Informe sobre el proceso de carga de datos en RapidMiner y análisis de hallazgos encontrados en un data set sobre características de Precios de la vivienda: predecir precios de venta y practicar ingeniería de características, RF y aumento de gradiente

Brayan David Pisuña Paillacho

Tecnología en Análisis de Sistemas Informáticos

Escuela Politécnica Nacional

brayan.pisuna@epn.edu.ec

***Resumen*- Este documento presenta en detalle el proceso a seguir para llevar a cabo la carga de datos, trasformación y depuración de los datos, generación y comprobación de modelos predicativos, y análisis de hallazgos en un data set, aplicando la herramienta RapidMiner.**

**El dataset muestra varias características de los precios de viviendas y aumento de gradientes registrados, donde se pueden encontrar diferentes características de los precios y que la mayoría de las veces son consideradas por los clientes al momento al adquirir un bien y mueble. A partir de esta data se desea validar que características apalancan de mejor manera la variable “House Price” que representa técnicas avanzadas de regresión, la misma que es definida como la variable objetivo que se va a predecir a través de un modelo.**

**El objetivo de este informe es elegir los atributos con mayor correlación con el precio de la vivienda y definir los modelos predictivos más eficientes a ser considerado a la hora de su elección. Además, esta actividad permitirá, predecir su precio de venta, conocimientos adquiridos respecto a minería de datos a través del uso de la herramienta RapidMiner.**

***Palabras Clave*- RapidMiner, minería de datos, modelos predictivos.**

1. INTRODUCCIÓN

El proceso e informe trata sobre la carga en RapidMiner y los hallazgos en un data set generado con métodos que hemos venido aprendiendo en el desarrollo de este semestre, los mismo serán utilizados para volcarlos a una base de datos en MongoDb que almacenara de forma local en la nube.

Deben considerarse en la implementación de modelo deben ser soportada directamente con los procesos estadísticos que nos permitirá medir cuantitativamente el nivel de influencia al momento de elegir precios de vivienda y aumento de gradiente.

1. DESARROLLO

Una vez realizado el análisis correspondiente a las indicaciones procederemos a utilizar las herramientas adecuadas para el respectivo proceso.

1. *Análisis exploratorio e ingesta de datos.*

Dentro de la información descarga de Kaggle, se observa los siguientes archivos:

Tabla I

DATASET FUENTE

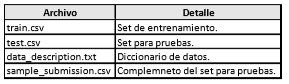


Tabla II

TRAIN.CSV FILTRADOS

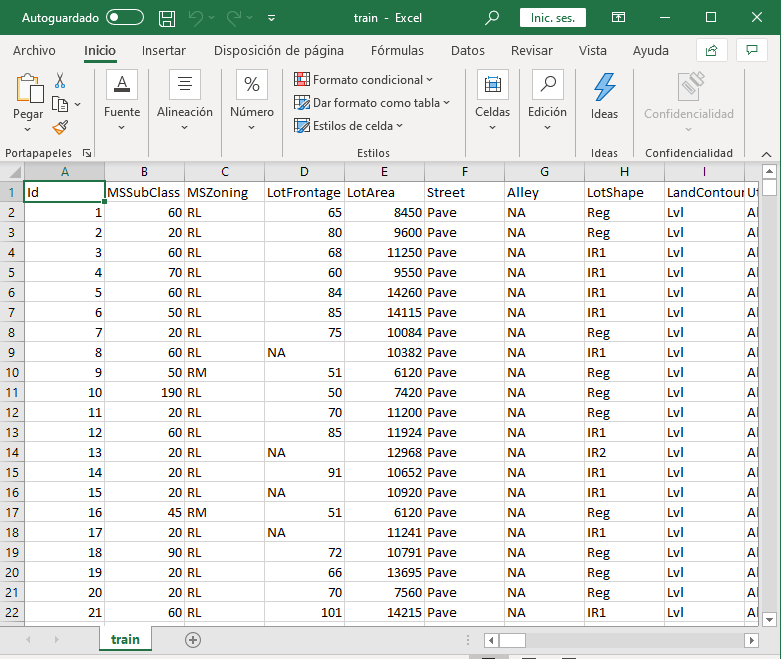
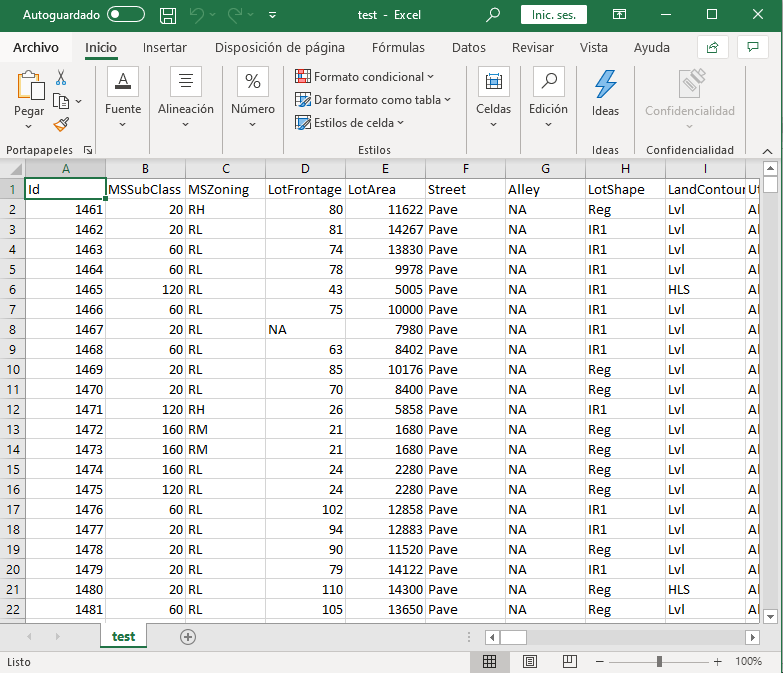


Tabla III

TEST.CSV FILTRADOS



Atributos principales tomados en cuenta para el desarrollo de la minería de datos.

* LotArea: tamaño de lote en pies cuadrados, se da de gran importancia en base a que recibe el valor que obtendrá la casa.
* Street: Nos brinda la información del tipo de carretera que se obtiene para acceder a las viviendas.
* ExterCond: Da la información y condición del material del exterior de la vivienda.
* OverallQual: material general y calidad de acabado da el valor agregado a las casas.
* SalePrice: El precio de venta de la propiedad en dólares, este será el variable objetivo que se va a predecir en el desarrollo.

Fundamentar la razón de los modelos seleccionados:

**Deep Learning:** La inteligencia artificial consiste en dejar a la maquina la posibilidad de resolver problemas que solo podían ser resueltos por los humanos.

El machine Learning es la capacidad de que el algoritmo aprenda por sí solo, y el Deep Learning es un tipo de algoritmo capaz de aprender por sí mismo.

**Decisión Tree:** Árbol de decisión o Decisión Tree Classification es un tipo de algoritmo de aprendizaje supervisado que se utiliza principalmente en problemas de clasificación, aunque funciona para variables de entrada y salida categóricas como continuas.

En esta técnica, dividimos la data en dos o más conjuntos homogéneos basados en el diferenciador más significativos en las variables de entrada. El árbol de decisión identifica la variable más significativa y su valor que proporciona los mejores conjuntos homogéneos de población. Todas las variables de entrada y todos los puntos de división posibles se evalúan y se elige la que tenga mejor resultado. [3]

1. *Desarrollo en RapidMiner*

Cargamos los dos data set tanto el Train como el Test ya filtrados como lo observamos en la [Tabla II] [Tabla III].

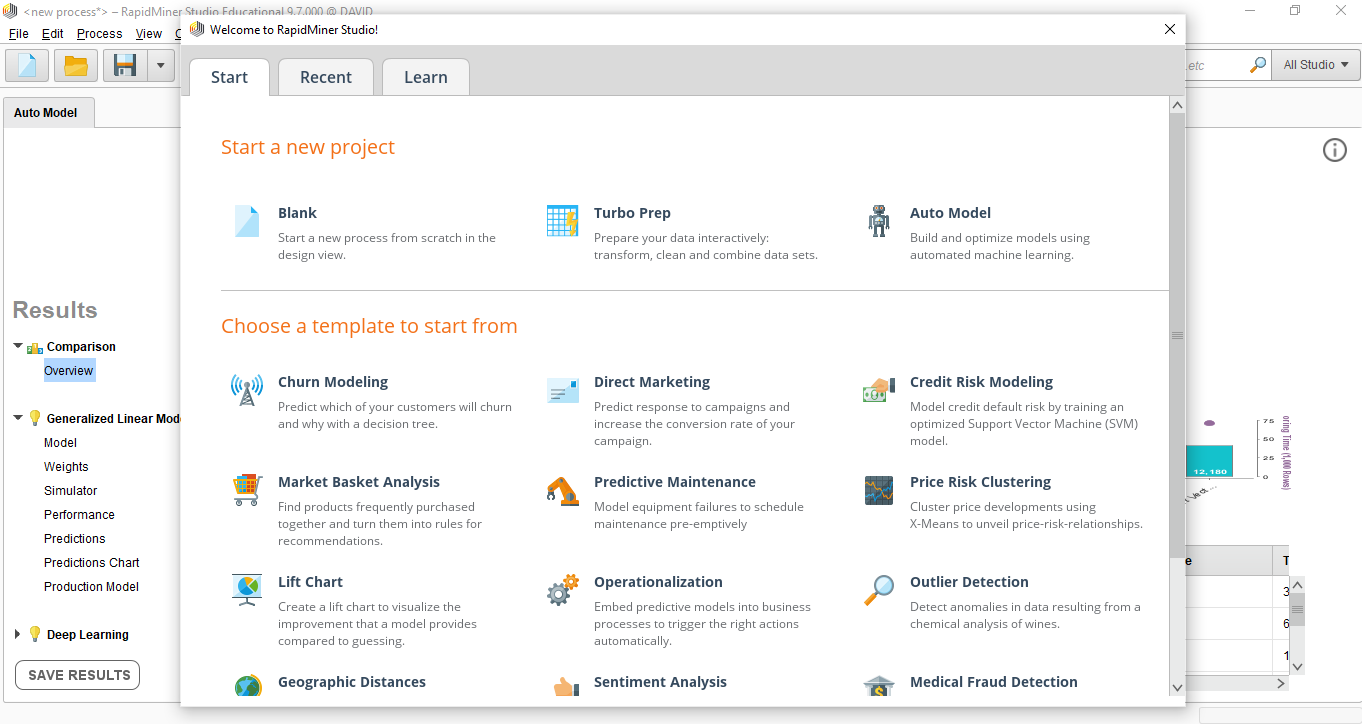


Figura Vista de Interfaz de RapidMiner

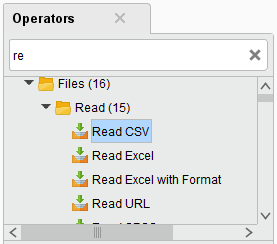


Figura Selección de operador CSV

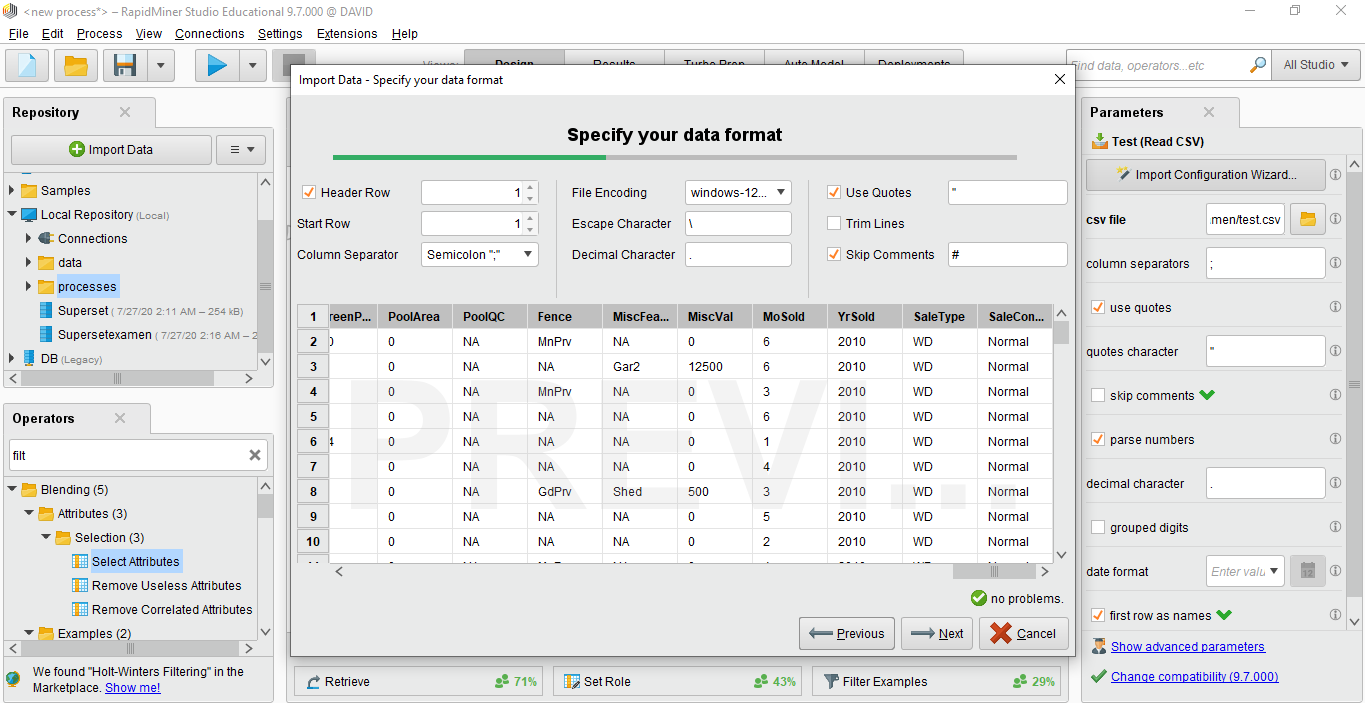


Figura 3 Carga de Dataset

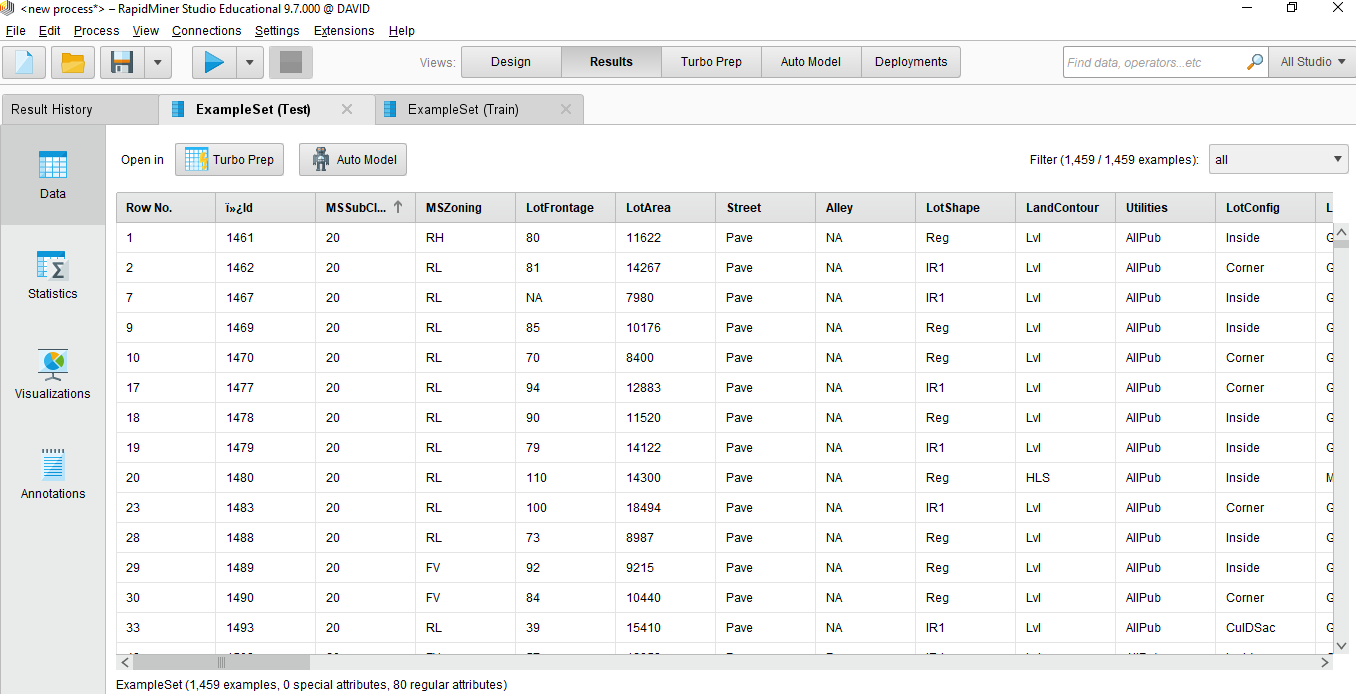


Figura Vista del Dataset ya cargado en RapidMiner

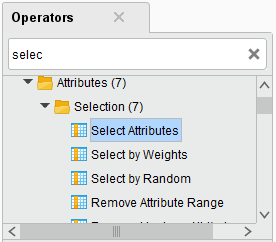


Figura Selección del Operador de Atributos a Escoger

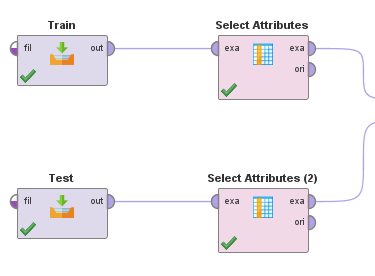


Figura Vista de los elementos ya cargados

Se usará operadores que se puedan utilizar al momento de utilizar nuestro dataset, y luego poder unificar y generar nuestro archivo CSV.

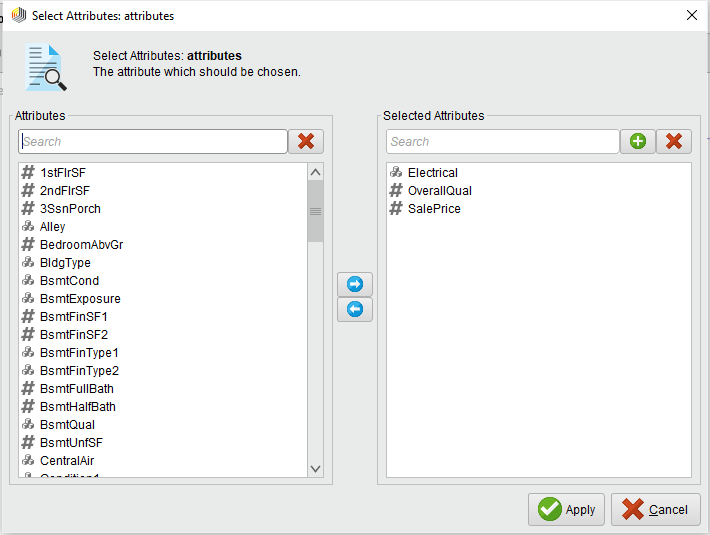


Figura Selección de los atributos para Train

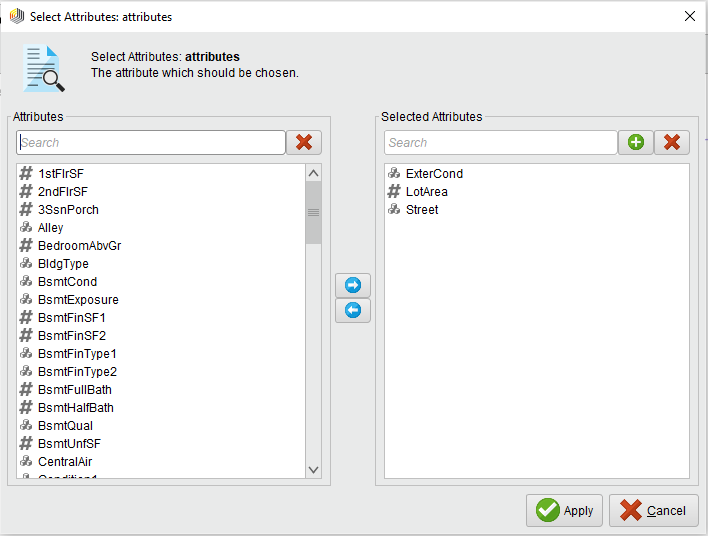


Figura Selección de los atributos Test

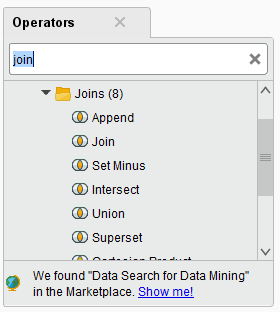


Figura Selección de los operadores Joins en este caso seleccionaremos el Superset

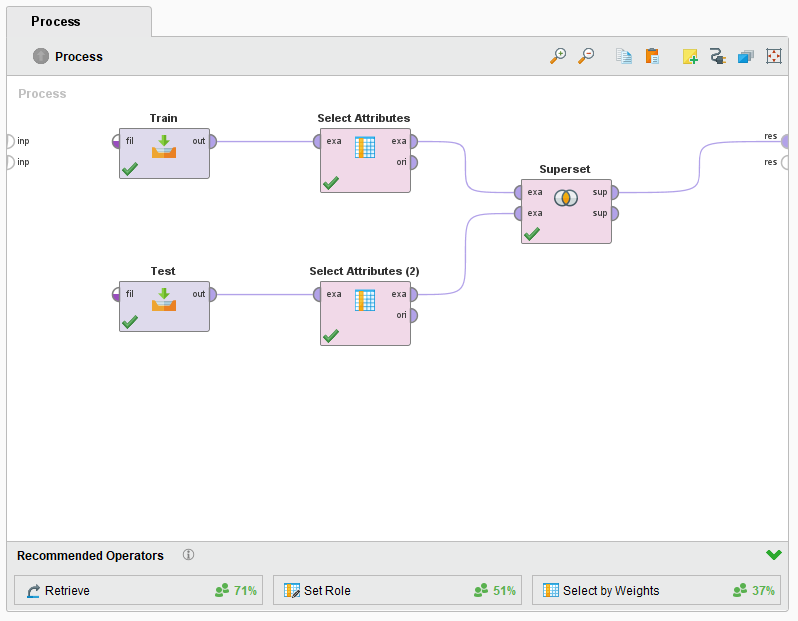


Figura Vista de como ya se encuentra corriendo RapidMiner

El operador Superset toma dos conjuntos y luego genera dos conjuntos de datos. Lo bueno es que los dos conjuntos de datos de salida contienen todas las columnas y sus nombres en cada conjunto. La única diferencia es que no todos los datos están contenidos en cada conjunto.

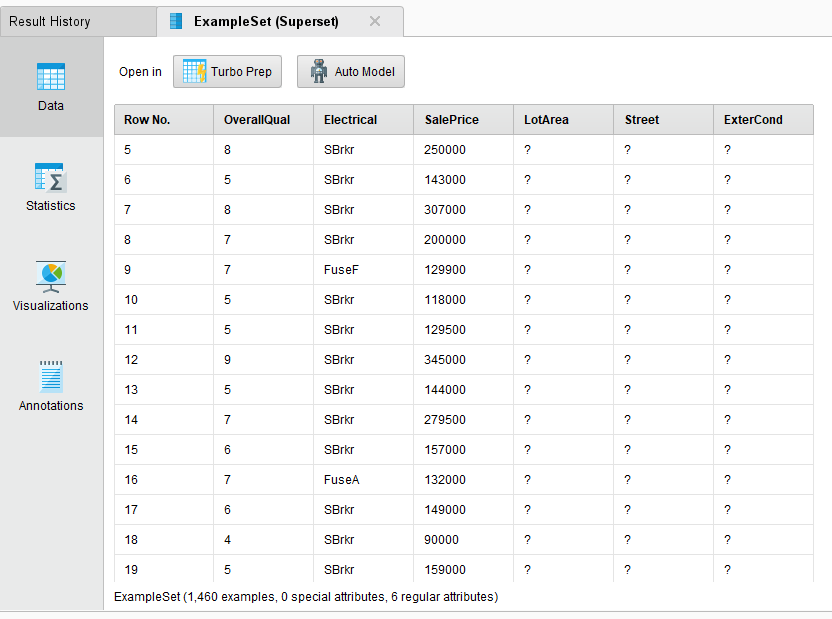


Figura Vista de nuestro SUPERSET ya ejecutándose

Una vez ya en ejecución nuestro proyecto procederemos a guardarlo para esto necesitaremos el operador Write CSV.

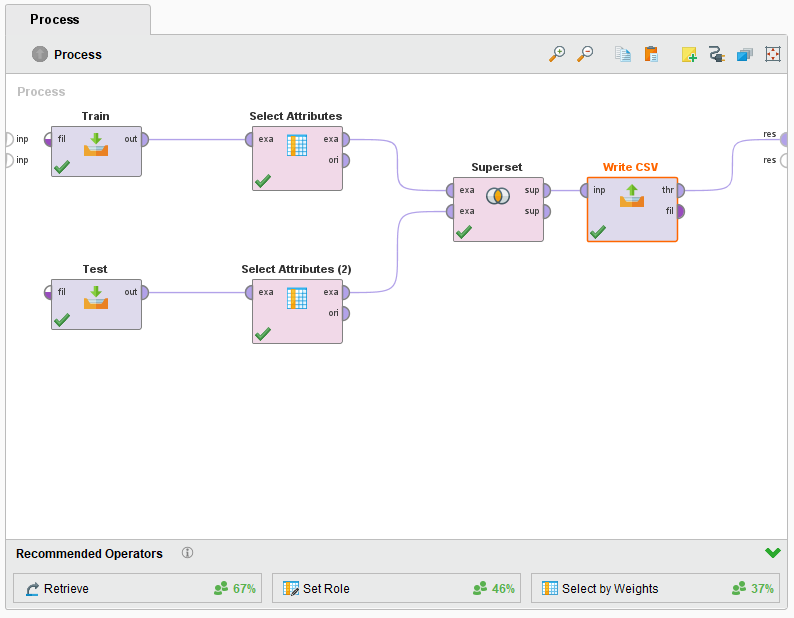


Figura Agregamos el operador WRITE CSV

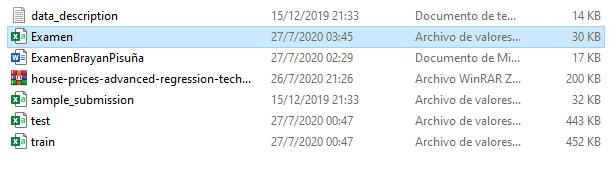


Figura Observación de que se guardó correctamente en la carpeta seleccionada

Una vez realizado todo el proceso de verificación, realizamos un AutoModel, la cual haremos una predicción respecto al campo “Sale Price”, esto permite verificar que posibilidad hay en el precio de la venta de la vivienda.

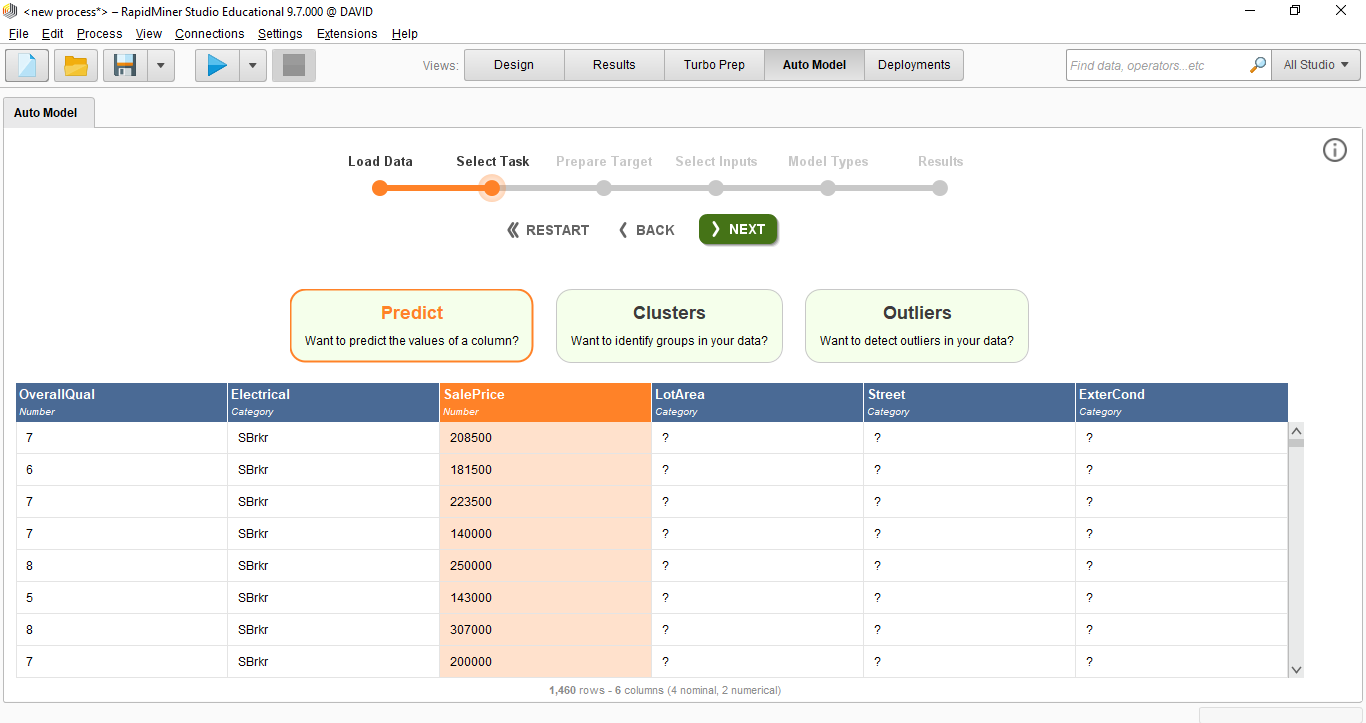


Figura Selección del campo "sale Price"

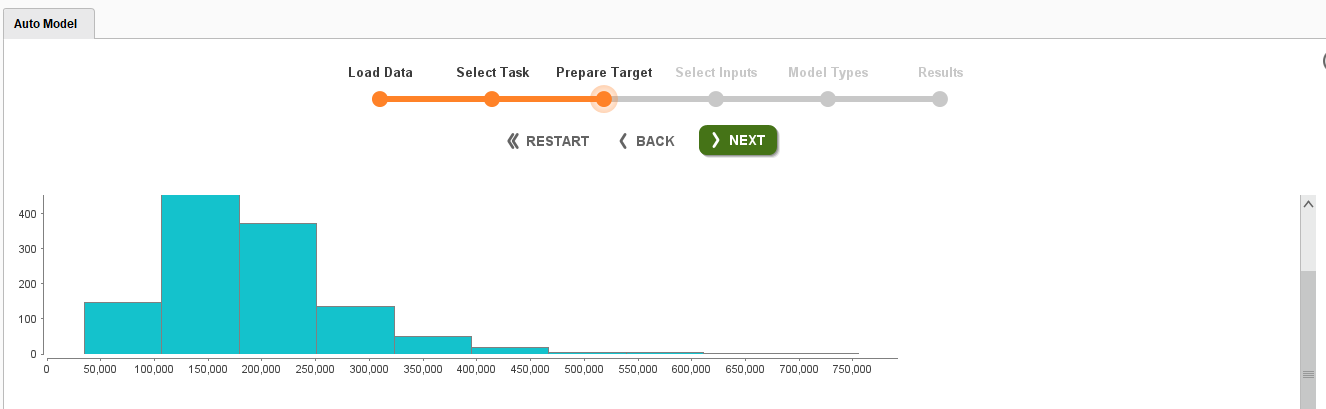


Figura Vista del Prepare Target

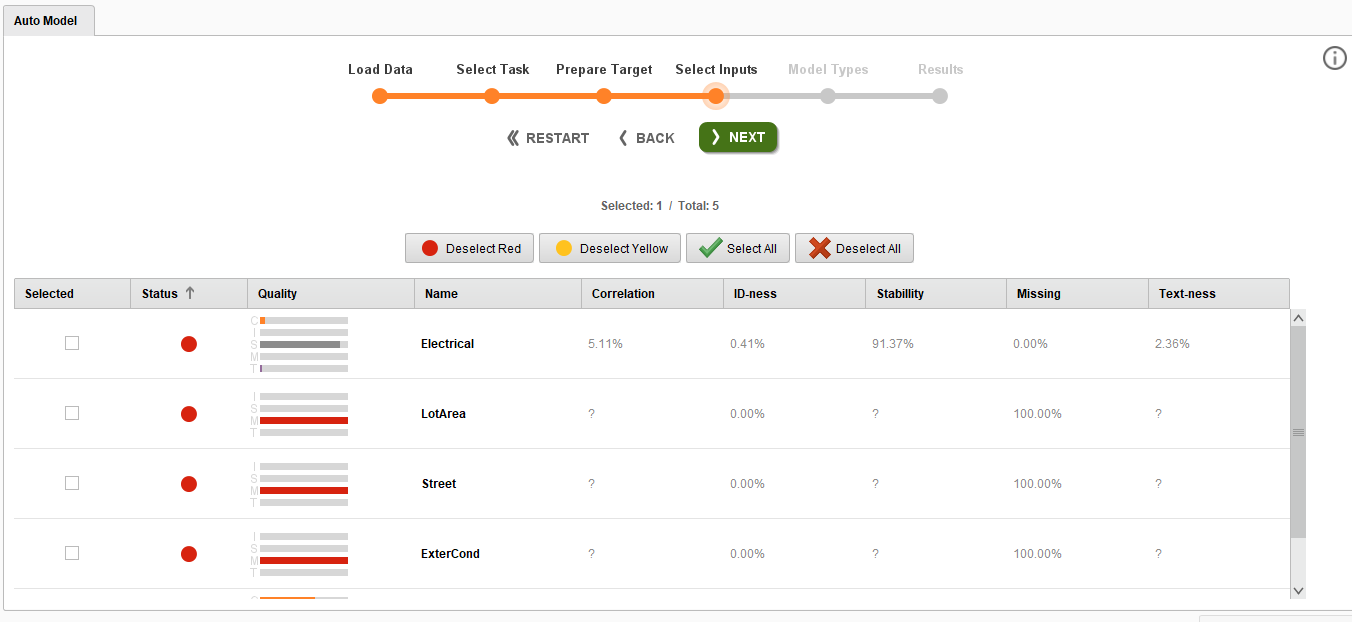


Figura Vista de Select Inputs

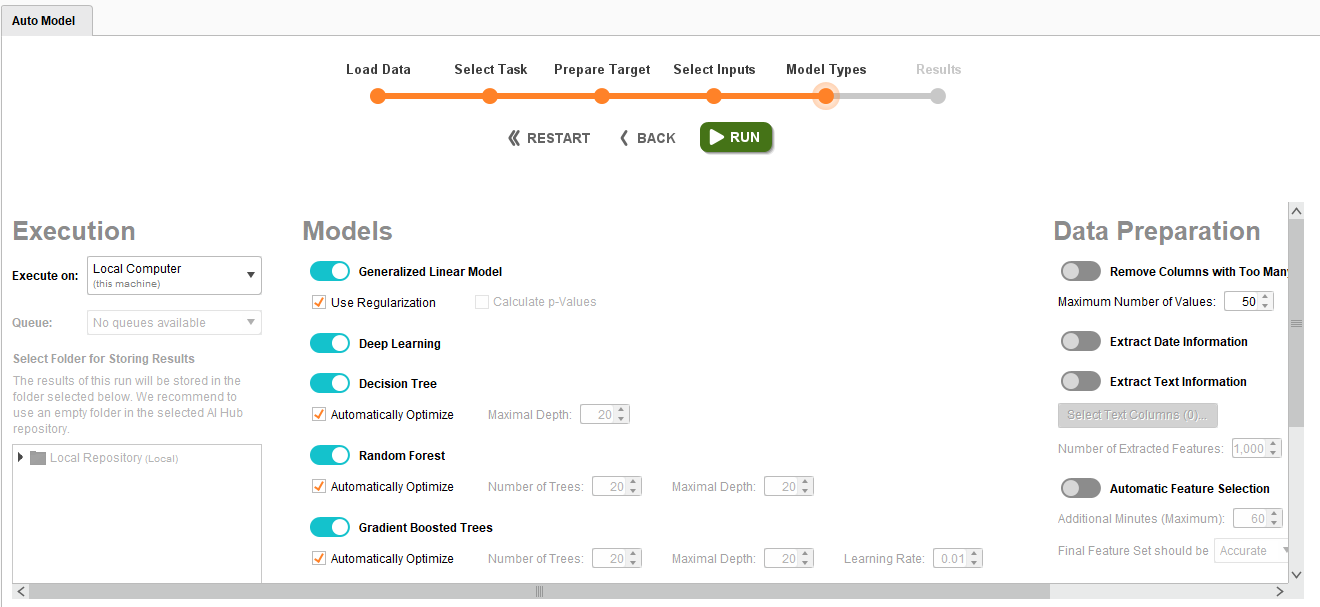


Figura Vista de Model Types

Una vez realizado todo este procedimiento esperamos a que me bote los resultados más favorables y procedemos al análisis de los algoritmos presentes.

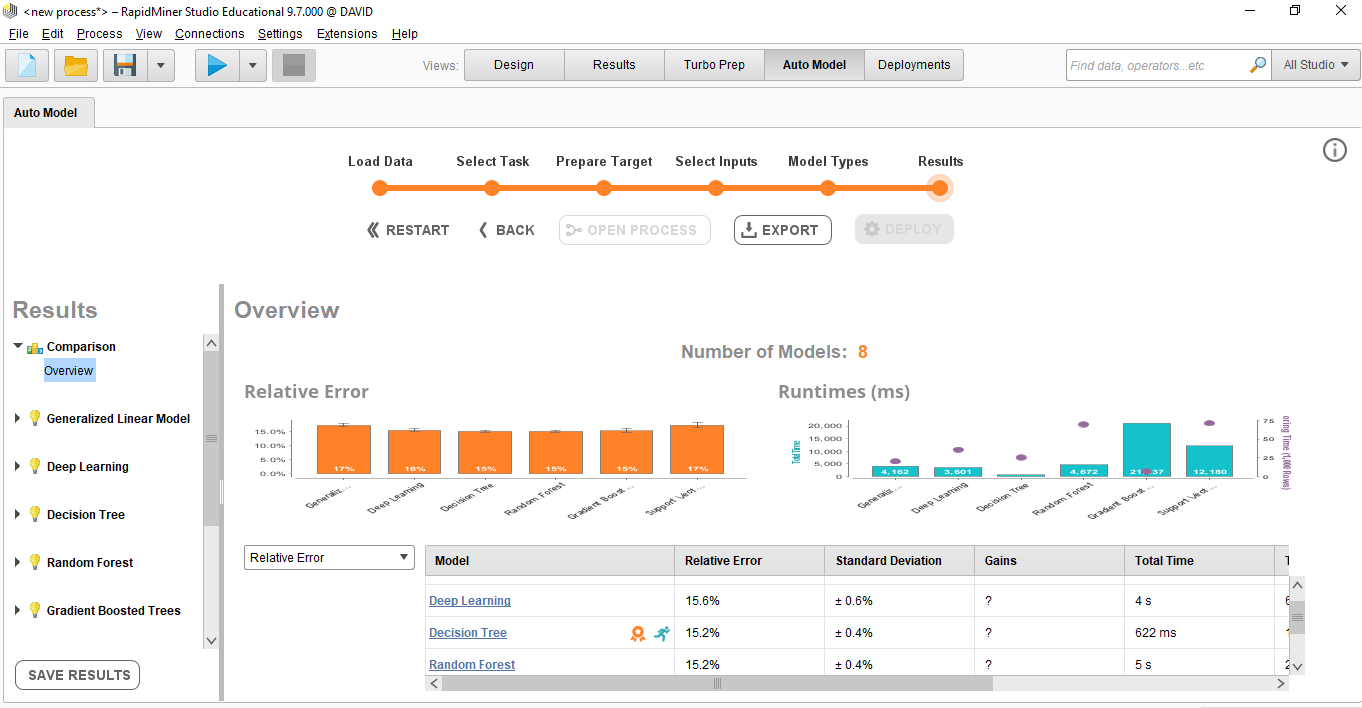


Figura Resultado del Auto-Model

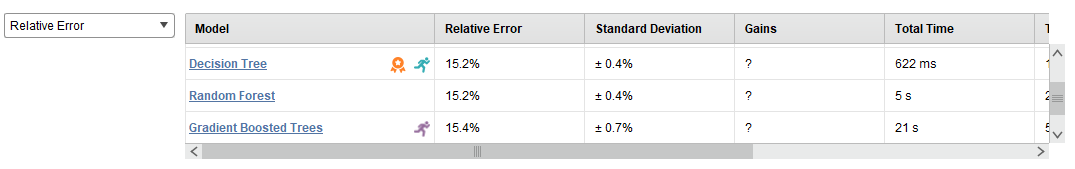


Figura Análisis del Algoritmo

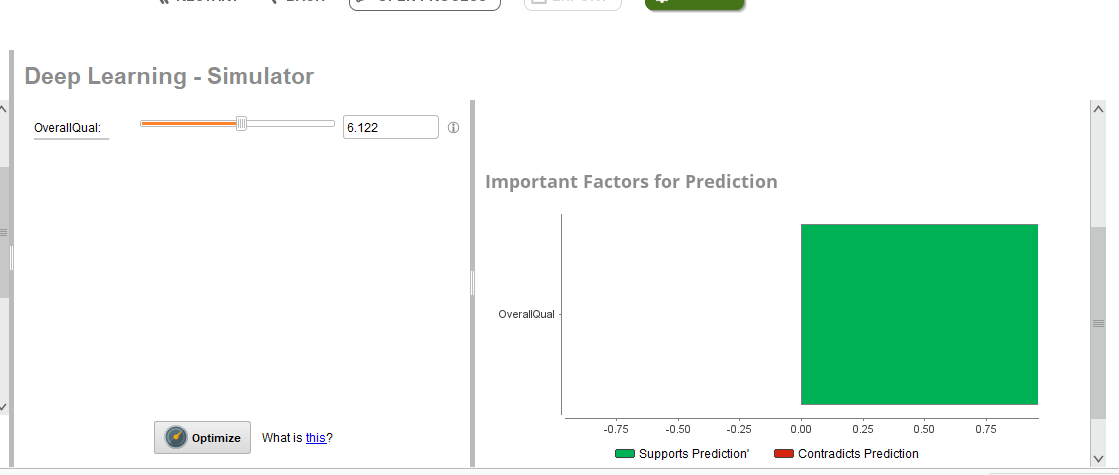


Figura Algoritmo Deep Learning

En este algoritmo nos podemos dar cuenta dentro de nuestros dataset que variables son tan esenciales para considerar y poder analizar en las viviendas y el valor adecuado para las casas, de acuerdo al atributo que utilizamos anterior mente nos pones que es más importante el material general y calidad de acabado da el valor agregado a las casas.

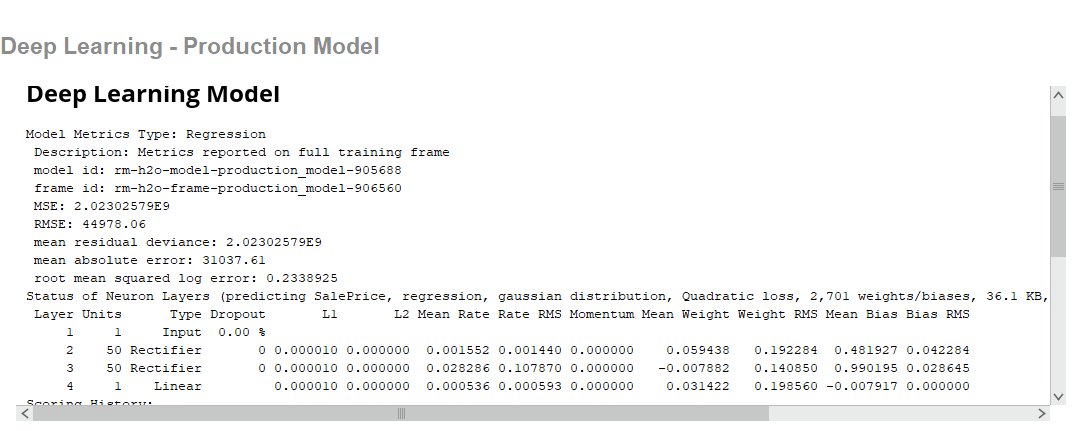


Figura Vista de Producction Model

**Algoritmo Tree**

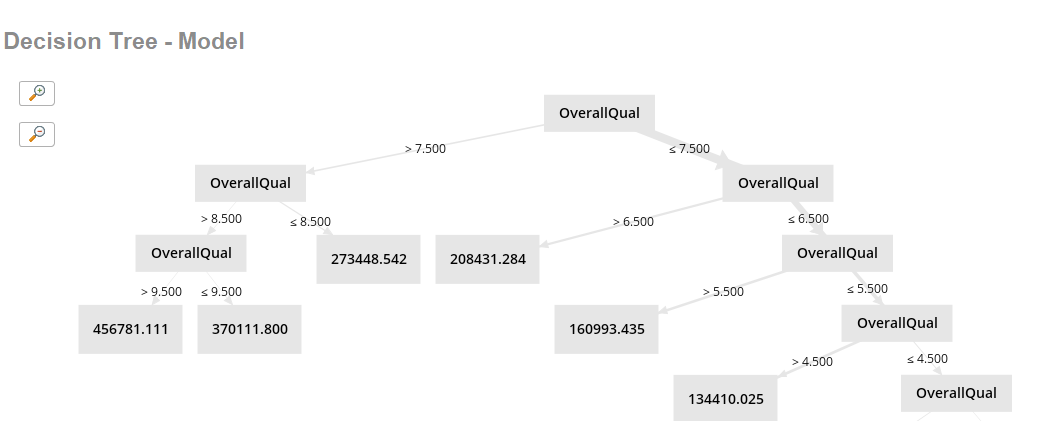


Figura Vista de OverallQual

En este algoritmo llegamos a la conclusión de que OverallQual, el material general y calidad de acabado da el valor agregado a las casas, donde nos indica que si el área es mayor a > 7500, el valor de la casa ira aumenta de acuerdo a su ubicación o en vez menorando.

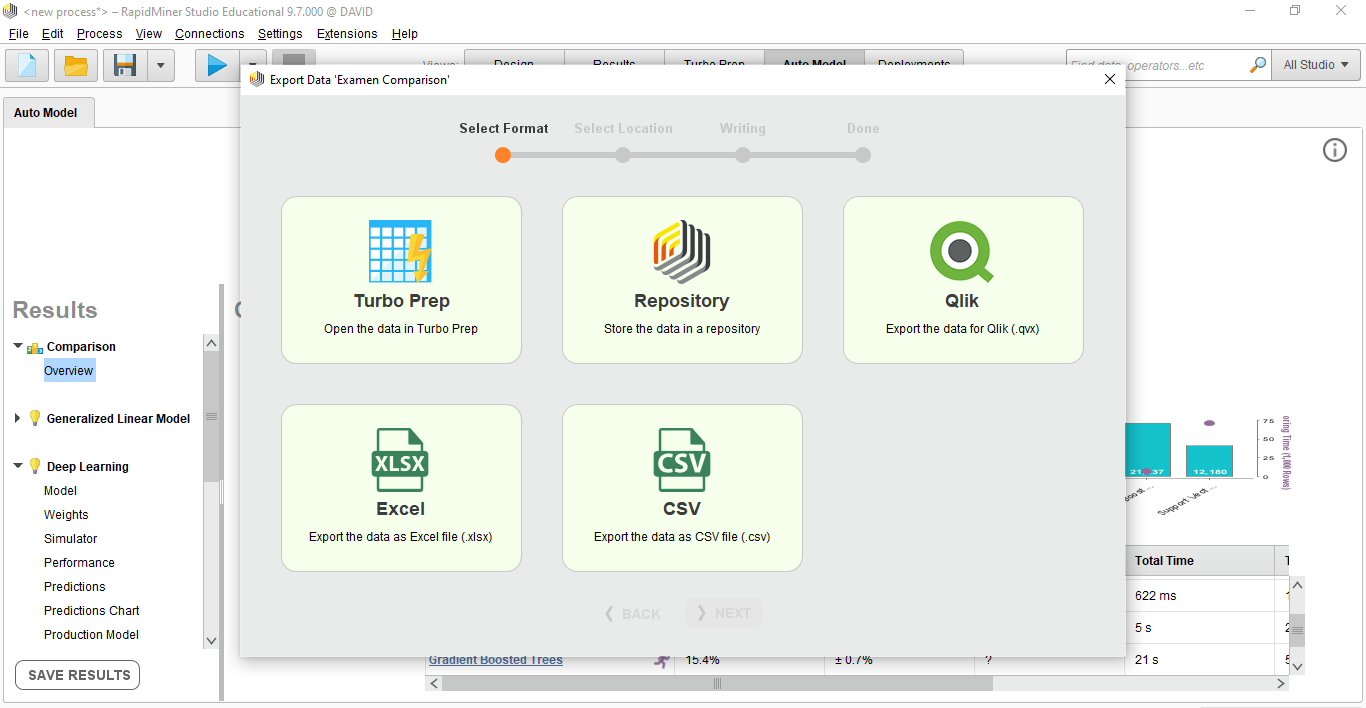


Figura Exportación del archivo en CSV

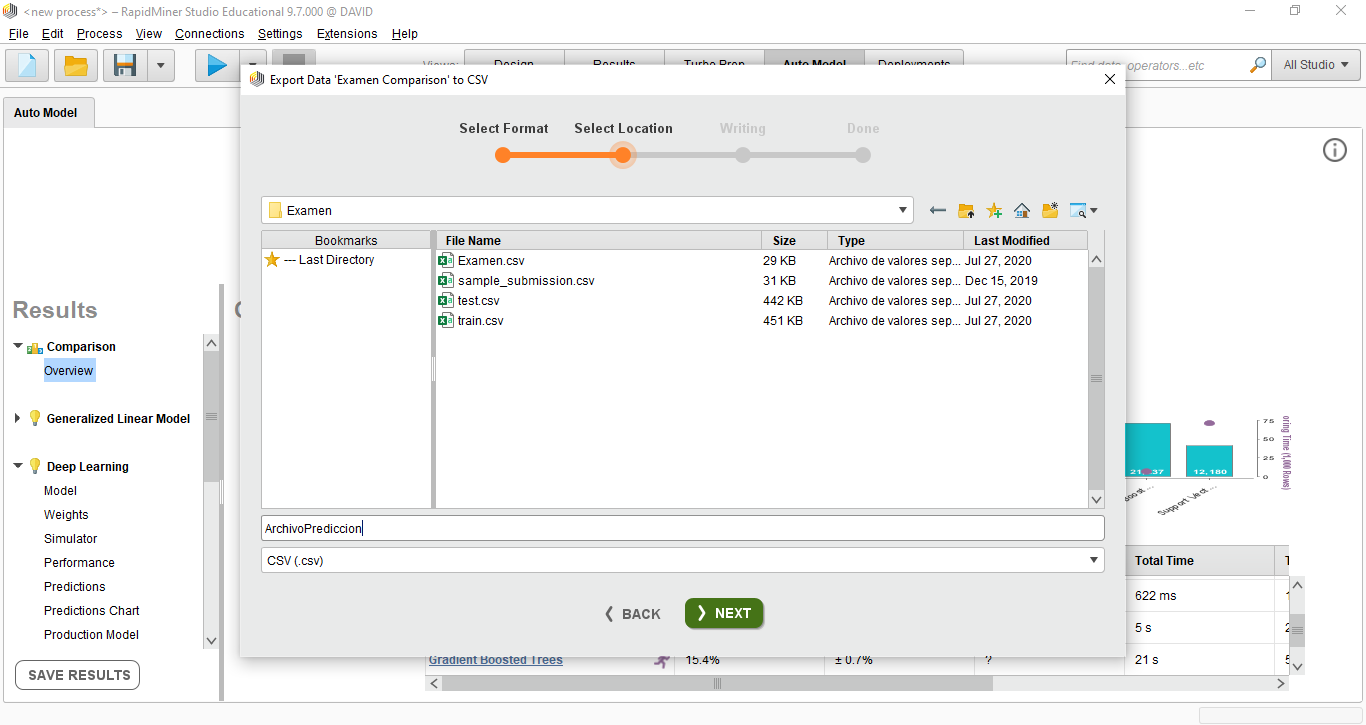


Figura Ubicación donde guardaremos el archivo

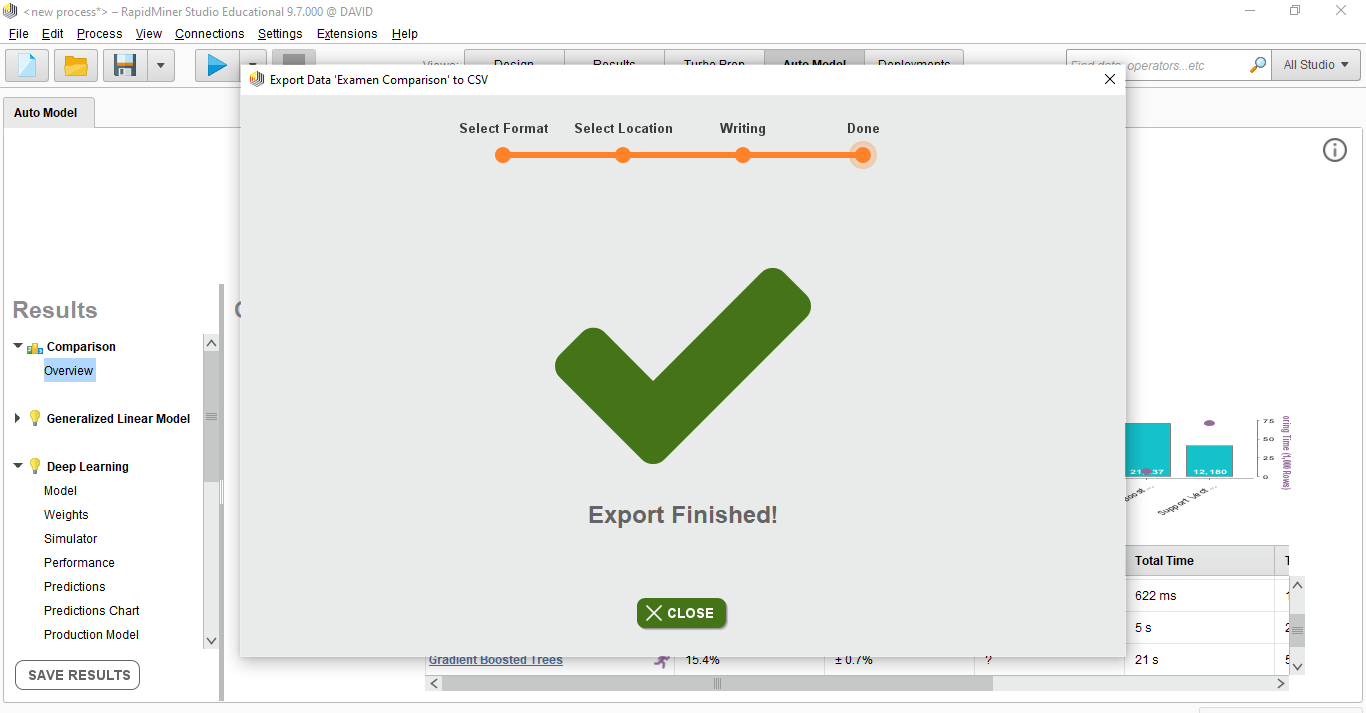
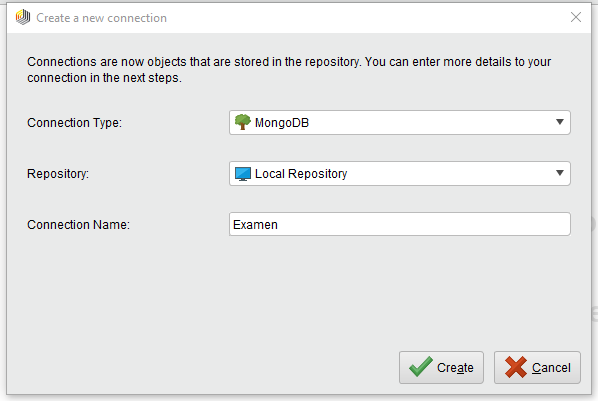
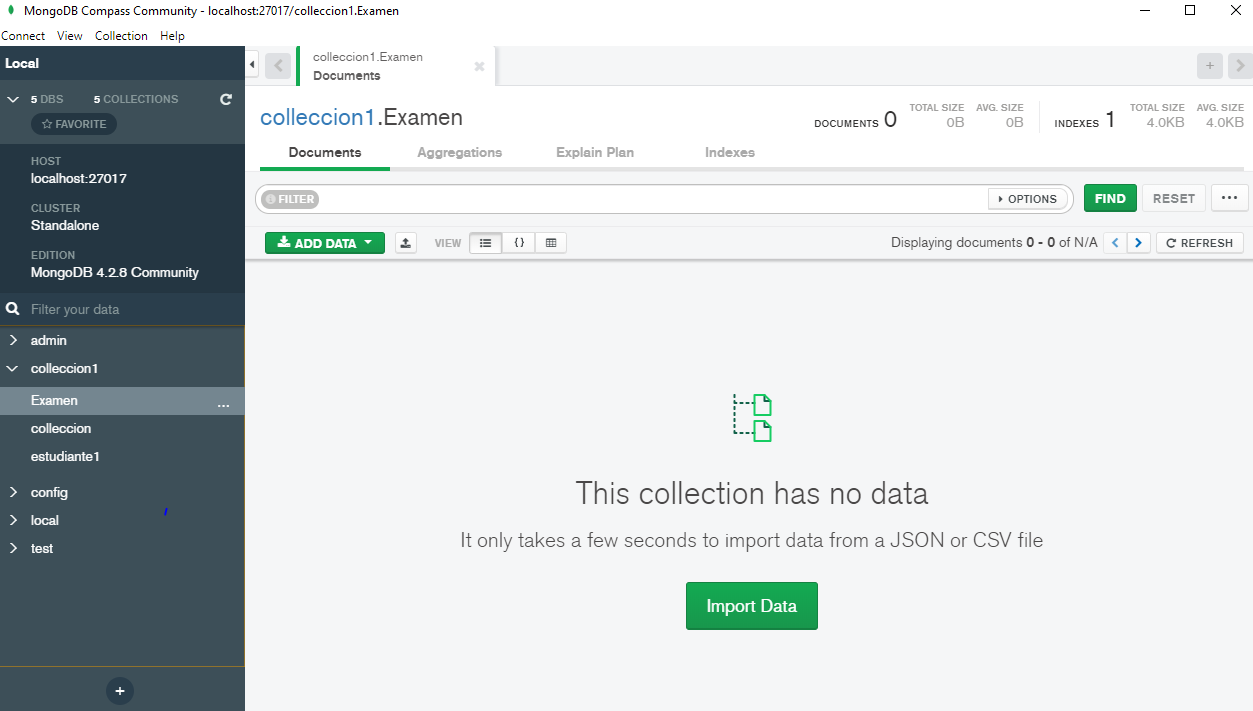
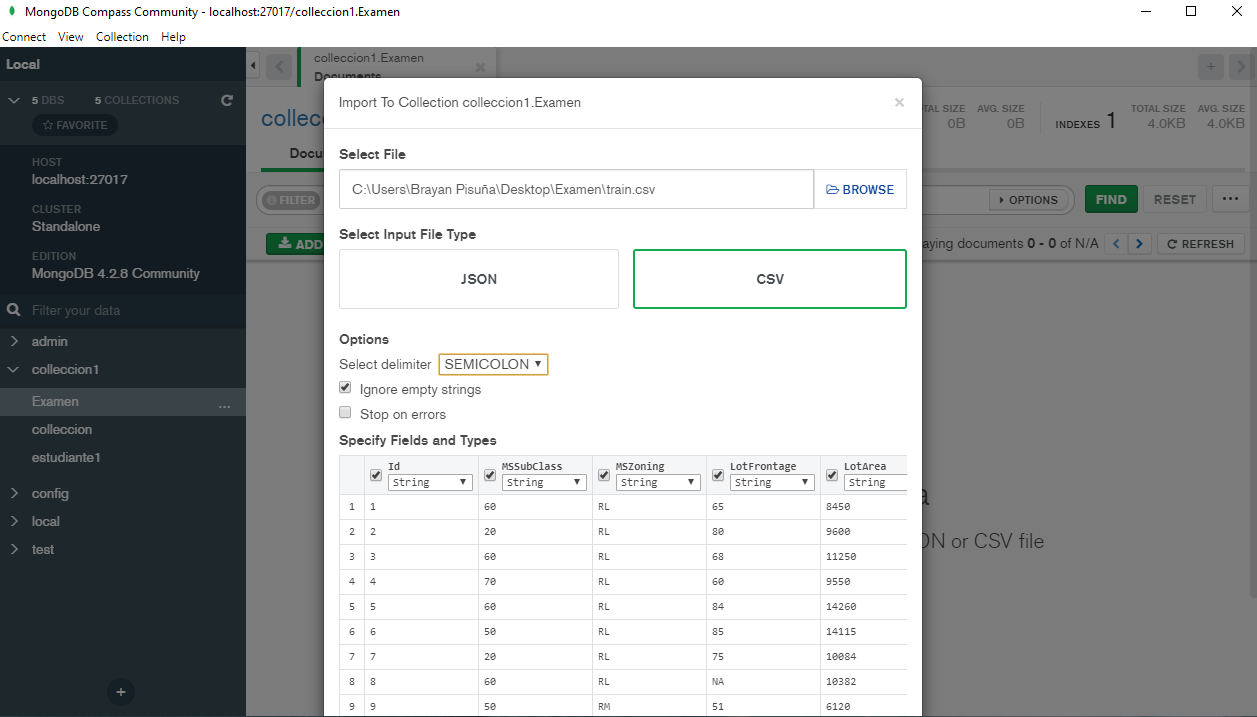


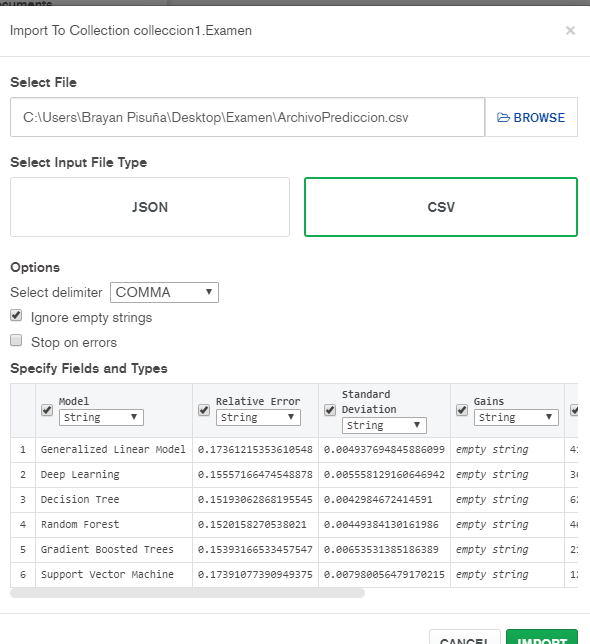
Figura Guardado correctamente

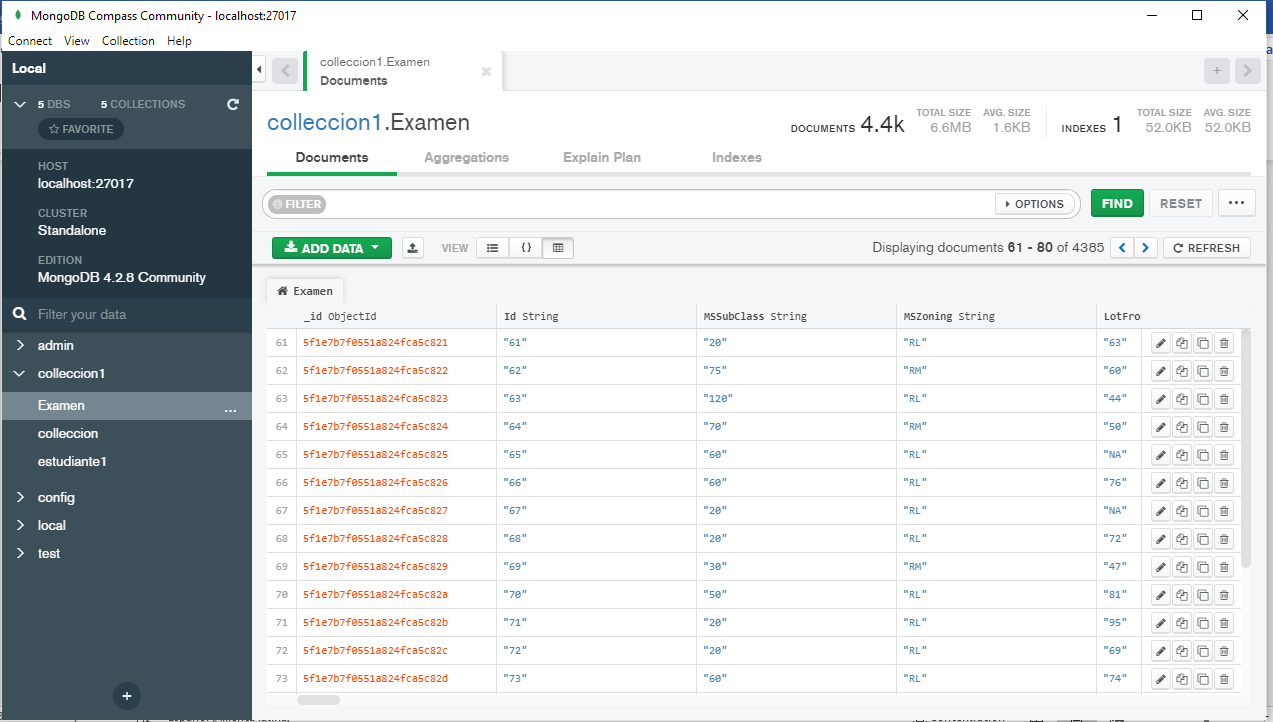
1. ***Creación de la base en MongoDb***











1. ENLACES

* **Cluster:** MongoDBmongodb+srv://brayanpisuna:brayanpisuna@cluster0.gpw2f.mongodb.net/test

1. CONCLUSIONES

En las ciencias económicas, RapidMiner tiene una aplicación muy interesante en temas similares a las series de tiempo, la creación de sistemas de decisión y el análisis de datos en grandes cantidades.

Normalmente se emplea mucho tiempo en minería de datos para pre-procesar los datos. RapidMiner ofrece varios operadores para leer datos de muchas fuentes diferentes y también operadores para procesar datos y facilitar el aprendizaje

En muchos casos no se puede definir un atributo objetivo y los datos deben ser agrupados automáticamente. Este procedimiento se denomina “Clustering”. RapidMiner soporta un amplio rango de esquemas de clustering que se pueden utilizar de la misma forma que cualquier otro esquema de aprendizaje.

1. REFERENCIAS

[1] «Kaggle,» 2020. [En línea]. Available: https://www.kaggle.com/c/house-prices-advancedregression-techniques/data. [Último acceso: 26 Julio 2020].

[2] «Rapidminer,» 2020. [En línea]. Available: <https://rapidminer.com/>

[3] «Rapidminer Documentacion,» 2020. [En línea]. Available: <https://rapidminer.com/blog/tips-tricks-different-ways-to-join-data/>