Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Título: "Tasación automatizada de inmuebles de CABA mediante técnicas de machine learning"

AUTOR: ING. GUIDO ESTEFANO

MENTOR: PHD. ROBERTO ABALDE

[MARZO, 2021]

Resumen

En la actualidad existen técnicas de aprendizaje automático que poseen una enorme aplicación en los contextos organizacionales. En el ámbito inmobiliario en particular, una problemática y necesidad recurrente consiste en fijar el precio de propiedades en venta. Sin dudas, la complejidad de dicha tarea recae en la diversidad de factores que es preciso considerar para establecer una cifra que se ajuste a la realidad del mercado.

En el presente trabajo se implementaron tres modelos de regresión basados en árboles, con el objetivo de predecir precios de inmuebles de dos ambientes ubicados en la región de la Ciudad Autónoma de Buenos Aires. Para lograrlo, fue necesario realizar ciertos tratamientos previos a la base de datos original, tales como: imputación de datos faltantes, estudio de valores atípicos, creación de nuevas variables explicativas a partir de técnicas de *Text Mining* y reducción de la dimensionalidad del problema mediante el procedimiento de Análisis de Componentes Principales (PCA). Luego de finalizar dicha etapa de preparación y limpieza de datos, se continuó con la optimización de los hiperparámetros de cada modelo detallado. Los modelos utilizados fueron los siguientes: *Decision Tree*, *Random Forest y Gradient Boosted Trees*. En todos los casos se aplicó la técnica de validación cruzada con 10 particiones, para evitar el fenómeno de sobreajuste. De los tres modelos aplicados, *Gradient Boosted Trees* fue aquel que reveló mejores resultados, destacándose con un RMSE (raíz del error cuadrático medio) de 4.506, un 47,49 % inferior al RMSE del modelo tomado como *Baseline*. Además, logró reducir en un 47,94 % el RMSE del modelo *Decision Tree* y en un 41,41 % el RMSE del modelo *Random Forest*. Asimismo, generó los mejores resultados en términos de error absoluto y error relativo.

Palabras claves: *Tasación de inmuebles, Decision Tree, Random Forest, Gradient Boosted Trees, Text Mining, N- Gramas, Análisis de Componentes Principales.*

Índice

Introducción

El sector inmobiliario es una de las principales actividades económicas que motorizan el nivel de actividad de una región. En este contexto, la gran oferta y demanda de inmuebles en la Ciudad de Buenos Aires es un hecho. La fijación del precio de una propiedad es una tarea sumamente compleja y es el punto de partida para dar inicio a cualquier proceso de compra – venta. Para lograr definir con precisión el valor de mercado de una propiedad se debe analizar un gran número de variables cuantitativas y cualitativas que influyen directamente en su precio final, tales como: zona geográfica, dimensiones, cercanía a zonas comerciales, comodidades, luminosidad, estado de la construcción, etcétera. Generalmente esta tarea, que insume gran cantidad de tiempo, es llevada a cabo por tasadores expertos en la materia.

En este trabajo se busca optimizar la tarea convencional de fijación del precio de una propiedad mediante una herramienta de valuación automatizada. Con dicha herramienta es posible conocer el valor de mercado de una propiedad al instante lo que a su vez permite comenzar rápidamente el proceso de venta con una noción real del valor o bien iniciar el proceso de compra sabiendo cuánto costará una propiedad con las características deseadas, y analizar comparativamente con la tasación realizada por un tasador o asesor inmobiliario. Además, un tasador de propiedades automático puede ser de gran utilidad en los siguientes escenarios: conocimiento de patrimonio, valoración de la herencia en sucesiones, certificaciones de estados de obra en préstamos para construir, ampliar o modificar inmuebles, peritaciones de valor en procesos judiciales, disolución de empresas, daciones de pagos y expropiaciones. En concreto, se aplicarán tres modelos predictivos de aprendizaje automático con el objetivo de tasar inmuebles de dos ambientes comprendidos en el rango de precios de 65.000 U$S a 95.000 U$S ubicados en la Ciudad Autónoma de Buenos Aires. Para alcanzar dicha meta, se utilizarán los siguientes modelos de regresión basados en árboles: *Decision Tree*, *Random Forest y Gradient Boosted Trees*.

Es razonable esperar que la ubicación geográfica, la superficie cubierta y las amenities de los inmuebles sean las variables de mayor relevancia a la hora de definir el correspondiente precio de mercado. En otro orden de ideas, es lógico pensar que se pueden obtener mejores resultados mediante la utilización del modelo *Gradient Boosted*

Trees ya que en este algoritmo, el entrenamiento de los árboles de decisión se hace de forma secuencial. Adicionalmente, para todos los modelos de regresión basados en árboles, se espera que las predicciones de los precios mejoren empleando un gran número de árboles de escasa profundidad o, por el contrario, un pequeño número de árboles de gran profundidad.

En el primer apartado, se explicará de manera detallada el modelo de negocios de una organización como Properati, reconocido portal inmobiliario de la región. Se profundizará sobre la recolección, almacenamiento, administración y seguridad de los datos de la empresa a fin de comprender su proceso de gestión integral de los datos. En el segundo apartado, se presentará un análisis completo de las variables de partida a fin de preparar los datos para la posterior aplicación de las distintas técnicas de regresión y optimización de sus hiperparámetros. Este análisis previo consistirá en el tratamiento de los valores faltantes y valores atípicos (*outliers*). Además, en virtud de crear nuevas variables explicativas, se buscarán distintas *amenities* de cada inmueble por medio de la aplicación de técnicas de *Text Mining*, las cuales permitirán procesar los campos que contienen detalles de la descripción de cada propiedad. Asimismo, se construirán nuevas variables métricas vinculadas con la distancia de cada inmueble a distintos puntos de interés de la ciudad. Cabe destacar también que antes de implementar los modelos predictivos se reducirá la dimensionalidad de la base de datos utilizando la técnica de Análisis de Componentes Principales (PCA). Finalmente, se avanzará con la aplicación de los tres modelos de regresión buscando mejorar el indicador RMSE (raíz del error cuadrático medio) del modelo *Baseline* tomado como referencia, siendo este aquel que predice sistemáticamente la media de los precios de las propiedades. Para seleccionar aquel modelo con mayor capacidad predictiva, se establecerán adicionalmente comparaciones entre las siguientes métricas: error absoluto y error relativo. Por último, en el tercer apartado, se expondrá el análisis necesario para llevar a producción el modelo predictivo que funcione como tasador automático, la metodología ágil seleccionada para implementar dicho proyecto dentro de una organización y finalmente, distintas visualizaciones de resultados en Power BI.

Apartado 1. Gestión de datos en contextos organizacionales

La gestión de datos se posiciona como un factor estratégico clave para potenciar la adecuada toma de decisiones en cualquier organización. En el presente apartado se versará sobre cómo se lleva a cabo la gestión de los datos en empresas dedicadas a ofrecer servicios inmobiliarios por Internet. En la primera sección del apartado se introducirá el caso del conocido portal inmobiliario Properati y su modelo de negocio de plataforma, tan popular en los tiempos que corren. En la sección subsiguiente, se ahondará sobre la recolección, almacenamiento, administración y seguridad de los datos de la compañía a fin de comprender su proceso de gestión integral de los datos. En la sección final del apartado, se introducirá la problemática detectada en el sector inmobiliario y se presentará una solución enfocada en la implementación de técnicas de aprendizaje automático.

1.1 Descripción de la organización y su modelo de negocio

En los últimos años el avance de la nueva tecnología e Internet han provocado un profundo impacto en la gran mayoría de las organizaciones. Hace tiempo el sector inmobiliario, no ajeno a este fenómeno, se ha visto revolucionado fuertemente con la aparición de los portales inmobiliarios. Estos nacieron como respuesta a una clara necesidad del mercado: brindar un espacio digital para publicar propiedades de múltiples fuentes distintas. Esencialmente se encargan de disponibilizar y reunir en un único sitio una acabada cantidad de inmuebles ofreciendo así un gran abanico de opciones para el mercado.

El modelo de negocios de los portales inmobiliarios es conocido técnicamente como modelo de negocios de plataforma, a diferencia de los llamados modelos lineales. Existen diversos tipos de modelos de plataforma entre los que se destacan: on-demand, marketplace, redes sociales, e-commerce, entre otros. Específicamente los portales inmobiliarios son marketpalces, es decir, plataformas online que conectan la oferta con la demanda inmobiliaria. Los compradores interesados tienen la posibilidad de encontrar el inmueble que desean de manera fácil y rápida; y las inmobiliarias y anunciantes particulares pueden publicar sus propiedades en alquiler o a la venta. De esta manera, estos portales funcionan como punto de encuentro entre los compradores y los vendedores y se financian normalmente a partir de la venta de publicidad y la

publicación de anuncios por empresas o particulares. Se trata de modelos de negocio multifacéticos que facilitan la interacción entre dos o más grupos interdependientes (Rahman, K. S., & Thelen, K., 2019). Para conseguir que estos modelos de negocio funcionen, se necesita generar una gran masa de personas a los dos lados de la plataforma (oferta y demanda) que sirvan de motor para impulsar un network effect[1] que haga que el uso de la plataforma se haga viral. (Johnson, Nicholas L., 2020)

Uno de los portales inmobiliarios más conocidos en nuestro país es Properati. Esta empresa nació en el año 2012 como una sociedad anónima de operaciones inmobiliarias online en Argentina[2]. Consiste en un sitio web que brinda a los usuarios el acceso y la utilización de diversos servicios y contenidos relacionados con la búsqueda de inmuebles, exhibiendo ofertas de ventas y alquileres de propiedades; suministrados principalmente por inmobiliarias, empresas constructoras, empresas intermediarias o relacionadas con los servicios de oferta pública de inmuebles para su venta y alquiler. Cabe mencionar que Properati no publica ni ofrece ningún inmueble propio, sino que acerca ofertas de venta y alquiler previamente concedidas por sus respectivos propietarios.

Resulta importante destacar que a diferencia de los esquemas tradicionales en donde las inmobiliarias pagan por colocar anuncios en los sitios, el modelo de negocios de Properati se basa en la venta de leads, es decir, contactos interesados. Esto es, se le cobra a la inmobiliaria cada vez que un usuario se contacta con ella a través de la plataforma web. El modelo de negocios se basa entonces en entregar contactos de calidad. Al pagar sólo por contacto recibido y no por banners o pop-ups, los incentivos entre el vendedor, el usuario y Properati quedan alineados, dando por resultado un sitio web libre de anuncios y con una interfaz amigable. Además, las inmobiliarias pueden acceder a planes pagos que les otorgan beneficios adicionales. Por ejemplo, un abono mensual o trimestral brinda la posibilidad de incorporar el logo corporativo en sus avisos, aumentar el ranking de los mismos dentro del sitio e incluso enviar insights sobre el perfil de los compradores. También, la plataforma online ofrece otros beneficios y descuentos especiales para inmobiliarias con las que firma acuerdos. Por este motivo, los precios de los planes varían de acuerdo al caso.

[1] El efecto red (“network effect” en inglés) es una ventaja competitiva que ocurre cuando el valor de un determinado bien o servicio

[2] En el año 2018 la firma de e-commerce OLX Group compró la empresa Properati con el objetivo de adentrarse en el mercado inmobiliario de Latinoamérica.

Actualmente Properati tiene presencia en Argentina, Colombia, Ecuador, Perú y Uruguay, y en todos los países concreta acuerdos con las inmobiliarias, agentes y constructoras regionales más importantes para publicar sus propiedades. Por otra parte, están disponibles las versiones Android e iOS de Properati, con funcionalidades especialmente diseñadas para dispositivos móviles. En lo que se refiere a cifras de mercado, Properati recibe 2,6 millones de visitas mensuales en toda Latinoamérica, genera 100 mil *leads* y cuenta con 5 mil agentes de bienes raíces y desarrolladores que publican en el sitio.

1.2 Gestión integral de datos

Al registrarse o navegar el sitio web de Properati el usuario brinda información personal, prestando su consentimiento libre, expreso e informado para que la misma sea recogida y almacenada directamente en la base de datos de Properati. Dicha base se encuentra protegida electrónicamente, gracias a los mecanismos de seguridad informática de protección de la información más completos y eficaces para mantenerla en total confidencialidad, conforme a lo indicado en la Ley Nº 25.326 de Protección de Datos Personales. La registración no es requisito para el acceso al sitio web y utilizar sus funciones básicas. El usuario que no completa el formulario de registro y accede al sitio es considerado como "Usuario Libre". Sin embargo, el acceso a determinadas funciones del sitio requiere la identificación del usuario mediante el llenado del formulario de registro en todos sus campos válidos y obligatorios, con información personal exacta, precisa y verdadera. Es responsabilidad exclusiva del usuario la veracidad, exactitud, vigencia y actualización de los datos personales ingresados. El usuario que complete el formulario de registro es considerado como "Usuario Registrado". Vale recalcar que Properati se reserva la voluntad de modificar sin necesidad de notificación previa de ningún tipo, qué funciones quedan abiertas al acceso sin registración y cuales son exclusivas para los usuarios registrados.

1.2.1 Recolección de datos

1. Datos provenientes de interacciones directas

Properati solamente recolecta los datos personales que el usuario envía de manera voluntaria y con una finalidad específica. En otras palabras, los datos personales se

reúnen exclusivamente si el usuario los proporciona, a través del registro, del llenado de formularios o de correos electrónicos, como parte de un pedido de mayor información, de consultas o solicitudes acerca del sitio y situaciones similares en las que el usuario haya elegido proveer esa información. Los datos serán utilizados para que los propietarios o empresas inmobiliarias que publiquen información en el sitio, puedan ser contactados vía telefónica o email por el cliente interesado en adquirir los bienes o servicios relacionados a la información publicada. Adicionalmente, los datos pueden ser usados para que Properati contacte a los usuarios a fin de brindarle información sobre cambios, novedades o requerirles información sobre la experiencia en el uso del sitio web. Esos datos personales no serán empleados con otra finalidad distinta que aquella para la que fueron recolectados en cada oportunidad, excepto en el caso que sea requerido de otro modo por la Ley. Exclusivamente si el usuario otorga permiso sus datos personales podrán ser compartidos con otras empresas subsidiarias o afiliadas a Properati. No se requiere que el usuario proporcione información personal como condición para usar el sitio web, a menos que sea necesario para proveerle información adicional que el mismo usuario solicite.

Algunos de estos datos pueden ser almacenados o procesados en ordenadores ubicados en otras jurisdicciones, cuyas leyes de protección de la información pueden diferir de la jurisdicción argentina. En tales casos, se garantiza que las protecciones adecuadas están o serán acordadas para requerir al procesador de datos de ese país que mantenga las protecciones adecuadas a la legislación argentina. En caso de que el usuario utilice los medios de comunicación de mensajes provistos por Properati, se podrá almacenar y acceder a dichos mensajes, hacer un seguimiento y remitir notificaciones relacionadas a dicho mensaje.

2. Datos recopilados automáticamente

Todo usuario, registrado o no, cuando navega por el sitio web deja una huella digital. Esta última está comprendida por diferentes datos, tales como: el nombre del proveedor de servicio de Internet que emplea el usuario, el sitio web que el usuario usó para vincularse con Properati, los sitios web de Properati que el usuario visitó, los sitios web que el usuario visitó desde Properati, la dirección IP (Internet Protocol) del usuario, el sistema operativo del dispositivo de acceso del usuario, la ubicación del usuario (esto dependerá de los permisos del dispositivo del usuario), la fecha y la duración de las

visitas a la página web. Dichos datos son utilizados periódicamente para fines estadísticos, manteniendo el anonimato de cada usuario individual de manera tal que la persona no pueda ser identificada. Asimismo estos datos, que constituyen la huella digital, se almacenan en servidores para posibilitar la conexión y por cuestiones de seguridad. Los procedimientos de seguridad experimentan revisiones continuas basadas en el desarrollo de nuevas tecnologías.

Los datos asociados con secuencias de clics del usuario son procesados y analizados para ofrecer contenido personalizado, por ejemplo, resultados más relevantes de búsquedas. Se utilizan también para determinar cuánto tiempo transita el usuario en la plataforma web y evaluar de qué modo navega por ella para comprender sus intereses. Por ejemplo, es posible sugerir al usuario cierto contenido en base al análisis de los contenidos que haya cliqueado con anterioridad. Además, las secuencias de clics son empleadas para monitorear y presentar informes sobre la eficacia de alguna campaña a socios comerciales de Properati y para el análisis comercial interno de la plataforma.

1.2.2 Almacenamiento de datos

Los datos personales que el usuario pudiera haber brindado podrán ser almacenados en servidores seguros durante el tiempo requerido para la finalidad para la cual pudieran haber sido solicitados por Properati. Para determinar el período correspondiente de conservación de los datos personales, se considera el volumen, la naturaleza y la sensibilidad de los datos personales, el riesgo potencial de daños por el uso no autorizado o la divulgación de los datos personales y los fines por los que Properati procesa los datos personales. Estos datos se utilizarán únicamente con el propósito de proporcionarle la información o evacuar la consulta que el usuario haya requerido o para otros propósitos para los cuales el usuario dio su consentimiento, salvo que la Ley estipule otra cosa.

El usuario como titular de los datos personales tiene la facultad de ejercer el derecho de acceso a los mismos en forma gratuita a intervalos no inferiores a seis meses, salvo que acredite un interés legítimo al efecto (Ley N°25326, 2000). Asimismo, el usuario tiene derecho a corregir los datos personales inexactos y a retirar el consentimiento que ha dado para su utilización a Properati solicitando que los mismos sean eliminados de los registros en los cuales se encuentren almacenados. La Agencia de Acceso a la

Información Pública es el órgano de Control de la Ley de datos personales y la misma tiene la atribución de atender las denuncias y reclamos que se interpongan con relación al incumplimiento de las normas sobre protección de datos personales.

1.2.3 Transmisión de datos

Previa aprobación del usuario mediante su explícita aceptación marcando las casillas correspondientes, Properati puede entregar a las inmobiliarias la información de contacto del usuario en relación a una propiedad en particular o a los parámetros de la búsqueda que dicho usuario hubiera realizado. Esto permite que las inmobiliarias se comuniquen con los usuarios que facilitaron sus datos de contacto. Cada inmobiliaria que resulte como oferente de propiedades a partir de las búsquedas realizadas en el sitio, declara estar habilitada para el ejercicio de la actividad inmobiliaria de acuerdo a los requisitos impuestos a esa profesión en cada jurisdicción. Properati no se hace responsable de la calidad, exactitud, fiabilidad, corrección y utilidad de los datos provistos por los oferentes de inmuebles. El contenido de la información será exclusiva responsabilidad de quienes la intercambian (remitente y destinatario).

Cabe destacar que Properati tiene la potestad de compartir información personal de los usuarios con sus proveedores externos de servicios, tales como Amazon Web Service, Microsoft Azure (quienes brindan infraestructura tecnológica, es decir, instalaciones de almacenamiento en la nube[3]) o proveedores de servicios de pago. Además, Properati utiliza proveedores que le brindan servicios de marketing y análisis de datos. En todos los casos, la información se transmite con un formato no identificable para monitorear y presentar informes sobre la eficacia de alguna campaña a socios comerciales de la firma y para el análisis comercial interno. Por otro lado, es posible que Properati difunda los datos personales de sus usuarios a las autoridades de cumplimiento de la ley, autoridades reglamentarias, gubernamentales o entidades públicas y otros terceros relevantes en cumplimiento de requisitos legales o normativos.

[3] El almacenamiento en la nube es un modelo de servicio en el cual los datos de un sistema de cómputo se almacenan, se administran, y se respaldan de forma remota, típicamente en servidores que están en la nube.

1.2.4 Seguridad de datos

Toda la información que se recibe del usuario se almacena en servidores seguros. Properati garantiza la implementación de medidas técnicas y organizativas adecuadas para proteger los datos personales de todos sus usuarios. Se evalúa continuamente la seguridad de la red y la adecuación del programa interno de seguridad de la información diseñado para: ayudar a mantener los datos seguros en caso de pérdida, acceso o divulgación accidentales o ilícitos; identificar riesgos razonablemente previsibles a la seguridad de la red; y minimizar riesgos de seguridad mediante evaluaciones de riesgo y verificaciones regulares. Vale la pena señalar que Properati asegura que todos los datos de pagos están cifrados con tecnología SSL (Secure Sockets Layer).

Por último, es importante mencionar que Properati lleva a cabo verificaciones de seguridad a sus proveedores de servicios externos. Se les solicita que respeten la seguridad de los datos personales de los usuarios y que los traten conforme a las normativas legales vigentes. Los proveedores no tienen permitido bajo ningún punto de vista utilizar los datos personales de los usuarios de Properati para sus propios fines.

1.3 Problemática detectada

En la era digital los datos representan un activo fundamental para cualquier organización y por ese motivo su correcta gestión resulta esencial. Los portales inmobiliarios, tales como Properati, administran, recogen y almacenan grandes volúmenes de datos correspondientes a propiedades de diversa índole, lo cual supone un verdadero desafío desde el punto de vista técnico. Sin embargo, este hecho también puede convertirse en una real oportunidad de negocio ya que mediante técnicas de Big Data y aprendizaje automático es posible transformar enormes cantidades de datos en valioso conocimiento que optimice la toma de decisiones. (Schmarzo, B., 2013).

Properati recoge miles de datos estructurados y no estructurados vinculados con anuncios de bienes raíces. En la mayoría de los casos, las inmobiliarias envían estos anuncios a través de sus sistemas de CRM (Customer Relationship Management) que se conectan a la plataforma a través de una API (Application Programming Interface) provista por Properati. Esta metodología le permite a las inmobiliarias realizar el seguimiento de todas las etapas del proceso de compra o venta de cada cliente y, además, publicar simultáneamente varias propiedades de manera automática en la

plataforma web. Esta formidable cantidad de datos se compone normalmente por: precio del inmueble, descripción de la propiedad, cantidad de habitaciones, dormitorios, baños, ubicación geográfica, amenities, imágenes del edificio, entre otras. El tener disponible esta inmensa cuantía de datos trae consigo una concreta e interesante oportunidad latente para Properati: ofrecer un servicio pago a los usuarios que resulte del análisis de los datos de las miles de propiedades enviadas por las inmobiliarias.

En el ámbito inmobiliario una necesidad muy recurrente consiste en la fijación del precio de una propiedad, la cual no es una tarea para nada sencilla. Esta resulta vital para garantizar la conformidad con el propietario del inmueble que desea vender su vivienda como así también con el comprador que anhela la adquisición de la misma. En este aspecto, el proceso de valuación convencional es complejo y requiere la intervención de un experto profesional que debe considerar una gran cantidad de variables relacionadas con el activo inmobiliario bajo análisis, como por ejemplo: zona geográfica, dimensiones, cercanía a zonas comerciales, comodidades, luminosidad, estado de la construcción, antigüedad, etcétera. El trabajo que lleva a cabo el agente tasador demanda demasiado tiempo y dinero, dos recursos sumamente escasos en los tiempos que corren.

En la actualidad las técnicas de *machine learning* se han convertido en un aliado esencial a la hora de abordar problemáticas que requieran la consideración y procesamiento de múltiples variables. Son muchos los campos de aplicación que tienen los modelos predictivos en el ámbito empresarial, y con el transcurso del tiempo se han logrado perfeccionar gracias a la mayor capacidad de procesamiento de grandes volúmenes de datos que ofrecen las computadoras actuales (Fawcett, T. & Provost, F., 2013). Es por eso que en este trabajo se propone reemplazar la figura del tasador profesional por un algoritmo de *machine learning* que realice una tasación automatizada de inmuebles a partir de una base de datos extraída de la plataforma online de Properati. Dentro del aprendizaje automático se puede distinguir dos grandes familias de algoritmos: modelos de aprendizaje supervisado y modelos de aprendizaje no supervisado. En la primera, los modelos son entrenados a partir de un conjunto de datos en el que los valores de la variable a predecir son conocidos. Por otro lado, en los modelos de aprendizaje no supervisado, el conjunto de datos empleado en el entrenamiento no contiene la respuesta, de modo que no existe un resultado a

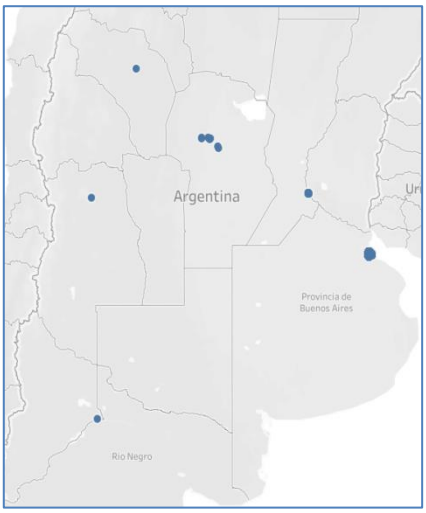
reproducir. Para llevar a cabo este trabajo de investigación se implementarán modelos de aprendizaje supervisado debido a que los precios de los inmuebles que se buscan predecir son valores perfectamente conocidos. Entre los claros beneficios que otorga una herramienta de valuación automatizada de inmuebles, se destacan los siguientes: agilizar notablemente el proceso de valuación y reducir los costos totales de la gestión puesto que se elimina la figura del tasador.

Apartado 2. Descripción metodológica

2.1 Recopilación de los datos

La base de datos utilizada para la elaboración del presente trabajo consta de 2.735 departamentos de dos ambientes pertenecientes a 45 barrios diferentes de la Ciudad Autónoma de Buenos Aires. Dicho dataset fue extraído de la página web **www.properati.com.ar** y el mismo contiene departamentos publicados durante el año 2019. La base se encuentra en formato .csv y posee datos de tipo multivariado. Los atributos del set de datos son mixtos ya que existen campos numéricos, de texto y fecha. Cabe destacar que la base presenta valores faltantes cuyo tratamiento será explicitado en las secciones siguientes del trabajo. En el Anexo 1 se presenta detalladamente la estructura del conjunto de datos. Todos los inmuebles disponen de dos habitaciones, un dormitorio y un baño.

2.2 Procesamiento y análisis de los datos

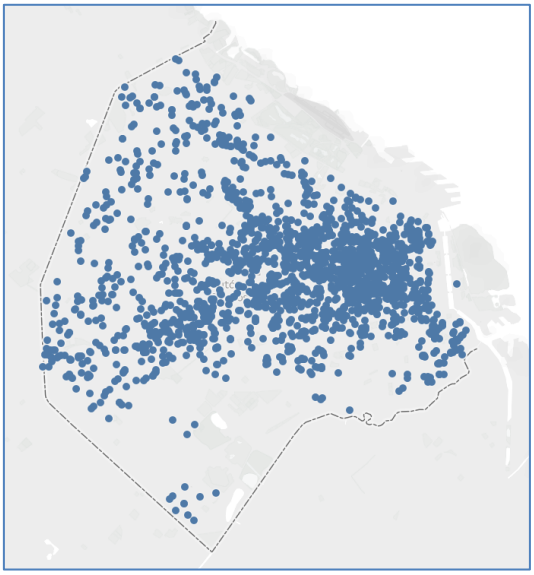


Antes de aplicar los algoritmos de machine de learning para predecir los precios de los inmuebles, se realizó un tratamiento previo de los datos para lo cual se emplearon los siguiente programas: **Excel** en su versión 2010, **Tableau Public** 2020.3, **R Studio** versión 4.0.1 (2020-06-06) y **Rapid Miner Studio Educational** versión 9.7.

En primer lugar, se inspeccionaron las ubicaciones geográficas de todos los inmuebles para garantizar que la totalidad de los mismos estén circunscriptos efectivamente a la zona de Capital

Federal. Los datos de Latitud y Longitud de cada departamento fueron de utilidad para volcar en un mapa la distribución geográfica de todos los inmuebles. Este análisis fue realizado utilizando la herramienta **Tableau Public** y se expone en la imagen del mapa de Argentina. Nótese que no todos los departamentos están ubicados en la Ciudad de Buenos Aires. Se encontraron 14 inmuebles ubicados en Santa Fe, 7 en Córdoba, 1 en La Rioja, 1 en Mendoza y 1 en Neuquén.

Los 24 registros detectados fueron removidos de la base de datos ya que no pertenecían a la zona bajo estudio, restando un total de 2.711 registros. La distribución final de las propiedades se puede visualizar en el siguiente mapa:



Fuente: elaboración propia

Posteriormente, se continuó con la búsqueda de registros duplicados. Se encontraron 667 instancias duplicadas las cuales fueron oportunamente eliminadas a fin de no darle mayor importancia a aquellas publicaciones de inmuebles repetidas, las cuales podrían distorsionar las futuras predicciones de precios. Una vez sustraídos los 667 registros duplicados, la base de datos permaneció con 2.044 registros útiles.

En lo que respecta al tratamiento de datos faltantes en los campos “sup_total” y “sup_cubierta”, se presentaron distintos casos que fueron resueltos con diversos enfoques. Las variantes encontradas fueron las siguientes:

- Datos faltantes en los campos “sup_total” y “sup_cubierta”.
- Dato faltante en el campo “sup_total” y dato presente en campo “sup_cubierta”.
- Dato faltante en el campo “sup_cubierta” y dato presente en campo “sup_total”.

En aquellos registros con datos faltantes en los campos “sup_total” y “sup_cubierta”, se procedió a inspeccionar los campos “título” y “descripción”. A partir de la información contenida en estos últimos campos, se completaron los valores faltantes de superficie de los departamentos. Cuando los campos de texto no alojaban los datos faltantes buscados, se implementó otro criterio de imputación. Se investigaron las superficies cuadradas más frecuentes de los departamentos del dataset y se descubrió que la mayoría de los departamentos eran de 35 m^2. Por esta razón, se decidió tomar dicho valor para completar aquellos registros sin información sobre la superficie cubierta del inmueble.

Para completar el campo de superficie total se tuvo presente la siguiente ecuación:

$$\boldsymbol{Superficie\ total = Superficie\ cubierta + Superficie\ no\ cubierta}$$

En donde: $\boldsymbol{Superficie\ no\ cubierta = Sup.balcón + Sup.Terraza + Sup.patio}$

Con el objetivo de identificar registros con superficie no cubierta, se crearon variables *dummies* para indicar la ausencia o presencia de “balcón”, “terraza” o “patio” en los campos “título” y “descripción” de cada inmueble. Para el cálculo de cada superficie se adoptaron los siguientes valores estándar: $3m^2$ (balcón), $10m^2$(terraza) y $4m^2$ (patio).

Por otro lado, en los registros con datos faltantes en el campo “sup_cubierta” y con datos presentes en el campo “sup_total”, se procedió según:

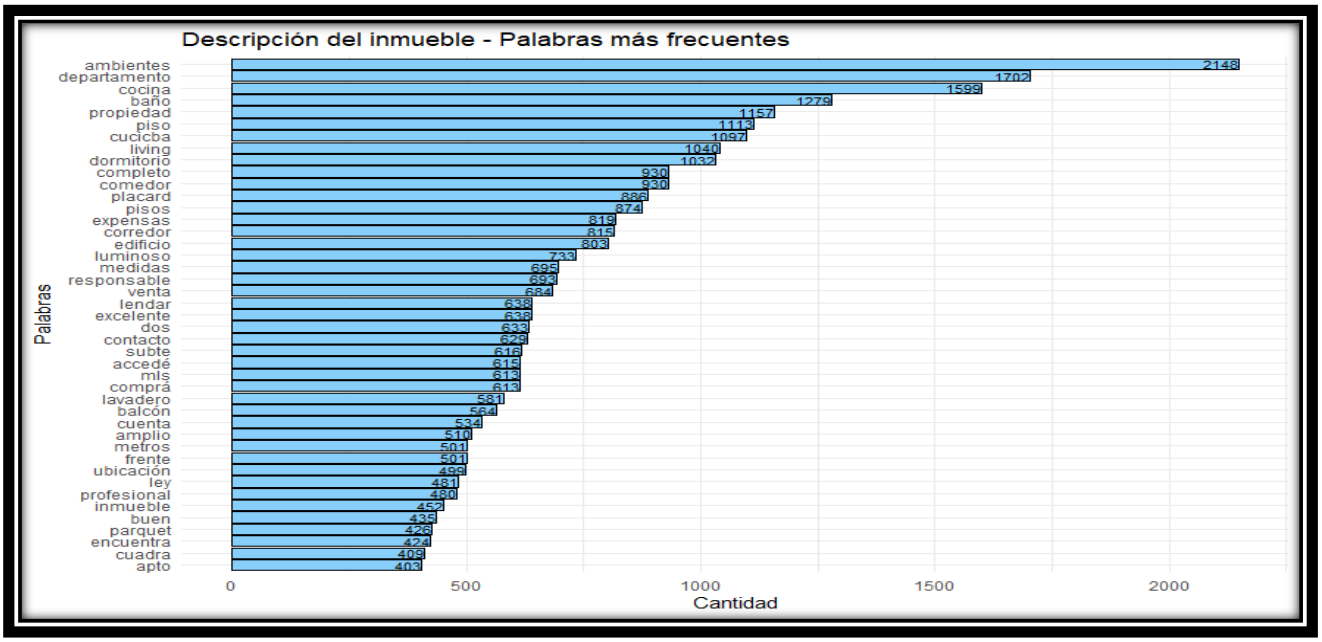
$$\boldsymbol{Superficie\ cubierta = Superficie\ total - Superficie\ no\ cubierta}$$

Mientras que en los casos en donde no se detectó la presencia de “balcón”, “terraza” o “Patio” se tomó: $\boldsymbol{Superficie\ cubierta = Superficie\ total}$

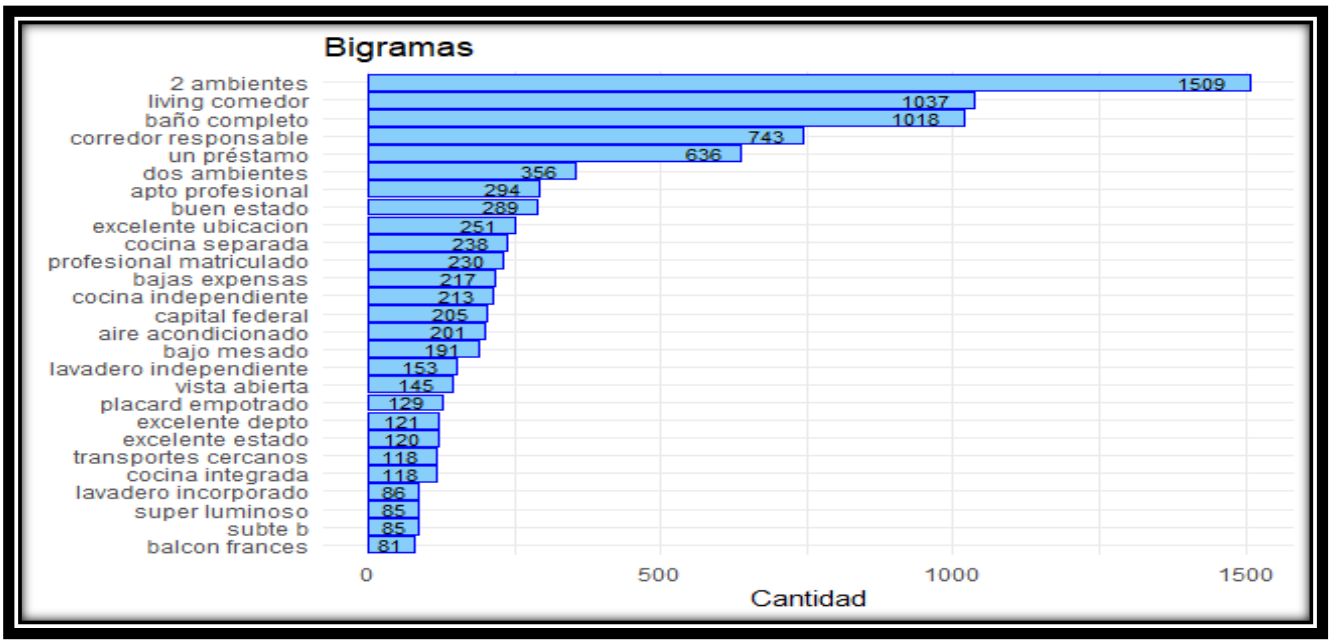
Además, todos los valores *outliers* en los campos “sup_total” y “sup_cubierta” fueron reemplazados según los criterios explicados anteriormente.

Luego de finalizar el tratamiento y preparación de los datos existentes, se continuó con la construcción de nuevos atributos. Por medio de un script ejecutado con el software **Rstudio**, se realizó el tratamiento de los campos de texto “título” y “descripción” de cada una de las propiedades. Se utilizó una biblioteca de *Text Mining* para llevar a cabo diferentes transformaciones en el texto: se convirtieron todas las letras mayúsculas en

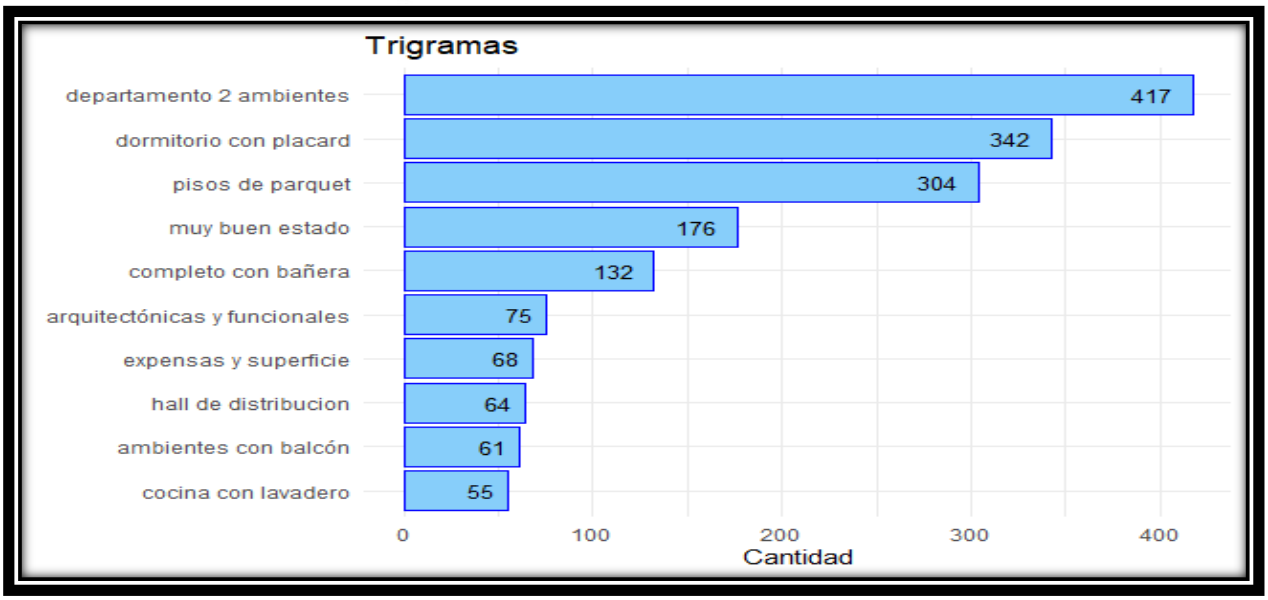
letras minúsculas, se eliminaron todos los espacios en blanco, signos de puntuación y números presentes en cada campo de texto y se removieron las ***stopwords***, es decir, palabras tales como pronombres, preposiciones, artículos, etc., que no aportaban información relevante en el análisis. Por otra parte, se implementó una técnica para detectar N-gramas en el contenido del texto. Con dicha técnica se logra extraer distintas combinaciones de palabras en las cadenas de texto que conforman la base.



Fuente: elaboración propia



Fuente: elaboración propia



Fuente: elaboración propia

Fuente: elaboración propia

Los gráficos de barra anteriormente expuestos fueron confeccionados a través de la librería **ggplot2,** mientras que las nubes de palabras se realizaron a través de las librerías **wordcloud** y **wordcloud2** disponible para el lenguaje de programación **R** mediante el *Entorno de Desarrollo Integrado* **Rstudio.** A partir del análisis anterior, se identificaron ciertas palabras y frases repetitivas relacionadas con las *amenities* de los departamentos. Este procesamiento del texto fue de gran utilidad para construir las siguientes variables *dummy* en la base de datos: luminoso, balcón, parquet, frente, contrafrente, piscina, cochera, parrilla, patio, terraza, baño completo, apto profesional y dormitorio con placard.

Los datos de latitud y longitud de cada inmueble fueron empleados para realizar distintos cálculos de distancia con respecto a ciertos puntos de interés de la ciudad y vías de acceso a transportes públicos. Se seleccionaron los siguientes 6 sitios de interés vinculados principalmente con diferentes zonas comerciales de la Ciudad Autónoma de Buenos Aires:

Sitios de interés	Latitud	Longitud
Patio Bulrich	-34,588799	-58,383965
Abasto Shopping	-34,603281	-58,410845
Distrito Arcos	-34,580624	-58,427692
Obelisco	-34,603695	-58,381538
Parque centenario	-34,606209	-58,435656
Plaza Italia	-34,581217	-58,420823

Se calculó la distancia existente entre cada punto de interés y cada departamento, totalizando seis nuevas variables incorporadas. Además, se incluyeron las siguientes variables en la base de datos:

- Distancia a la boca de subte más cercana (en kilómetros).
- Distancia a la universidad más cercana. (en kilómetros).
- Precio por metro cuadrado de cada uno de los barrios.

Para fabricar las variables anteriormente mencionadas, se calculó la distancia del inmueble a cada boca de subte o universidad y se tomó la distancia mínima de todas las calculadas. Las distribuciones de las 393 universidades y 379 bocas de subte se pueden visualizar en los mapas de las Figuras 1 y 2 respectivamente.

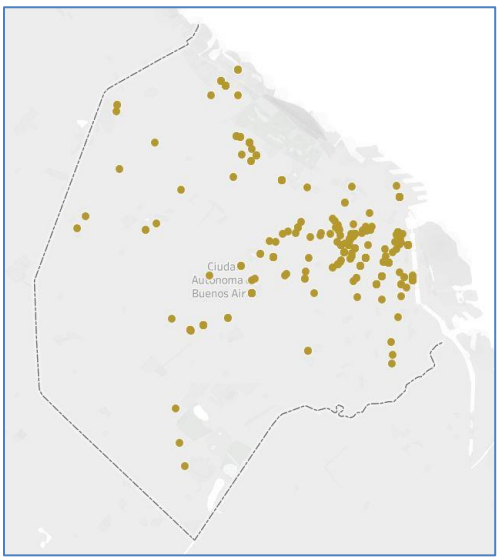


Figura 1

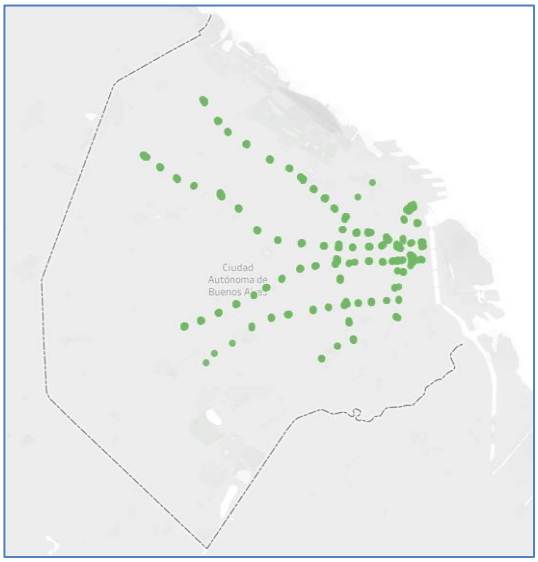


Figura 2

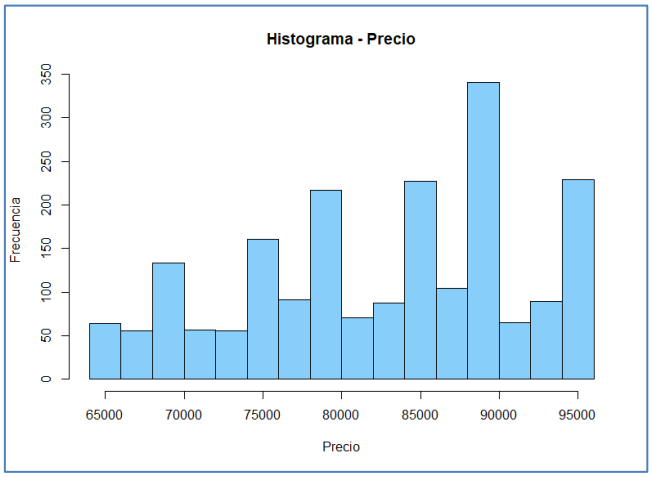
A continuación se muestran ciertas medidas correspondientes a las variables métricas continuas de la base de datos.

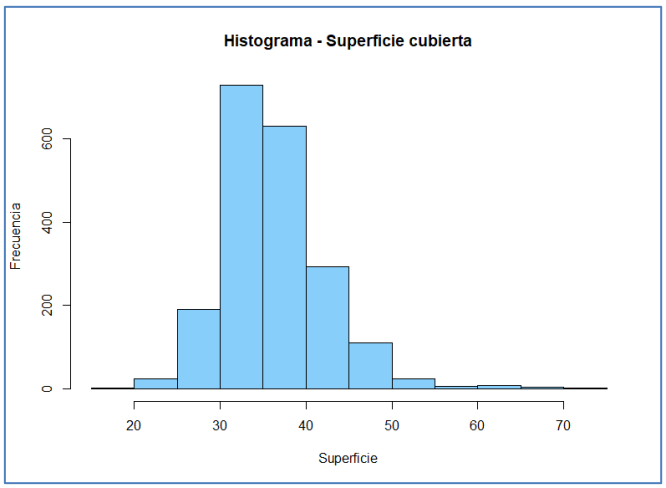
Medidas	Sup_total	Sup_cubierta	Precio	Precio barrio m2
Mín.	20	19	65.000	1.217
Máx.	86	74	95.000	5.715
Media	39,07	37,04	82.974,12	2.507
Mediana	38	36	85.000	2.600
Moda *	35	35	95.000	1.900
Varianza	55,58	36,47	73.669.661,51	3.356.769,2
Desvío	7,46	6,04	8.583,10	1.832,15

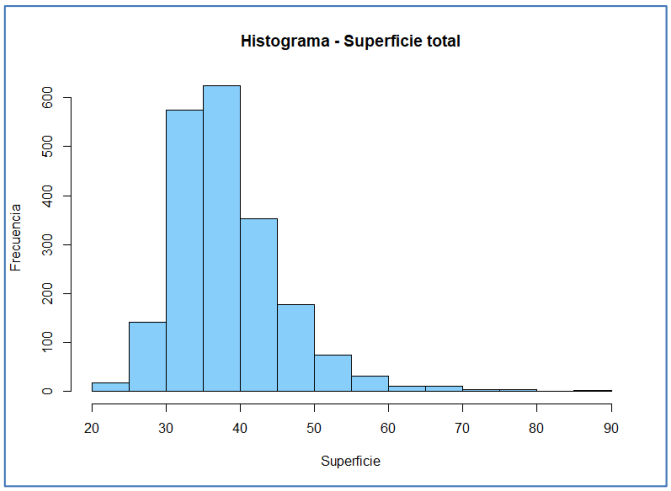
Medidas	Patio Bulrich	Abasto Shopping	Distrito Arcos	Parque Centenario	min_bocas	min_univ	Obelisco	Plaza Italia
Mín.	0,36	0,07	0,21	0,03	0,01	0,01	0,00	0,07
Máx.	14,65	11,75	13,17	10,24	5,74	5,32	14,26	13,31
Media	5,77	3,91	5,03	3,83	0,85	0,78	5,28	5,02
Mediana	5,07	3,16	4,73	3,54	0,42	0,50	4,50	4,57
Moda *	-	-	-	-	-	-	-	-
Varianza	10,02	7,21	4,81	4,03	1,13	0,69	11,77	5,30
Desvío	3,16	2,68	2,19	2,01	1,06	0,83	3,43	2,30

*Moda: Las variables relacionadas con cálculos de distancia no poseen moda ya que todos los valores de los datos son distintos y por lo tanto únicos.

Se presentan los histogramas de las variables Precio, Superficie cubierta y Superficie total para complementar las cifras anteriormente expuestas:







Además, la base de datos contiene tres atributos de fecha: Pub_inicio, Pub_fin y Pub_creada. La variable Pub_inicio tiene un rango entre el 08/04/2019 y el 31/12/2019 mientras que la variable Pub_fin posee un rango entre el 11/04/2019 y el 30/06/2020 (la variable Pub_creada posee las mismas fechas que la variable Pub_inicio).

Enseguida se expone en detalle la frecuencia de aparición de cada una de las variables *dummy* creadas a partir de las técnicas de *Text Mining*:

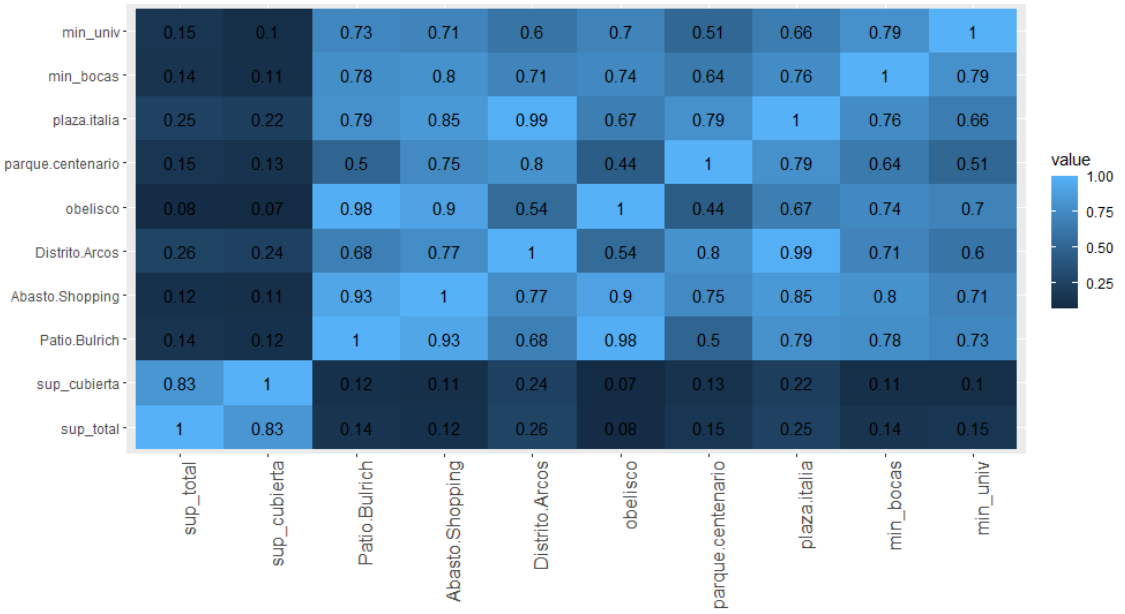
Campo	Frecuencia	Porcentaje
Luminoso	574	28,08 %
Balcón	671	32,83 %
Terraza	191	9,34 %
Patio	140	6,84 %
Parquet	361	17,66 %
Frente	569	27,84 %
Contrafrente	287	14,04 %
Cochera	104	0,05 %
Parrilla	71	3,47 %
Piscina	72	3,52%
Baño completo	1.018	49,81 %
Apto profesional	294	14,38 %
Dormitorio con placard	342	16,73 %

Por otro lado, la base de datos consta de inmuebles de 55 barrios distintos, siendo el más frecuente Balvanera (201 registros) y Agronomía el de menor frecuencia (2 registros).

2.3 Reducción de la dimensionalidad de la base

A fin de disminuir la cantidad de variables cuantitativas de la base de datos, se implementará la técnica de Análisis de Componentes Principales (PCA). Dicha técnica se aplicará sobre 10 variables métricas cuyos coeficientes de correlación de Pearson se presentan a continuación.

Matriz de Correlación de Pearson



Antes de aplicar PCA será necesario corroborar que realmente existe correlación entre las variables métricas originales, es decir, que haya multicolinealidad entre las variables. De no ser así, no tendría sentido alguno proseguir con este método de análisis multivariado pues si las variables originales ya son incorrelacionadas, no sería necesario encontrar nuevas variables linealmente independientes.

Un valor bajo del determinante de la matriz de correlaciones indica alta multicolinealidad. En este caso se puede apreciar que es muy cercano a 0, lo que sugiere un alto nivel de colinealidad en el conjunto de variables involucradas en la matriz:

```
> det(mat_cor)
[1] 1.698012e-09
```

Otro método para evaluar si la correlación entre las variables analizadas es lo suficientemente fuerte como para justificar la implementación de PCA, es el test de Barlett. Este test plantea como hipótesis nula que la matriz de correlaciones es igual a

matriz identidad, en otras palabras, que las variables son linealmente independientes. Entonces, si se logra rechazar dicha hipótesis se concluirá que las variables bajo análisis están correlacionadas entre sí.

```
> library(psych)
> cortest.bartlett(mat_cor,n=46)
$chisq
[1] 824.5805

$p.value
[1] 6.570517e-144

$df
[1] 45
```

Dado que el "p value" obtenido acusa un valor muy bajo entonces se puede rechazar la hipótesis nula lo cual permite decir que las variables presentan correlación entre sí. Los análisis previamente abordados habilitan a la aplicación del método de PCA. Al emplear dicha técnica sobre el set de variables métricas originales, se consiguen los siguientes resultados:

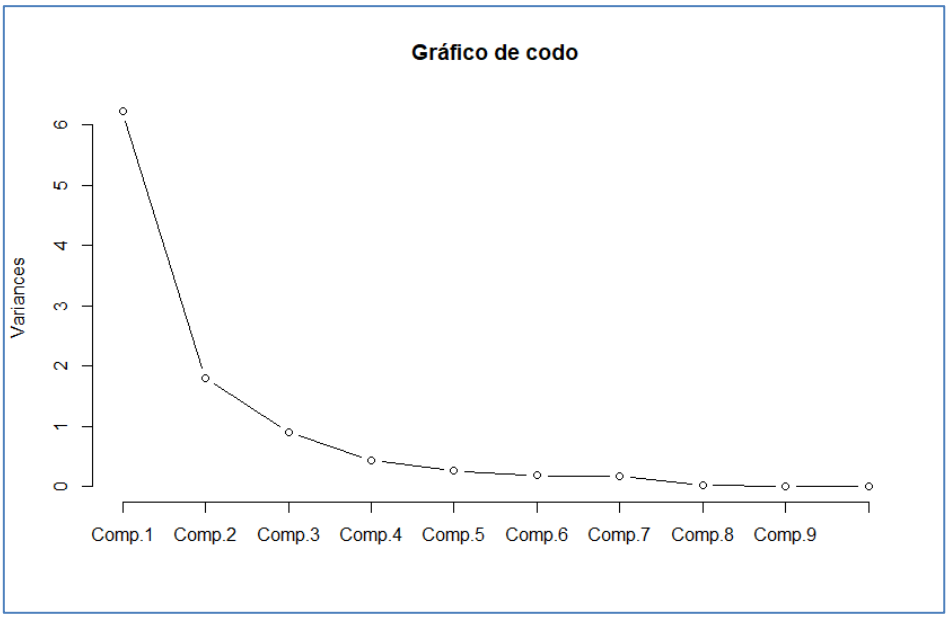
```
> #Análisis de Componentes Principales
> pca=princomp(base,cor=TRUE,scores=TRUE)
> pca
Call:
princomp(x = base, cor = TRUE, scores = TRUE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
2.49649814 1.34044859 0.94785884 0.66258141 0.51433489 0.42810741 0.40641318 0.13253066
    Comp.9    Comp.10
0.04685790 0.02227198

 10  variables and  2044 observations.
> summary(pca)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5    Comp.6     Comp.7
Standard deviation     2.4964981 1.3404486 0.94785884 0.66258141 0.51433489 0.4281074 0.40641318
Proportion of Variance 0.6232503 0.1796802 0.08984364 0.04390141 0.02645404 0.0183276 0.01651717
Cumulative Proportion  0.6232503 0.8029305 0.89277418 0.93667559 0.96312963 0.9814572 0.99797439
                           Comp.8       Comp.9      Comp.10
Standard deviation     0.132530663 0.0468578998 2.227198e-02
Proportion of Variance 0.001756438 0.0002195663 4.960409e-05
Cumulative Proportion  0.999730830 0.9999503959 1.000000e+00
```

De las 10 posibles componentes principales, se seleccionará únicamente 2 componentes para reducir la dimensionalidad de la base, con esta cantidad se asegura retener un 80,2% de la información original lo cual resulta sumamente representativo. Este número se puede obtener recordando que la varianza simboliza la retención de la información y que los autovalores de la matriz de correlación son los que contienen dicha varianza.

$$Información\ retenida = \frac{AVA_1 + AVA_2}{\sum_{i=1}^{10} AVA_i} = \frac{6{,}23 + 1{,}79}{10} = 0{,}802$$

AVA = autovalor de la matriz de correlación

En el gráfico presentado a continuación, denominado gráfico de codo o de sedimentación (*scree plot*), se logra visualizar la información retenida según la cantidad de componentes principales:



Las 2 componentes principales elegidas tienen entonces la siguiente composición:

$$z_1 = -0.1007X1 - 0.089X2 - 0.365X3 + 0.350X4 - 0.339X5 - 0.308X6 - 0.374\,X7 - 0.361X8 - 0.354\,X9 - \ 0.323\,X10$$

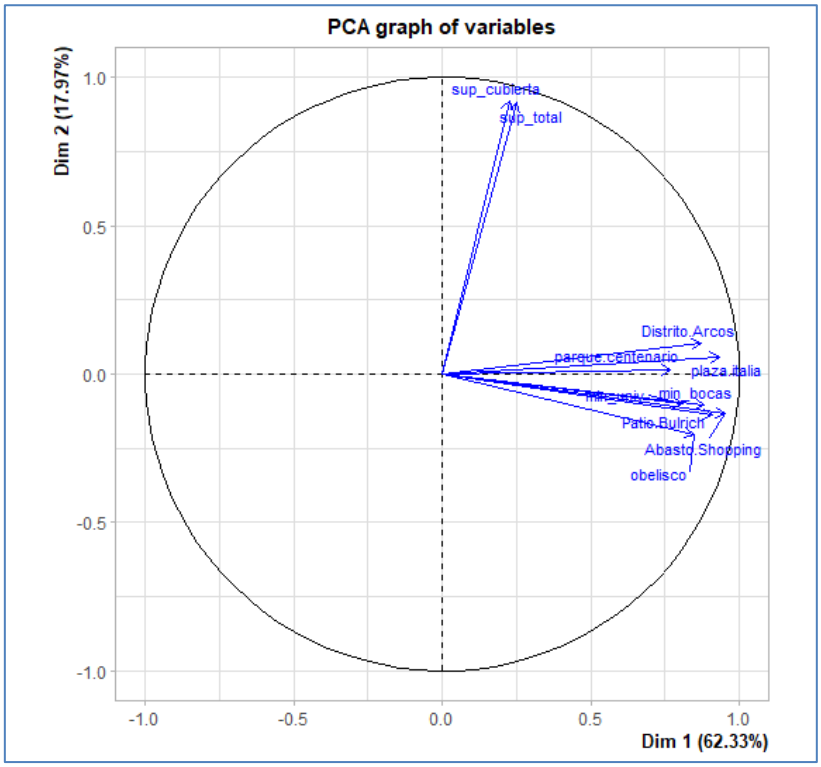
$$z_2 = 0.682\,X1 + 0.686\,X2 - 0.104\,X3 - 0.101\,X4 + 0.082\,X5 - 0.151\,X6 + 0.012\,X7 + 0.044\,X8 - 0.078\,X9 - \ 0.075\,X10$$

```
> eigen(mat_cor)$vectors[,1:2]
              [,1]        [,2]
 [1,] -0.10076632  0.68170068
 [2,] -0.08993771  0.68627813
 [3,] -0.36527346 -0.10378762
 [4,] -0.38173053 -0.10107633
 [5,] -0.35051267  0.08214928
 [6,] -0.33981232 -0.15127453
 [7,] -0.30793550  0.01243953
 [8,] -0.37397748  0.04397897
 [9,] -0.35359287 -0.07778082
[10,] -0.32337528 -0.07447805
```

Cada coeficiente que acompaña a cada variable Xi es el escalar del autovector asociado al autovalor correspondiente de la matriz de correlaciones. A continuación se muestran los scores o puntuaciones que se obtuvieron para las tres componentes principales seleccionadas (se exponen las puntuaciones correspondientes a los primeros 10 registros):

	Comp.1	Comp.2
[1,]	-1.920231421	0.1449311165
[2,]	-1.138436668	2.8414454263
[3,]	-1.275688570	1.6606156525
[4,]	1.029354516	1.9295850129
[5,]	5.246739824	1.0163790055
[6,]	4.851458993	-1.3292002074
[7,]	4.678522968	-0.5281574081
[8,]	3.327315721	-0.9574102554
[9,]	3.462153719	-0.3645105502
[10,]	4.666286716	-0.2977264918

Las puntuaciones son los valores de las componentes principales correspondientes a cada observación. Dichas puntuaciones serán los datos que sustituirán a los valores de las 10 variables originales en la base de datos. En el siguiente gráfico se puede visualizar como es la composición de las dos componentes principales escogidas.



2.4 Modelos predictivos aplicados

Se implementaron tres modelos predictivos de regresión, a saber: *Decision Tree, Random Forest y Gradient Boosted Trees*. Mediante ellos se buscó predecir el precio de las propiedades, es decir, la variable *Precio_inmueble.* Con todos los modelos se persiguió superar el valor del RMSE (raíz del error cuadrático medio) obtenido por el modelo tomado como *Baseline*. Cabe destacar, que dicho modelo es aquel que siempre predice la media de los valores de los precios de las propiedades, en este caso, 82.974,12 USD lo cual implica un RMSE igual a 8.581 USD.

El primero modelo aplicado fue ***Decision Tree*** y se utilizaron las siguientes variables explicativas: pub_inicio, pub_fin, barrio, precio_m2, luminoso, balcón, patio, terraza, parque, frente, contrafrente, cochera, parrilla, piscina, lavadero, baño_completo, apto_profesional, dormitorio_con_placard y las dos componentes principales comp.1 y comp.2.

Para el método ***Random Forest*** se tomaron las siguientes variables independientes: pub_inicio, pub_fin, barrio, precio_m2, luminoso, balcón, patio, terraza, parque, frente, cochera, parrilla, baño_completo, apto_profesional, y las dos componentes principales comp.1 y comp.2.

Para el modelo ***Gradient Boosted Tree*** se emplearon las siguientes variables explicativas: pub_inicio, pub_fin, barrio, precio_m2 luminoso, balcón, patio, terraza, parque, frente, contrafrente, parrilla, piscina, apto_profesional, dormitorio_con_placard y las dos componentes principales comp.1 y comp.2.

2.5 Métricas de evaluación

Para realizar el entrenamiento y prueba de la base de datos se aplicó la técnica de Validación Cruzada utilizando 10 particiones, lo cual se llevó a cabo con el operador *Cross Validation* disponible en el programa **Rapid Miner**. Esto se hizo para evitar el sobreajuste de los modelos y alcanzar mayor poder de generalización en los mismos. Las métricas que se utilizaron para comparar los modelos fueron las siguientes: RMSE, Error Absoluto y Error Relativo. Los resultados se expondrán mediante la presentación de una tabla y gráfico.

Los parámetros de los modelos fueron optimizados gracias al operador *Optimize Parameters* del programa de minería de datos **Rapid Miner**. Para el método *Decision Tree* fueron optimizados simultáneamente los cuatro parámetros: *Maximal depth, Minimal leaf size, Minimal size for Split y Number of prepruining alternatives*. Se aplicó pruning con confidence = 0.1 y prepruning con minimal gain = 0.01.

Para el modelo *Random Forest* se optimizó en primer lugar el parámetro *number of trees* y luego se continuó con el ajuste del parámetro *maximal depth*. Con los valores obtenidos se prosiguió con la optimización de los parámetros: *minimal leaf size, minimal size for split y number of prepruining alternatives*. Se aplicó pruning con *confidence* = 0.1 y *prepruning* con *minimal gain* = 0.01.

Finalmente, para el caso del modelo *Gradient Boosted Trees* se comenzó con el ajuste del parámetro *number of trees* y luego se avanzó con la optimización del parámetro *number of bins, learning rate y maximal depth*. A continuación se exponen los resultados obtenidos al ejecutar la optimización del último de los parámetros, en donde se configuró una profundidad mínima de 4 y máxima de 200 con 20 particiones, resultando un total de 21 combinaciones posibles.

Optimize Parameters (Grid) (21 rows, 3 columns)

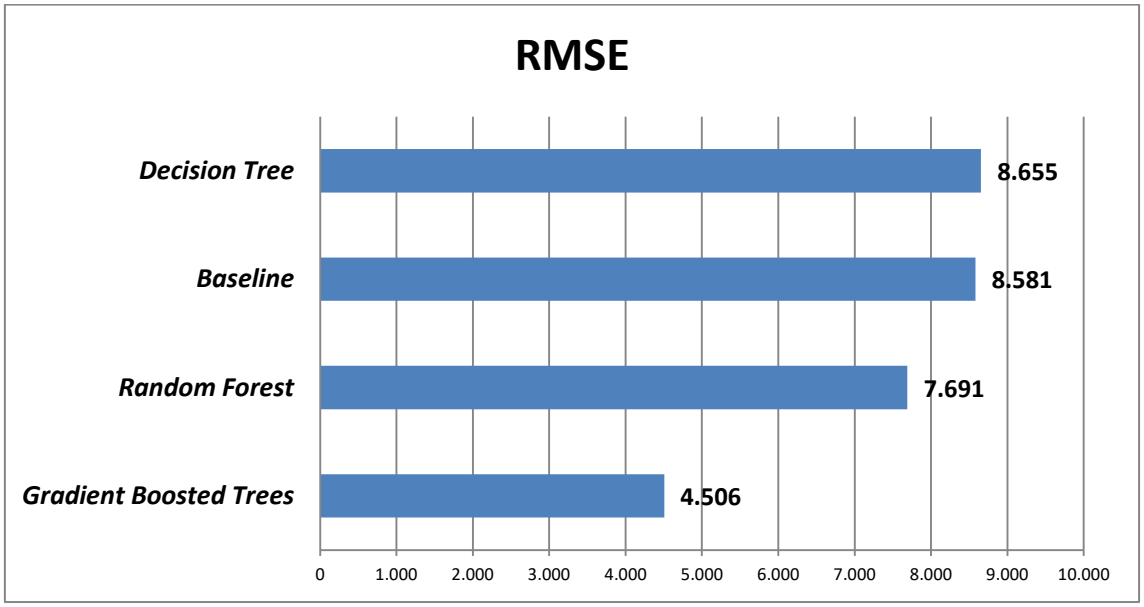
iteration	Gradient Boosted Trees.maximal_depth	root_mean_squared_error ↑
1	4	4553.651
11	102	4889.925
6	53	4889.925
12	112	4889.925
7	63	4889.925
3	24	4889.925
8	73	4889.925
13	122	4889.925
4	33	4889.925
9	82	4889.925
14	131	4889.925
5	43	4889.925
10	92	4889.925
15	141	4889.925
16	151	4889.925
19	180	4889.925
20	190	4889.925
17	161	4889.925
18	171	4889.925
21	200	4889.925
2	14	4889.985

2.6 Resultados

Los resultados arrojados por los distintos modelos se revelan en la siguiente tabla:

Modelo de regresión	RMSE	Error absoluto	Error relativo
Decision Tree	8654.817 +/- 391.485	7340.519 +/- 370.322	9.12% +/- 0.53%
Random Forest	7691.154 +/- 362.224	6336.691 +/- 389.537	7.85% +/- 0.54%
Gradient Boosted Trees	4506.600 +/- 482.828	2986.670 +/- 290.465	4.00% +/- 0.44%

Para establecer un análisis comparativo de los modelos aplicados con respecto al modelo *Baseline*, se construyó el siguiente gráfico:



Como puede observarse en el gráfico anterior, no todos los modelos aplicados superaron la métrica **RMSE** del modelo tomado como *Baseline*. En particular, el método *Decision Tree* brindó un modesto **RMSE** de 8.655, no pudiendo superar a la cifra obtenida por el *Baseline*. La técnica *Random Forest*, por su parte, logró un **RMSE** de 7.691, lo que significa una mejora del 10,37% con respecto al modelo tomado como *Baseline*. Sin embargo, el modelo con mayor exactitud en las predicciones fue *Gradient Boosted Trees*, presumiendo un **RMSE** de 4.506, con lo cual consiguió mejorar el *Baseline* en un 47,49 %. En el **Anexo II** se presentará una tabla con una descripción de los modelos con sus respectivos parámetros optimizados y métricas.

Para el cálculo de la métrica **RMSE** del modelo *Baseline*, se consideró la siguiente expresión:

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{precio\ real - precio\ pronosticado})^{\mathbf{2}}}$$

Siendo:

N = cantidad de registros

Precio pronosticado = Precio promedio de los inmuebles = 82.974,12

Apartado 3. Implementación

3.1 Puesta en producción del modelo

Video 1: https://drive.google.com/file/d/1E8Ska4MYQ9bVFMvwJsMgx2auTeYzetKs/view

Video 2: https://drive.google.com/file/d/1CtGbaRgz38LnpT4qAGcaBXm_hsHJJKC-/view

Video 3: https://drive.google.com/file/d/1FnNpTRYvz_Qzctd3k9CHOF3aaxB8OxTt/view

Video 4: https://drive.google.com/file/d/1NmENwK8bzXrZVm5jGf6cJSyclXskgUmc/view

Video 5: https://drive.google.com/file/d/1dMlhlVY3JmkcloYbl36qkByrIYF8DOYi/view

Video 6: https://drive.google.com/file/d/1wAhPIFvXUbQhHiHCRrHa4uWB1XFMa4Mn/view

Video 7: https://drive.google.com/file/d/146SsM9BpJnAQMEgX1cVcmnxvGOafeDNr/view

Video 8: https://drive.google.com/file/d/1hVlI8jjagzAAUBVksl5IjjNDS0y3T3t6/view

Video 9: https://drive.google.com/file/d/1bvBtlOSIOBXiTfd4FdoyTKbI1aQbT_x1/view

Video 10: https://drive.google.com/file/d/11OUM4FBJ8JUzFItEUOiazl-Be_umxUmu/view

Video 11: https://drive.google.com/file/d/1wBXKs06s0N_9RY15C4b7EFvmA1syounY/view

3.2 Metodologías ágiles para la implementación de proyectos de aprendizaje automático

3.2.1 Contexto y arquitectura disponible en la organización

El proyecto de aprendizaje automático se implementará en el contexto de una empresa de internet la cual consiste en una plataforma online en donde principalmente inmobiliarias y anunciantes particulares pueden publicar sus propiedades a la venta. Se trata de un portal web de propiedades que le permite al usuario final encontrar rápidamente su futura vivienda. En lo que respecta a la arquitectura disponible, la empresa cuenta únicamente con Azure Data Factory para capturar todos los datos

provenientes de las distintas inmobiliarias y con el producto Azure Data Lake para almacenar dichos datos.

3.2.2 Problema del negocio a solucionar

La fijación del precio de una propiedad no es una tarea sencilla y enmarca el punto de partida para dar inicio a cualquier proceso de compra – venta. Efectuar una correcta valuación de bienes raíces es vital para garantizar la conformidad con el propietario del inmueble que desea vender su vivienda como así también con el comprador que anhela la adquisición de la misma. En este aspecto, el proceso de valuación convencional requiere la intervención de un experto profesional que debe considerar una gran cantidad de variables relacionadas con el activo inmobiliario bajo análisis. El trabajo que lleva a cabo el agente tasador demanda tiempo y dinero, dos recursos sumamente escasos en los tiempos que corren.

En la actualidad, las técnicas de machine learning se convirtieron en un aliado esencial a la hora de abordar problemas que requieran la consideración y procesamiento de múltiples variables. En este caso, la implementación de un modelo predictivo de precios de propiedades le permitirá a todo aquel interesado en vender o comprar una vivienda conocer su precio de manera totalmente automática y sencilla sin necesitar la intervención de un agente tasador. Por lo tanto, los claros beneficios que otorga esta herramienta son agilizar al máximo la tarea de valuación y reducir los costos totales asociados a la gestión de dicha tasación inmobiliaria puesto que se elimina la figura del experto tasador.

3.2.3 Planteamiento de la metodología

La metodología de trabajo elegida para implementar el proyecto será la metodología Agile y dentro de ella se optará por la metodología Scrum. Esta presenta características que resultan sumamente adecuadas para proyectos innovadores, tales como proyectos asociados a modelos de aprendizaje automático en donde el entendimiento de las problemáticas abordadas se descubre y perfecciona a medida que se avanza en el proyecto. La metodología Agile es fundamental para lograr un lanzamiento y llegada al

mercado más rápida del producto final, en este caso, un tasador automático de propiedades. Además, se consigue refinar la solución planteada en cada ciclo iterativo, también conocido como sprint.

De esta forma, propone generar desarrollos iterativos e incrementales que brindan propuestas de alto valor agregado para el usuario final, lo cual es sumamente perseguido en toda organización. Se logra pensar en iniciativas mucho más precisas, inversiones más fiables, que los equipos comprendan mejor y más profundamente las soluciones propuestas y conozcan claramente lo que quieren alcanzar. Esta metodología otorga flexibilidad, velocidad de respuesta y adaptación ante los cambios que presenta el mercado ya que a lo largo de todo el proyecto se trabaja para que los esfuerzos estén totalmente alineados con las necesidades de los clientes.

Es importante destacar que la metodología Agile no se aplica de igual forma en todas las organizaciones. El modo en que se aplique dependerá fuertemente de la cultura organizacional de la empresa en cuestión, es decir, de cómo sean las características de las personas que componen el ecosistema de trabajo.

Para este proyecto en particular, se propone un esquema de trabajo en el que participarán los siguientes roles multidisciplinarios:

- **Product Owner**: es la persona responsable de asegurar que el equipo aporte valor al negocio. Se deberá ocupar de ser el nexo entre los potenciales clientes, los stakeholders y el equipo de trabajo que lleva a cabo las distintas tareas del proyecto. Deberá seleccionar un subconjunto de actividades del *Backlog* inicial para trabajar en cada uno de los diferentes *sprints*. Será quién priorice, gestione y redefina (en caso de ser necesario) las tareas contenidas en el *Backlog*.
- **Scrum Master**: es quien debe asegurar que se cumplan los valores y pilares de la metodología *Scrum* durante todos los eventos del proyecto. Debe organizar la sesión de planificación al inicio de cada sprint, donde hará fluir la conversación para que *Product Owner* y equipo de trabajo puedan definir el objetivo del *sprint* a través de la selección de los elementos del *Product Backlog* que más valor aportan a los clientes. Su función principal es eliminar obstáculos que se presenten en el proyecto.
- **Data Scientist**: Entre sus principales tareas se encontrarán el tratamiento, preparación, análisis exploratorio y estadístico de las bases de datos asociadas al problema de negocio.
- **Data Engineer**: será el responsable de capturar la totalidad de los datos provenientes de todas inmobiliarias y otras fuentes de información. Desarrollar,

gestionar, controlar y mantener los procesos ETL necesarios para garantizar que la ingesta de datos se ejecute de manera correcta. Asegurar la integración de todos los orígenes de datos.

- **Machine Learning Engineer**: encargado de entrenar, desarrollar el modelo de aprendizaje automático y optimizar sus hiperparámetros. Además, será quien deba llevar a producción y operacionalizar el modelo predictivo.
- **Big Data Architect**: este rol deberá disponibilizar un ambiente de software apropiado para que puedan trabajar correctamente el Data Engineer, el Data Scientist y el Machine Learning Engineer.
- **Developers front-end**: deberán desarrollar la interfaz gráfica necesaria para que los usuarios finales puedan visualizar e interactuar con el modelo predictivo (ingresar los datos del inmueble o propiedad que desean valuar).
- **QA (Tester):** Controlar la calidad del producto desarrollado y detectar posibles fallas de funcionamiento.

El *Product Backlog* inicial del proyecto será el siguiente:

- Recolección de datos de distintas inmobiliarias y orígenes de datos.
- Análisis exploratorio de datos (EDA).
- Feature Engineering y Selección de predictores del modelo.
- Entrenamiento de los modelos predictivos y optimización de los hiperparámetros.
- Análisis y comparación de métricas de regresión.
- Desarrollo de la interfaz gráfica para el usuario final.
- Puesta en producción y operacionalización del tasador automático.

Antes de dar comienzo a cada *sprint*, el *Product Owner* deberá elegir un subconjunto de tareas que considere más conveniente trabajar en dicho *sprint*.

Este proyecto constará de 10 *sprints* y cada uno de ellos tendrá una duración de 4 semanas. El *sprint* es continuo, es decir, su duración no debe cambiar mientras está en marcha el desarrollo del tasador automático, se puede interpretar como una medida de ritmo de trabajo constante a lo largo del tiempo. La duración de un *sprint* está determinada por el periodo mínimo en que un equipo de desarrollo puede generar valor a través de un incremento determinado. El resultado del sprint es un producto (o incremental) potencial que puede ser entregado.

Dentro de cada *sprint* y durante todos los días se deberán llevar a cabo las llamadas reuniones *daily*, en las que cada integrante del equipo expondrá qué tareas hizo ayer, cuales hará el día de la fecha y, por último, indicar si existe algún impedimento o bloqueante que dificulte el avance en determinada actividad que tenga asignada o que

esté bajo su responsabilidad. Estas reuniones se realizarán por la mañana, a fin de definir el contexto para el resto del día de trabajo y durarán como máximo 15 minutos. Esto hace que la reunión sea breve y se traten puntos importantes. Al enfocarse en lo que cada miembro del equipo hizo ayer y hará hoy, el equipo ganará una visión general de lo que se ha realizado y aquello que falta por realizar.

Al finalizar cada *sprint* tendrán lugar las reuniones de *review* y tendrán una duración de 4 horas. Estas reuniones son necesarias para revelar al cliente los distintos avances en los que se trabajó durante todo el *sprint*. El *Product Owner* se encarga de organizar e invitar al evento tanto al cliente como a todo el equipo Scrum. Durante las reuniones de *review* se realizarán las demostraciones a los usuarios del tasador automático y se recolectará feedback del cliente para enfocarse en los objetivos del negocio.

Por último, al terminar el *sprint* y luego de la reunión de *review*, se llevará adelante la reunión de retrospectiva cuyo objetivo será que el equipo reflexione en conjunto sobre lo trabajado durante las últimas 4 semanas e identifique posibles oportunidades de mejora para el próximo *sprint*. Estas ceremonias tendrán una duración de 2 horas y son fundamentales para evaluar en qué aspectos el equipo puede mejorar y por otra parte destacar todo aquello en lo que se trabajó correctamente para mantenerlo en el futuro.

Para que el proyecto resulte exitoso será necesario que el tasador automático sea estable en el tiempo y no presente fallas de funcionamiento. Además, es pertinente que las predicciones del modelo implementado sean precisas, lo cual significa que los valores arrojados por el tasador no difieran significativamente de los valores reales de las propiedades. Esto se medirá a través de la métrica RMSE la cual es muy utilizada en modelos de regresión.

Otro aspecto a tener en cuenta es que la interfaz gráfica debe ser amigable, intuitiva y fácil de utilizar para el usuario. Para lograr esto, los desarrolladores estarán muy atentos a las necesidades y deseos que manifieste el cliente en las reuniones de *review*.

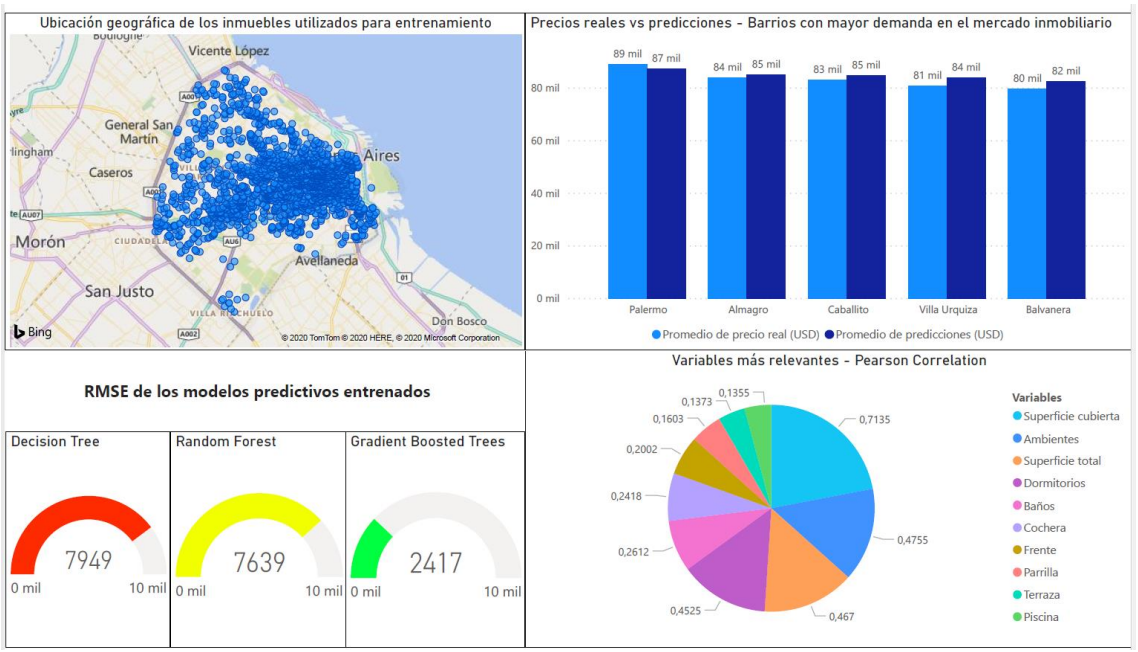
Por otro lado, se definirá un indicador para medir el nivel de uso semanal que tendrá el tasador automático dentro del portal web de inmuebles. Esto permitirá conocer la demanda del producto desarrollado. En la plataforma web se habilitará una encuesta para completar por los usuarios que utilicen el tasador automático. Con ella se recolectará información acerca del nivel de satisfacción del producto y se brindará una pregunta de respuesta abierta para que el usuario sugiera posibles mejoras.

Es fundamental monitorear periódicamente la distribución de los datos nuevos que reciba el modelo predictivo cuando este sea llevado a producción. En la medida que la distribución de los de entrenamiento difiera de los datos "del mundo real", el tasador automático no arrojará predicciones precisas. El equipo deberá prestar especial atención a este aspecto para que el modelo sea confiable.

Los usuarios del proyecto serán aquellas personas interesadas en valuar un inmueble con el objetivo de venderlo o comprarlo. El principal actor para la toma de decisiones será el *Product Owner* y el *Scrum Master* quienes deberán liderar el proyecto desde su inicio hasta su fin.

3.3 Visualizaciones de resultados

A continuación se exponen gráficos realizados mediante la herramienta Power BI:



Conclusiones

Los resultados conseguidos en este trabajo permitieron visualizar ciertos hallazgos que merecen ser destacados. De los tres modelos de aprendizaje automático basados en árboles de decisión, aquel que mejor desempeño obtuvo fue *Gradient Boosted Trees*. Dicho modelo logró reducir en un 47,49 % el RMSE del modelo *Baseline*, en un 47,94% el RMSE del modelo *Decision Tree* y en un 41,41 % el RMSE del modelo *Random Forest*. También obtuvo las cifras más bajas en términos de error absoluto y error relativo.

En lo que respecta al proceso de ajuste de los parámetros de cada modelo, el mismo se llevó a cabo por medio del operador *Optimize Parameters* de Rapid Miner. Por otro lado, a fin de evitar el sobreajuste de los modelos implementados y garantizar su robustez en cuanto a capacidad de generalización, se utilizó la técnica de validación cruzada operando con 10 pliegues.

Cabe señalar que todos los modelos de regresión presentaron mejores resultados al ser configurados para trabajar con una gran cantidad de árboles y escasa profundidad de los mismos. Otro punto a destacar es que no todos los modelos funcionaron de igual forma con las mismas variables explicativas. Además, la creación de nuevas variables mejoró en mayor medida el desempeño de los modelos que la optimización de los hiperparámetros. La aplicación de la técnica de Componentes Principales fue de gran utilidad para reducir la dimensionalidad de la base de datos original, permitiendo transformar 10 variables métricas correlacionadas en 2 nuevas variables incorrelacionadas.

Como próximos pasos en futuros trabajos, sería interesante utilizar la técnica de ensamble de modelos, el cual busca incrementar la exactitud a partir de la combinación de las predicciones de múltiples modelos predictivos. Por otro lado, se podrían predecir precios de inmuebles de distintas características y zonas geográficas.

Para concluir, la ventaja notoria que brinda un tasador de propiedades automático como el implementado en este trabajo, es agilizar todas las tareas involucradas en el proceso convencional de fijación de precios de un inmueble. Esta herramienta logra reducir los costos totales de la gestión integral de dicho proceso ya que se elimina la figura del profesional tasador.

Referencias Bibliográficas

Aldas, J., & Uriel, E. (2017). *Análisis multivariante aplicador con R.* Alfacentauro. Madrid, España.

Atlassian Agile Couch *¿Qué es scrum?* https://www.atlassian.com/es/agile/scrum

Big data inmobiliario: *¿Cómo y por qué usarlo en una inmobiliaria?* https://inmogesco.com/blog/big-data-inmobiliario/

Cuadras, C.M. () *Nuevos Métodos de Análisis Multivariante*. CMC Editions. Barcelona, España.

Eibe, F. & Witten. I. (2016) "*Data Mining: Practical Machine Learning Tools and Techniques".*

Gujarati, Damodar N., & Porter Dawn C. (2009). *Econometría*. Mc Graw Hill.

Fawcett, T. & Provost, F. (2013). Data Science and its relationship to Big Data and data-driven decision making. Data Science and Big Data.

Fundamentos y conceptos básicos de Scrum. https://www.scrum.org/resources/blog/fundamentos-y-conceptos-basicos-de-scrum

Hao Peng, Jianxin Li, Member, IEEE, Zheng Wang, Renyu Yang, Member, IEEE, Mingzhe Liu, Mingming Zhang, Philip S. Yu, Fellow, IEEE, and Lifang He, Member, IEEE. "*Lifelong Property Price Prediction: A Case Study for the Toronto Real Estate Market"*. [Submitted on 12 Aug 2020] Disponible en inglés en https://arxiv.org/abs/2008.05880

Johnson, Nicholas L. (2020) ¿What are Network Effects? https://www.applicoinc.com/blog/network-effects/

La tecnología en el sector inmobiliario ya es un must para todos (2018) https://nuovit-homes.es/tecnologia-en-el-sector-inmobiliario/

Ley N°25326. Protección de datos personales (2000).

Marco De Nadai and Bruno "*LepriThe economic value of neighborhoods: Predicting real estate prices from the urban environment."* (Tue, 7 Aug 2018) Disponible en inglés en https://arxiv.org/abs/1808.02547

Mayer, T., & Cyment , A. (2014). *Por Un Scrum Popular: Notas para una Revolucion Agil*. UK.

Properati (2020). Política de privacidad. https://www.properati.com.ar/

Proyectos Ágiles ¿Qué es Scrum? https://proyectosagiles.org/que-es-scrum/

© RapidMiner GmbH . (2020). rapidminer.com. Obtenido de https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html

Rahman, K. S., & Thelen, K. (2019). The rise of the platform business model and the transformation of twenty-first-century capitalism. Politics & Society, 47(2), 177-204.

Shashi Bhushan Jha, Radu F. Babiceanu, Vijay Pandey, Rajesh Kumar Jha, "*Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study"* (Wed, 17 Jun 2020). Disponible en ingles en https://arxiv.org/abs/2006.10092

Schmarzo, B. (2013). *Big Data: Understanding how data powers big business.* Indianapolis: John Wiley & Sons.

Vijay Kotu, Bala Deshpande, PhD. (2014) *"Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner"*

Anexo

1. Descripción de la base de datos

Nombre del atributo	Tipo de dato	Descripción
id	Número entero (integer)	Registro de la publicación del inmueble
Pub_inicio	Fecha	Fecha de inicio de la publicación
Pub_fin	Fecha	Fecha de fin de la publicación
Latitud	Número decimal (float)	Latitud de la ubicación del inmueble
Longitud	Número decimal (float)	Longitud de la ubicación del inmueble
Barrio	Texto (string)	Barrio del inmueble
Sup_total	Número entero (integer)	Superficie total
Sup_cubierta	Número entero (integer)	Superficie cubierta

Título	Texto (string)	Título de la publicación
Descripción	Texto (string)	Descripción de la publicación
Precio	Número entero (integer)	Precio del inmueble
Luminoso	Número entero (integer)	Indica presencia o ausencia de luminosidad
Balcón	Número entero (integer)	Indica presencia o ausencia de balcón
Patio	Número entero (integer)	Indica presencia o ausencia de patio
Terraza	Número entero (integer)	Indica presencia o ausencia de terraza
Parquet	Número entero (integer)	Indica presencia o ausencia de parquet
Frente	Número entero (integer)	Indica presencia o ausencia de frente
Contrafrente	Número entero (integer)	Indica presencia o ausencia de contrafrente
Cochera	Número entero (integer)	Indica presencia o ausencia de cochera
Parrilla	Número entero (integer)	Indica presencia o ausencia de parrilla
Piscina	Número entero (integer)	Indica presencia o ausencia de piscina
Baño completo	Número entero (integer)	Indica presencia o ausencia de piscina
Apto profesional	Número entero (integer)	Indica presencia o ausencia de piscina
Dormitorio con placard	Número entero (integer)	Indica presencia o ausencia de piscina
Dist_patio_bulrrich	float	Distancia a Patio Bulrich
Dist_abasto_shopping	float	Distancia a Abasto Shopping
Dis_distrito_arcos	float	Distancia a Distrito Arcos
Dist_obelisco	float	Distancia a Obelisco
Dist_parque_centenario	float	Distancia a Parque Centenario
Dist_plaza_italia	float	Distancia a Plaza Italia
Min_bocas	float	Distancia a la boca de subte más cercana
Min_univ	float	Distancia a la universidad más próxima

2. Modelos utilizados y parámetros

Decision Tree

Parámetro	Valor
Maximal depth	9
Minimal leaf size	51
Minimal size for Split	151

Number of prepruning alternatives	51
Confidence	0.1
Minimal gain	0.01

Random Forest

Parámetro	Valor
Number of trees	1208
Maximal depth	23
Confidence	0.1
Minimal gain	0.01
Minimal leaf size	2
Minimal size for Split	4
Number of prepruning alternatives	70

Gradient Boosted Trees

Parámetro	Valor
Number of trees	713
Maximal depth	4
Min rows	10
Min Split improvement	1.0E-5
Number of bins	20
Learning rate	0.01
Sample rate	1

Apéndice

1. Código de programación en R

Script Minería de Texto

```
install.packages("tm")

library("tm")

#Creación del corpus

corpus=Corpus(VectorSource(Base))
```

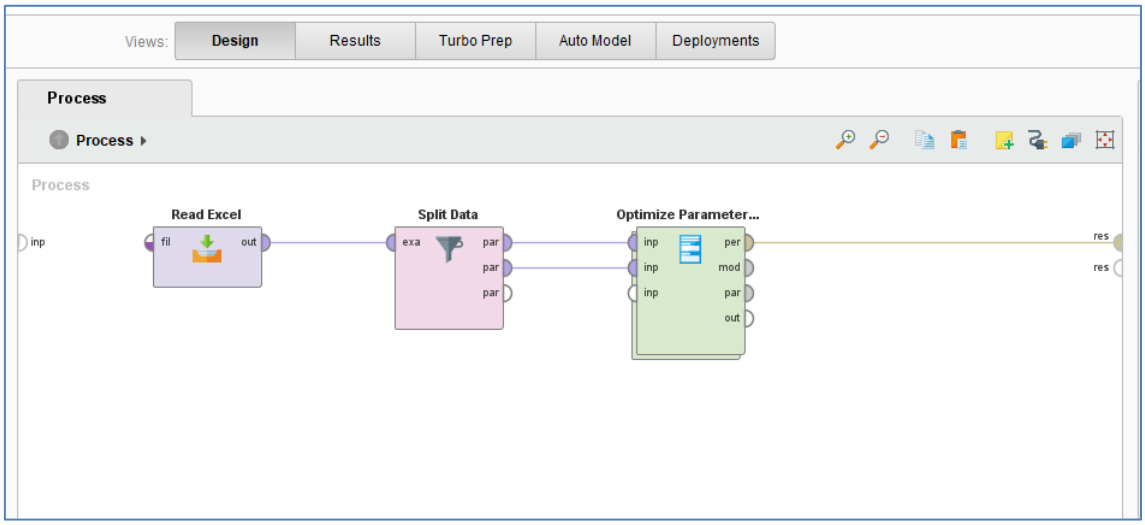
```
inspect(corpus)

#Transformación en letras minúsculas.

nube=tm_map(corpus,tolower)

#Eliminación de los espacios en blanco.

nube=tm_map(nube,stripWhitespace)

#Eliminación de las puntuaciones.

nube=tm_map(nube,removePunctuation)

#Eliminación de números presentes en el texto.

nube=tm_map(nube,removeNumbers)

#Inspección de las stopwords cargadas por defecto en la librería.

stopwords("spanish")

#Eliminación de las 308 stopwords predefinidas dentro de la librería.

nube=tm_map(nube,removeWords,stopwords("spanish"))

#Creación de una matriz de términos.

tdm=TermDocumentMatrix(nube)


#Inspección de palabras más frecuentes.

findFreqTerms(tdm,lowfreq=400)

m=as.matrix(tdm)

v=sort(rowSums(m),decreasing=TRUE)

df=data.frame(word=names(v),freq=v)

#Construcción de diagramas de barra con las palabras más frecuentes

library(ggplot2)

 ggplot(df[1:43,],aes(x=reorder(word,freq), y=freq)) +

 geom_bar(stat = "identity", color = "black", fill = "#87CEFA") +

 geom_text(aes(hjust = 1, label = freq),size=3) +

 coord_flip()+
```

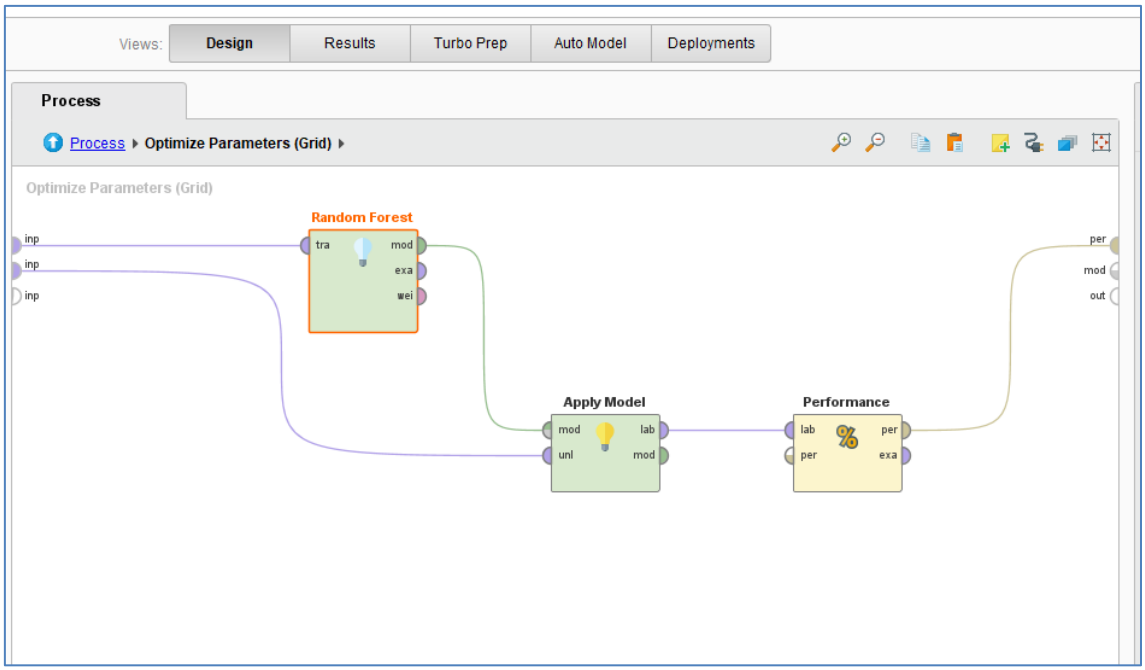
```
 labs(title = "Descripción del inmueble - Palabras más frecuentes",  x = "Palabras", y = "Cantidad")+theme_minimal()

 #Confección de nubes de palabras.

 install.packages("wordcloud2")

 library(wordcloud2)

 wordcloud2(df,size=0.5,rotateRatio=0.7,color='random-dark')

#Confección de histogramas

hist(base$precio,main="Precio",xlab="Precio",ylab="Frecuencia",col="#87CEFA")

hist(base$sup_total,xlab="Superficie total",ylab="Frecuencia",col="#87CEFA")

hist(base$sup_cubiert,xlab="Superficie cubierta",ylab="Frecuencia",col="#87CEFA")

#Ngramas

library(tidytext)

library("tidyr")

library(stringr)

library(dplyr)

unnest_tokens(tbl=data,output=Resultado, input=data_minusc,token = "ngrams", n = 2) %>% count(Resultado, sort = TRUE)

#Matriz de correlación

mat_cor=round(cor(base),2)

melted_cormat <- melt(mat_cor)

library(ggplot2)

ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +

 geom_tile() +

 geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +

 theme(

   axis.title.x = element_blank(),

   axis.title.y = element_blank())+

 theme(axis.text.x = element_text(angle = 90, vjust = 1,
```

size = 12, hjust = 1))

2. Optimización de Parámetros

En las siguientes visualizaciones se puede apreciar cómo se procedió para optimizar los parámetros del modelo *Random Forest* en particular. Se actuó de igual forma para el resto de los modelos, utilizando el operador *Optimize Parameters* de **Rapid Miner**.





3. Validación Cruzada

En esta sección, se puede observar cómo se llevó a cabo la conexión de cada operador para implementar la técnica de Validación Cruzada con 10 pliegues, utilizando el operador *Cross Validation* de **Rapid Miner**. En este caso se aprecia que el modelo entrenado es *Gradient Boosted Trees*.