Examen Bimestral

Jessica N. Amoguimba, Andrea L. Quishpe
Análisis de Sistemas Informáticos
Escuela Politécnica Nacional
jessica.amoguimba@epn.edu.ec, andrea.quishpe01@epn.edu.ec

Resumen – El presente documento es relacionado datos de predicción de los precios de venta de viviendas, ayudándose de atributos varios (Precio de venta, año de construcción, año de remodelación, LotArea, Garage, Neightborhood,etc) con los cuales realizan un análisis con las diferentes variables que se recopilado con el paso del tiempo, con este análisis obtenido usaremos el auto modelamiento que Rapid Miner nos proporciona para escoger y aplicar los mejores algoritmos que adaptaremos a los datos que tenemos; seguido de esto la construiremos manualmente el modelo basándonos en los algoritmos que seleccionamos esto permitió la evaluación del modelo, finalmente se logra obtener un modelo final de predicción con la opción "Automodelado" que nos permite tener un conjunto de predicciones que ayudarán a la toma de decisiones en el negocio.

I. INTRODUCCIÓN

Es importante hoy en día contar con una herramienta que nos permita predecir ciertos procesos o comportamientos, en el sector empresarial. En este caso se puede implementar el automodelado con el fin de predecir precios de casas, en base a factores como el tipo de casa, área o localidad donde se encuentra, tamaño, etc.

RapidMiner es una solución que facilita el autoservicio de análisis predictivo permitiendo una avanzada analítica empleando solamente drag and drop y opcionalmente la generación de código, beneficiando de esta forma a empresas

II. DESARROLLO

Como primer paso usaremos dos "Read Excel" los cuales permiten leer los dataset previamente filtrados y transformados en archivo .xlsx, también usaremos dos "Select Attributes" para seleccionar los atributos que consideramos importantes para este caso, a la vez utilizamos "Union" este ayuda a unificar los dataset , una vez hecho esto nos percatamos que cada dataset posee un número diferente de atributos y que además en uno de los dataset no posee el atributo "SalesPrice" por lo que en los siguientes pasos explicaremos la solución para ello. Y finalmente usaremos "Write CSV" para que una vez realizado los procesos mencionados anteriormente este se escriba y genere un archivo CSV.

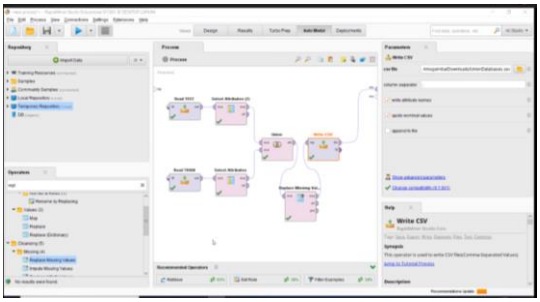


Fig. 1 Diagrama del proceso

Escogemos los atributos de cada tabla, en este caso la tabla train previo a realizar la unión, e ignoramos el ID.

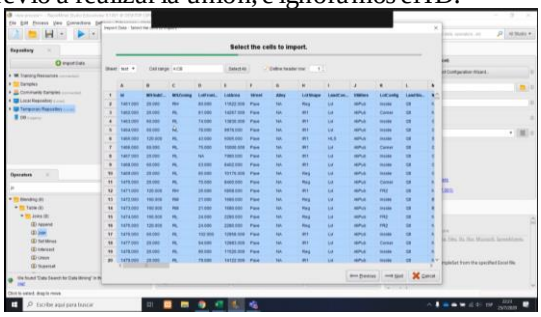
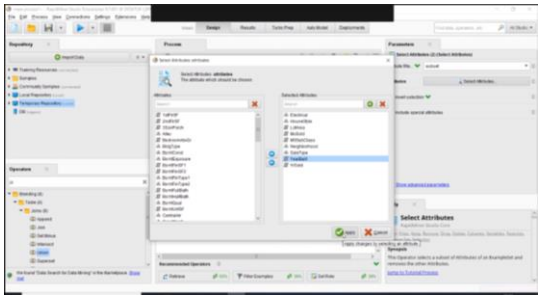
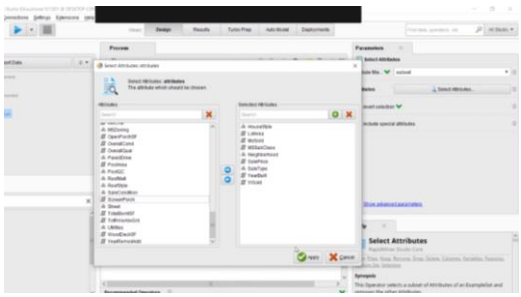


Fig. 2 Selección de atributos



Repetimos el mismo paso aplicado a la tabla Test.

Documento recibido el 26 de julio de 2020.



A continuación se muestran la unión de ambas tablas con los respectivos atributos escogidos: Electrical, HouseStyle, LotArea, MoSold, MSSubClass, Neighborhood, SaleType, YearBuilt,YrSold y SalePrice.

Atributos escogidos por ser considerados los más relevantes al momento de adquirir una casa.

Como se puede observar en la siguiente imagen, al unir dos datasets (train y test) se producen valores nulos por las diferencias de datos que se encuentran en cada uno.

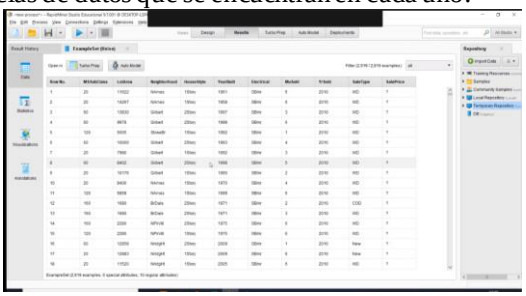


Fig. 3 Unión de datasets

Para corregir esto, utilizamos el algoritmo Replace Missing Values el cual generará valores aleatorios en los atributos que se encuentren vacíos.

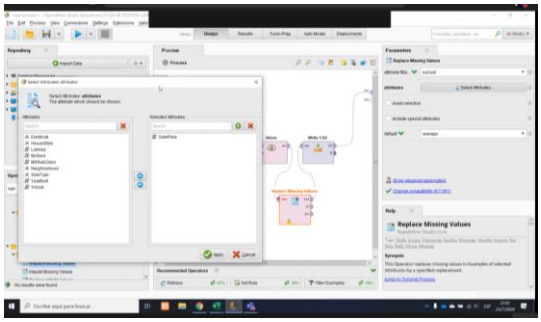


Fig. 4 Reemplazo de valores nulos

Observamos el resultado de la aplicación del algoritmo en la columna SalePrice.

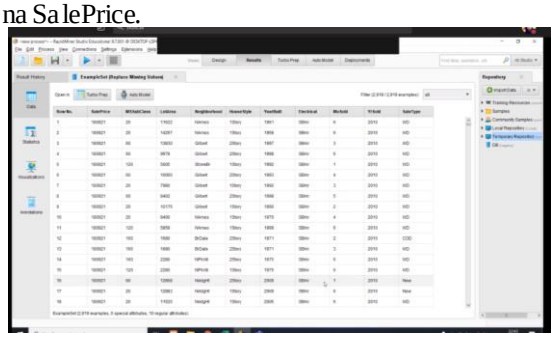


Fig. 5 Tabla uniónde dos datasets

Una vez corregido el dataset, procedemos al AutoModelado.

Escogemos lo que deseamos hacer, en este caso Predecir, y seleccionamos la columna que entablará relación con los demás atributos.

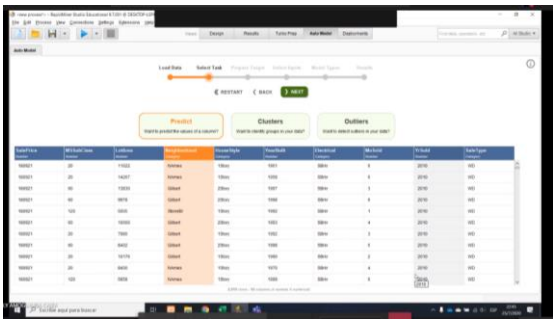


Fig. 6 Automdelado

Se observa un bosquejo de cómo los datos se relacionarán con los atributos escogidos. Tomando en cuenta que en el eje X se presentarán los atributos seleccionados y en el eje Y los atributos con los que será relacionado.

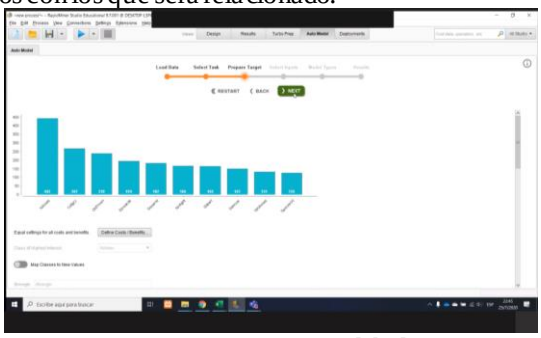


Fig. 7 Proceso automodelado

Escogemos los atributos más relevantes (LotArea, HouseStyle, YearBuilt, YrSold, SaleType, SalePrice) para relacionarlos con el atributo Neighborhood

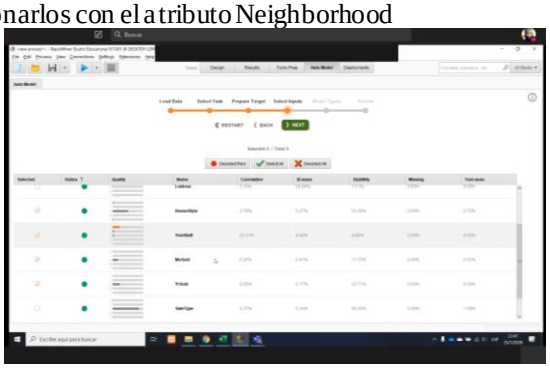


Fig. 8 Escoger atributos para automodelado

Escogemos los modelos que queremos procesar en el automodelado y damos clic en RUN.

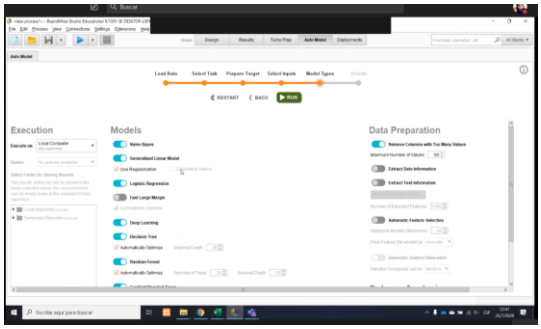


Fig. 9 Finalización de automodelado

Esperamos mientras se completan los modelados previamente escogidos. Esto puede tardar varios minutos.

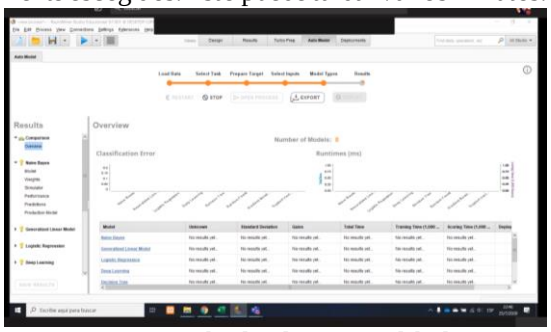


Fig. 10 Resultados de automodelado

III. RESULTADOS

Una vez transcurrido el tiempo, se puede observar la finalización de cada proceso entre los detalles se pueden observar la clasificación Error, Standard Deviation, Gains, Total time.

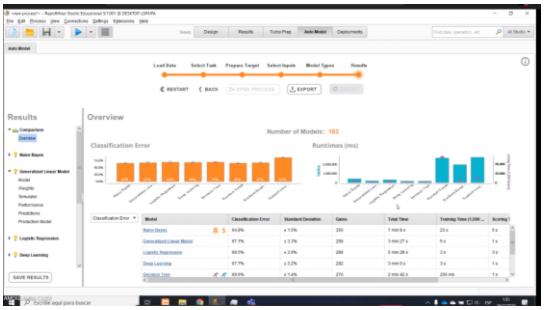
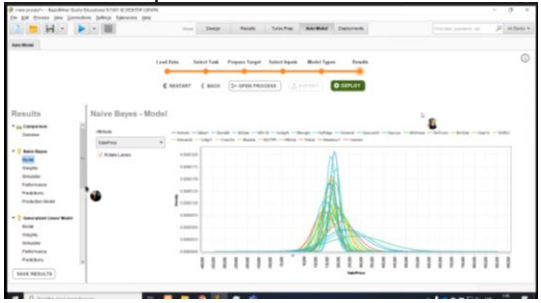


Fig. 11 Resumen de procesos de automodelado

Ya terminado el proceso, podemos ir navegando por los modelos que se generaron, en este caso podemos ver las curvas generadas en relacion con el precio de venta y el barrio en el que más ventas posee en este caso resulta ser "NAmes"



Por otro lado en Random Forest nos permite ver el modelo que se genera a raiz de los datos seleccionados.

Genera un árbol con los datos randomicos de "YearBuilt" relacionado con "Neightborhood" esto permite interpretar de mejor manera la información generada por RapidMiner.

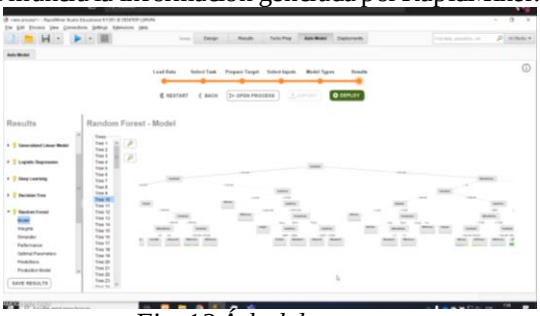


Fig. 12 Árbol de proceso

Otra de las formas de generar modelos es Gradien booster tree el cual también genera un modelo de árbol con la diferencia que este utiliza datos de la misma fuente.

Fig. 13 Árbol Gradient Booster

Si nos dirigimos al apartado General podemos ver Estadísticas evaluadas, la siguiente gráfica nos parecio la mas importante ya que en esta se muestra de forma clara mediante un gráfico de barras las ventas generadas en cada "Neightborhood" y es esta también resulta que NAmes es el barrio con más ventas es decir un 15.18% en comparación con los demás barrios.

Fig. 14 Estadísticas

Generamos un archivo CSV de los resultados obtenidos el cual se guardara localmente para tenerlo de respaldo de lo procesado por RapidMiner.

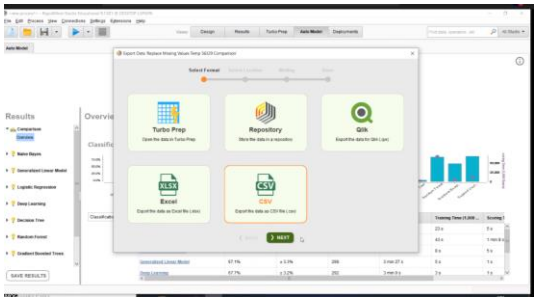


Fig. 15 Creacion de archivo CSV

Damos clic en Next, y si todo esta correcto y seguro del lugar donde se lo va a guardar nos saldrá que la exportación del archivo fue realizada con éxito.

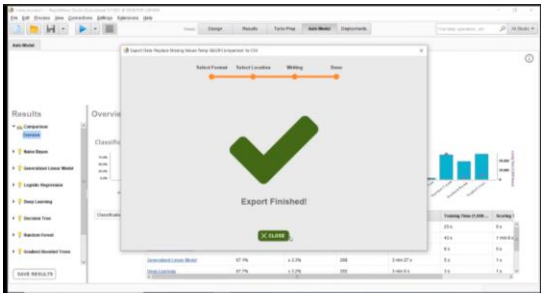


Fig. 16 Guardamos el archivo

RapidMiner permite generar un modelo automático del proceso que deseemos, a continuacion podemos observar un modelo del proceso de automodelado ayudandose de un sin número de algoritmos los cuales ayudan al momento de generar los resultados.

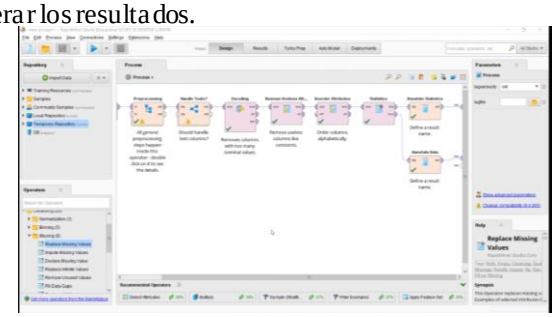


Fig. 17 Generación automática del modelo de un proceso

Al momento de dar RUN se generan los resultados numéricos en los cuales se ve los nombres del barrio, el estilo de casa el mes en el que se vendio más casas, el precio de venta el año de la construcción de la casa y el año de la venta del mismo.

Fig. 18 Correr el proceso

IV. Conexión MongoDB

Para la conexión con MongoDB, ingresamos a la aplicación y creamos un database.

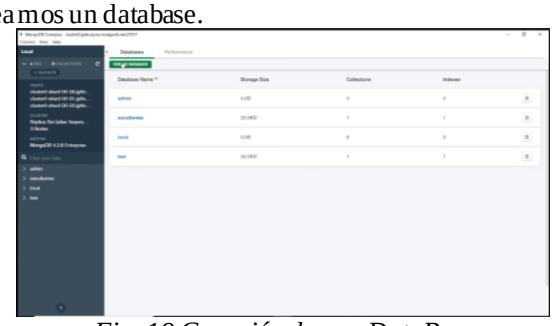


Fig. 19 Creación de una DataBase

La DATABASE se llamará examen1 y la colección que creamos amoguimbaquishpe y resultadosmodelamiento.

Usar esta cadena de conexión para visualizar de mejor manera:

mongodb+srv://jeka:jeka2020@cluster0.ijy6n.azure.mongodb.net/examen1

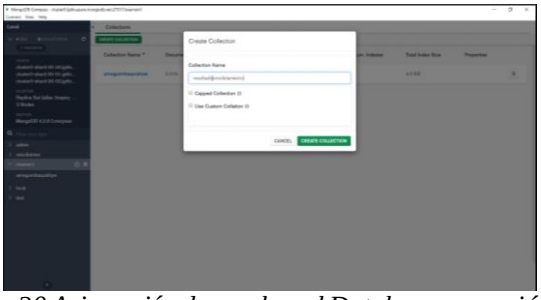


Fig. 20 Asignación de nombre al Database y creaciónde colección

Seleccionamos los archivos CSV que se generaron en RapidMiner.

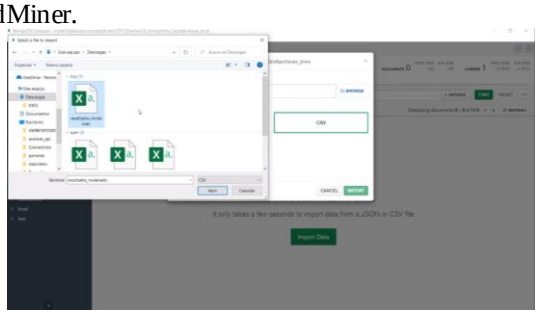
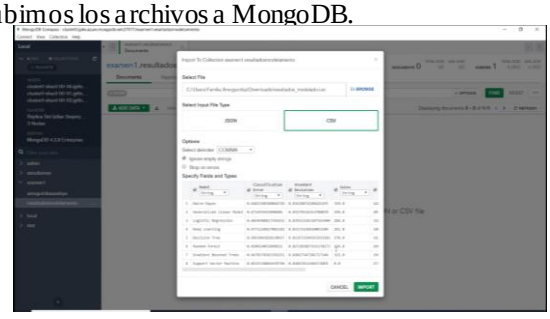


Fig. 21 Añadir CSVal Database

Subimos los archivos a MongoDB.



Se observan ya los archivos subidos en formato CSV.

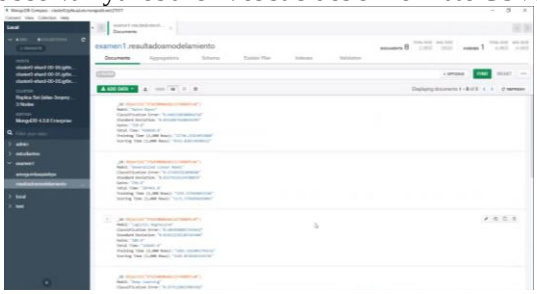


Fig. 22 CSV cargados al Database

V. Conexión a MySQL

Para la conexión con MySQL exportaremos los archivos generado por RapidMiner. Damos clic en exportar archivos CSV y damos clic en continuar.

Fig. 23 Conexión a MySQL

Una vez realizado el paso anterior se observan los resultados agregados a MySQL los cuales se subieron con éxito.

Fig. 24 Carga de archivos CSV

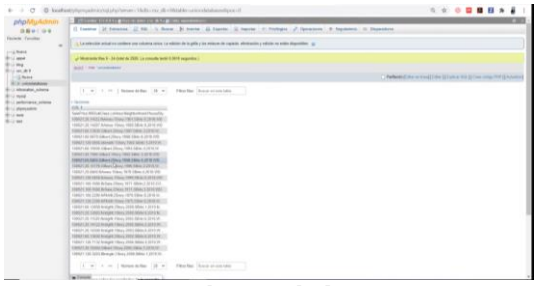


Fig. 25 Archivos subidos a MySQL

VI. CONCLUSIÓN

- El aplicar auto-modelamientos de predicción a la información es una gran ventaja, ya que acelera la producción de prototipos y su desarrollo, así como también se usa para saber cuan factible puede ser el desarrollo de un proyecto.