RapidMiner y MongoDB

Samanta Michelle Gómez Jácome, Edgar Gabriel Ibujés Gómez
Escuela de Formación de Tecnólogos
Escuela Politécnica Nacional
Dirección Postal.
samanta.gomez@epn.edu.ec, edgar.ibujes@epn.edu.ec

***Resumen*- Con el trabajo desarrollado, se muestra el uso potencial dentro de la academia del datamining como una herramienta de análisis muy importante.**

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.

I. INTRODUCCIÓN

Este informe trata sobre el proceso de carga de datos en RapidMiner y los hallazgos en un dataset generado con métodos que se han venido aprendiendo durante el primer bimestre, los mismo que serán utilizados para volcarlos a una base de datos en MongoDB que se almacenará de forma local y en la nube.

II. RECURSOS

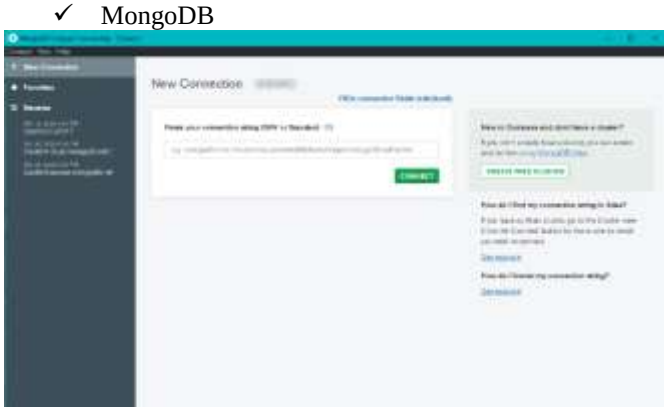
✓ RapidMiner

Pluggins para RapidMiner

- MongoDB

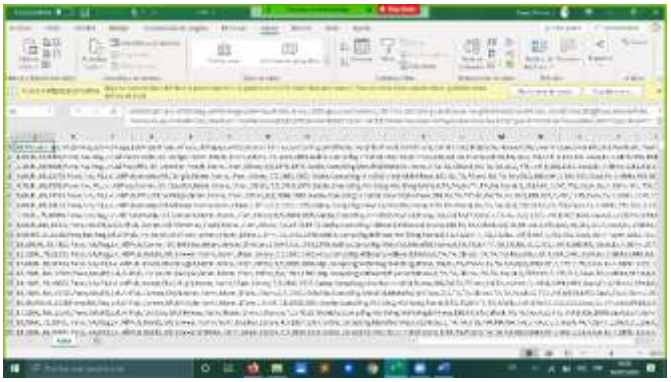
- Weka

✓ MongoDB

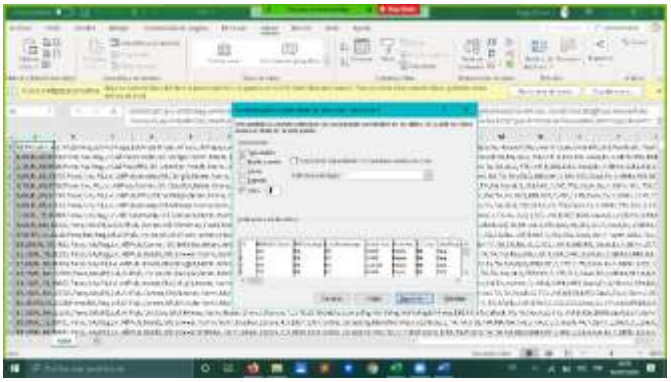


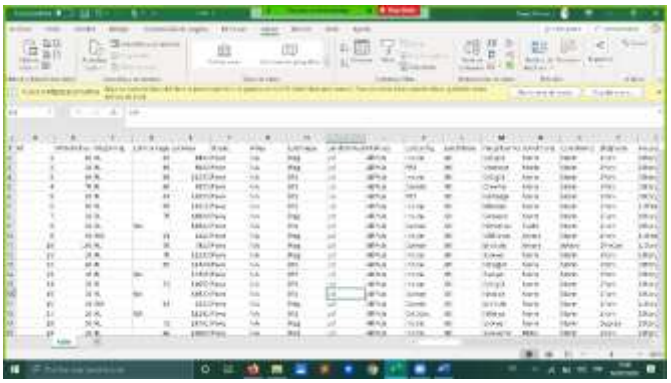
✓ Datasets

- Train

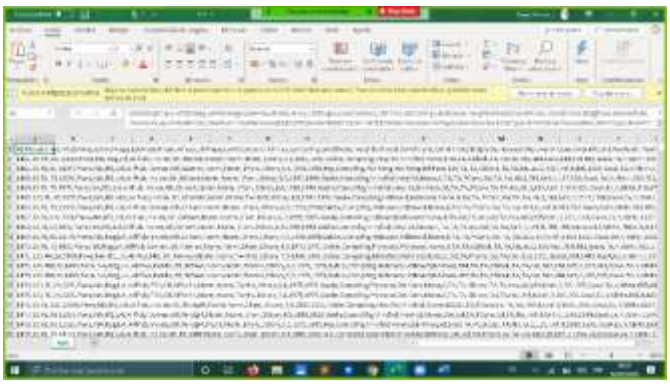


Filtrar datos para convertir en un archivo xlsx.

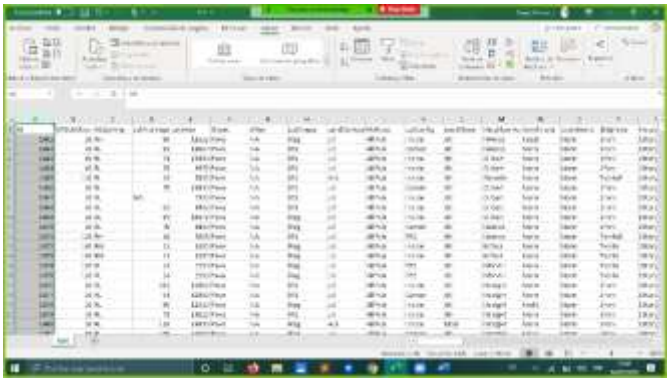




- Test



Filtrar datos para convertir en un archivo xlsx.



Atributos tomados en cuenta para el desasrrollo de la minería de datos.

- SalePrice: el precio de venta de la propiedad en dólares, se le consideró ya que es la variable objetivo que se va a predecir.
- LotFrontage: pies lineales de calle conectados a la propiedad, tomado en cuenta por la localidad de las casas.
- LotArea: tamaño del lote en pies cuadrados, es importante para el valor que recibe la casa.
- Street: tipo de acceso por carretera nos permite ver la accesibilidad a la casa.
- Neighborhood: ubicaciones físicas dentro de los límites de la ciudad de Ames.

- OverallQual: material general y calidad de acabado da el valor agregado a las casas.
- OverallCond: calificación de condición general se considera importante para la descripción de la casa.
- ExterQual: calidad del material exterior nos muestra la primera impresión de la casa para darle valor.
- ExterCond: condición actual del material en el exterior considerado después de su construcción según su mantenimiento.
- Heating: tipo de calefacción en la actualidad es un plus de venta de casas ya que aun no poseen con frecuencia.

Fundamentar la razón de los modelos seleccionados:

Deep Learning : La inteligencia artificial consiste en dejar a la máquina la posibilidad de resolver problemas que sólo podían ser resueltos por los humanos. O más bien, problemas fáciles de resolver por los humanos pero difíciles para las máquinas.

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

Por lo tanto este método resuelve de manerá rápida y eficiente los factores que serán predecidos, facilitando así la comprensión de las prediciones por factores de importancia.

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.

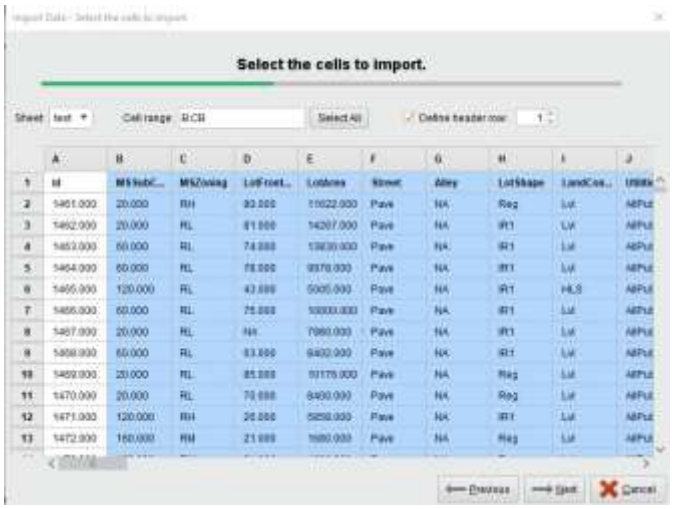
Nos facilita de manera más visual el caso de uso y las desiciones que se puedan tomar para poder determinar una desición puntual a futuro.

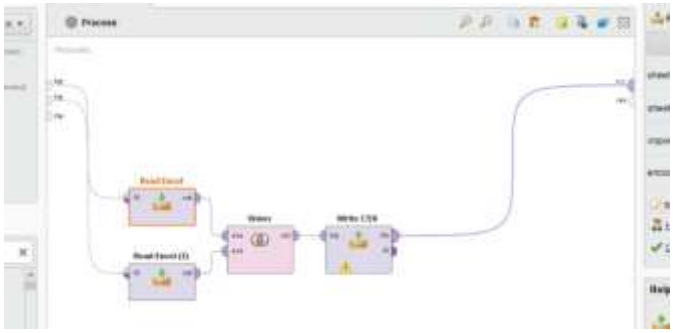
Gradient Boosted Trees : Es una técnica de aprendizaje automático utilizado para el análisis de la regresión y para problemas de clasificación estadística, el cual produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión.

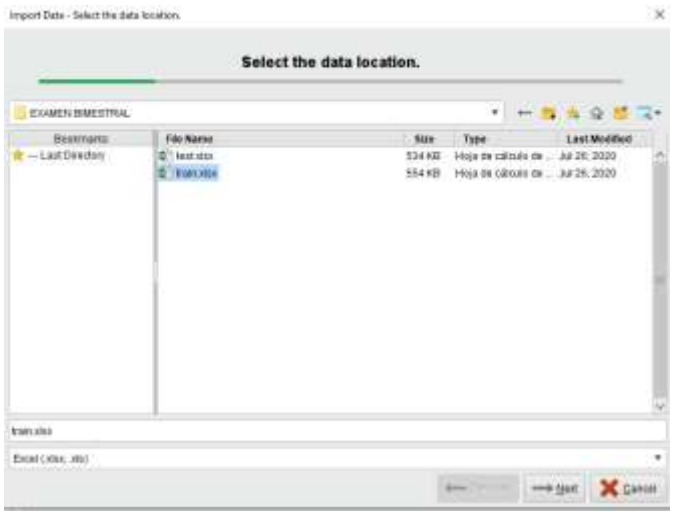
La rapidéz o el cambio parcial que se deduce por un gradiente descendiente de las variables expuestas.

III. DESARROLLO EN RAPIDMINER

Cargamos los dos dataset tanto del test como train ya filtrados los datos y transformados a archivos xlsx.

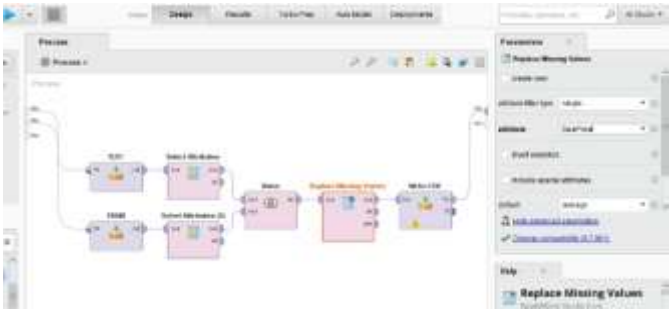




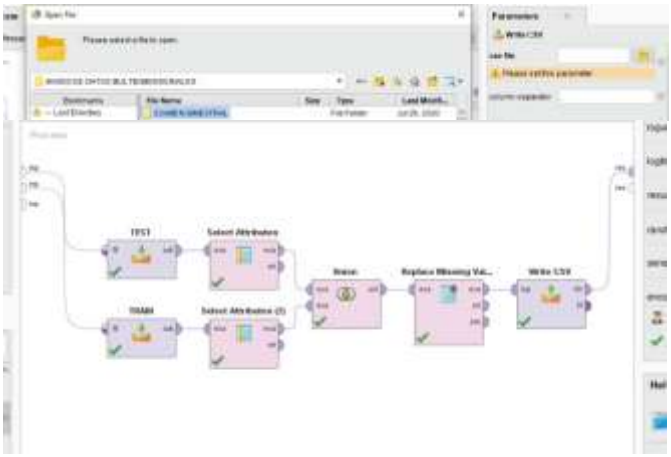


Elegimos los campos que deseamos analizar en los dos datasets..

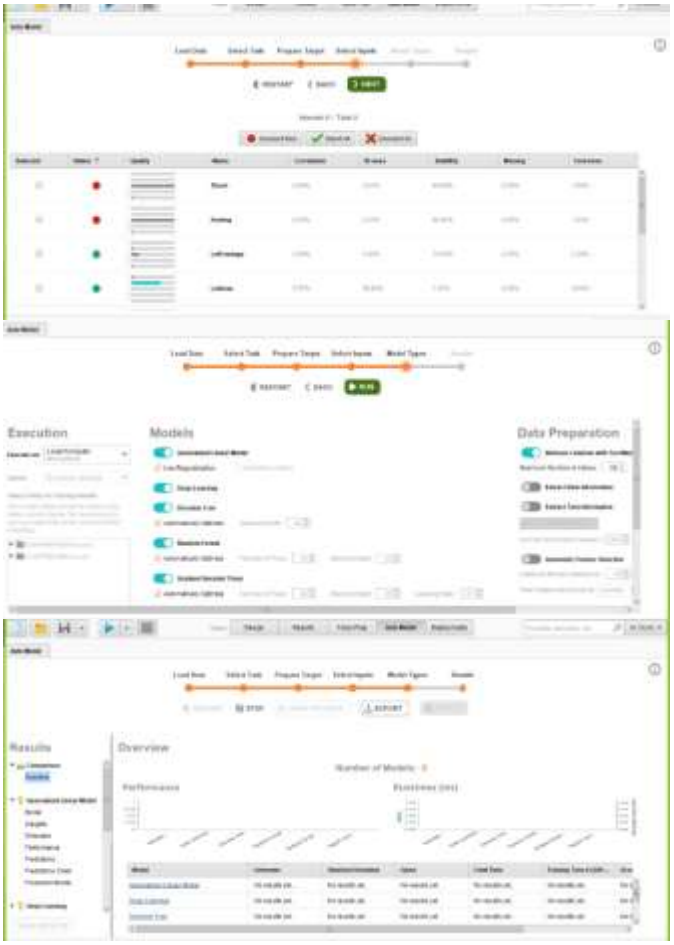
Usamos operadores para que se pueda leer los datasets, luego seleccionamos atributos para proceder a unificar y poder generar un archivo csv.



Le damos un nombre al archivo generado



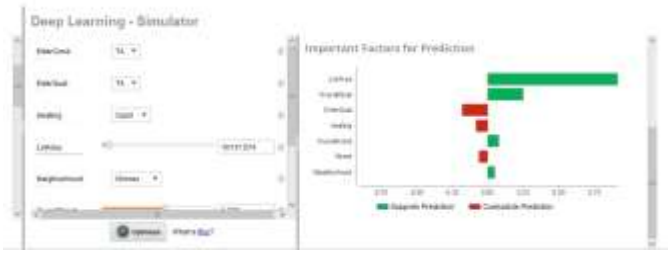
Creamos un automodelo con el archivo csv. Haciendo una predicción con la variable “sale Price”



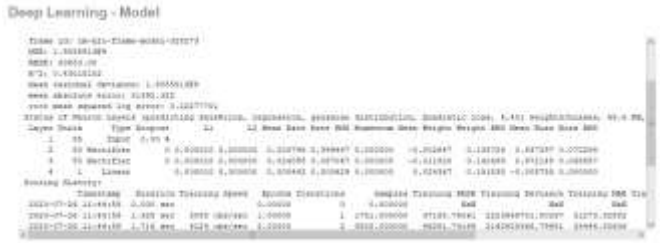
Resultados del auto-modelo

Analisis de algoritmos

Algoritmo Deep Learning



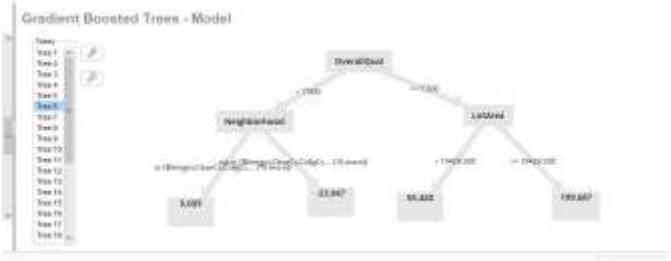
En este algoritmo podemos darnos cuenta dentro de nuestros datasets que variables son mas importantes para considerar la predicción del valor de venta de las casas, en este caso nos muestra con mas seguridad la predicción por área de lotes ya que conlleva un 80% de certeza, luego con el material y acabados, las condiciones generales y el barrio, y por el contario vemos que tiene menor influencia los acabados exteriores, la calefacción, y la calle en donde se ubica la casa.



- Algoritmo Decision Tree

En conclusión según los datasets utilizados.- dentro de los límites de la ciudad de Ames se considera una posible venta según el material general usado en la construcción en determinado barrio y la calidad de acabados interiores que tiene cada casa por tamaño del lote en pies cuadrados.
Por ejemplo en el barrio NAmes para predecir la venta de las construcciones por pies cuadrados subdividimos en material usado con la calidad de acabados interiores, donde nos dice que si el aréa es mayor a 7.500 el valor de venta de la casa se establecerá de acuerdo al porcentaje que le corresponda respecto al radio total que ocupa en su localidad.

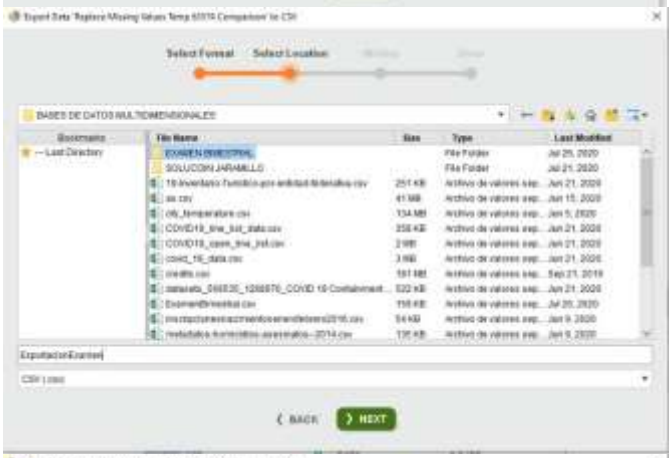
- Algortmo gradient boosting o Potenciación del gradiente

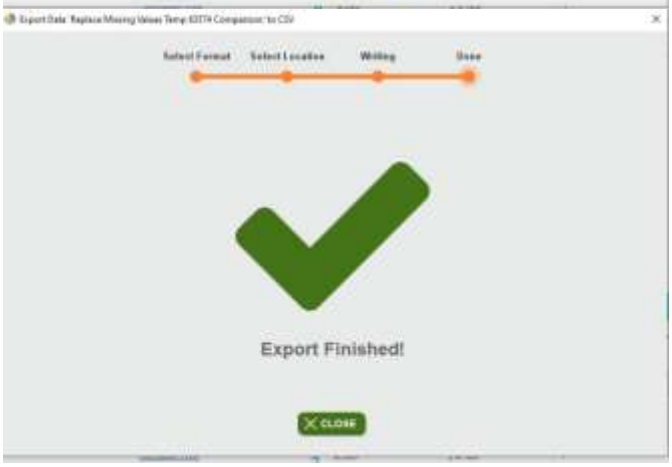


IV. DESARROLLO EN MONGODB

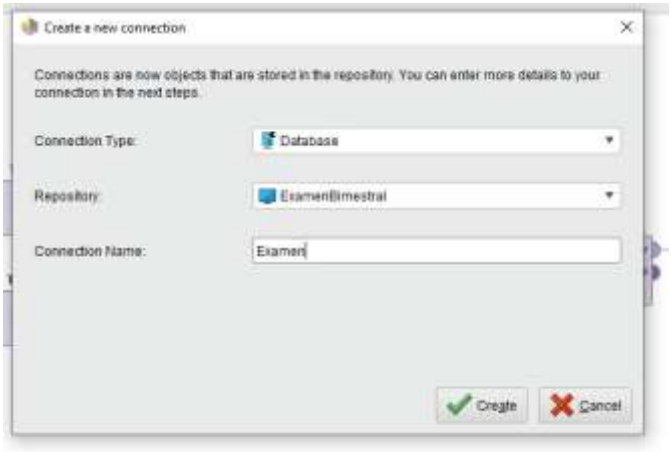
Para fijar un precio de acabado de interiores nos podemos guiar de acuerdo al barrio y el área de lotes considerando que si existe lotes menores a 7.500 pies cuadrados vamos a obtener porcentajes de perdidas, por otro lado los lotes que son equivalentes o mayores a 7500 pies cuadrados podemos predicir que obtendremos ganancias en un rango porcentual aproximado de 19429.50.

Exportamos el archivo csv.

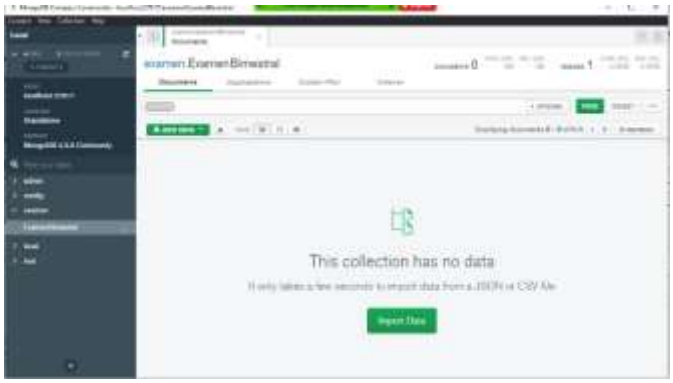




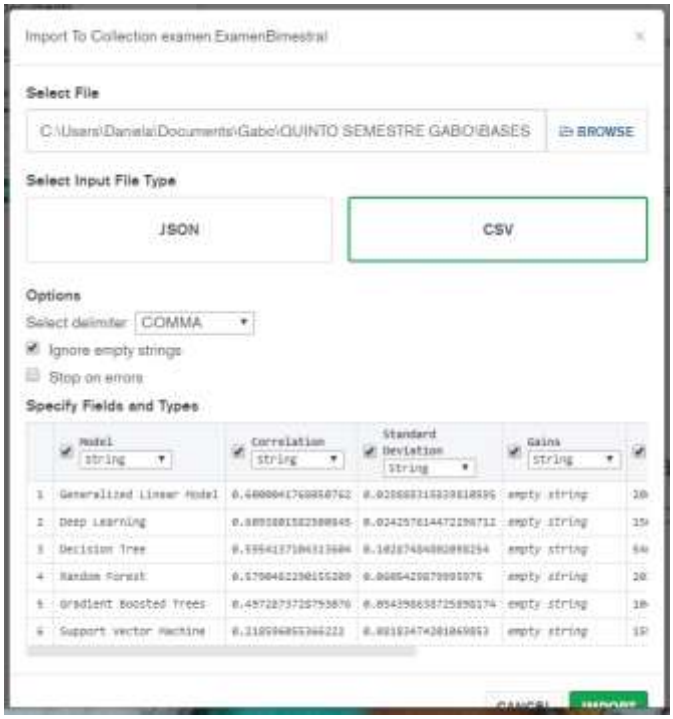
Creamos la conexión local en MongoDB

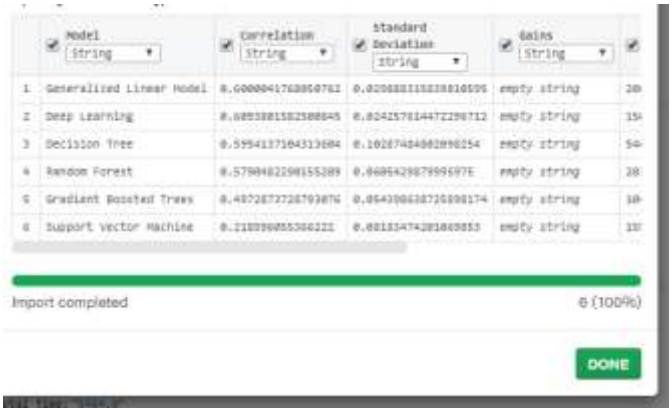


Y creamos la nueva base de datos

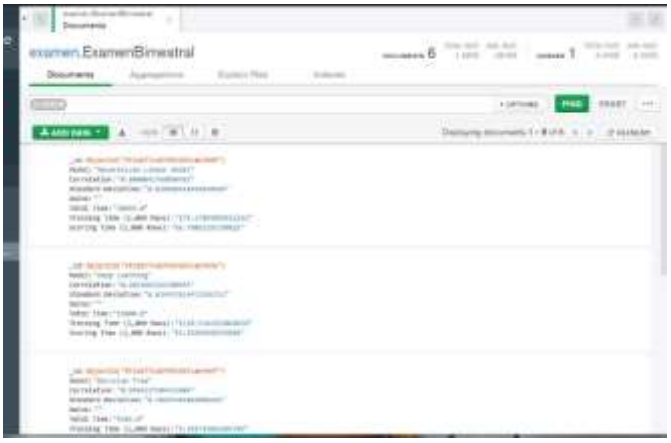


Importamos el archive csv que obtuvimos.





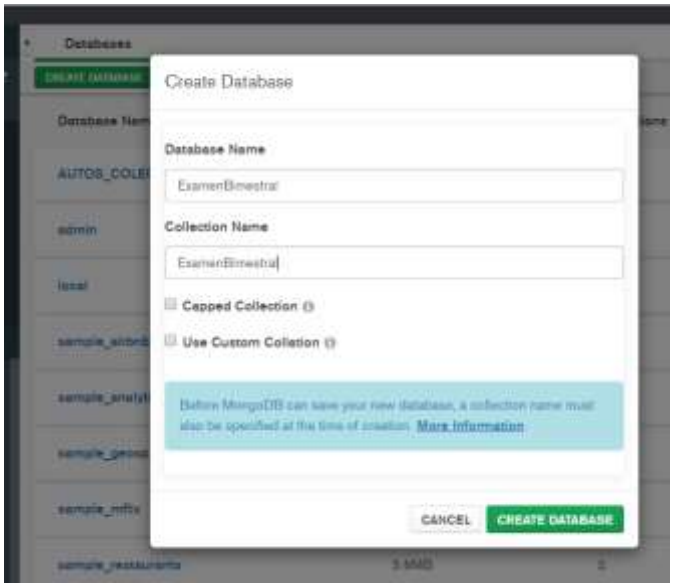
Mostramos la base importada

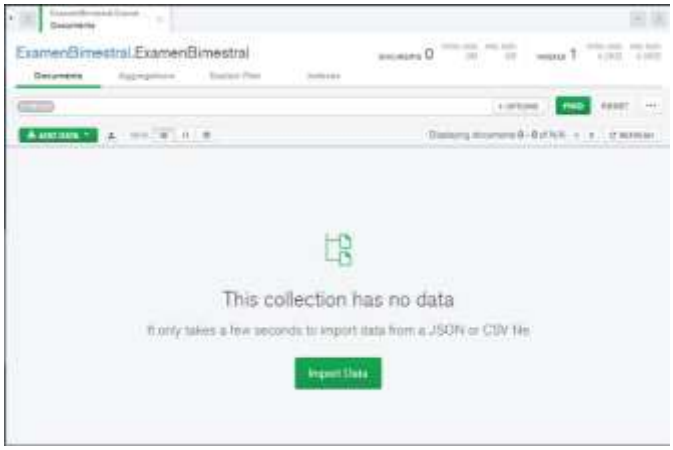


V. CARGA EN LA NUBE

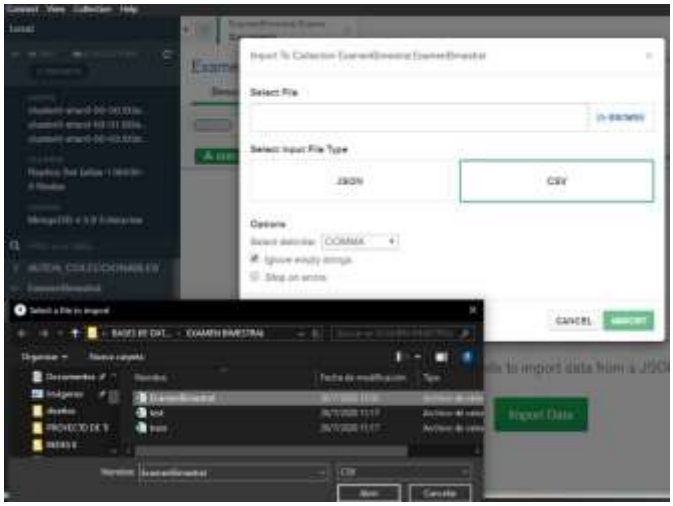
Ubicamos el cluster para poder agregar la base.

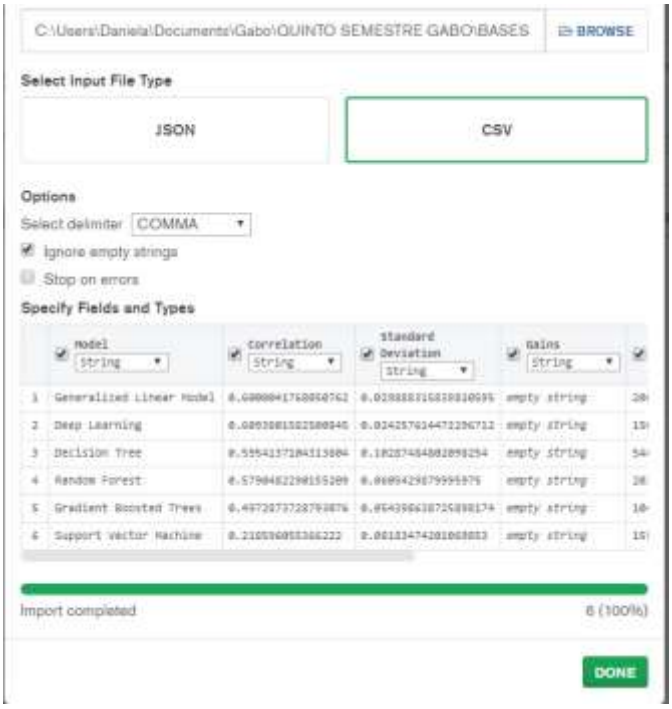
Creamos la base de datos

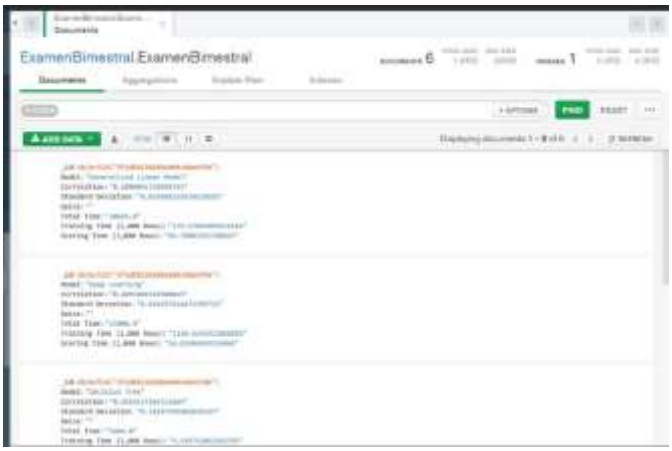




Importamos el archive dataset







VI. ENLACES

Cluster:
mongodb+srv://gabo2580:gabo2580@cluster0.l92ev.mongodb.net/ExamenBimestral

Video Gabriel Ibujés: https://youtu.be/4E1GAirNm38

Video Samanta Gómez: https://youtu.be/cdRgs8bedSk

VII. CONCLUSIONES

Al usar RapidMiner se puede concluir que otro elemento importante para mejorar la capacidad predictiva del modelo es determinar en qué grado participan o qué relevancia tienen para la predicción cada una de las variables a usar

Se pueden utilizar estas técnicas para segmentar datos en función de su tendencialidad o para encontrar cambios y alteraciones significativas en las decisiones a futuro, a estas técnicas las conocemos como algoritmos, y en RapidMiner, estos algoritmos son los que nos agilitan el proceso de toma de decisiones con cualquier tipo de dato

REFERENCIAS

[1] J. Díaz-Verdejo, "Ejemplo de bibliografía", En Actas de las XI Jornadas de Ingeniería Telemática, vol. 1, n. 1, pp. 1-5, 2013.