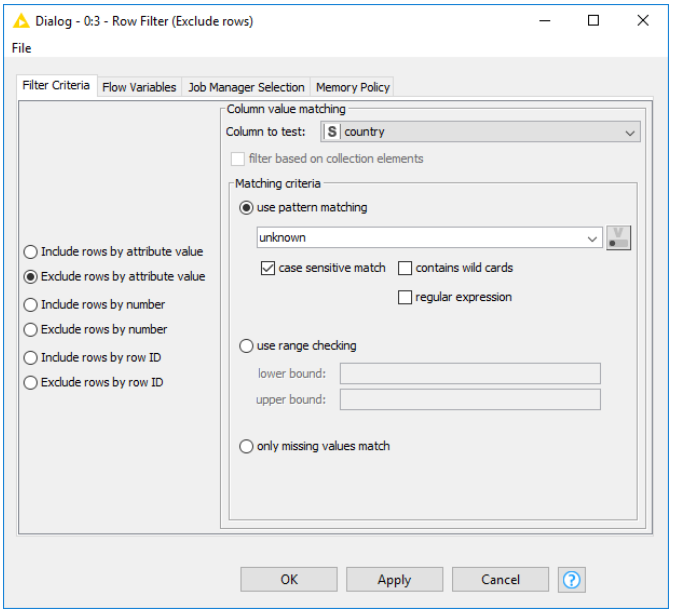


Figura 10. Configuración del nodo Row filter

- Se da clic en *Apply* y en *OK* y después se ejecuta el nodo.

Ahora que los datos se han filtrado, pasemos a la visualización de datos:

- Se busca los nodos *Stacked Area Chart (JavaScript)* y *Pie / Donut Chart (JavaScript)* en el repositorio de nodos, y se agregan al flujo de trabajo, ambos nodos se conectan al último nodo del flujo de trabajo es decir al nodo *Row Filter*.
- Se abre el cuadro de diálogo de configuración del nodo *Stacked Area Chart (JavaScript)*. Se selcciona la columna "date" como columna del eje x, como se muestra en la figura 11.

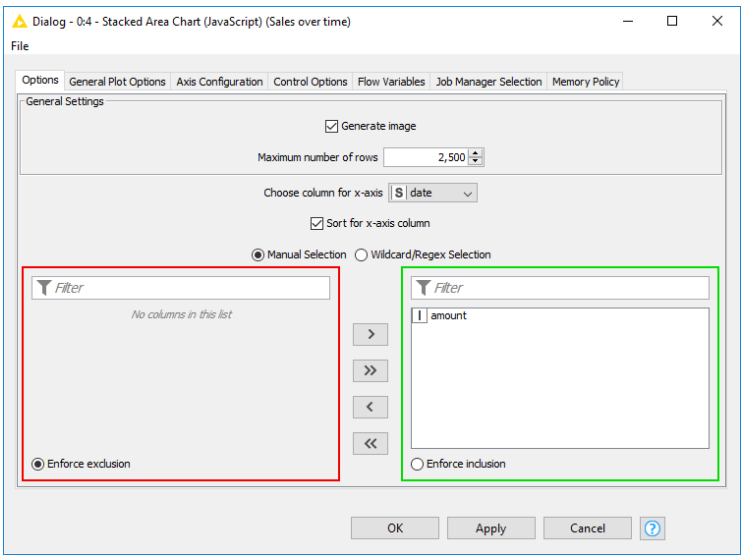


Figura 11. Configuración del nodo Stacked Area Chart (JavaScript)

- Ahora se abre el cuadro de diálogo de configuración del nodo *Pie/Donut Chart (JavaScript)* y se selecciona "country" como columna de categoría, "Sum" como método de agregación y "amount" como columna de frecuencia para el gráfico circular. Las opciones de configuración se muestran en la figura 12.

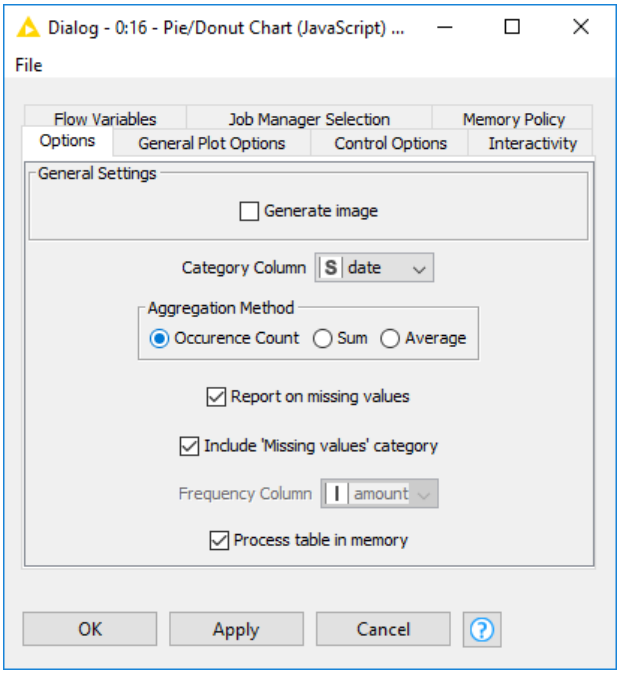


Figura 12. Configuración del nodo Gráfico circular / de anillos (JavaScript)

El flujo de trabajo ha finalizado y el siguiente paso es ejecutar todo el flujo de trabajo para ver el resultado. Esto se puede hacer dando clic en el botón *"Execute all executable nodes"* en la barra de herramientas que se muestra en la figura 13.

Figura 13. Para ejecutar todos los nodos ejecutables desde la barra de herramientas

O también se pueden ejecutar todos los nodos seleccionando los últimos nodos del flujo de trabajo, y haciendo clic con el botón derecho en la selección y luego clic en *Execute*.

Para inspeccionar la vista de salida interactiva de un nodo basado en JavaScript:

- Se elige la opción *Execute* and *Open Views* para un nodo basado en JavaScript como se muestra en la figura 14

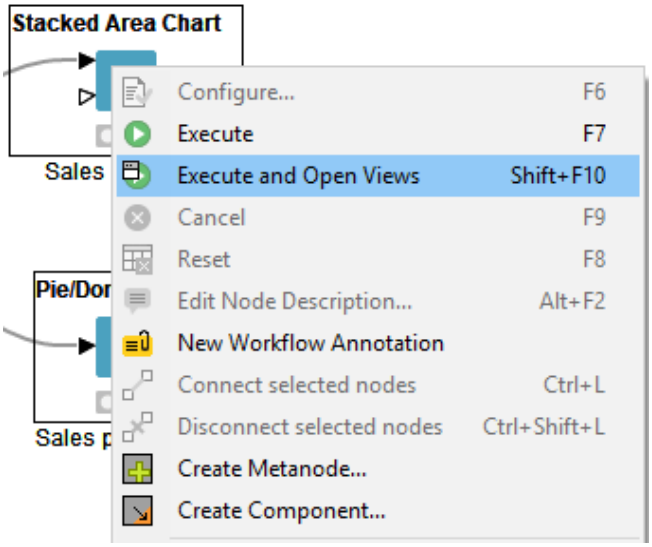


Figura 14. Ejecución y apertura de la vista interactiva JavaScript

- O bien, una vez que se ejecuta un nodo, haga clic derecho en el nodo y seleccione *Interactive View* como se muestra en la figura 15.

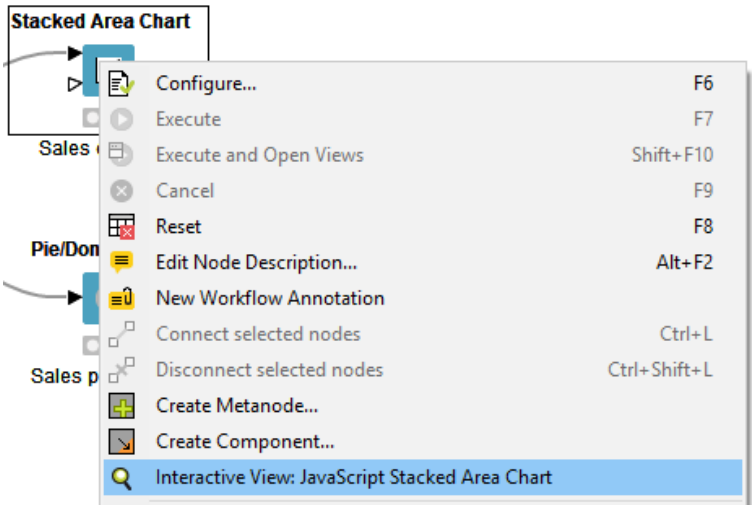


Figura 15. Apertura de la vista interactiva JavaScript de un nodo ejecutado

Hemos terminado el flujo de trabajo para este ejemplo sencillo.

Para cambiar los colores de los gráficos e insertar un nodo entre dos nodos (opcional)

Generalmente, el gráfico circular o *pie chart* usa colores predeterminados para diferentes países en los datos del ejemplo anterior. Con el nodo *Color Manager* se puede asignar a los países colores distintos a los predeterminados que se ven en la figura 5. Los colores **deben asignarse antes** de construir el gráfico, por lo que, ahora ya teniendo nuestro flujo de trabajo terminado, se tendrá que agregar el nodo *Color Manager* en medio del flujo de trabajo.

Para agregar el nodo *Color Manager:*

- Se arrastra el nodo *Color Manager* desde el repositorio de nodos y se suelta en el flujo de trabajo justo en medio del nodo el nodo *Row Filter* y el nodo *Pie Donut Chart*. La conexión roja, como se muestra en la figura 16, significa que el nodo en el flujo de trabajo está listo para aceptar el nuevo nodo cuando se suelte el ratón.

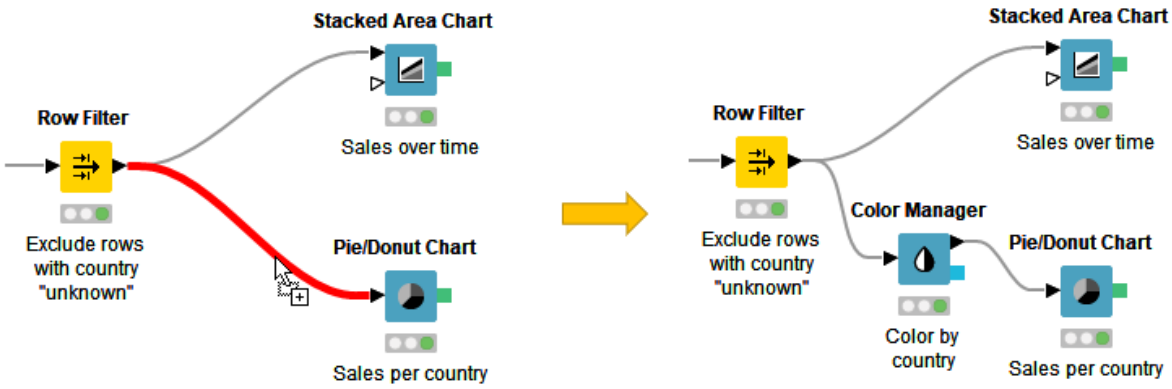


Figura 16. Cómo insertar un nodo entre dos nodos en un flujo de trabajo.

3. TÉCNICAS DE MINERÍA DE DATOS APLICADAS AL MARKETING ANALÍTICO PARA PREDECIR EL COMPORTAMIENTO DE LOS CONSUMIDORES

3.1 Introducción al marketing analítico

El marketing analítico es una disciplina basada en estadística y mercadotecnia que busca encontrar patrones en los datos de marketing con el fin de aumentar el conocimiento procesable que se puede utilizar en estrategias y campañas de marketing.
El marketing analítico también es fundamental para comprender el impacto de las campañas publicitarias y predecir las tendencias de mercado, el comportamiento y las preferencias de los consumidores, y optimizar la experiencia del usuario (UX) para impulsar las ventas. El marketing analítico emplea estadísticas, modelos predictivos y aprendizaje automático para revelar información y así crear estrategias efectivas.

El marketing analítico beneficia tanto a los especialistas en marketing como a los consumidores. El marketing analítico permite a los especialistas en marketing lograr un mayor retorno de la inversión en inversiones de marketing al comprender qué tiene éxito y que no, también adquieren un mayor conocimiento de la marca o del producto frente a sus competidores. En cuanto a los consumidores, el marketing analítico también asegura que los consumidores vean una mayor cantidad de anuncios personalizados y dirigidos que hablan de sus necesidades e intereses específicos, en lugar de recibir publicaciones masivas que tienden a molestar.
Los datos de marketing se pueden analizar utilizando una variedad de métodos y modelos según los KPI que se midan. Por ejemplo, el análisis del conocimiento de la marca se basa en diferentes datos y modelos que el análisis de conversiones.

La importancia del marketing analítico

En el panorama del marketing moderno, los análisis precisos son más importantes que nunca. Los consumidores se han vuelto muy selectivos a la hora de elegir los medios de marca con los que interactúan y los medios que ignoran. Si las marcas quieren captar la atención del comprador ideal, deben confiar en la analítica para crear anuncios específicos basados en intereses individuales, en lugar de asociaciones demográficas más amplias. Esto permitirá que los equipos de marketing publiquen el anuncio correcto, en el momento correcto, en el canal correcto para mover a los consumidores por el embudo de ventas. Vivimos en una era de datos y estos datos soy muy accesibles para las empresas. Las empresas pueden acceder a muchos teléfonos inteligentes para acceder a datos sobre los hábitos de ejercicio de los usuarios, sus patrones de sueño e incluso sus registros médicos.

En internet, las empresas pueden utilizar archivos de texto conocidos como cookies y otras herramientas similares para recopilar información sobre sus clientes. Las empresas pueden obtener casi todo lo que necesitan saber sobre sus clientes, desde qué tipo de productos compran sus clientes hasta qué grupos de edad tienden a visitar un sitio web con más frecuencia. Las empresas pueden desglosar estos datos, hasta el nivel individual si eso es lo que necesitan. Al final, no son los datos lo que importa, sino lo que se puede hacer con ellos. Hoy en día, como ya he mencionado, la mayoría de las empresas tienen acceso a los datos de los clientes y a las herramientas de análisis web. La diferencia está en si esas empresas saben hacer uso efectivo de esos datos o no.

¿Cómo utilizan las compañías el marketing analítico?

Según una investigación realizada por la compañía McKinsey, las organizaciones que utilizan marketing analítico y aprovechan los datos de comportamiento del cliente superan a sus pares en un 85 por ciento en el crecimiento de las ventas y en más del 25 por ciento en el margen bruto de ganancias. Hoy en día Amazon, Netflix y Google son algunas de las compañías que han declarado abiertamente que han construido sus respectivos imperios en torno a un núcleo de datos y al análisis del comportamiento del cliente.

3.2 Predicción del abandono o "churn prediction"

La pérdida de clientes (también conocida como deserción de clientes o abandono del cliente) se refiere a cuando un cliente (jugador, suscriptor, usuario, etc.) cesa su relación con una empresa. Dicho esto, entonces concluimos que la predicción de abandono significa detectar qué clientes es probable que cancelen una suscripción a un servicio. Es una predicción fundamental para muchas empresas porque, a menudo, adquirir nuevos clientes cuesta más que retener a los existentes. Una vez que las empresas puedan identificar a los clientes que están en riesgo de cancelación, pueden saber qué acción de marketing tomar para cada cliente individualmente y así maximizar las posibilidades de que el cliente permanezca. Comprender lo que mantiene a los clientes comprometidos es un conocimiento extremadamente valioso, ya que puede ayudar a desarrollar estrategias de retención y a implementar prácticas operativas destinadas a evitar que los clientes salgan por la puerta.

La predicción de la deserción o del abandono es una realidad para cualquier negocio de suscripción. Las técnicas de modelado de predicción de abandono intentan comprender los comportamientos y atributos precisos de los clientes que señalan el riesgo y el momento en que los clientes se van. Predecir y prevenir la pérdida de clientes representa una enorme fuente de ingresos potencial para todas las empresas.

CASO PRÁCTICO DE APLICACIÓN DE MINERÍA DATOS A UNA COMPAÑIA TELEFÓNICA PARA PREDECIR SI UN CLIENTE RENOVARÁ SU CONTRATO O NO.

3.3 Brightstar Corporation

Brightstar Corp. www.brightstar.com es una corporación privada estadounidense fundada en 1997. Proporciona distribución y servicios inalámbricos globales, sirviendo a fabricantes de dispositivos móviles, operadores inalámbricos y minoristas. Brightstar ofrece distribución de dispositivos y accesorios, protección y seguros para teléfonos móviles y productos digitales móviles. En 2019, Forbes nombró a Brightstar como una de las "Mejores empresas medianas de Estados Unidos". Brightstar fue fundada por Marcelo Claure en 1997. En abril de 2011 Brightstar Corp adquirió eSecuritel, un proveedor de servicios de seguros de telefonía celular con sede en Alpharetta, Georgia. En febrero de 2014, Brightstar Corp. completó la adquisición de 20:20 Mobile, un proveedor de telefonía móvil en Europa. En junio de 2018, Brightstar adquirió Next Wireless Group, un vendedor en línea de teléfonos inteligentes usados. En abril de 2020, la empresa adquirió WeFix. En septiembre de 2021, la compañía anunció que estaba cambiando de marca para reflejar una nueva dirección de servicios, el nuevo nombre de la compañía será "Likewize". La sede de la empresa se encuentra en Dallas, Texas. El lema de la empresa en su página web es “Simplificamos el mundo inalámbrico, haciendo que la tecnología móvil sea accesible para todos. Nos ocupamos de cada etapa del ciclo de vida de un dispositivo para nuestros clientes, desde el momento en que se fabrica hasta el momento en que es hora de intercambiarlo y volver a comercializarlo, poniéndonos en el corazón del ecosistema inalámbrico”

3.4 Antecedentes, definición del problema y objetivos

Antecedentes

El estudio de la deserción o abandono de clientes *(churn)* es un área en la que las empresas invierten grandes recursos año tras año. Siempre, con la intención de poder saber de antemano, si un cliente decidirá pasar de su empresa a la competencia. La predicción de abandono es un problema importante en las compañías de dispositivos móviles y en la industria de las telecomunicaciones. En general y en casi todos los casos los clientes abandonan un a proveedor de servicios telefónicos en busca de mejores servicios o tarifas.

En particular, en el área de las telecomunicaciones, se ha vuelto cada vez más necesario estudiar el abandono o deserción de los clientes, dada la alta competitividad que se está desarrollando a nivel mundial. Según un estudio realizado por Telco System, en la industria de las telecomunicaciones de 2008 a 2010, la deserción de los clientes fue de un 30% anual aproximadamente.

Definición del problema

Con el enorme aumento en el número de clientes que utilizan los servicios telefónicos, supongamos que la división de marketing de Brightstar quiere atraer a más clientes nuevos y quiere evitar también la terminación del contrato de los clientes existentes (reducir la tasa de abandono). Para que Brightstar amplíe su clientela, es decir incrementar su tasa de crecimiento (número de nuevos clientes) debe de mantener, o en el mejor de los casos, reducir su tasa de abandono (número de clientes existentes que deciden no renovar su contrato).

Una alta tasa de abandono afectará negativamente las ganancias de Brightstar e impedirá su crecimiento. Dado que el costo de adquirir nuevos clientes es mucho más alto al de retener a sus clientes existentes, Brightstar está seguro que un modelo de predicción que indique si el cliente abandonará su contrato o no, proporcionará claridad a la empresa y de esta forma Brightstar podría enfocarse en ese segmento de clientes con el fin de retenerlos.

Objetivos

De lo anterior se desprende que la pérdida de clientes es un problema de la industria y donde se hace necesario aplicar herramientas avanzadas que permitan predecir y describir de alguna manera, qué clientes tienen mayor potencial de riesgo para no renovar su contrato telefónico con Brightstar.

El objetivo de este caso práctico es construir modelos de minería de datos a partir de los datos históricos de Brightstar para estimar la probabilidad de deserción de los clientes.

Este es un problema de clasificación y nuestro objetivo entonces es tomar un conjunto de datos históricos, donde podemos ver quién renovó contrato y quién no y mediante la construcción de modelos de clasificación predecir qué cliente abandonará.

3.5 Técnicas de clasificación utilizadas en este caso práctico

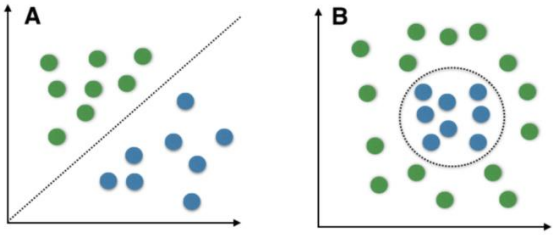
1. Árbol de decisión (Decision Tree)

El árbol de decisión crea modelos de clasificación o regresión en forma de estructura de árbol. Se divide un conjunto de datos en subconjuntos cada vez más pequeños mientras que al mismo tiempo se desarrolla un árbol de decisión asociado de forma incremental. El resultado final es un árbol con nodos de decisión y nodos de hoja. Un nodo de decisión (por ejemplo, *Outlook*) tiene dos o más ramas (por ejemplo, sunny, *overcast* y *rainy*). El nodo de hoja (por ejemplo, *Play Golf*) representa una clasificación o decisión. El nodo de decisión más alto en un árbol que corresponde al mejor predictor llamado nodo raíz. Los árboles de decisión pueden manejar datos tanto categóricos como numéricos.

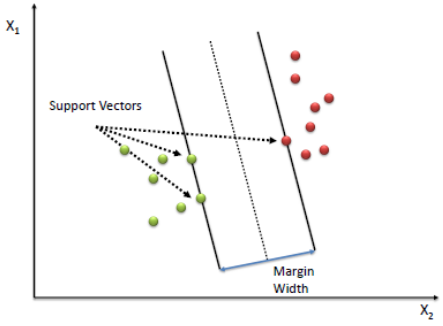
El algoritmo central para construir árboles de decisión fue llamado ID3 por J. R. Quinlan, que emplea una búsqueda minuciosa de arriba hacia abajo a través del espacio de posibles ramas sin retroceso. ID3 utiliza entropía y ganancia de información para construir un árbol de decisiones. En el modelo ZeroR no hay predictor, en el modelo OneR intentamos encontrar el mejor predictor único, NaiveBayes, por ejemplo, incluye todos los predictores que utilizan la regla de Bayes y los supuestos de independencia entre los predictores, pero el árbol de decisiones incluye todos los predictores con los supuestos de dependencia entre predictores.

2. Máquina de vectores de soporte (SVM)

Máquina de vectores de soporte o Support Vector Machine(SVM) es un algoritmo supervisado relativamente simple que se utiliza para clasificación y / o regresión. Es más preferido para la clasificación, pero a veces también es muy útil para la regresión. Básicamente, SVM encuentra un hiperplano que crea un límite entre los tipos de datos. En el espacio bidimensional, este hiperplano no es más que una línea. En SVM, trazamos cada elemento de datos en el conjunto de datos en un espacio N-dimensional, donde N es el número de características / atributos en los datos. Entonces, con esto, se debe haber entendido que, de manera inherente, SVM solo puede realizar una clasificación binaria (es decir, elegir entre dos clases). Sin embargo, existen varias técnicas que se pueden utilizar para problemas de varias clases.



Una máquina de vectores de soporte (SVM) realiza la clasificación al encontrar el hiperplano que maximiza el margen entre las dos clases. Los vectores (casos) que definen el hiperplano son los vectores de soporte



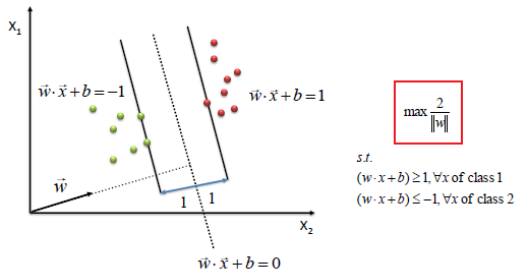
Pasos del Algoritmo

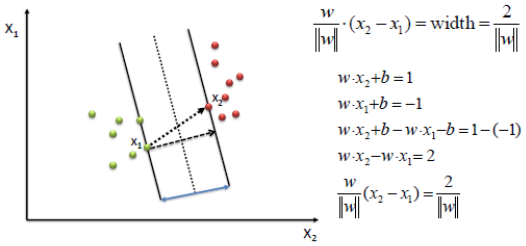
1.Definir un hiperplano óptimo: maximizar el margen

2. Amplía la definición anterior para problemas separables no linealmente: tiene un término de penalización para clasificaciones erróneas.

3. Mapea datos a un espacio de alta dimensión donde es más fácil clasificar con superficies de decisión lineales: reformula el problema para que los datos se mapeen implícitamente en este espacio.

Para definir un hiperplano óptimo necesitamos maximizar el ancho del margen (w).

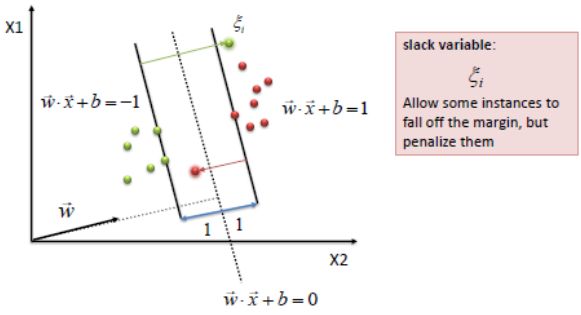




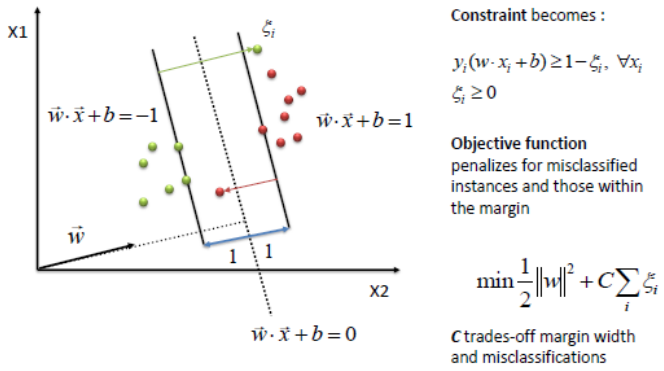
Hallamos w y b resolviendo la siguiente función objetivo usando programación Cuadrática.

$$\min \frac{1}{2}\|w\|^2$$
$$s.t.\ y_i(w\cdot x_i+b)\geq 1,\ \forall x_i$$

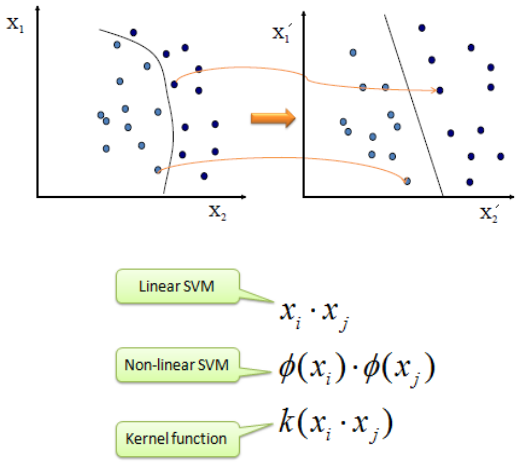
La belleza de SVM es que, si los datos son linealmente separables, existe un valor mínimo global único. Un análisis SVM ideal debería producir un hiperplano que separe completamente los vectores (casos) en dos clases que no se superponen. Sin embargo, puede que no sea posible una separación perfecta, o puede resultar en un modelo con tantos casos que el modelo no clasifica correctamente. En esta situación, SVM encuentra el hiperplano que maximiza el margen y minimiza las clasificaciones erróneas



El algoritmo intenta mantener la variable de holgura en cero mientras maximiza el margen. Sin embargo, no minimiza el número de clasificaciones erróneas (problema NP-completo) sino la suma de las distancias desde los hiperplanos marginales.



La forma más sencilla de separar dos grupos de datos es con una línea recta (1 dimensión), un plano plano (2 dimensiones) o un hiperplano N-dimensional. Sin embargo, hay situaciones en las que una región no lineal puede separar los grupos de manera más eficiente. SVM maneja esto usando una función de kernel (no lineal) para mapear los datos en un espacio diferente donde no se puede usar un hiperplano (lineal) para hacer la separación. Significa que una función no lineal se aprende mediante una máquina de aprendizaje lineal en un espacio de características de alta dimensión, mientras que la capacidad del sistema está controlada por un parámetro que no depende de la dimensionalidad del espacio. Esto se llama truco del núcleo, que significa que la función del núcleo transforma los datos en un espacio de características de mayor dimensión para hacer posible la separación lineal.



Se mapean los datos en un nuevo espacio, luego se toma el producto interno de los nuevos vectores. La imagen del producto interno de los datos es el producto interno de las imágenes de los datos. A continuación, se muestran dos funciones del núcleo.

Polynomial

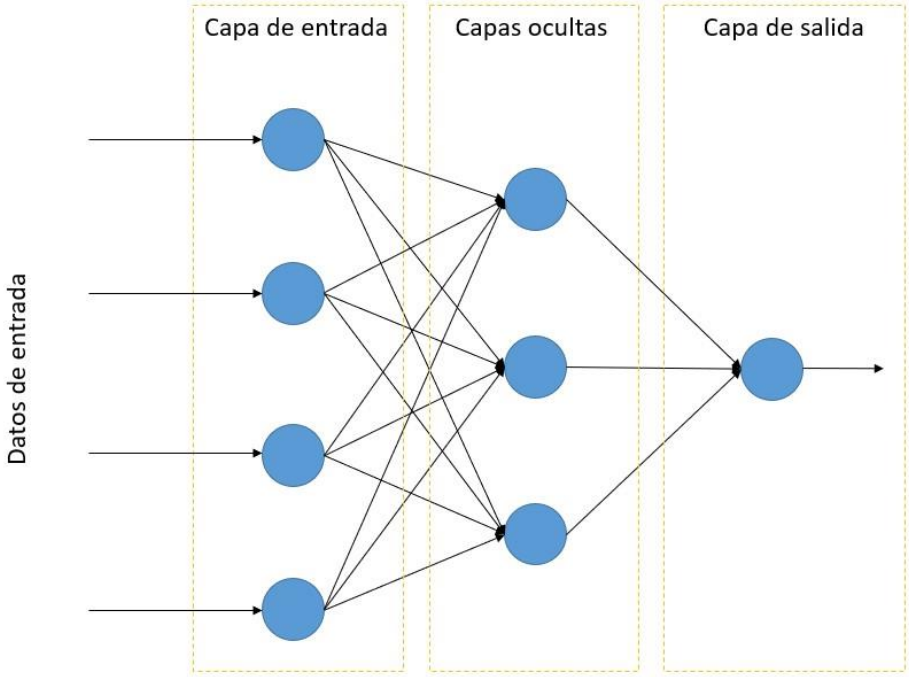
$$k(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i . \mathrm{x}_j)^d$$

Gaussian Radial Basis function

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{\left\|\mathrm{x}_i - \mathrm{x}_j\right\|^2}{2\sigma^2}\right)$$

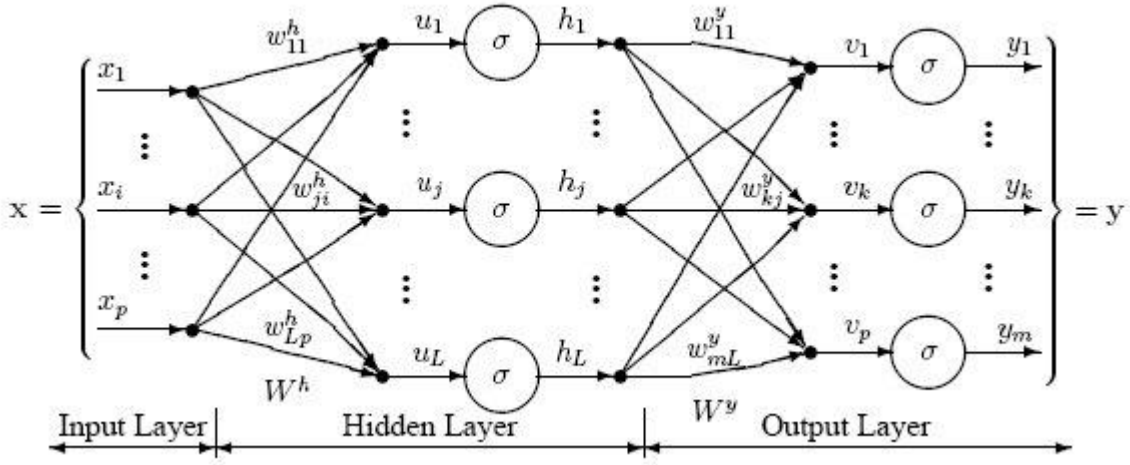
3. Perceptrón Multicapa o Multilayer Perceptron (MLP)

El **perceptrón multicapa** es una **red neuronal artificial (RNA)** formada por múltiples capas, de tal manera que tiene capacidad para resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón (también llamado perceptrón simple). En el primer caso cada salida de una neurona de la capa "i" es entrada de todas las neuronas de la capa "i+1", mientras que en el segundo cada neurona de la capa "i" es entrada de una serie de neuronas (región) de la capa "i+1".



El modelo de red neuronal de perceptrón multicapa

El siguiente diagrama ilustra una red de perceptrones **con tres capas**:



Esta red tiene una **capa de entrada** (a la izquierda) con tres neuronas, **una capa oculta** (en medio) con tres neuronas y una **capa de salida** (a la derecha) con tres neuronas.

Hay una neurona en la capa de entrada para cada variable predictora. En el caso de las variables categóricas, se utilizan N-1 neuronas para representar las N categorías de la variable.

Capa de entrada: Constituida por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento. Se presenta un vector de valores de variables predictoras ($x_1...x_p$) a la capa de entrada. La capa de entrada (o procesamiento antes de la capa de entrada) estandariza estos valores para que el rango de cada variable sea de -1 a 1. La capa de entrada distribuye los valores a cada una de las neuronas en la capa oculta. Además de las variables predictoras, hay una entrada constante de 1.0, llamada sesgo que alimenta a cada una de las capas ocultas; el sesgo se multiplica por un peso y se agrega a la suma que ingresa a la neurona.

Capa oculta: Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores. Al llegar a una neurona en la capa oculta, el valor de cada neurona de entrada se multiplica por un peso (w_{ji}), y los valores ponderados resultantes se suman para producir un valor combinado u_j. La suma ponderada (u_j) alimenta a una función de transferencia, σ, que genera un valor h_j. Las salidas de la capa oculta se distribuyen a la capa de salida. Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores

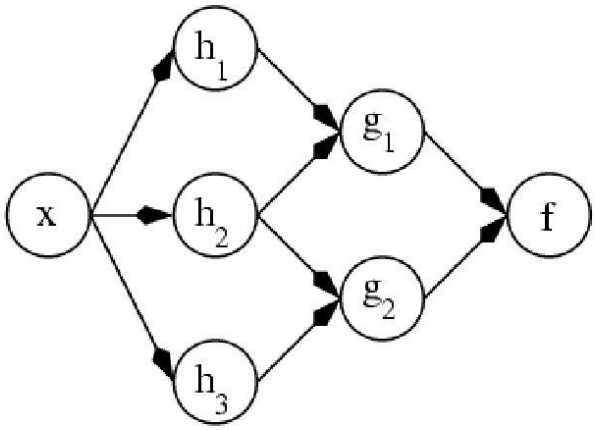
Capa de salida: Neuronas cuyos valores de salida se corresponden con las salidas de toda la red. Al llegar a una neurona en la capa de salida, el valor de cada neurona de la capa oculta se multiplica por un peso (w_{kj}) y los valores ponderados resultantes se suman para producir un valor combinado v. La suma ponderada (v) alimenta a una función de transferencia, σ, que genera un valor y_k. Los valores de y son las salidas de la red.

Si se está realizando un análisis de regresión con una variable objetivo-continua, entonces hay una sola neurona en la capa de salida y genera un solo valor y. Para problemas de clasificación con variables objetivo-categóricas, hay N neuronas en la capa de salida que producen N valores, uno para cada una de las N categorías de la variable objetivo.

Arquitectura de perceptrón multicapa

El diagrama de red que se mostró arriba es una red neuronal de perceptrón de alimentación hacia adelante (feed-forward), completamente conectada, de tres capas. "Totalmente conectado" significa que la salida de cada entrada y neurona oculta se distribuye a todas las neuronas en la siguiente capa. "Feed forward" significa que los valores solo se mueven de las capas de entrada a las capas ocultas a las de salida; no se retroalimentan valores a las capas anteriores (una red recurrente permite retroalimentar los valores).

Todas las redes neuronales tienen una capa de entrada y una capa de salida, pero el número de capas ocultas puede variar. Aquí hay un diagrama de una red de perceptrones con dos capas ocultas y cuatro capas totales:



Red de perceptrones con dos capas ocultas y cuatro capas totales

Cuando hay más de una capa oculta, la salida de una capa oculta se alimenta a la siguiente capa oculta y se aplican pesos separados a la suma de cada capa.

Entrenamiento de redes de perceptrones multicapa

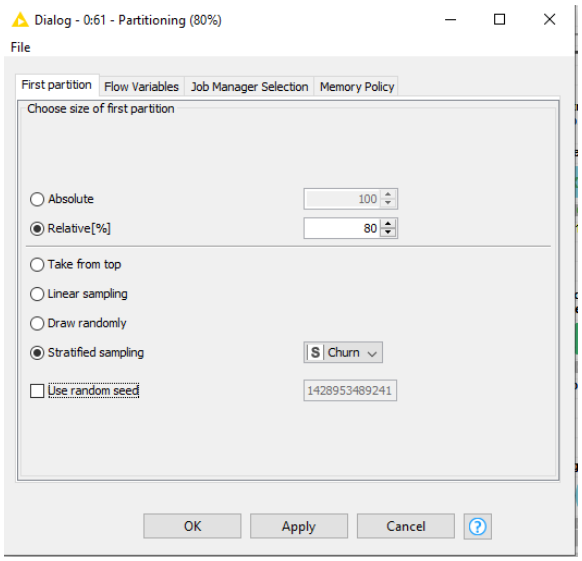
El objetivo del proceso de entrenamiento es encontrar el conjunto de valores de peso que harán que la salida de la red neuronal coincida con los valores objetivo-reales lo más cerca posible. Hay varios aspectos involucrados en el diseño y entrenamiento de una red de perceptrones multicapa:

• Seleccionar cuántas capas ocultas usar en la red.
• Decidir cuántas neuronas utilizar en cada capa oculta.
• Encontrar una solución óptima a nivel mundial que evite los mínimos locales.
• Convergencia hacia una solución óptima en un período de tiempo razonable.
• Validación de la red neuronal para probar el sobreajuste.

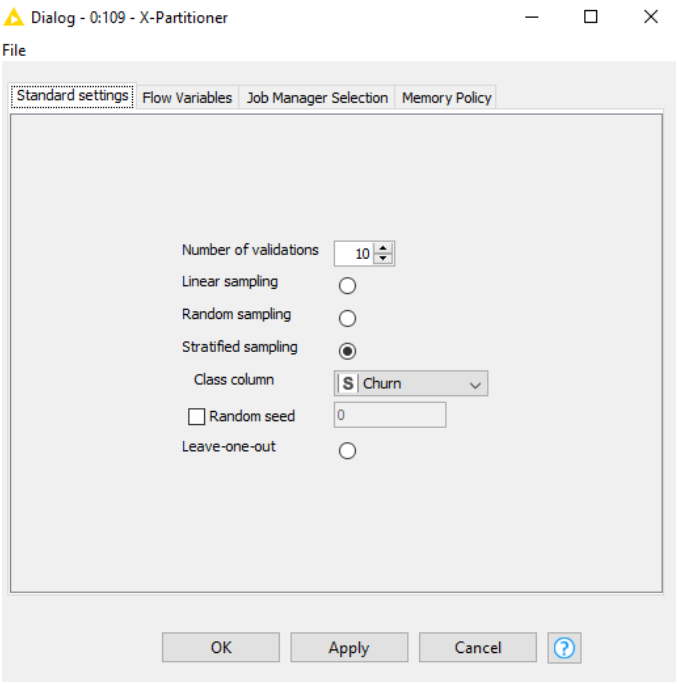
3.6 Tipos de validación utilizados en este caso práctico

En terminaos generales Knime ofrece dos tipos de validación: **validación simple** y **validación cruzada**. Es decir, Knime ofrece dos nodos generales (se pueden agregar más) de validación, cada uno de estos nodos tiene sus propias configuraciones.

- Nodo de Knime para **validación simple** (*Partitioning node*) con diferentes configuraciones como: Take from top, Linear samplig, Draw randonly y Stratified sampling. También ofrece la opción de escoger el porcentaje de partición entre el conjunto de prueba y el conjunto de entrenamiento.



- Nodo de Knime para **validación cruzada** (*CrossValidation node*) con diferentes configuraciones como: Número de validaciones o pliegues, Linear sampling, Random samplig, Stratified samplig y Leave one-out.



¿Cómo decidir el valor de k para validación cruzada?

El valor de k se elige de modo que cada grupo de pruebas de muestras de datos sea lo suficientemente grande como para ser estadísticamente representativo del conjunto de datos más amplio. **Un valor de k = 10 es muy común** en el campo del aprendizaje automático aplicado y se recomienda si tiene dificultades para elegir un valor para su conjunto de datos. Si se elige un valor para k que no divide uniformemente la muestra de datos, entonces un grupo contendrá el resto de los ejemplos. Es preferible dividir la muestra de datos en k grupos con el mismo número de muestras, de modo que la muestra de puntajes de habilidad del modelo sea equivalente.

Para este caso práctico **solo utilizaremos el nodo de validación cruzada con K=10 para todos los casos que aplique.**

Se utilizarán estos tres tipos de validaciones cruzadas en este caso práctico:

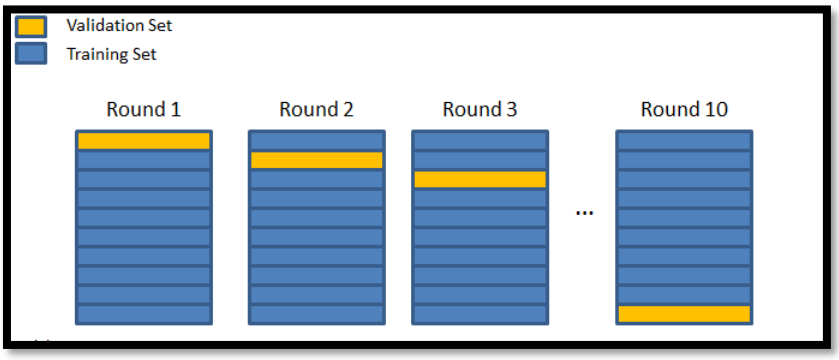
- **Random Sampling K-fold Cross Validation**
- **Stratified K-fold Cross Validation**
- **Leave One Out Cross Validation**

3.7 Validación cruzada

La validación cruzada es un método estadístico que se utiliza para estimar el rendimiento (o la precisión) de los modelos de aprendizaje automático. Se utiliza para proteger contra el sobreajuste en un modelo predictivo, particularmente en un caso donde la cantidad de datos puede ser limitada. En la validación cruzada, se realiza un número fijo de pliegues (o particiones) de los datos, se ejecuta el análisis en cada pliegue y luego se promedia la estimación del error general.

Existen diferentes tipos de técnicas de validación cruzada, pero el concepto general sigue siendo el mismo:

• Para dividir los datos en varios subconjuntos
• Mantenga un juego a la vez y entrene al modelo en el juego restante
• Modelo de prueba en espera
• Repita el proceso para cada subconjunto del conjunto de datos



Principio general para los diferentes tipos de validación cruzada

Tipos de validación cruzada

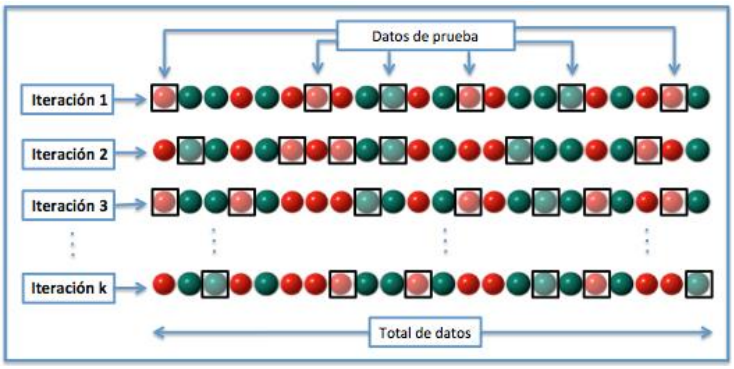
Existen diferentes tipos de métodos de validación cruzada y podrían clasificarse en dos categorías amplias: **métodos no exhaustivos** y **exhaustivos**.

- ### Métodos no exhaustivos de validación cruzada

Los métodos de validación cruzada no exhaustivos, como sugiere el nombre, no calculan todas las formas de dividir los datos originales.

1. Random sampling K-fold Cross Validation

En la validación cruzada aleatoria el bloque de validación se escoge aleatoriamente, repitiéndose el proceso k veces (siendo tanto el tamaño de cada bloque como el número k cifras arbitrarias). La ventaja de este método es que el número de iteraciones (k) no depende del tamaño de los sub-bloques que se consideren. La desventaja es que no es posible asegurar que todas las muestras van a ser consideradas como parte de los conjuntos de entrenamiento o de validación (pues habrá muchas que no sean escogidas nunca), y también que habrá muestras que puedan ser consideradas en más de una iteración.



Random sampling K-fold Cross Validation

2. Stratified K-fold Cross Validation

Primero que nada se debe aclarar que **la validación cruzada estratificada** (***Stratified K-fold Cross Validation*****) no es la misma que la conocida validación cruzada K-fold** (***K-fold Cross Validation*****)** son similares pero existe una diferencia.

Antes de profundizar en la validación cruzada estratificada, es importante conocer el muestreo estratificado. El muestreo estratificado es una técnica de muestreo en la que las muestras se seleccionan en **la misma proporción** (dividiendo la población en grupos llamados 'estratos' en función de una característica) tal como aparecen en la población.

La implementación del concepto de **muestreo estratificado en la validación cruzada asegura que los conjuntos de entrenamiento y prueba tengan la misma proporción de la característica de interés que en el conjunto de datos original**. Hacer esto con la variable de destino asegura que el resultado de la validación cruzada sea una aproximación cercana al error de generalización.

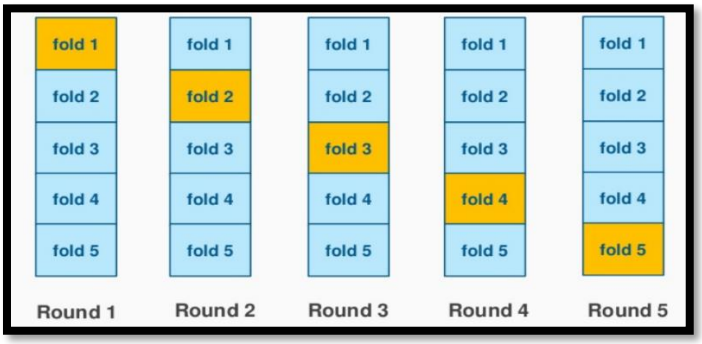
Diferencias entre K-fold Cross Validation y Stratified K-fold Cross Validation

K-fold Cross Validation mezcla aleatoriamente los datos y luego los divide en pliegues, esto hace que se probable que obtengamos pliegues altamente desequilibrados que pueden

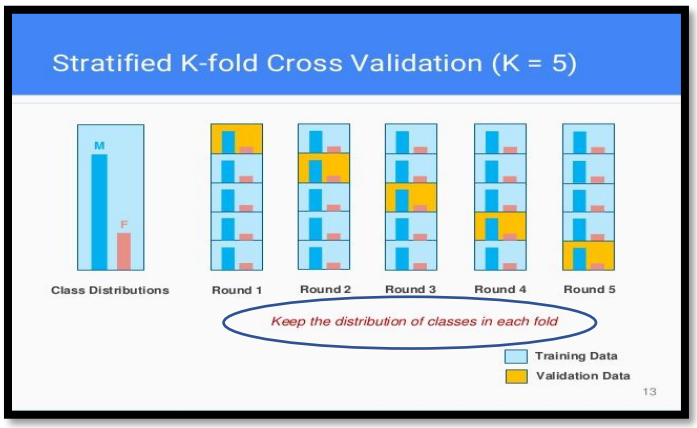
causar que nuestro entrenamiento sea sesgado. Por ejemplo, podemos obtener de alguna manera un pliegue que tenga una mayoría perteneciente a una clase (digamos positivo) y solo unos pocos como clase negativa. Esto sin duda arruinará nuestro entrenamiento y para evitarlo hacemos pliegues estratificados utilizando la estratificación.

Stratified K-fold Cross Validation reorganiza los datos para garantizar que cada pliegue sea un buen representante del todo. La división de datos en pliegues puede regirse por criterios como asegurar que cada pliegue tenga la misma proporción de observaciones con un valor categórico dado, como el valor de resultado de la clase. Esto se llama validación cruzada estratificada. Por ejemplo, en un problema de clasificación binaria como el nuestro dónde cada clase comprende el 50% de los datos, es mejor organizar los datos de manera que en cada pliegue, cada clase comprenda aproximadamente la mitad de las instancias.
La validación cruzada implementada mediante muestreo estratificado garantiza que la proporción de características de interés sea la misma en los datos originales, el conjunto de entrenamiento y el conjunto de prueba. Esto asegura que ningún valor esté sobre o infrarrepresentado en los conjuntos de entrenamiento y prueba, lo que proporciona una estimación más precisa del rendimiento / error.

Diferencias entre K-fold Cross Validation y Stratified K-fold Cross Validation



K-fold Cross Validation



Stratified K-fold Cross Validation

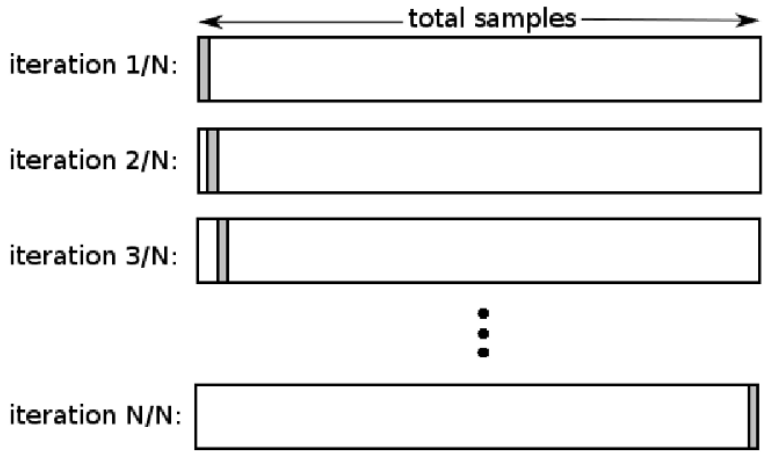
- **Métodos exhaustivos de validación cruzada**

Los métodos exhaustivos de validación cruzada son métodos de validación cruzada que aprenden y prueban todas las formas posibles de dividir la muestra original en un conjunto de entrenamiento y validación.

3. Leave One Out Cross Validation (LOOCV)

Leave One Out Cross Validation es una variación simple de otro tipo de validación cruzada llamada *Leave-P-Out* solo que el valor de *p* se establece como uno. Esto hace que el método sea mucho menos exhaustivo, ya que ahora para *n* puntos de datos y $p = 1$, tenemos *n* número de combinaciones.

Este enfoque deja 1 punto de datos fuera de los datos de entrenamiento, es decir, si hay *n* puntos de datos en la muestra original, entonces, se usan *n-1* muestras para entrenar el modelo y *p* puntos se usan como el conjunto de validación. Esto se repite para todas las combinaciones en las que la muestra original se puede separar de esta manera, y luego se promedia el error para todos los ensayos, para obtener la efectividad general. El número de combinaciones posibles es igual al número de puntos de datos en la muestra original *n*.



Representation of leave one out cross validation

3.8 Matriz de confusión

La matriz de confusión es una herramienta fundamental a la hora de evaluar el desempeño de un algoritmo de clasificación, ya que dará una mejor idea de cómo se está clasificando dicho algoritmo, a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación. Así se puede comprobar si el algoritmo está clasificando mal las clases y en qué medida.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Las matrices de confusión contienen información acerca de los valores reales y las clasificaciones predichas hechas por cualquier sistema de clasificación.

Funcionamiento

El desempeño de un sistema es usualmente evaluado usando los datos en dicha matriz. La siguiente tabla muestra la matriz de confusión para un clasificador en dos clases:

		Clasificador	
		Negativos	Positivos
Valores reales	Negativos	a	b
	Positivos	c	d

En dicha tabla:

- ***a*** es el número de predicciones **correctas** de que un caso es **negativo**.
- ***b*** es el número de predicciones **incorrectas** de que un caso es **positivo**, o sea la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I.
- ***c*** es el número de predicciones **incorrectas** de que un caso es **negativo**, o sea la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II.
- ***d*** es el número de predicciones correctas de que un caso es positivo.

Han sido definidos varios términos estándar para medir el desempeño de un clasificador en cualquier rama donde se apliquen sistemas de clasificación:

- La Exactitud (*Accuracy*) es la proporción del número total de predicciones que fueron correctas:

$$Ac = \frac{a+d}{a+b+c+d}$$

- La Razón de Verdaderos Positivos (*TPR-True Positive Rate*), a veces también denominada *Recall*, es la proporción de casos positivos que fueron correctamente identificados:

$$TPrate = \frac{d}{c+d}$$

- La Razón de Falsos Positivos (*FPR-False Positive Rate*) es la proporción de casos negativos que han sido incorrectamente clasificados como positivos:

$$FPrate = \frac{b}{a+b}$$

- La Razón de Verdaderos Negativos (*TNR-True Negative Rate*) es la proporción de casos negativos que han sido correctamente clasificados

$$TNrate = \frac{a}{a+b}$$

- La Razón de Falsos Negativos (*FNR-False Negative Rate*) es la proporción de casos positivos que fueron incorrectamente clasificados como negativos:

$$FNrate = \frac{c}{c+d}$$

- La precisión (P, en inglés, también *Precision*) es la proporción de casos predichos positivos que fueron correctos

$$P = \frac{d}{b+d}$$

Frecuentemente son utilizados también los términos siguientes:

- **Sensibilidad** (Se,del inglés,*Sensitivity*) como sinónimo de *TPR* porque es la capacidad del clasificador de ser "sensible" a los casos positivos. Note que *1-Se=FNR*
- **Especifidad** (*Sp*, del inglés *Specificity*) como sinónimo de *TNrate*, porque puede dar una medida de la especificidad del test para marcar los casos positivos. Note que *1-Sp =FPR*

Si un clasificador puede variar determinados parámetros puede lograrse incrementar los *TP* a costa de incrementar los *FP* o viceversa. En otras palabras, se desea una alta sensibilidad con una gran especificidad (o equivalentemente una reducida *FPR*)

IMPLEMENTACIÓN DEL PROBLEMA EN KNIME

3.9 Análisis de los datos

Para predecir la probabilidad de que sus clientes actuales no renueven su contrato, Brighstar necesita datos de clientes anteriores con su historial de deserción. En el momento de renovar los contratos, algunos clientes lo hicieron y otros no, es decir, desertaron. Estos ejemplos de clientes pasados, tanto los que abandonaron como los que no, se pueden utilizar para entrenar un modelo para predecir cuáles de los clientes actuales están en riesgo de deserción.

El conjunto de datos incluye datos de contratos de servicio telefónico para 3,333 clientes. Esta base de datos recoge un conjunto de variables que ofrecen información relacionada con el contrato de los clientes como la provincia en donde se contrató el teléfono, tipos de llamadas, tipo de planes contratados, reclamaciones…etc. Utilizaremos estos datos para la construcción de modelos predictivos. Los datos contienen, entre varios atributos, un campo llamado ***churn***, que será nuestra variable a predecir.

churn = 0 indica un contrato renovado;
churn = 1 indica un contrato no renovado.

Este es un problema de clasificación binaria. Y nuestro objetivo entonces es tomar un conjunto de datos históricos, donde podemos ver quién renovó contrato y quién no y mediante la construcción de modelos de clasificación predecir qué cliente abandonará (churn = 1) y qué cliente no lo hará (churn = 0).

Es decir: *attr 1, attr 2, …, attr n => churn (0/1)*

La lista de variables y su descripción que se utilizan en el conjunto de datos para este caso práctico se muestra en la siguiente tabla:

Esta es nuestra variable (clase) a predecir.

Notemos que el tipo de dato es *integer* porque los valores que contiene son 1 y 0. Por lo cual tendremos que convertir esta variable de *integer* a *string*.

Nota: el conjunto de datos no tiene valores nulos (*missing values*)

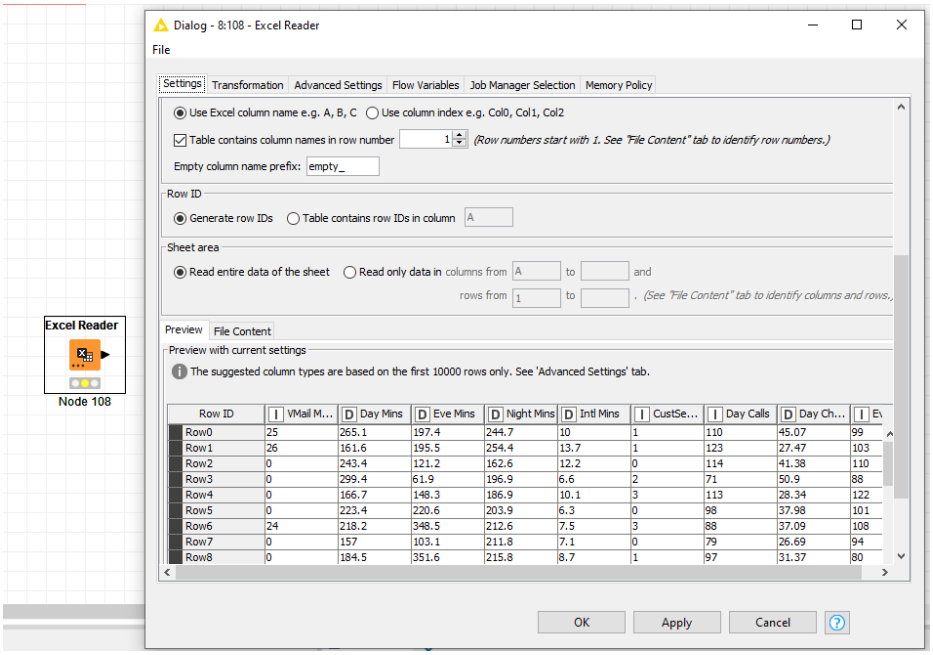
Attribute	Data type	Description
Account Length	Integer	Account length
VMail Message	Integer	Sending voice messages
Day Mins	Double	Minutes in the day
Eve Mins	Double	Minutes in the afternoon
Night Mins	Double	Minutes at night
Intl Mins	Double	International minutes
CustServ Calls	Integer	Consumer Service Calls
Churn	Integer	Churn-0 customer remained with contract Churn-1 Customer quit contract
Int'l Plan	Integer	International Plan
VMail Plan	Integer	Voice Message Plan
Day Calls	Integer	Calls per day
Day Charge	Double	Charge per day
Eve Calls	Integer	Afternoon calls
Eve Charge	Double	Charge for the call in the afternoon
Night Calls	Integer	Calls at night
Night Charge	Double	Call charge at night
Intl Calls	Integer	International Calls
Intl Charge	Double	International call charge
State	String	Status, Place
Area Code	Integer	Area Code
Phone	Integer	Phone No.

Resumen de nuestro caso práctico

Objetivo:	El objetivo de este caso práctico es construir un modelo de minería de datos a partir de los datos históricos de la compañía Brightstar para predecir qué cliente abandonará el contrato (churn=1) y que cliente no lo abandonará (churn=0).
Tipo de aprendizaje:	Supervisado
Tipo de problema:	Clasificación
Técnicas de Mineria de datos	Árbol de decisión Máquina de vectores de soporte (SVM) Perceptrón Multicapa (MLP)
Variable a predecir	churn
Número total de columnas	21
Número total de renglones	3333
Métodos de validación	Random sampling Cross Validation Stratified K-fold Cross Validation Leave One Out Cross Validation
Valor de k	k=10
Softawre utlizado	KNIME Analytics Platform

PASO 1: Carga del fichero en Knime

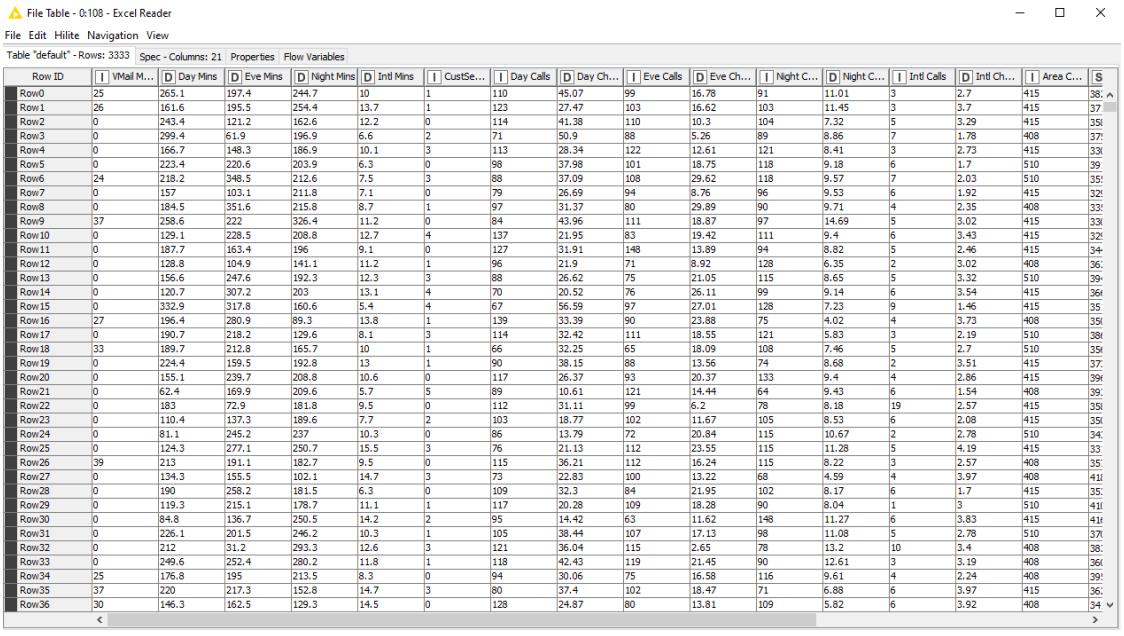
Para cargar el fichero en Knime simplemente hay que arrastrar el fichero .cvs o .xml al área de flujo trabajo de Knime o bien buscando el nodo *Read File o Excel Reader* en el repositorio de nodos. Se verifica que los datos sean leídos correctamente.



Row ID	VMail M...	Day Mins	Eve Mins	Night Mins	Intl Mins	CustSe...	Day Calls	Day Ch...	Ev
Row0	25	265.1	197.4	244.7	10	1	110	45.07	99
Row1	26	161.6	195.5	254.4	13.7	1	123	27.47	103
Row2	0	243.4	121.2	162.6	12.2	0	114	41.38	110
Row3	0	299.4	61.9	196.9	6.6	2	71	50.9	88
Row4	0	166.7	148.3	186.9	10.1	3	113	28.34	122
Row5	0	223.4	220.6	203.9	6.3	0	98	37.98	101
Row6	24	218.2	348.5	212.6	7.5	3	88	37.09	108
Row7	0	157	103.1	211.8	7.1	0	79	26.69	94
Row8	0	184.5	351.6	215.8	8.7	1	97	31.37	80

PASO 2: Se verifica que los datos se hayan cargado correctamente

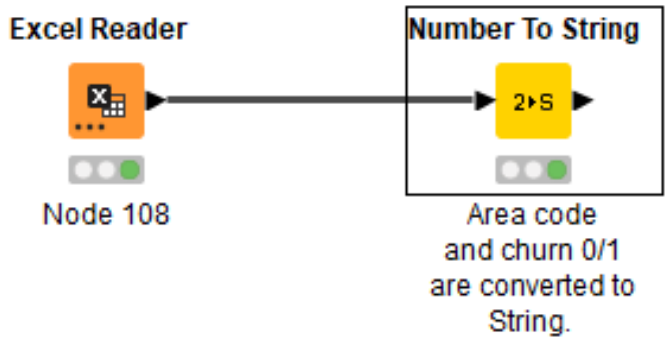
En el nodo de *Excel Reader* se da clic derecho *Execute --> File Table* para visualizar toda la tabla cargada

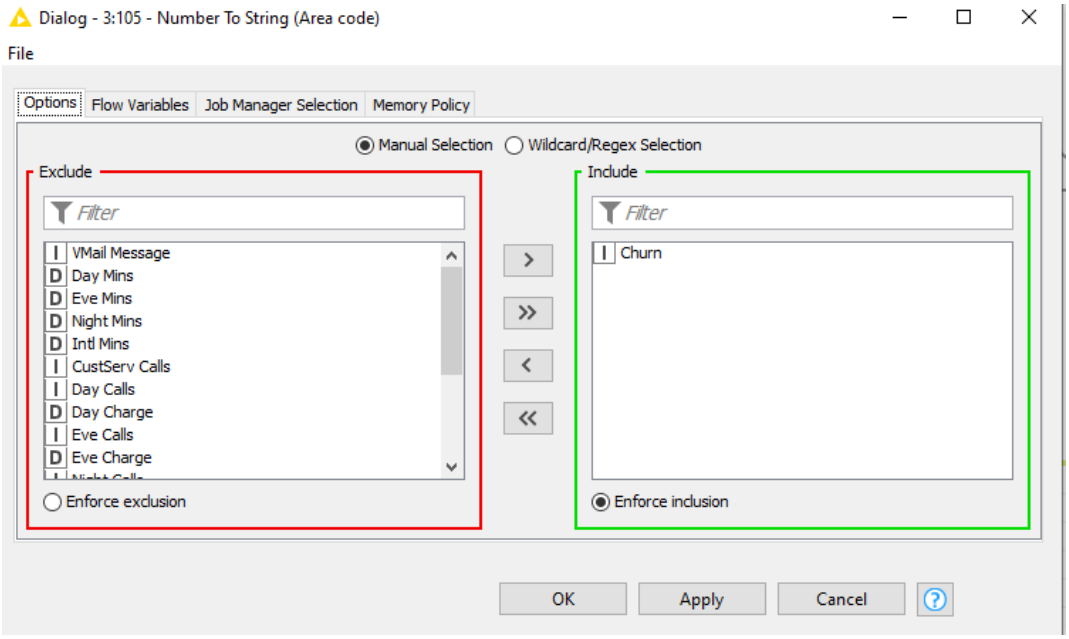


Row ID	VMail M...	Day Mins	Eve Mins	Night Mins	Intl Mins	CustSe...	Day Calls	Day Ch...	Eve Calls	Eve Ch...	Night C...	Night C...	Intl Calls	Intl Ch...	Area C...	S
Row0	25	265.1	197.4	244.7	10	1	110	45.07	99	16.78	91	11.01	3	2.7	415	38:
Row1	26	161.6	195.5	254.4	13.7	1	123	27.47	103	16.62	103	11.45	3	3.7	415	37
Row2	0	243.4	121.2	162.6	12.2	0	114	41.38	110	10.3	104	7.32	5	3.29	415	35
Row3	0	299.4	61.9	196.9	6.6	2	71	50.9	88	5.26	89	8.86	7	1.78	408	37
Row4	0	166.7	148.3	186.9	10.1	3	113	28.34	122	12.61	121	8.41	3	2.73	415	33
Row5	0	223.4	220.6	203.9	6.3	0	98	37.98	101	18.75	118	9.18	6	1.7	510	39
Row6	24	218.2	348.5	212.6	7.5	3	88	37.09	108	29.62	118	9.57	7	2.03	510	35
Row7	0	157	103.1	211.8	7.1	0	79	26.69	94	8.76	96	9.53	6	1.92	415	32
Row8	0	184.5	351.6	215.8	8.7	1	97	31.37	80	29.89	90	9.71	4	2.35	408	33
Row9	37	258.6	222	326.4	11.2	0	84	43.96	111	18.87	97	14.69	5	3.02	415	33
Row10	0	129.1	228.5	208.8	12.7	4	137	21.95	83	19.42	111	9.4	6	3.43	415	32
Row11	0	187.7	163.4	196	9.1	0	127	31.91	148	13.89	94	8.82	5	2.46	415	34
Row12	0	128.8	104.9	141.1	11.2	1	96	21.9	71	8.92	128	6.35	2	3.02	408	36
Row13	0	156.6	247.6	192.3	12.3	3	88	26.62	75	21.05	115	8.65	5	3.32	510	39
Row14	0	120.7	307.2	203	13.1	4	70	20.52	76	26.11	99	9.14	6	3.54	415	36
Row15	0	332.9	317.8	160.6	5.4	4	67	56.59	97	27.01	128	7.23	9	1.46	415	35
Row16	27	196.4	280.9	89.3	13.8	1	139	33.39	90	23.88	75	4.02	4	3.73	408	35
Row17	0	190.7	218.2	129.6	8.1	3	114	32.42	111	18.55	121	5.83	3	2.19	510	38
Row18	33	189.7	212.8	165.7	10	1	66	32.25	65	18.09	108	7.46	5	2.7	510	35
Row19	0	224.4	159.5	192.8	13	1	90	38.15	88	13.56	74	8.68	2	3.51	415	37
Row20	0	155.1	239.7	208.8	10.6	0	117	26.37	93	20.37	133	9.4	4	2.86	415	39
Row21	0	62.4	169.9	209.6	5.7	5	89	10.61	121	14.44	64	9.43	6	1.54	408	39
Row22	0	183	72.9	181.8	9.5	0	112	31.11	99	6.2	78	8.18	19	2.57	415	35
Row23	0	110.4	137.3	189.6	7.7	2	103	18.77	102	11.67	105	8.53	6	2.08	415	35
Row24	0	81.1	245.2	237	10.3	0	86	13.79	72	20.84	115	10.67	2	2.78	510	34
Row25	0	124.3	277.1	250.7	15.5	3	76	21.13	112	23.55	115	11.28	5	4.19	415	33
Row26	39	213	191.1	182.7	9.5	0	115	36.21	112	16.24	115	8.22	3	2.57	408	35
Row27	0	134.3	155.5	102.1	14.7	3	73	22.83	100	13.22	68	4.59	4	3.97	408	41
Row28	0	190	258.2	181.5	6.3	0	109	32.3	84	21.95	102	8.17	6	1.7	415	35
Row29	0	119.3	215.1	178.7	11.1	1	117	20.28	109	18.28	90	8.04	1	3	510	41
Row30	0	84.8	136.7	250.5	14.2	2	95	14.42	63	11.62	148	11.27	6	3.83	415	41
Row31	0	226.1	201.5	246.2	10.3	1	105	38.44	107	17.13	98	11.08	5	2.78	510	37
Row32	0	212	31.2	293.3	12.6	3	121	36.04	115	2.65	78	13.2	10	3.4	408	38
Row33	0	249.6	252.4	280.2	11.8	1	118	42.43	119	21.45	90	12.61	3	3.19	408	36
Row34	25	176.8	195	213.5	8.3	0	94	30.06	75	16.58	116	9.61	4	2.24	408	39
Row35	37	220	217.3	152.8	14.7	3	80	37.4	102	18.47	71	6.88	6	3.97	415	36
Row36	30	146.3	162.5	129.3	14.5	0	128	24.87	80	13.81	109	5.82	6	3.92	408	34

PASO 3: Se cambia la variable churn de entero (*integer*) a cadena (*string*)

En el repositorio de nodos se busca el nodo *Number To String*, se conecta este nodo al primero nodo y se configura, seleccionado variable churn únicamente, después se ejecuta el nodo.

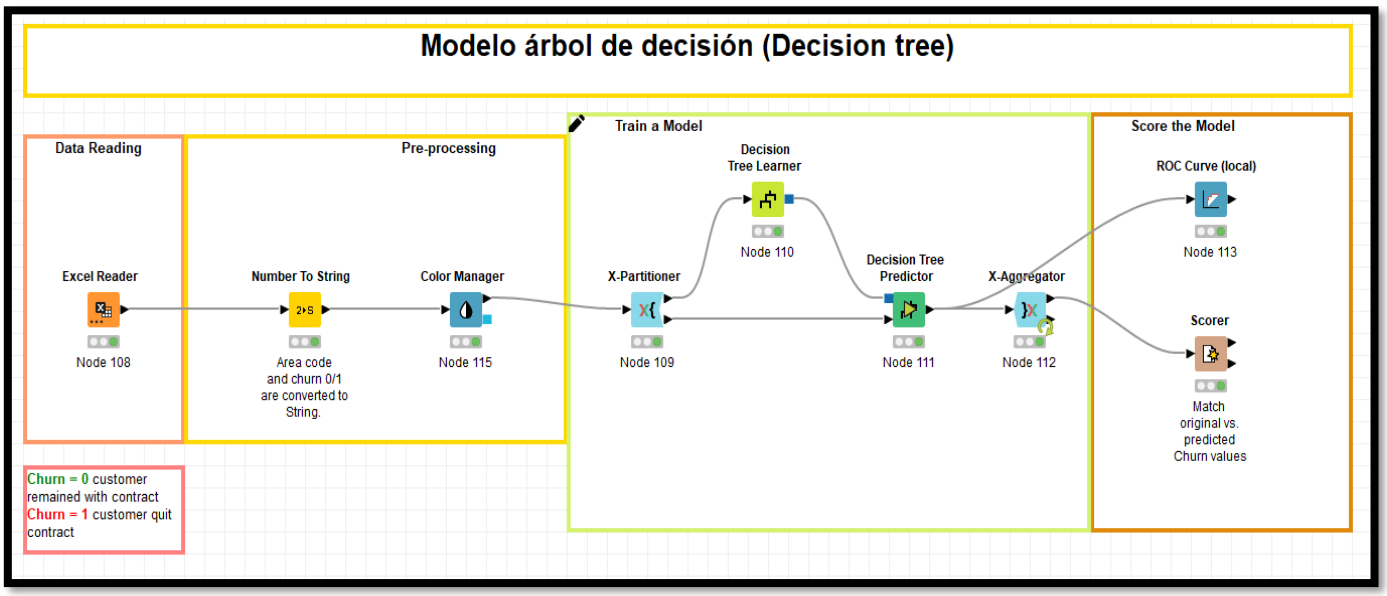




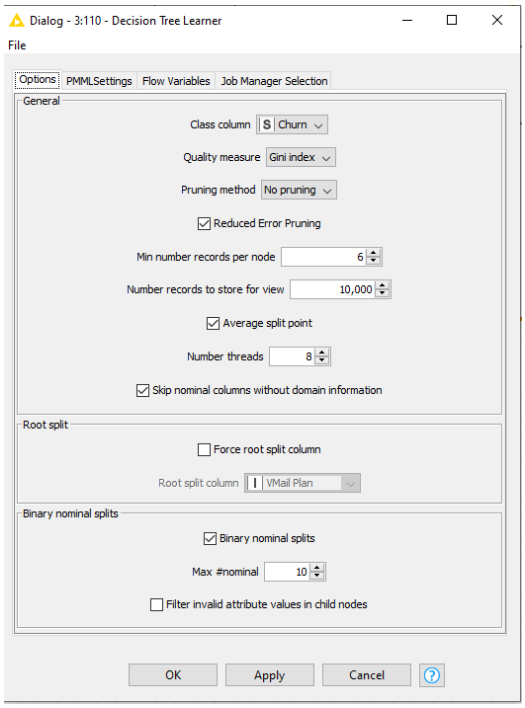
IMPLEMENTACIÓN DE MODELOS DE CLASIFICACIÓN EN KNIME

3.10 Método de árbol de decisión (Decision tree)

La siguiente figura muestra el flujo de trabajo en Knime para este método.

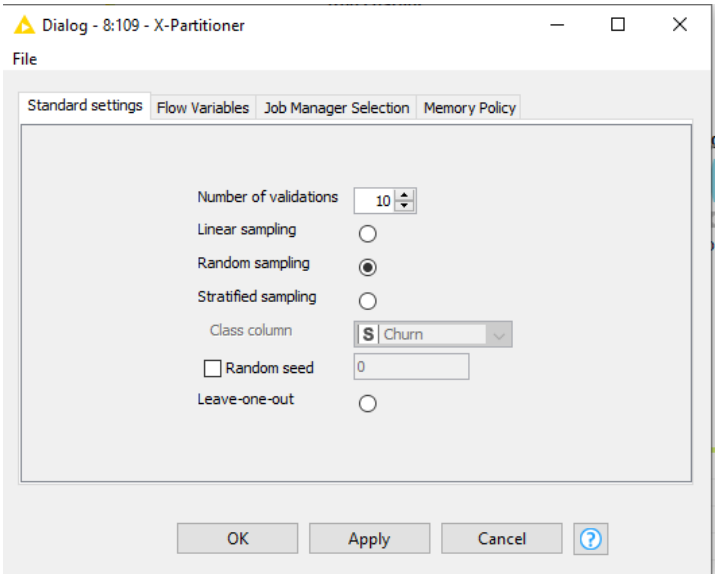


Configuración para el nodo de árbol de decisión

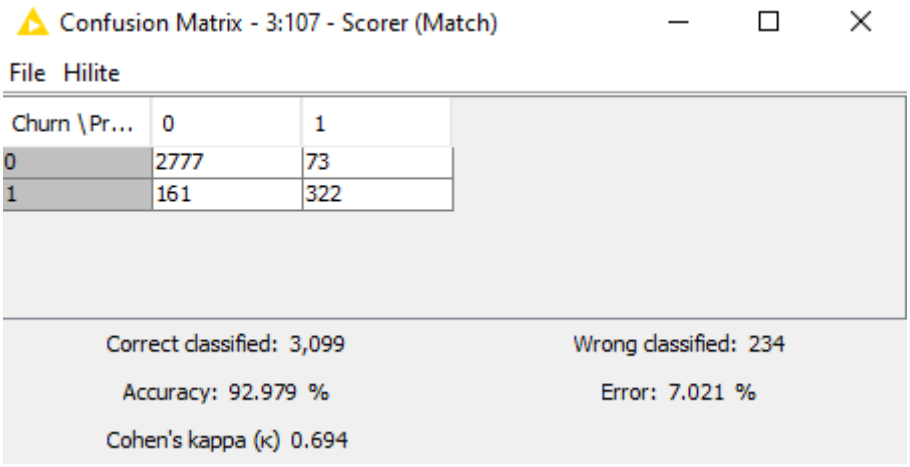


Nota: el algoritmo para el árbol de decisión que genera automáticamente Knime para este nodo es el algoritmo C4.5 desarrollado por Ross Quinlan.

1. Aplicando Random sampling Cross Validation (k=10)



Estos fueron los resultados para la matriz de confusión



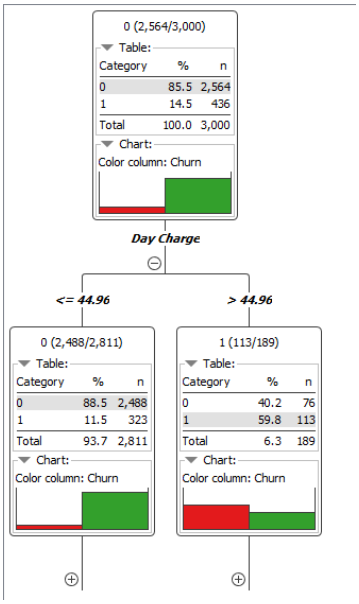
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
0	2777	161	322	73	0.974	0.945	0.974	0.667	0.96	?	?
1	322	73	2777	161	0.667	0.815	0.667	0.974	0.733	?	?
Overall	?	?	?	?	?	?	?	?	?	0.93	0.694

La matriz de confusión muestra que para este modelo 3,099 clientes se clasificaron correctamente mientras que 234 se clasificaron erróneamente. El nivel general de *accuracy* es de **92.979%**

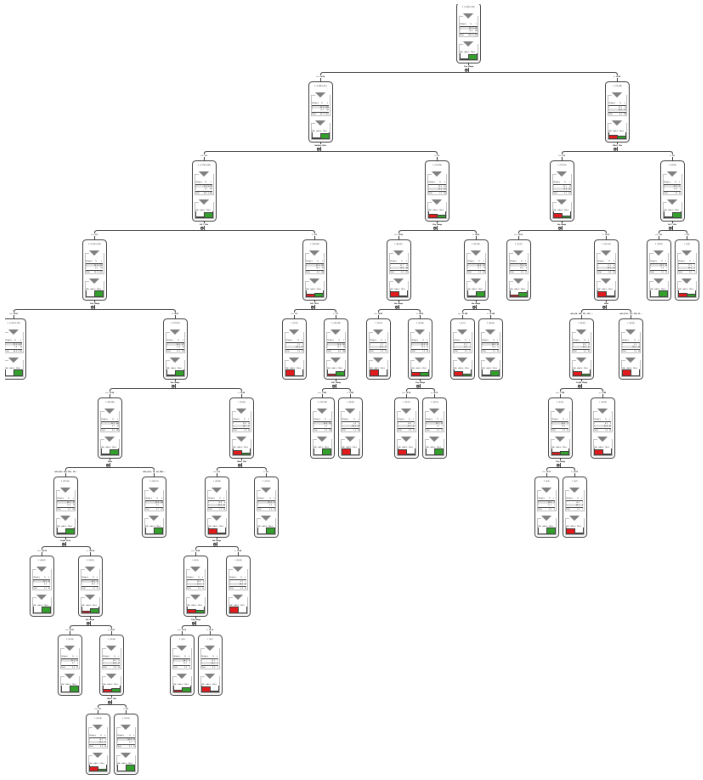
Árbol de decisión generado con este método

Churn=0 (cliente que renueva contrato - **verde**)
Churn=1 (cliente que no renueva contrato - **rojo**)

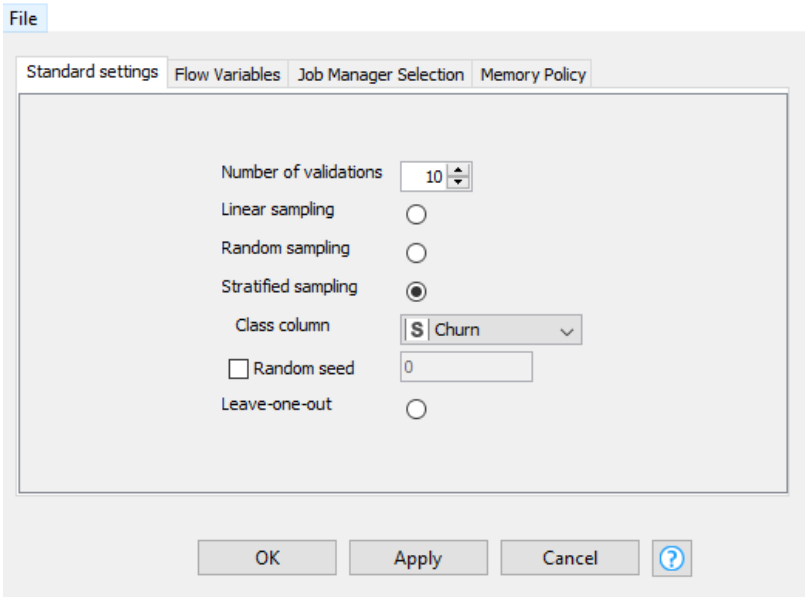
Primeros nodos del árbol de decisión



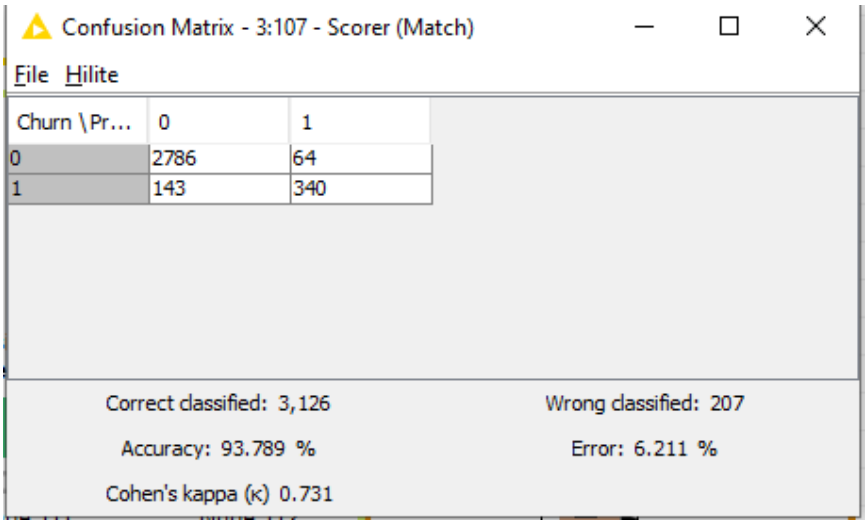
Todo el árbol de decisión



2. Aplicando Stratified K-fold Cross Validation (k=10)



Estos fueron los resultados para la matriz de confusión



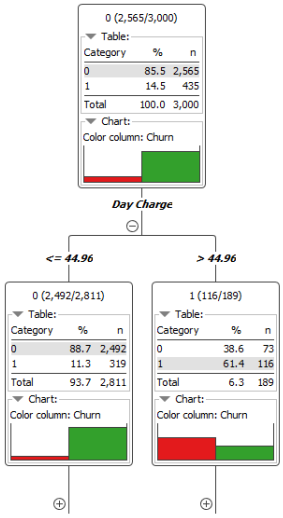
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
0	2786	143	340	64	0.978	0.951	0.978	0.704	0.964	?	?
1	340	64	2786	143	0.704	0.842	0.704	0.978	0.767	?	?
Overall	?	?	?	?	?	?	?	?	?	0.938	0.731

La matriz de confusión muestra que para este modelo 3,126 clientes se clasificaron correctamente mientras que 207 se clasificaron erróneamente. El nivel general de *accuracy* es de **93.789%**

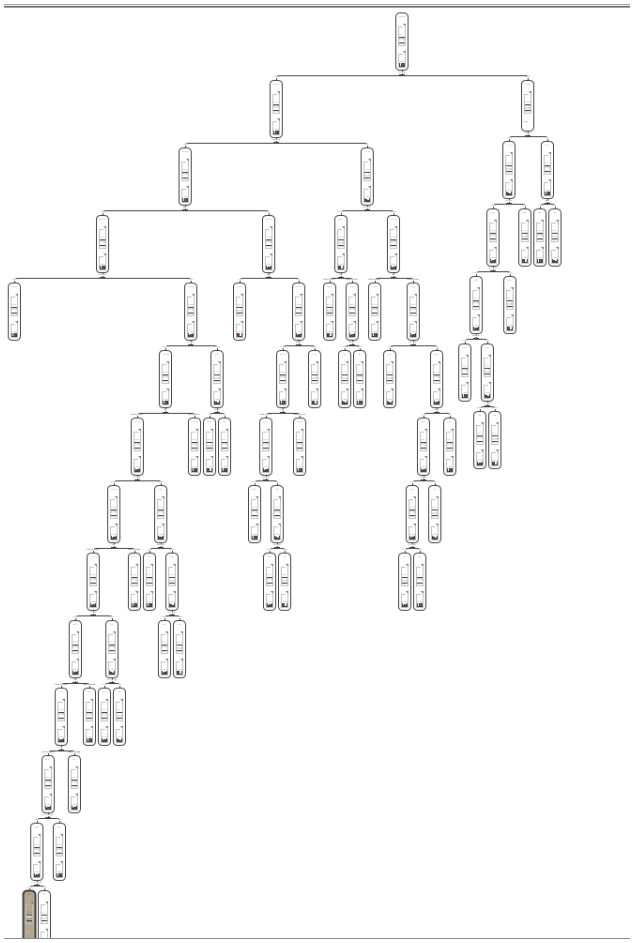
Árbol de decisión generado con este método

Churn=0 (cliente que renueva contrato - **verde**)
Churn=1 (cliente que no renueva contrato - **rojo**)

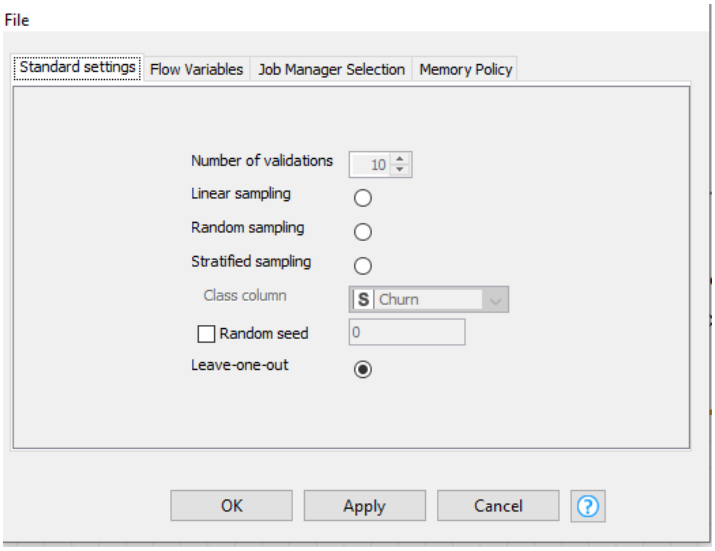
Primeros nodos del árbol de decisión



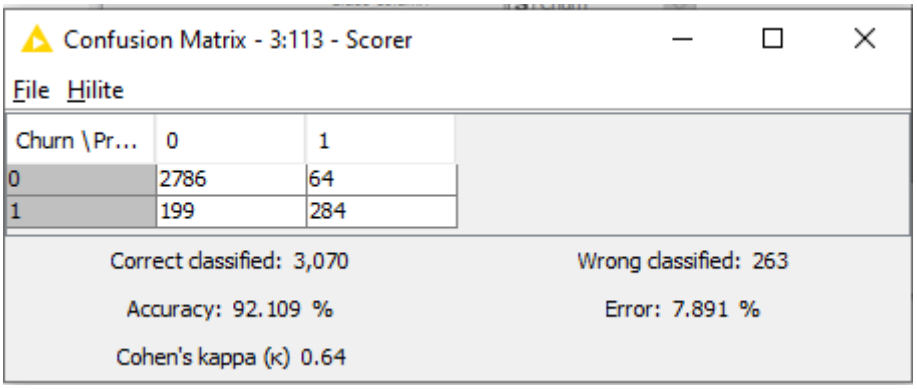
Todo el árbol de decisión



3. Aplicando Leave One Out Cross Validation



Estos fueron los resultados para la matriz de confusión



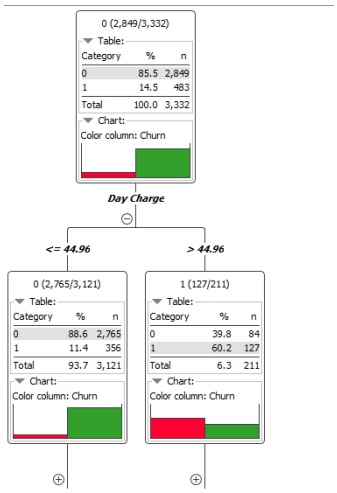
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen'...
0	2786	199	284	64	0.978	0.933	0.978	0.588	0.955	?	?
1	284	64	2786	199	0.588	0.816	0.588	0.978	0.684	?	?
Overall	?	?	?	?	?	?	?	?	?	0.921	0.64

La matriz de confusión muestra que para este modelo 3,070 clientes se clasificaron correctamente mientras que 263 se clasificaron erróneamente. El nivel general de *accuracy* es de **92.109%**

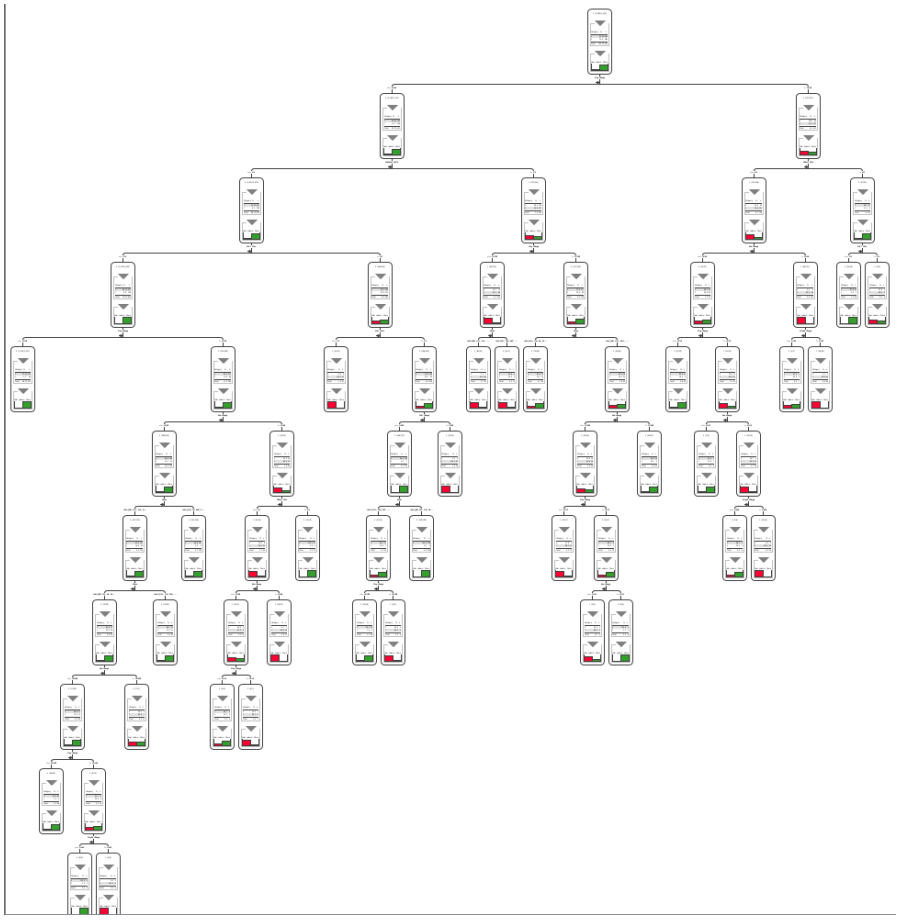
Árbol de decisión generado con este método

Churn=0 (cliente que renueva contrato - **verde**)
Churn=1 (cliente que no renueva contrato - **rojo**)

Primeros nodos del árbol de decisión

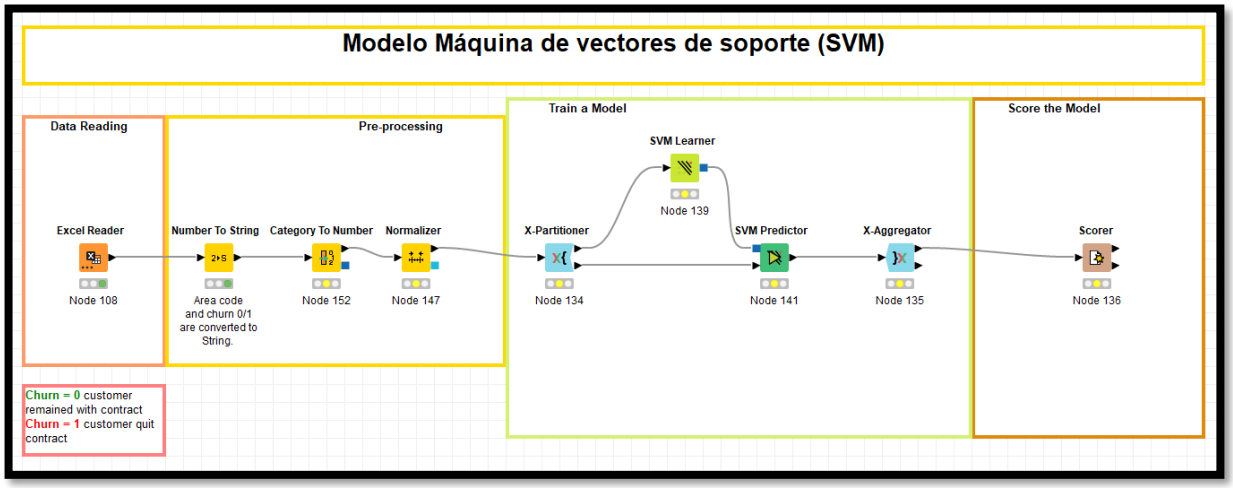


Todo el árbol de decisión

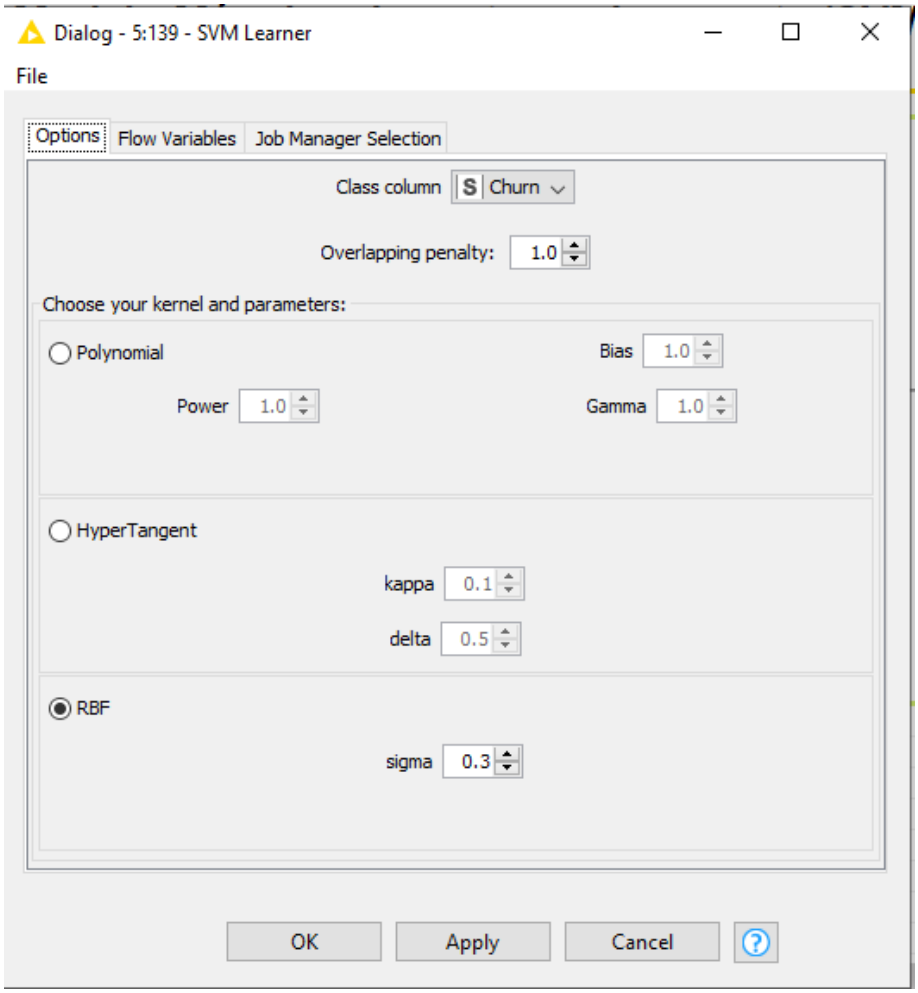


3.11 Método de Máquina de vectores de soporte (SVM)

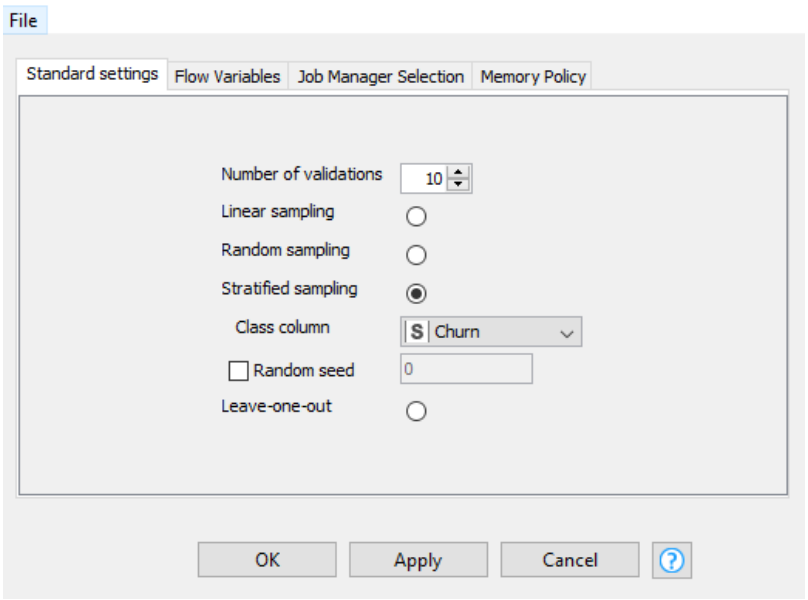
La siguiente figura muestra el flujo de trabajo en Knime para este método.



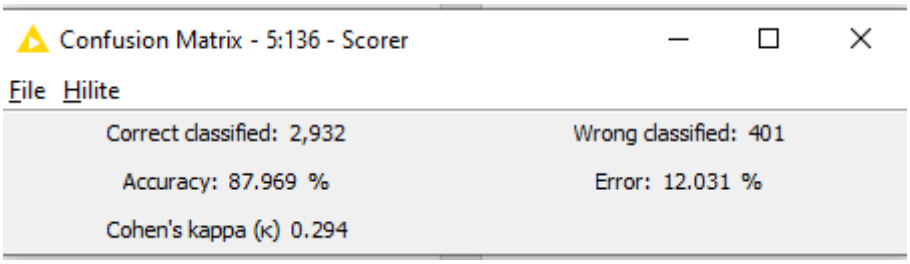
Configuración para el nodo de Máquina de vectores de soporte (SVM)



1. Aplicando Random sampling Cross Validation (k=10)

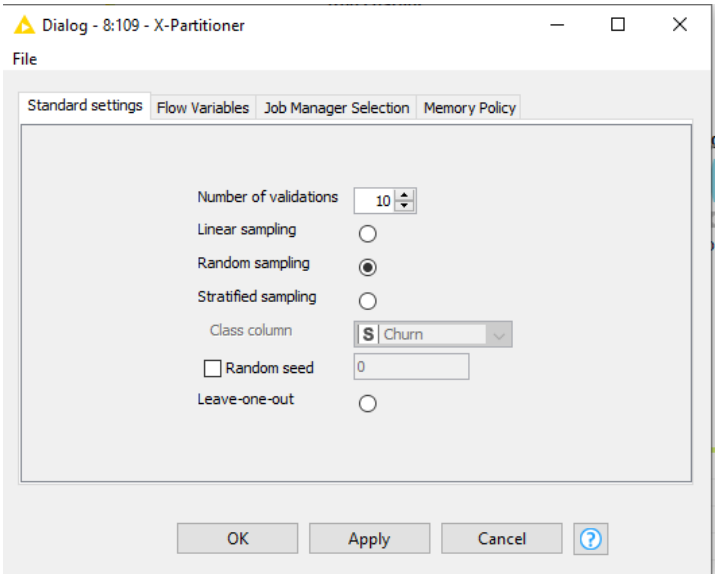


Estos fueron los resultados para la matriz de confusión

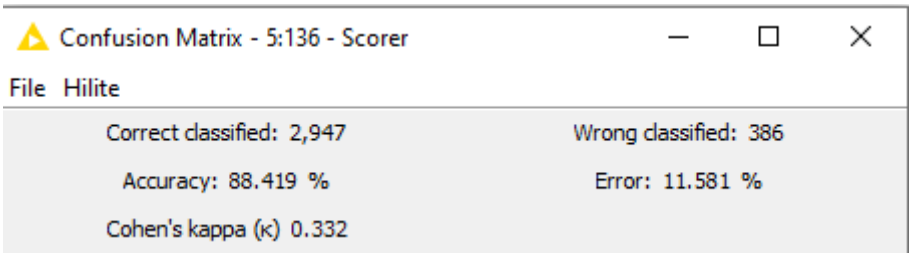


La matriz de confusión muestra que para este modelo 2,932 clientes se clasificaron correctamente mientras que 401 se clasificaron erróneamente. El nivel general de *accuracy* es de **87.969 %**

2. Aplicando Stratified K-fold Cross Validation (k=10)

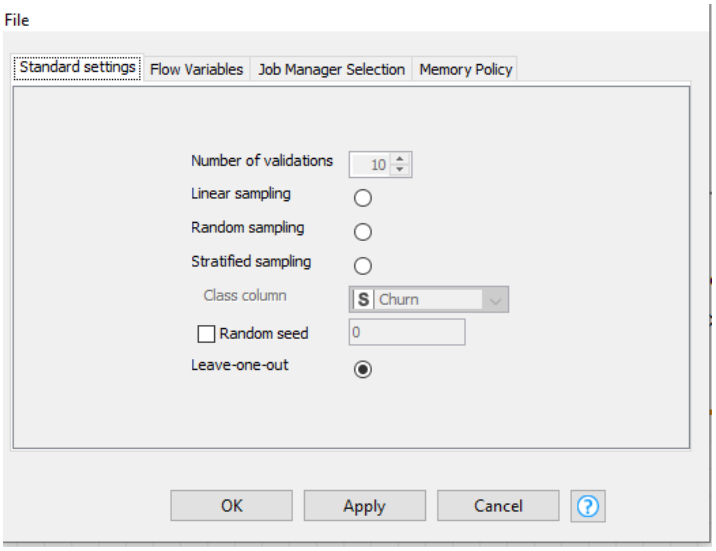


Estos fueron los resultados para la matriz de confusión

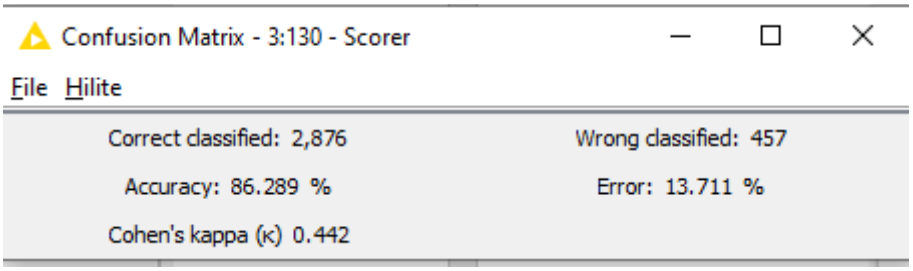


La matriz de confusión muestra que para este modelo 2,947 clientes se clasificaron correctamente mientras que 386 se clasificaron erróneamente. El nivel general de *accuracy* es de **88.419 %**

3. Aplicando Leave One Out Cross Validation



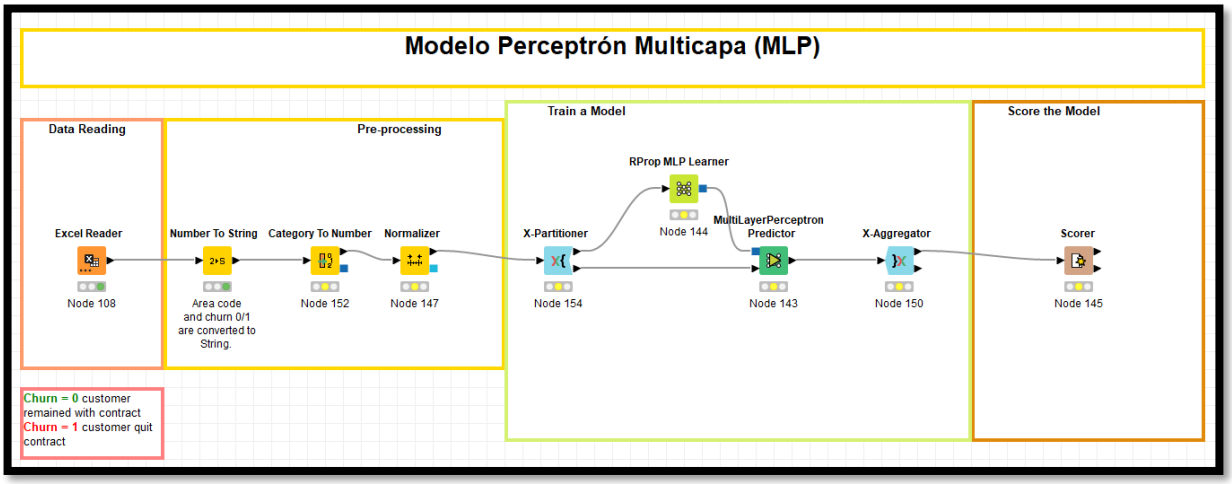
Estos fueron los resultados para la matriz de confusión



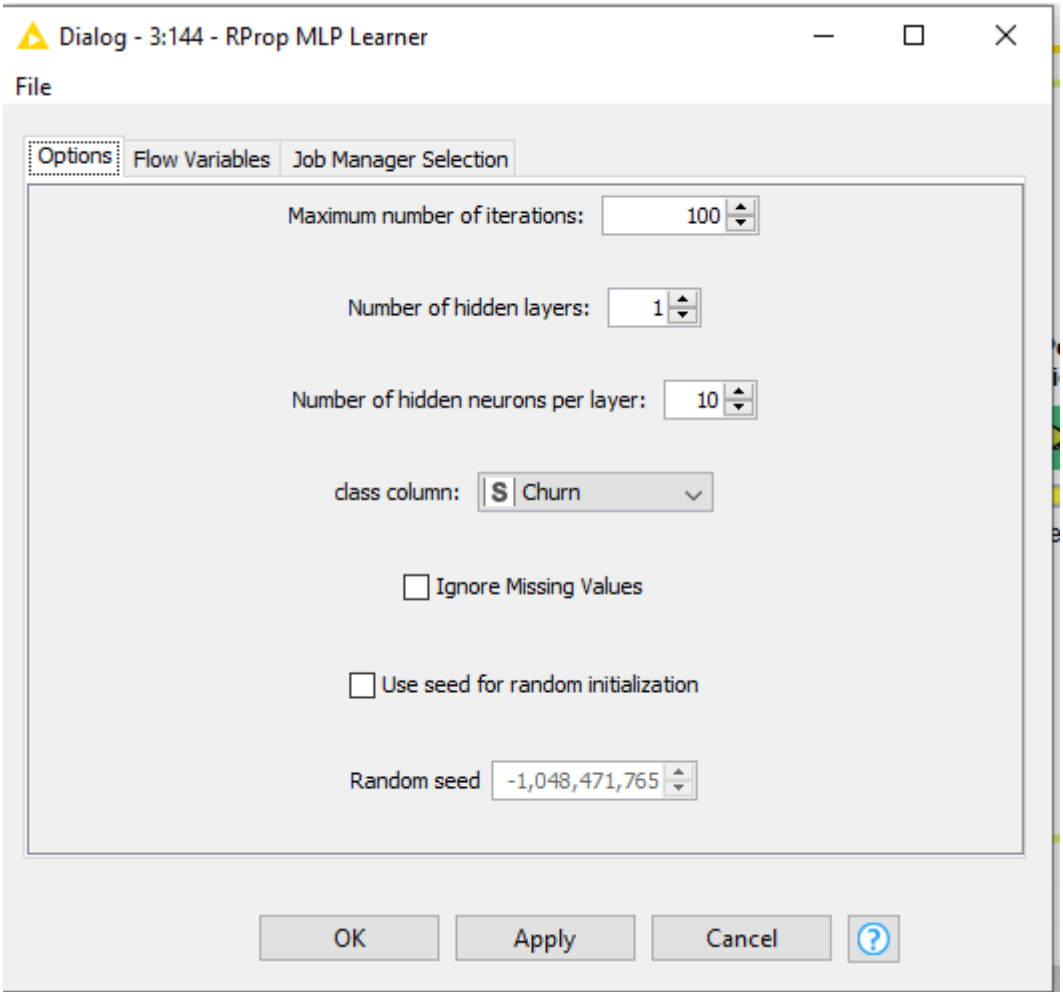
La matriz de confusión muestra que para este modelo 2,876 clientes se clasificaron correctamente mientras que 457 se clasificaron erróneamente. El nivel general de *accuracy* es de **86.289 %**

3.12 Método de Perceptrón Multicapa (MLP)

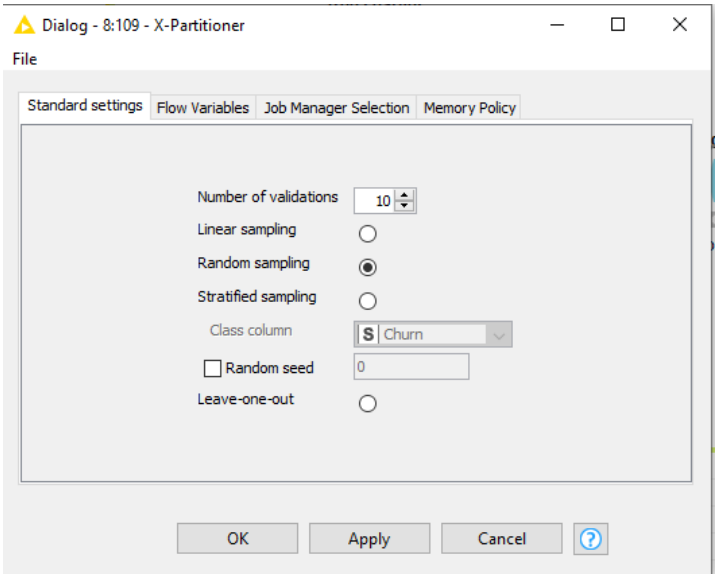
La siguiente figura muestra el flujo de trabajo en Knime para este método.



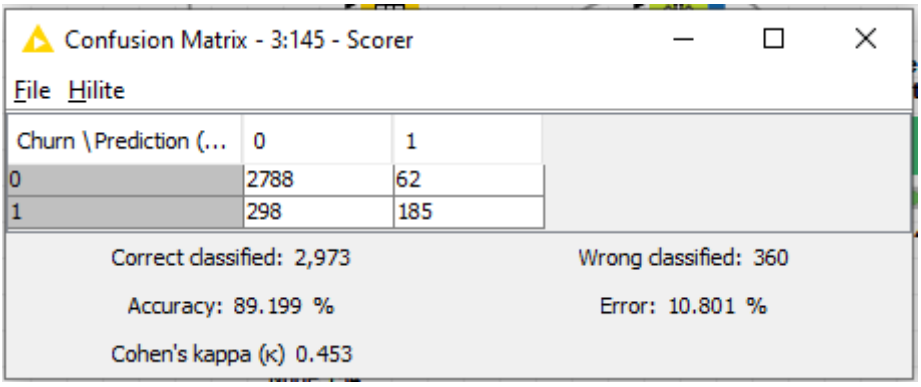
Configuración para el nodo de Perceptrón Multicapa (MLP)



1. Aplicando Random sampling Cross Validation (k=10)

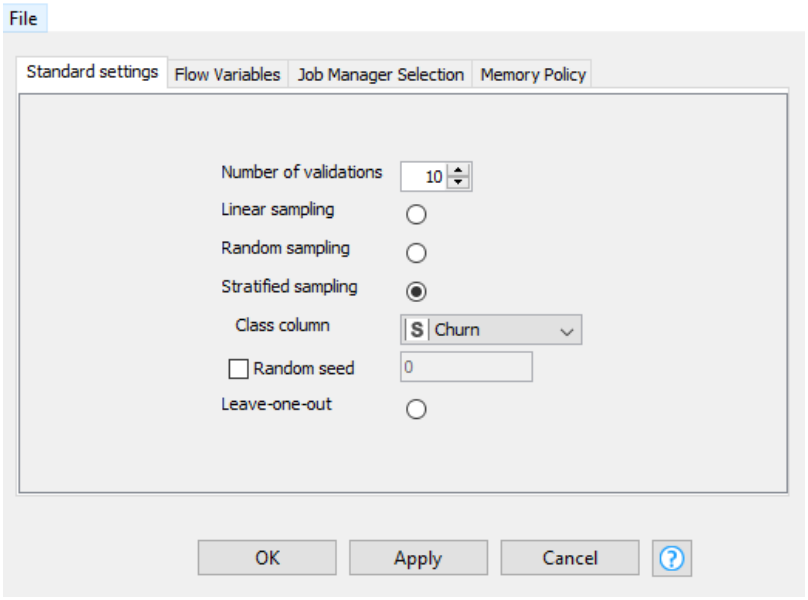


Estos fueron los resultados para la matriz de confusión

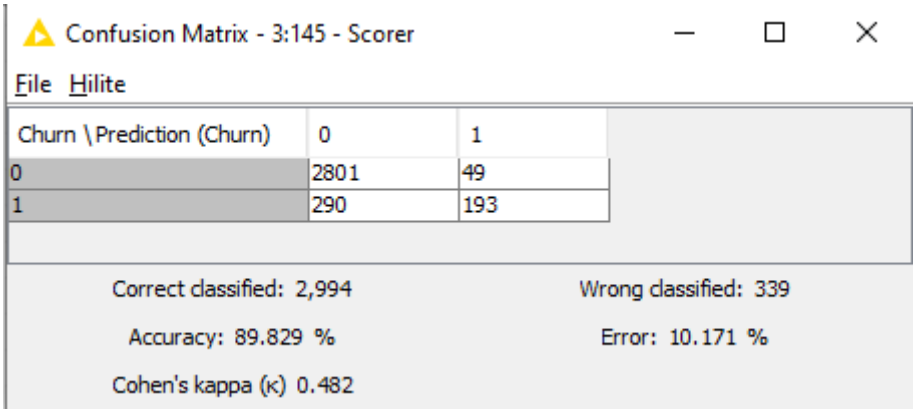


La matriz de confusión muestra que para este modelo 2,973 clientes se clasificaron correctamente mientras que 360 se clasificaron erróneamente. El nivel general de *accuracy* es de **89.199%**

2. Aplicando Stratified K-fold Cross Validation (k=10)

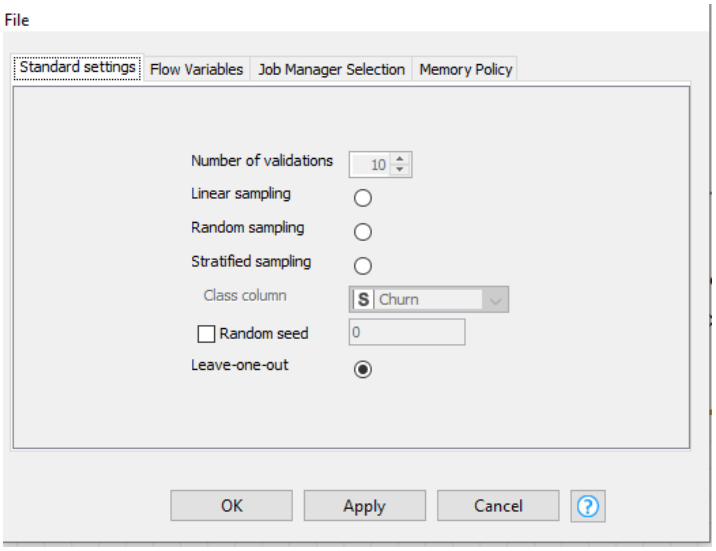


Estos fueron los resultados para la matriz de confusión

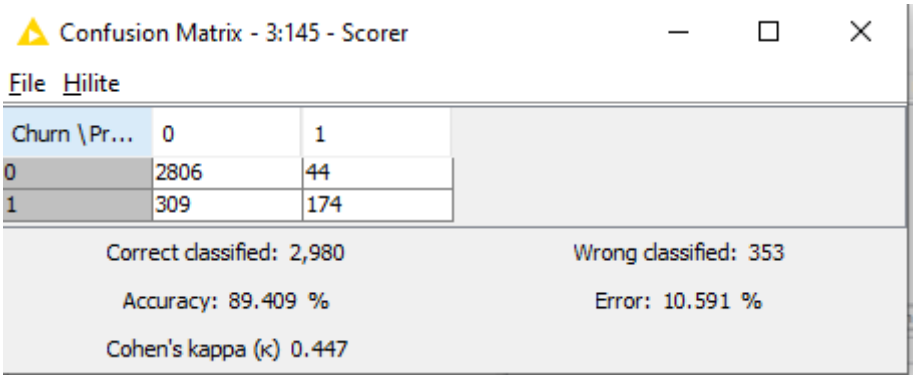


La matriz de confusión muestra que para este modelo 2,994 clientes se clasificaron correctamente mientras que 339 se clasificaron erróneamente. El nivel general de *accuracy* es de **89.829%**

3. Aplicando Leave One Out Cross Validation



Estos fueron los resultados para la matriz de confusión



La matriz de confusión muestra que para este modelo 2,806 clientes se clasificaron correctamente mientras que 353 se clasificaron erróneamente. El nivel general de *accuracy* es de **89.409 %**

3.13 Evaluación de resultados

Los resultados obtenidos en los anteriores procesos de clasificación se resumen en la siguiente tabla:

		Total de instancias en dataset:3,333				
Clasiificador	**Test de Prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Accuracy (%)**	**Índice Kappa**	**Error (%)**
Árbol de decisión	Random sampling Cross Validation (k=10)	3,099	234	**92.979**	0.694	7.021
Árbol de decisión	Stratified K-fold Cross Validation (k=10)	3,126	207	**93.789**	0.731	6.211
Árbol de decisión	Leave One Out Cross Validation	3,070	263	**92.109**	0.640	7.891
Máquina de vectores de soporte (SVM)	Random sampling Cross Validation (k=10)	2,932	401	**87.969**	0.294	12.031
Máquina de vectores de soporte (SVM)	Stratified K-fold Cross Validation (k=10)	2,947	386	**88.419**	0.332	11.581
Máquina de vectores de soporte (SVM)	Leave One Out Cross Validation	2,876	457	**86.289**	0.442	13.711
Perceptrón Multicapa (MLP)	Random sampling Cross Validation (k=10)	2,973	360	**89.199**	0.453	10.810
Perceptrón Multicapa (MLP)	Stratified K-fold Cross Validation (k=10)	2,994	339	**89.829**	0.482	10.171
Perceptrón Multicapa (MLP)	Leave One Out Cross Validation	2,980	353	**89.409**	0.447	10.591

De la tabla podemos concluir que el clasificador árbol de decisión presenta valores superiores al 90% de instancias correctamente clasificadas para todos sus modos de prueba. Máquina de vectores de soporte (SVM) supera el 86% de instancias correctamente clasificadas para todos sus modos de prueba y Perceptrón Multicapa (MLP) supera el 89 % de instancias correctamente clasificadas para todos sus modos de prueba.

En relación a los diferentes modos de prueba, se pudo determinar que Stratified K-fold Cross Validation es el que presenta mejores resultados en todos los clasificadores. Recordemos que Stratified K-fold Cross Validation asegura que los conjuntos de entrenamiento y prueba tengan la misma proporción de la característica de interés que en el conjunto de datos

original. Esto asegura que ningún valor esté sobre o infrarrepresentado en los conjuntos de entrenamiento y prueba, lo que proporciona una estimación más precisa del rendimiento y una aproximación cercana al error de generalización.

En cuanto al modo de prueba que presenta peores resultados es el Leave One Out Cross Validation, con una sola excepción para el clasificador Perceptrón Multicapa (MLP), ya que ahí Leave One Out Cross Validation obtuvo mejor resultado que Random sampling Cross Validation, pero en los otros dos clasificadores Leave One Out Cross Validation fue el método de prueba que obtuvo los peores resultados. Cabe mencionar que el método Leave One Out Cross Validation, si bien es fácil de implementar, presenta una enorme desventaja que es su coste computacional. El tiempo requerido para ejecutar Leave One Out Cross Validation fue de casi cuatro veces más que para ejecutar los otros dos métodos de prueba. El tiempo de ejecución para Random sampling Cross Validation y Stratified K-fold Cross Validation fue muy similar.

En relación al tipo de clasificador que mejores resultados ofrece, concretamente para los valores de los índices Kappa, fue árbol de decisión para todos sus modos de prueba, seguido por Perceptrón Multicapa (MLP) y por último Máquina de vectores de soporte (SVM).

Finalmente, es el clasificador Máquina de vectores de soporte (SVM) el que peores resultados ha tenido. Aun así, los resultados obtenidos por este clasificador no son del todo malos ya que supera el 86% de instancias correctamente clasificadas en todos sus modos de prueba.

3.14 Conclusiones

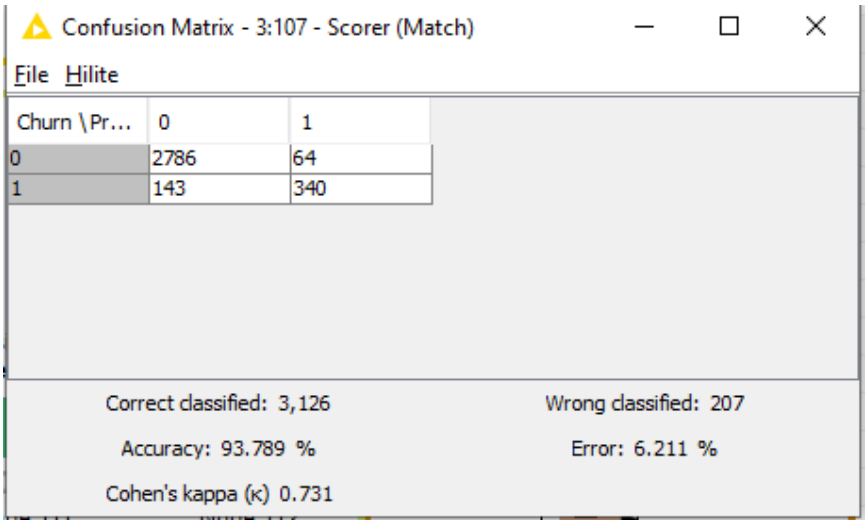
En este caso práctico, se presentó un experimento para comparar tres modelos de clasificación de minería de datos, aplicados a un conjunto de datos de abandono o de no renovación del contrato de servicios de una compañía telefónica. Después de evaluar estos modelos de minería de datos utilizando el software KNIME, se encontró que el modelo de árbol de decisión con el modo de prueba Stratified K-fold Cross Validation es el más adecuado para este problema de predicción de abandono (*churn prediction*).

En conclusión, se ha podido diseñar con éxito un modelo de clasificación basado en un árbol de decisión que permite clasificar la posibilidad de que un cliente abandone su contrato de servicio de telefonía. Además de la gran precisión que este clasificador ofrece para este caso en específico, un árbol de decisión tiene la ventaja de ser fácil de interpretar y permite al usuario, de forma rápida y sencilla, determinar si un cliente, dado un conjunto de atributos que define su comportamiento histórico, está en riesgo de abandono del servicio.
Por lo tanto, la técnica de clasificación seleccionada para el diseño del modelo predictivo propuesto en este caso práctico es el **Árbol de decisión - Stratified K-fold Cross Validation**.

MODELO MÁS ADECUADO PARA ESTE TRABAJO:
Árbol de decisión - Stratified K-fold Cross Validation.

		Total de instancias en dataset:3,333				
Clasiificador	**Test de Prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Accuracy (%)**	**Índice Kappa**	**Error (%)**
Árbol de decisión	Stratified K-fold Cross Validation (k=10)	3,126	207	**93.789**	0.731	6.211

Estos fueron los resultados para la matriz de confusión



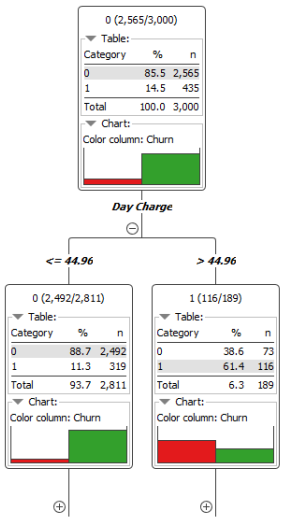
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
0	2786	143	340	64	0.978	0.951	0.978	0.704	0.964	?	?
1	340	64	2786	143	0.704	0.842	0.704	0.978	0.767	?	?
Overall	?	?	?	?	?	?	?	?	?	0.938	0.731

La matriz de confusión muestra que para este modelo 3,126 clientes se clasificaron correctamente mientras que 207 se clasificaron erróneamente. El nivel general de *accuracy* es de **93.789%**

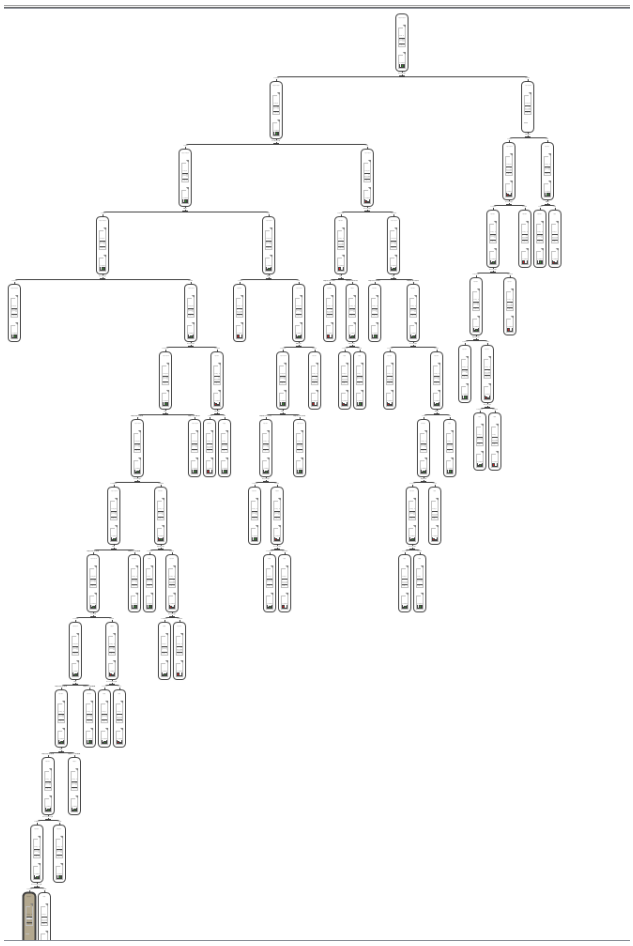
Árbol de decisión generado para: **- Stratified K-fold Cross Validation**.

Churn=0 (cliente que renueva contrato - **verde**)
Churn=1 (cliente que no renueva contrato - **rojo**)

Primeros nodos del árbol de decisión

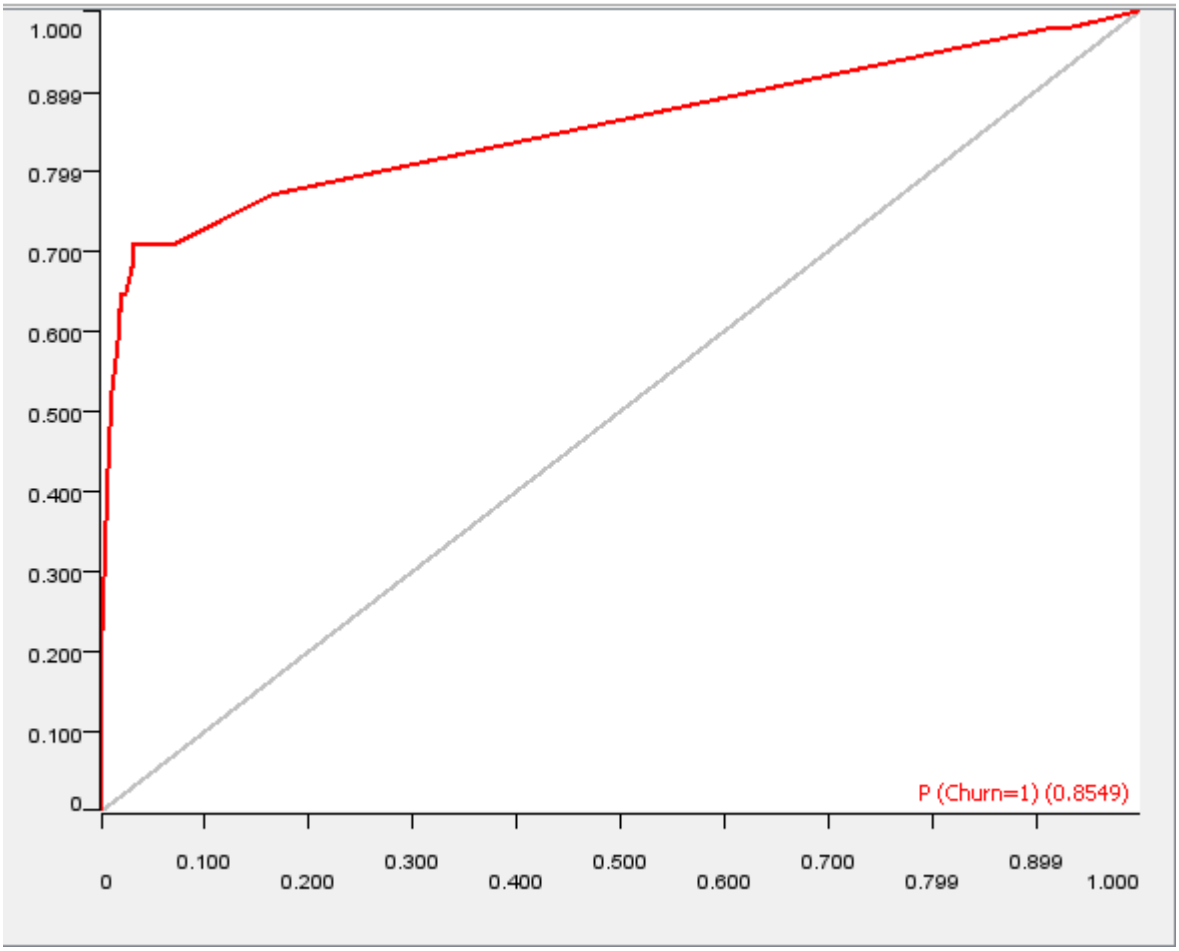


Todo el árbol de decisión



Curva ROC para: **Árbol de decisión - Stratified K-fold Cross Validation**.

Curva ROC **AUC=0.8549**



El gráfico de la curva ROC para este modelo de árbol de decisión, presenta un AUC de 0.8549 lo que valida el modelo como bueno, el nivel general de *accuracy* es de **93.789%**

3.15 BIBLIOGRAFÍA

- B. Kröse and P. van der Smagt. An introduction to Neural Networks. None, (1996). 135 P.
- Berthold, Michael R.; Cebron, Nicolas; Dill, Fabian; (16 November 2009). "KNIME - the Konstanz information miner" (PDF). ACM SIGKDD Explorations Newsletter.
- Coussement, Kristof, and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert systems with applications, vol. 34, pp. 313-327, (2008).
- Dickey.Introduction to Predictive Modeling with Examples. Obtained from SAS Global. (2012).
- Foxal, G., "Foundations of Consumer Behaviour Analysis" Marketing Theory,(2020). pp 95–199
- Haro S., 2017. Técnicas de clasificación en minería de datos y software Orange Canvas. Aplicación con datos meteorológicos. Trabajo fin de máster. Universidad de Granada.
- Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. (12) 11373.
- Lalwani, Praveen; Mishra, Manas Kumar; Chadha, Jasroop Singh; Sethi, Pratyush (2021). "Customer churn prediction system: a machine learning approach". Computing.pp.37-73
- Maimon, O., Rokach, L. Introduction to Knowledge Discovery and Data Mining. 2nd Edition. New York: Springer, (2010). p.1-15.
- Mehta, Nick; Steinman, Dan; Murphy, Lincoln (2016-02-16). Customer Success: How Innovative Companies Are Reducing Churn and Growing Recurring Revenue. John Wiley & Sons. p. 84.
- Morgan, Neil A., and Lopo Leotte Rego. "The value of different customer satisfaction and loyalty metrics in predicting business performance." Marketing Science 25, no. 5 (2006): pp. 426-439
- N. K. Kasabov. Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. MIT Press, Cambridge, MA, USA, 1st edition, (1996). 550 pp.
- Rokach, L., Maimon, O. (2008). Data Mining with Decision Trees: Theory and Applications. World Scientific.
- Shaaban, E., Helmy, Y., Khedr, A. and Nasr, M., 2012. A proposed churn prediction model. International Journal of Engineering Research and Applications, vol. 2, pp. 693-697, (2012).
- Stanton, William J (1984). Fundamentals of marketing. McGraw-Hill
- Venkatesan, Rajkumar, Paul Farris and Ron Wilcox (2014), Cutting Edge Marketing Analytics: Real World Cases and Datasets for Hands On Learning. Pearson/FT Press, NY, NY.
- Wei, Chih-Ping, and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach, Expert systems with applications, vol. 23, pp. 103-112, (2002).
- Xia, G.E. and Jin, W.D., 2008. Model of customer churn prediction on support vector machine. Systems Engineering-Theory & Practice, vol. 28, pp. 71- 77, (2008).