INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE
Departamento de Matemáticas y Física

Desarrollo tecnológico y generación de riqueza sustentable

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)
Modelos de Predicción en Empresas y Gobierno Mediante Aprendizaje Estadístico

PAP 4J07 - Modelos de Predicción en Empresas y Gobierno Mediante Aprendizaje Estadístico

“Sistema de recomendación de propiedades usando KNN”

PRESENTAN
Ing. en Sistemas Computacionales. José Ignacio González Cárdenas
Ing. en Nanotecnología. Alejandro González Meléndrez

Profesor PAP: Pablo Dávalos de la Peña.
Tlaquepaque, Jalisco, diciembre de 2017

ÍNDICE

Contenido

REPORTE PAP

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son una modalidad educativa del ITESO en la que el estudiante aplica sus saberes y competencias socio-profesionales para el desarrollo de un proyecto que plantea soluciones a problemas de entornos reales. Su espíritu está dirigido para que el estudiante ejerza su profesión mediante una perspectiva ética y socialmente responsable.

A través de las actividades realizadas en el PAP, se acreditan el servicio social y la opción terminal. Así, en este reporte se documentan las actividades que tuvieron lugar durante el desarrollo del proyecto, sus incidencias en el entorno, y las reflexiones y aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

En este reporte se presentan los resultados obtenidos en la realización del sistema de recomendaciones de propiedades para el sitio web de la empresa Casas y Terrenos. Para la realización del sistema se usaron los datos proporcionados por la empresa acerca de las propiedades registradas, las búsquedas en el sitio web y las conversiones hechas por usuarios. Los datos fueron pre-procesados para su uso y después se hizo un cruce de información con el código de área geoestadística básica (AGEB) del Instituto Nacional de Estadística y Geografía (INEGI) para añadir variables sociodemográficas las cuales fueron reducidas por análisis de componentes principales (PCA, por sus siglas en inglés) y de esta manera poder generar el modelo clasificador usando K-vecinos más cercanos (KNN, por sus siglas en inglés).

1. Introducción

1.1. Objetivos

Este proyecto pretende desarrollar un algoritmo para la recomendación personalizada de las propiedades vistas por un usuario en el sitio web de la empresa Casas y Terrenos.

1.2. Justificación

El desarrollo de este proyecto permitirá a los usuarios de Casas y Terrenos obtener resultados en la página web que se adapten de acuerdo a sus gustos y necesidades. Esto permitirá que los usuarios encuentren una propiedad, ya sea para comprar o rentar, de una manera más fácil.

1.3 Antecedentes

Para la realización del algoritmo la empresa Casas Y Terrenos contactó en ocasiones anteriores al profesor PAP, Pablo Dávalos, pero no pudo llevarse a cabo ya que las bases de datos no se encontraban en condiciones para trabajar con ellas. Debido a esto, se trabajó en la limpieza de dichas bases de datos y en la agregación de información relevante externa obtenida del Censo de Población y Vivienda 2010 del INEGI referente a la información sociodemográfica de las viviendas en un nivel de agregación geográfica de áreas geoestadísticas básicas (AGEB).

1.4. Contexto

En la actualidad varios sitios web, como Netflix, Amazon o Ebay, ofrecen recomendaciones personalizadas a sus usuarios, sin embargo, estos métodos de inteligencia artificial todavía no son utilizados a gran escala por empresas en México. La empresa Casas y Terreno desea ofrecer recomendaciones a sus clientes para tener una ventaja competitiva.

2. Desarrollo

2.1. Sustento teórico y metodológico

Según Segaran, un algoritmo de filtro colaborativo trabaja buscando un grupo grande de personas y buscando un conjunto pequeño con gustos similares a los tuyos. Busca otras cosas que los otros usuarios gustan y combinan los gustos para crear una lista ordenada de recomendaciones.

Una de las desventajas es que el filtro colaborativo basado en usuarios requiere un traslape significativo entre los gustos de los usuarios, para decidir cuáles personas son similares.

Un sistema recomendador basado en contenido, según Lops et. all., trata de recomendar ítems similares a los que el usuario ha gustado en el pasado. El sistema recomendador analiza el conjunto de descripciones de ítems previamente gustados por un usuario, construye un modelo o perfil de los intereses del usuario. El perfil es una representación estructurada de los intereses de los usuarios, adaptada para recomendar nuevos ítems.

James et. all., explica en qué consiste un clasificador *K-nearest neighbors* (KNN). Dado un entero positivo K y una observación x, el clasificador primero identifica los K puntos en el conjunto de entrenamiento que son más cercanos a x, representados por N0. Después estima la probabilidad condicional de que pertenezca a la clase j de acuerdo a la fracción de puntos en N0 cuyos valores de respuesta sean igual a j.

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

Después KNN aplica la regla de Bayes y clasifica la observación de pruebas x a la clase con la mayor probabilidad.

En el recomendador realizado en este proyecto utilizamos KNN para obtener las casas con las propiedades más cercanas, las cuales serán recomendadas.

2.2. Planeación y seguimiento del proyecto

- Descripción del proyecto

Para la realización del proyecto se inició por la obtención de las bases de datos de las plataformas, la primera relacional, en la tecnología MariaDB; la segunda NoSQL, usando la tecnología mongoDB. Fue necesario el acceso al servidor donde se encontraban almacenados los datos por medio de llaves SSH privadas y la creación de usuarios.

La limpieza, pre-procesamiento, análisis exploratorio y aplicación de modelo clasificador fue realizado en un shell interactivo IPython que añade funcionalidades extra al modo interactivo incluido con Python, como resaltado de líneas y errores mediante colores, una sintaxis adicional para el shell, autocompletado mediante tabulador de variables, módulos y atributos; entre otras funcionalidades. Se utilizó un kernel de Ipython en Jupyter Notebook para trabajar realizar las actividades.

- Plan de trabajo

La descripción de actividades realizadas a lo largo del semestre de Otoño 2017 se detallan a continuación:

- Semana 1, 2 y 3 (14 de Agosto – 2 de Septiembre):

En el transcurso de las primeras tres semanas del semestre se comenzó trabajando en la práctica de manejo de bases de datos tales como hacer análisis exploratorio (identificar variables correlacionadas, agrupación en subgrupos, entre otras), manejo de paqueterías en Ipython para la lectura de tablas (paquetería pandas), así como generar gráficos para mejorar el entendimiento de los datos de manera simple y rápida. Las bases de datos

públicas utilizadas fueron obtenidas de la página web www.kaggle.com. Tiempo estimado: 48 horas.

- Semana 4, 5 y 6 (4 de Septiembre – 23 de Septiembre):

Una vez que se tuvieron las bases de datos con el análisis exploratorio se procedió a aplicar un modelo predictivo de regresión o clasificación, según fuera el caso. Para poder lograr esto primero se estuvo trabajando con la limpieza de las bases de datos (identificar y tratar valores faltantes, eliminar variables innecesarias). Después de haber limpiado los datos se pudo continuar con la aplicación del modelo predictivo. Entre los principales modelos utilizados se encuentran Random forest, Gradient boosting, Lasso, Support vector machines y redes neuronales multicapa. Tiempo estimado: 48 horas.

- Semana 7 y 8 (25 de Septiembre – 7 de Octubre):

En la semana 7 se hizo una visita al Instituto de Información Estadística y Geográfica del Estado de Jalisco (IIEG) para platicar acerca de los objetivos que tienen y obtener acceso a los datos necesarios para trabajar. Los datos que se nos compartieron fueron cifras de empleo, exportación e importación y sobre inversión extranjera del estado de Jalisco. En lo que restó de la semana 7 y la semana 8 se estuvo en el análisis exploratorio de los datos para conocer la metodología a seguir para la aplicación de modelos predictivos de machine learning. Tiempo estimado: 32 horas.

- Semana 9 y 10 (9 de Octubre – 21 de Octubre):

Para la semana 9 se obtuvo una cita con la empresa Casas y Terrenos para que, al igual que con el IIEG, nos hablaran acerca de los objetivos les gustaría se lograran con los datos para el fin de semestre. Se obtuvo acceso a los datos de las propiedades publicadas en la página web www.casasyterrenos.com, así como los datos de conversiones hechas por usuarios desde el 15 de Febrero de 2016 hasta el 15 de Julio de 2017.

Durante la semana 10 se trabajó en el análisis exploratorio de dichos datos para saber el procedimiento que se tendría que llevar a cabo y los problemas que se tendrían que resolver y se comenzó por la limpieza de los datos, así como la selección de los campos importantes que describen a las propiedades. Tiempo estimado: 32 horas.

- Semana 11 (23 de Octubre – 28 de Octubre):

En el transcurso de la semana 11 se continuó con el pre-procesamiento de las bases de datos para dejarlas aptas para la aplicación del modelo clasificador. Tiempo estimado: 16 horas.

- Semana 12, 13 y 14 (30 de Octubre – 18 de Noviembre):

En las semanas 12, 13 y 14 se trabajó en la aplicación del modelo clasificador KNN para las propiedades de las cuales se contaba con código AGEB así como los campos que describen las características de las mismas. Tiempo estimado: 48 horas.

- Semana 15 (20 de Noviembre – 25 de Noviembre):

En la última semana se presentó el resultado final del clasificador KNN a la empresa Casas y Terrenos en sus instalaciones. Esto con el fin de recibir retroalimentación de su parte así como hacerles llegar las recomendaciones con las cuales se puede mejorar el modelo recomendador o construir otro con distinto enfoque. Tiempo estimado: 16 horas.

El resumen de las actividades realizadas durante el semestre se muestra en la Tabla 1.

Tabla 1. Cronograma de actividades de la semana 1 a la 15.

	Semana														
Actividad	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Análisis exploratorio de bases de datos públicas															
Construcción de modelos de predicción															
Exploración de datos IIEG															
Exploración de datos Casas y Terrenos															
Pre-procesamiento de bases de datos															
Análisis exploratorio de bases de datos															
Construcción de modelo de predicción															
Presentación de avances Casas y Terrenos															

- Desarrollo de propuesta de mejora

Los desarrollos de las actividades descritas en la sección de Plan de trabajo ya se encuentran descritas en la misma, así como en la sección de Resultados.

3. Resultados del trabajo profesional

1. Obtención de los datos

La información de la empresa Casas y Terrenos se encuentra contenida en dos tipos de bases de datos. La primera relacional, en la tecnología MariaDB; la segunda NoSQL, usando la tecnología mongoDB. En la base de datos relacional se encuentran las propiedades a analizar y más bases de datos que ayudan a las reglas de negocio. En mongoDB se encuentran datos sobre las búsquedas hechas por los usuarios y las propiedades convertidas. Hacer una conversión significa pedir los datos acerca de una propiedad. En la Figura 1 se muestra el esquema de la base de datos de propiedades.

#	Name
1	prp_id_propiedad
2	prp_id_inmobiliaria
3	prp_id_usuariozp
4	prp_id_promotor
5	prp_proposito
6	prp_fechah_alta
7	prp_fechah_status
8	prp_status
9	prp_vigencia_desde
10	prp_vigencia_hasta
11	prp_recomendada
12	prp_destacada
13	prp_cve_catastral
14	prp_comision
15	prp_horario
16	prp_comentario_status
17	prp_publicar_domicilio
18	prp_topografia
19	prp_observaciones
20	prp_observaciones_inmob
21	prp_id_ubicacion
22	prp_activa
23	prp_fecha_actualizacion
24	prp_id_paquete_activo
25	prp_sello_confianza
26	prp_sello_destacadas
27	prp_nueva
28	prp_sello_pro
29	prp_url
30	prp_habitaciones
31	prp_banos
32	prp_terreno
33	prp_construccion
34	prp_estacionamiento
35	prp_publicada
36	prp_tipo
37	prp_id_estado
38	prp_id_municipio
39	prp_id_colonia
40	prp_estado
41	prp_municipio
42	prp_colonia
43	prp_costov
44	prp_costor
45	prp_costot
46	prp_costom2
47	prp_id_moneda
48	prp_latitud
49	prp_longitud
50	prp_frente
51	prp_fondo
52	prp_descripcion
53	prp_calle
54	prp_numero_ext
55	prp_numero_int
56	prp_referencia
57	prp_edad
58	prp_servicio
59	prp_niveles
60	prp_medios_bannos
61	prp_crm

Figura 1. Esquema de la base de datos de propiedades

Existen muchos problemas de inconsistencia de datos y valores faltantes en las bases de datos relacionales. Gracias a la aportación de un alumno PAP, Pedro Martínez, pasado pudimos obtener una base de datos 'limpia', con pocos valores faltantes, y con una columna extra, el AGEB (Área Geoestadística Básica). La Figura 2 muestra la base de datos de propiedades ya sin valores faltantes y con la columna agregada de AGEB.

	ID	cos_moneda	cos_costov	prp_propos	prp_id_ubi	prp_habita	prp_banos	prp_terren	prp_constr	prp_estaci	prp_tipo	prp_latitu	prp_longit
0	57	1	650000.0	1	8763	3	2.0	73.44	80.00	1	18	20.762697	-103.427492
1	1743	1	5400000.0	1	9158	3	3.5	390.00	402.00	2	18	20.694279	-103.438895
2	5716	1	1078000.0	1	9769	2	1.0	167.78	95.67	2	18	20.739950	-103.359667
3	7144	1	1550000.0	1	33410	3	2.5	170.00	200.00	2	18	20.545405	-103.472929
4	7145	1	1450000.0	1	33719	3	2.5	181.00	200.00	2	18	20.545044	-103.473186

POSTALCODE	MUN_NAME	SETT_NAME	SETT_TYPE	prp_m2	prom_cos_c	prom_prp_c	prom_prp_b	prom_prp_e	prom_prp_m
45138	ZAPOPAN	365	2	8125.000000	1460000.0	129.0	3.0	2.0	11317.829457
45110	ZAPOPAN	663	3	13432.835821	5400000.0	402.0	3.5	2.0	13432.835821
45188	ZAPOPAN	315	2	11267.900073	1780000.0	180.0	2.0	2.0	10581.649832
45645	TLAJOMULCO_DE_ZUNIGA	680	7	7750.000000	2120000.0	195.0	2.5	2.0	12294.867318
45645	TLAJOMULCO_DE_ZUNIGA	680	7	7250.000000	2120000.0	195.0	2.5	2.0	12294.867318

prom_prp_m	prom_cos_p	prom_cos_m	prom_prp_p	CVE_AGEB
11317.829457	1460000.0	3500000.0	1460000.0	6508
13432.835821	3019500.0	2475000.0	5400000.0	3984
10581.649832	1780000.0	3430000.0	1780000.0	0890
12294.867318	2570000.0	2190000.0	2120000.0	0477
12294.867318	2570000.0	2190000.0	2120000.0	0477

Figura 2. Ejemplo de propiedades en la base de datos relacional ya limpia.

Utilizando el código del AGEB se hizo un cruce de información con el censo de población y vivienda 2010, en el que se encuentran más de 192 variables sociales, económicas y demográficas.

La base de datos NoSQL también tenía algunos valores faltantes, como los datos de promotor o algunas características de las propiedades. En la Figura 3 se puede apreciar el esquema de la base de datos de conversiones.

Key	Value	Type
(1) {_id : 57630c7ea573b9d03e114371}	{ 12 fields }	Document
_id	57630c7ea573b9d03e114371	ObjectId
nombre	Pedro	String
telefono	122345234	String
email	prodriguez@quid.com.mx	String
php_session	bhqmdp4qmgrj34067adtolcna4	String
ip	187.162.202.172	String
referrer	http://www.cyt.mx/detalle-terreno/venta-vallarta+la+patria-zapopan-jalisco-450	String
user_agent	Mozilla/5.0 (Macintosh; Intel Mac OS X 10.10; rv:47.0) Gecko/20100101 Firefox/47.	String
prp_nueva	Propiedad	String
conversion_tipo	SOLICITAR CITA PROPIEDAD	String
propiedad	{ 20 fields }	Object
time	2016-06-16T20:30:53.761Z	Date

Figura 3. Ejemplo de la base de datos de conversiones.

Para poder visualizar los campos en los que había valores faltantes se realizó un histograma. La Figura 4 muestra el histograma con la frecuencia de valores faltantes por cada campo.

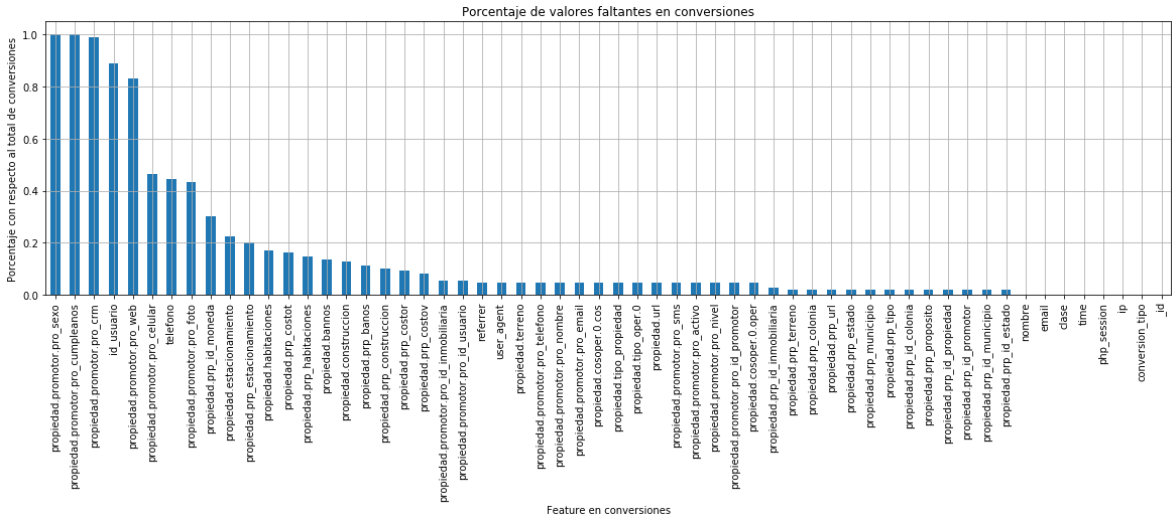

Figura 4. Porcentaje de valores faltantes en base de datos de conversiones.

2. Análisis exploratorio de la base de datos de búsquedas y conversiones.

Búsquedas

De un total de 3,321,288 búsquedas en mapa solo 368,434 son de sesiones individuales, lo cual representa aproximadamente un 11% del total de búsquedas. En la Figura 5 se muestra la frecuencia de búsquedas que realizan los usuarios por sesión individual; se puede notar como la mayoría de los usuarios realizan entre 1 y 5 búsquedas por sesión.

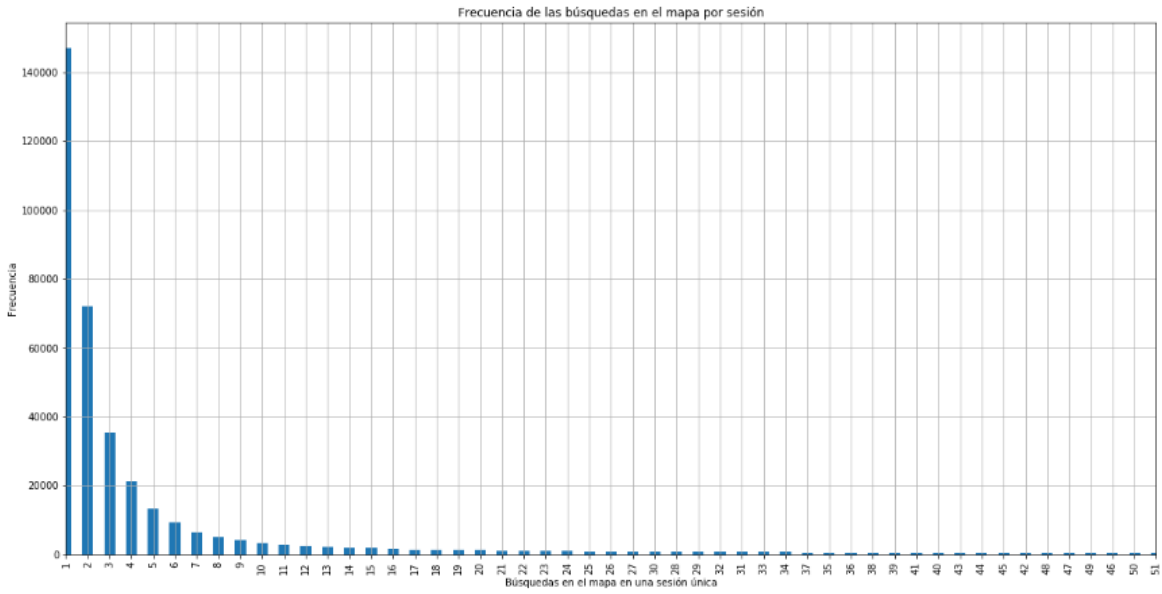

Figura 5. Frecuencia de búsquedas en el mapa por sesión única.

La mayor parte del mercado de la página web se encuentra en el estado de Jalisco, aproximadamente el 95% de las búsquedas pertenecen al estado de Jalisco. Esto se observa en la Figura 6, además de que hay otros estados en donde los usuarios también buscan propiedades.

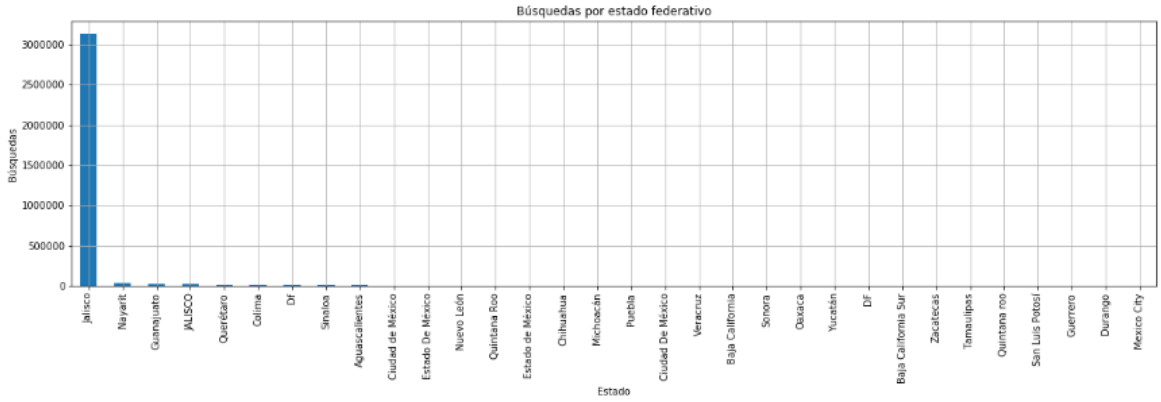

Figura 6. Búsquedas por estado federativo.

Las búsquedas del estado de Jalisco muestran que las principales colonias al hacer la primera búsqueda se remiten a los municipios de la zona metropolitana de Guadalajara (ZMG). En la Figura 7 se muestran las colonias con mayor frecuencia de búsquedas.

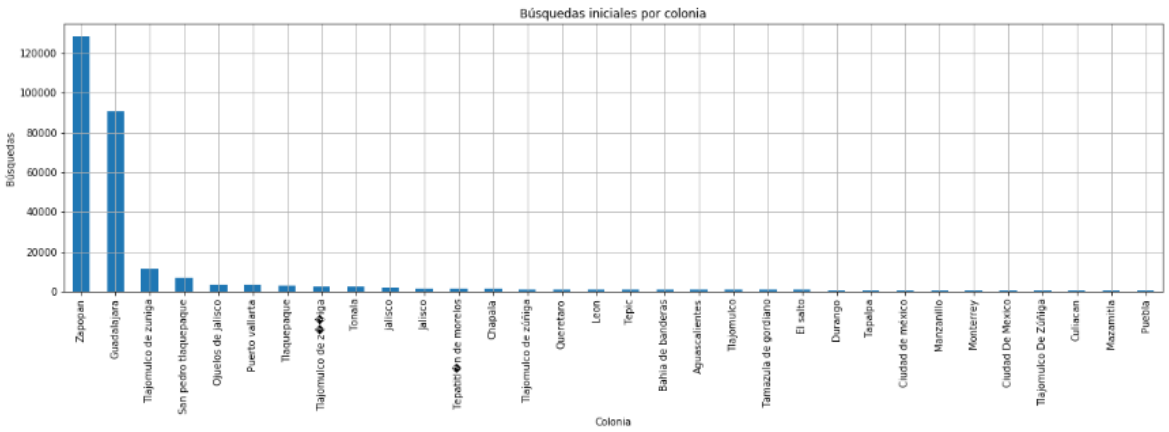

Figura 7. Municipios más populares en las búsquedas.

Sin embargo, una vez que el usuario mueve de posición el mapa se genera una nueva búsqueda. La Figura 8 muestra que las colonias más populares son ciudad granja, la estancia, chapalita, centro, bugambilias y providencia.

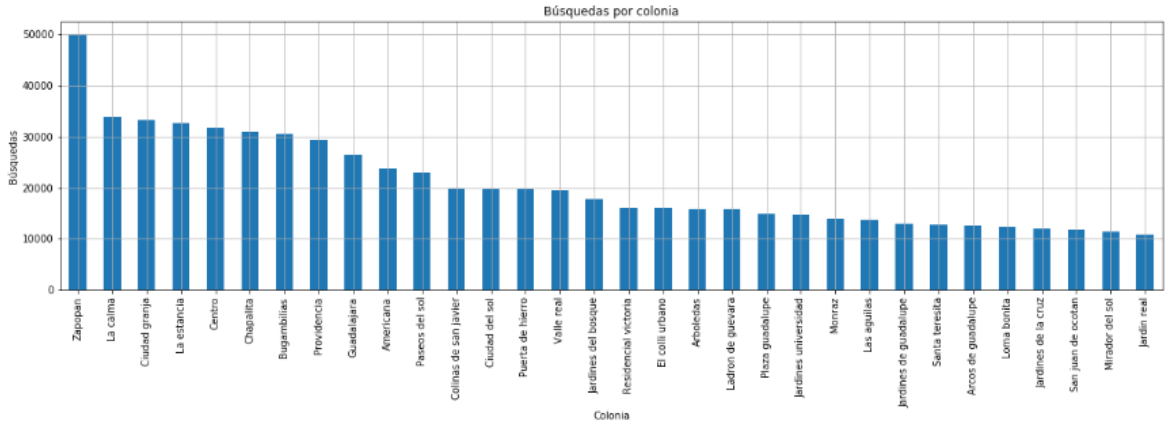

Figura 8. Colonias más populares.

De igual manera, los municipios más populares son Zapopan y Guadalajara con mayor frecuencia de resultados contenidos en sus límites. En la Figura 9 se muestran los resultados ordenados por mayor frecuencia.

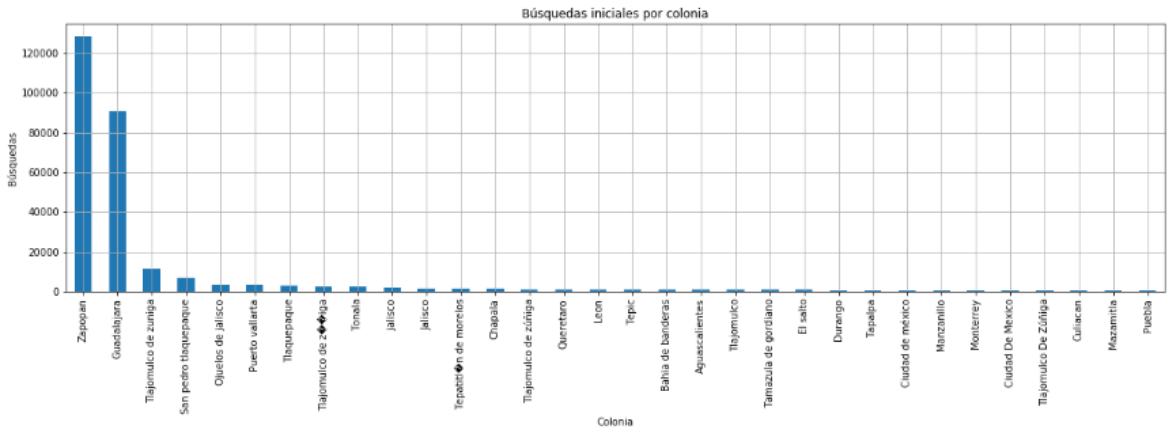

Figura 9. Municipios más populares.

Con las coordenadas de los resultados de búsquedas se graficaron en un mapa de la República Mexicana para poder observar la densidad de búsquedas. En la Figura 10 se puede observar que efectivamente la mayoría de búsquedas son realizadas en Jalisco.

Figura 10. Búsquedas en la República Mexicana, cada punto azul es una búsqueda.

Para la ZMG se realizó el mismo procedimiento el cual se muestra en la Figura 11. Se puede notar que las búsquedas tienen mayor agrupamiento en la parte oeste de la ciudad.

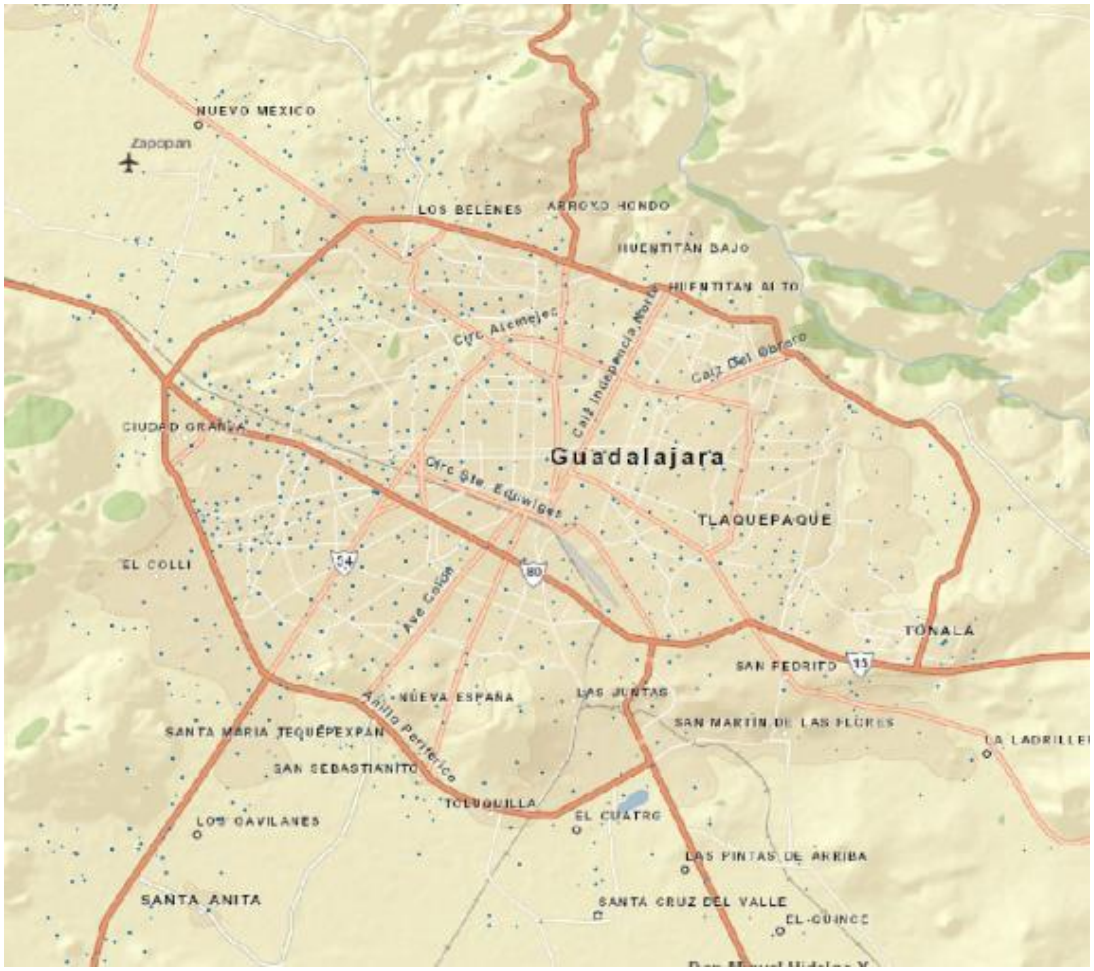

Figura 11. Búsquedas en la ZMG, cada punto azul es una búsqueda.

Uso de filtros.

Al hacer una búsqueda hay filtros para segmentar propiedades, hay filtros de precio, habitaciones, baños, estacionamiento, etc. Cerca del 95% de las personas que utilizan las búsquedas no utilizan los filtros ya que los campos en la base de datos estaban con los valores por default.

Conversiones

Al analizar los datos de conversiones pudimos notar que al ~95% de los usuarios solo les interesa ver los datos de la propiedad y no enviarla o solicitar la ficha de desarrollo. En la Figura 12 se observa que el porcentaje de usuarios que solicitan ver los datos de propiedad corresponde a casi el 90% del total.

	Tipo de conversión	%
0	VER DATOS PROPIEDAD	89.497891
1	VER DATOS DESARROLLO FICHA	2.533121
2	SOLICITAR CITA PROPIEDAD	2.360143
3	VER DATOS DESARROLLO PROTOTIPO	1.728684
4	CALCULADORA PROPIEDAD	1.641649
5	FAVORITO PROPIEDAD	1.104872
6	ENVIAR AMIGO PROPIEDAD	0.779310
7	SOLICITAR CITA DESARROLLO FICHA	0.197012
8	SOLICITAR CITA DESARROLLO PROTOTIPO	0.088128
9	FAVORITO DESARROLLO FICHA	0.025491
10	ENVIAR AMIGO DESARROLLO	0.016752
11	FAVORITO DESARROLLO PROTOTIPO	0.013838
12	ENVIAR AMIGO DESARROLLO PROTOTIPO	0.012746
13	VER DATOS DESARROLLO	0.000364

Figura 12. Tipos de conversiones.

Continuando con los datos de conversiones se pudo notar que aproximadamente el 61.15% de los visitantes solo hace una conversión por sesión única. En la Figura 13 se observa el número de conversiones contra la frecuencia de cada una. Los valores atípicos corresponden a las sesiones de inmobiliarias que pueden llegar a realizar más de 200 conversiones.

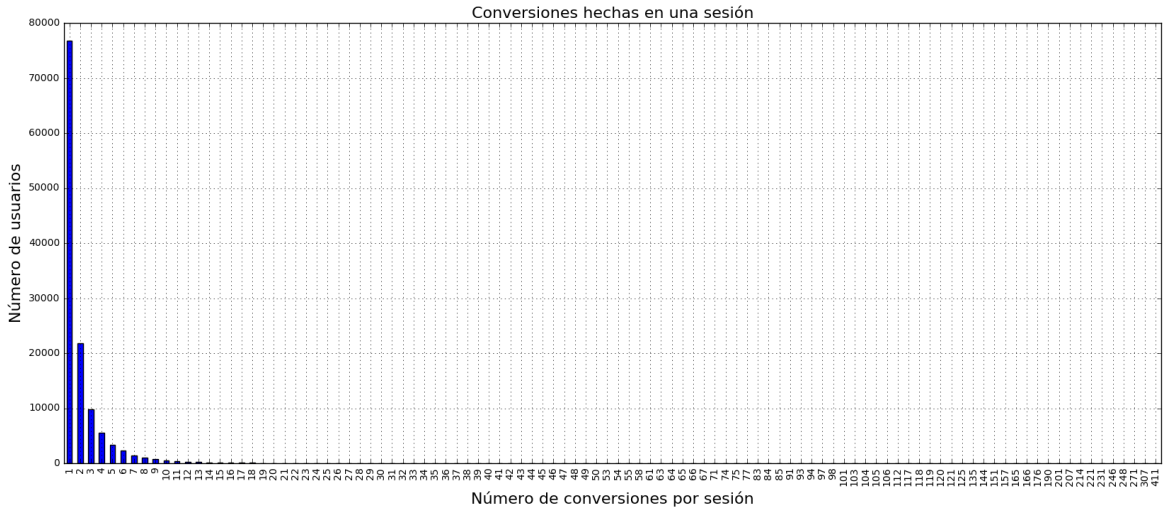

Figura 13. Número de conversiones por sesión.

3. Integración de datos sociodemográficos del INEGI.

El censo de población y vivienda 2010 expone una selección de indicadores sobre las características sociodemográficas de la población y las viviendas de las localidades urbanas del país, desagregados hasta un nivel de área geoestadística básica y manzana. Se realizó un cruce de información entre estos indicativos y la base de datos de propiedades.

De los 192 indicadores encontrados en el censo se seleccionaron manualmente 42 indicadores que se consideraron más relevantes.

4. Modelo predictivo.

Una vez finalizado el análisis exploratorio y la limpieza de los datos se procedió a realizar el modelo clasificador. Para la selección de las características de las propiedades se usó un criterio arbitrario para elegir las que se consideraron más importantes en la elección de una propiedad por un usuario. La Figura 14 muestra las características finales que se usaron.

	ID	cos_costov	prp_habita	prp_banos	prp_terren	prp_constr	prp_estaci	prp_latitu	prp_longit	MUN_NAME	AGEB
0	57	650000.0	3	2.0	73.44	80.00	1	20.762697	-103.427492	ZAPOPAN	6508
1	1743	5400000.0	3	3.5	390.00	402.00	2	20.694279	-103.438895	ZAPOPAN	3984
2	5716	1078000.0	2	1.0	167.78	95.67	2	20.739950	-103.359667	ZAPOPAN	0890
3	7144	1550000.0	3	2.5	170.00	200.00	2	20.545405	-103.472929	TLAJOMULCO_DE_ZUNIGA	0477
4	7145	1450000.0	3	2.5	181.00	200.00	2	20.545044	-103.473186	TLAJOMULCO_DE_ZUNIGA	0477

Figura 14. Características de la propiedad usadas en el modelo clasificador.

En la selección de las características sociodemográficas del INEGI se escogieron 42 variables de los rubros de población, migración, discapacidad, escolaridad, economía, salud, situación conyugal, religión y viviendas. Cuando se tuvieron seleccionadas las 42 variables de INEGI se hizo una reducción de las mismas por un análisis de componentes principales, el cual proyecta un espacio de dimensión N, 42 en este caso, a otro de menor dimensión con diferentes valores de vectores representativos. Se utilizó una reducción a 6 variables o vectores representativos ya que como muestra la Figura 15, las primeras 6 componentes explicaban aproximadamente el 65% de la varianza en las características del INEGI. Esto se hizo con el fin de que el modelo clasificador no tomara el código AGEB como la variable de mayor peso al momento de seleccionar características similares entre propiedades.

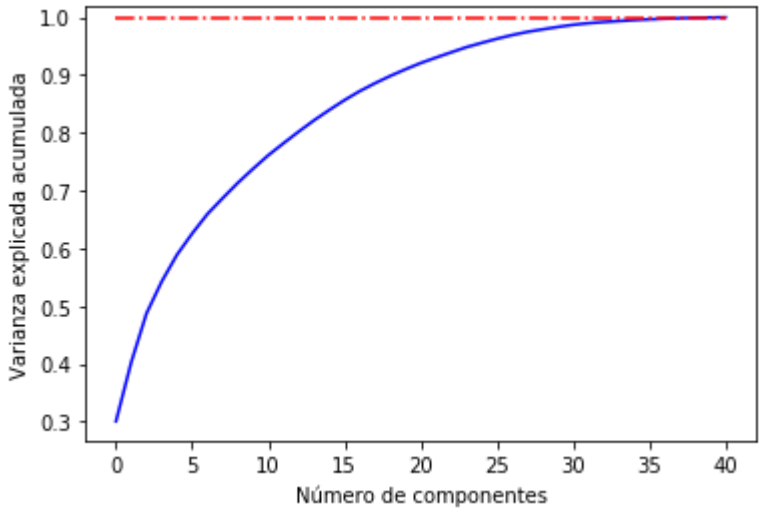

Figura 15. Gráfica de la varianza acumulada de las características del INEGI.

Con los 6 componentes restantes se hizo el cruce de información con las características de las propiedades según el AGEB correspondiente a cada propiedad.

La aplicación del modelo clasificador KNN consiste en obtener la distancia euclidiana entre las propiedades y aquellas con características similares son las que presentan menor distancia. En la Figura 16 se muestra un resultado aleatorio de 9 recomendaciones de propiedades similares a la primera propiedad. Se puede observar que las características de las propiedades (como baños, costo y habitaciones) son muy parecidas o casi iguales, mientras que el código AGEB se repite en solo dos propiedades además de la original.

	MUN_AGEB	ID	cos_costov	prp_habita	prp_banos	prp_terren	prp_constr	prp_estaci	prp_latitu	prp_longit
6543	120_4094	191957	1395000.0	3	3.0	93.0	115.0	2	20.755626	-103.376728
6544	120_4094	1265169	1900000.0	3	2.0	138.0	138.0	2	20.749160	-103.377141
6029	120_3698	1205416	1490000.0	3	2.5	102.0	142.0	2	20.749185	-103.372087
6030	120_3698	1345814	1700000.0	3	3.0	138.0	150.0	0	20.748598	-103.376739
6545	120_4094	1307416	4950000.0	4	5.0	330.0	325.0	2	20.749160	-103.377141
9959	120_6705	1169131	930000.0	3	2.5	90.0	101.0	2	20.773909	-103.443452
9955	120_6705	529685	930000.0	3	2.5	90.0	101.0	2	20.779366	-103.442765
2726	120_2219	952151	1750000.0	3	2.5	105.0	160.0	2	20.754958	-103.366598
9362	120_6508	1354089	1460000.0	3	3.0	108.2	129.0	2	20.759505	-103.430958
9365	120_6508	1354139	1460000.0	3	3.0	99.2	129.0	2	20.760127	-103.430936

Figura 16. Ejemplo de recomendación de propiedades.

En la Figura 17 se muestra otro ejemplo aleatorio del resultado del clasificador en donde esta vez el código AGEB se repite en 6 de las 9 propiedades recomendadas. Esto se debe a que la propiedad referencia puede ser parte de un complejo de propiedades en donde las características de todas son las mismas.

	MUN_AGEB	ID	cos_costov	prp_habita	prp_banos	prp_terren	prp_constr	prp_estacl	prp_latitu	prp_longit
3316	120_2488	929685	1200000.0	3	2.0	168.0	147.0	1	20.607138	-103.435730
3321	120_2488	988084	1200000.0	3	2.0	168.0	147.0	1	20.607138	-103.435722
3318	120_2488	952376	1200000.0	3	2.0	168.0	147.0	1	20.607158	-103.435697
3320	120_2488	984195	1200000.0	3	2.0	168.0	147.0	1	20.607028	-103.434560
3319	120_2488	971001	1155000.0	3	2.0	97.0	163.0	1	20.607028	-103.434560
3325	120_2488	1261719	1200000.0	3	2.5	99.0	99.0	1	20.607028	-103.434560
7200	120_4501	1393594	1400000.0	3	2.0	90.0	130.0	1	20.599728	-103.434560
3323	120_2488	1018162	1650000.0	3	2.5	210.0	200.0	1	20.607961	-103.432404
7195	120_4501	1144548	1350000.0	3	2.5	85.0	126.0	2	20.599728	-103.434560
7198	120_4501	1304606	1350000.0	3	2.5	92.0	125.0	2	20.599728	-103.434560

Figura 17. Ejemplo de recomendación de propiedades.

5. Recomendaciones en la empresa para mejorar la calidad de los datos y el modelo predictivo.

Como parte de las recomendaciones a la empresa Casas y Terrenos que podemos observar en base a los análisis que se hicieron de los datos es necesario que se implemente el guardado de la información sobre las propiedades que ve un usuario al usar el sitio, ya que esto permitiría tener un perfil del usuario basándose en las propiedades que visita y con ellas poder implementar un filtro colaborativo como se mencionaba en la sección 2.1.

Otra de las recomendaciones es que las sesiones de los usuarios deberían de ser únicas por navegador para poder tener un historial amplio sobre el tiempo de las búsquedas que se hacen o propiedades que visitan durante un intervalo de tiempo mayor a una sesión.

Por último, es necesario ajustar el peso que se da a las variables sociodemográficas y a las características de la propiedad para conocer cuáles parámetros se adaptan mejor o son más adecuados para la recomendación de propiedades similares.

4. Reflexiones del alumno o alumnos sobre sus aprendizajes, las implicaciones éticas y los aportes sociales del proyecto

- Aprendizajes profesionales

José Ignacio:

Durante la experiencia del proyecto PAP puedo distinguir varias competencias profesionales adquiridas. Un grupo de ellas se desarrolló alrededor de la necesidad de entender lo que mi aporte requería solucionar, y de cómo comunicar el trabajo realizado. Estas importantes competencias fueron desarrolladas a través de las reuniones, reportes y documentación generada. Es importante para mi vida profesional saber entender las necesidades de los clientes, saber comunicar mis resultados y cómo es que satisfacen las necesidades vistas. En el caso de las competencias técnicas, puedo mencionar que no tenía experiencia previa en el área de análisis de datos, pero bajo la tutela de Pablo Dávalos pude adquirir los conocimientos necesarios para llevar a cabo el proyecto de aplicación personal.

Alejandro:

A lo largo del transcurso del PAP de modelos de predicción para empresas pude aprender acerca de diferentes temas. La parte que más pude sacar provecho fue en la del aprendizaje de aplicaciones en materia financiera, ya que al ser un PAP multidisciplinario, mayormente enfocado a la carrera de ingeniería financiera, aprendí de diversos temas que en mi carrera no son tocados. Otro de los mayores aprendizajes fue acerca de la teoría detrás de los modelos de regresión o clasificación de machine learning puesto que de esta manera puedo saber el origen de los algoritmos que día a día podría utilizar.

- Aprendizajes sociales

José Ignacio:

Una de las cosas que sinceramente no me sorprendió, pero es preciso recalcar, es que no hay un interés por parte del mercado inmobiliario de reglamentar la construcción en ciertas zonas de la ciudad, aunque estas no tengan la infraestructura vial necesaria para soportar la cantidad de personas que se migrarán a estas zonas residenciales. Ignorar este problema nos afecta a todos, en términos

de tráfico vial, contaminación del aire, etc. Y por el contrario, solamente beneficia algunos pocos. Sería ideal que el Gobierno del Estado sea expuestos a los conocimientos generados sobre explosión demográfica y de viviendas, generados en este proyecto. Pero el objetivo de la empresa Casas y Terrenos dista de ser un objetivo social. En una de las reuniones donde se expuso el análisis de las búsquedas de los usuarios, hubo una intención explícita de la empresa de "incluso vender el análisis a nuestros clientes". Siendo sus clientes las principales inmobiliarias de Guadalajara.

Alejandro:

En la parte social puedo asegurar que este proyecto se trata del principio de una gran iniciativa que podría generar información valiosa para diferentes sectores. Al conocer los lugares en los que más personas buscan viviendas se podrían mejorar los servicios como el transporte para hacer más eficiente la movilidad o en cambio se podría investigar las razones por las cuáles no se buscan propiedades en ciertos sectores de la ciudad como podría ser inseguridad o deficiencia de servicios básicos. También es importante para los mismos usuarios ya que así podrían encontrar una vivienda que sea adapte a lo que verdaderamente están buscando y no lo que se trata de hacerles llegar por el bombardeo de publicidad y grandes compañías inmobiliarias.

- Aprendizajes éticos

José Ignacio:

El principal debate ético en el contexto de este proyecto gira alrededor de la privacidad de los usuarios cuando hacen una búsqueda y ven propiedades en el sitio web de la empresa. Sería una práctica adecuada si se le hiciera la pregunta explícita al usuario si está de acuerdo con el tratamiento que se hace de sus datos y posteriores objetivos que se logran con los mismos. En la Unión Europea se implementó nuevas regulaciones, *Cookie Law,* con el objetivo de proteger la privacidad de los usuarios y darles la opción de permitir o no el uso de sus datos. Para ello la ley obliga a los sitios web a obtener el consentimiento explícito de los

visitantes para guardar y utilizar cualquier información obtenida a través de una computadora, teléfono inteligente o tableta.

Alejandro:

Una de las primeras decisiones que tomé fue el unirme al proyecto de Casas y Terrenos en lugar de trabajar en el proyecto de IIEG. Esta decisión conllevó a que trabajara en un modelo clasificador de propiedades y no con datos numéricos que los compañeros financieros ya tienen mayor experiencia trabajando.

Esta experiencia me invita a ejercer mi profesión en el ámbito de la programación puesto que pude adquirir muchas habilidades las cuales son necesarias en diversas empresas para el manejo de sus datos.

- Aprendizajes en lo personal

José Ignacio:

Durante el desarrollo de este proyecto pude darme cuenta que el área de ciencia de datos e inteligencia de negocios es de gran aporte a las organizaciones e incluso podría enfocar mi carrera profesional a este rubro con el soporte que aportan mis conocimientos técnicos en sistemas computacionales. Todavía existen muchas compañías que no entienden el valor que pueden generar si se aplica análisis sus datos.

Alejandro:

En el aprendizaje personal me queda el saber lo que quiero como mi plan de vida. Me sirvió para darme cuenta que quiero pertenecer al sector trabajador con un empleo que me pueda ayudar a desarrollar mis habilidades y donde pueda avanzar con proyectos de emprendimiento.

5. Conclusiones

José Ignacio:

El objetivo principal del proyecto de crear recomendaciones se cumplió, con las propiedades similares que nos entrega el modelo KNN, que no solo recomienda propiedades con características similares, también que se encuentren en una zona socioeconómica similar. No se pudo llegar a recomendaciones personalizadas por

la falta de información de la empresa sobre las propiedades que ve un usuario. El siguiente paso para la empresa es que sus desarrolladores integren el sistema de recomendaciones a su código base actual.

Alejandro:

Al iniciar el semestre se tenía como objetivo desarrollar un algoritmo de recomendación de propiedades el cual pudo ser cumplido de acuerdo a las expectativas que se tenían. Una de las mejoras que se proponen es que la empresa guarde la información necesaria para que se pueda desarrollar un sistema colaborativo de recomendaciones para poder dar aún mejores resultados que los actuales. En siguientes proyectos se puede seguir trabajando con el sistema basado en KNN para su implementación en la página web de la empresa.

6. Bibliografía

[1] Segaran, T. (2007) *Programming Collective Intelligence*. O'Reilly Media.

[2] Lops, P., de Gemmis, M., & Semeraro, G. (2011) *Content-based Recommender Systems: State of the Art and Trends.* En: Ricci F., Rokach L., Shapira B., Kantor P. (eds) Recommender Systems Handbook. Springer, Boston, MA. 73-105.

[3] James, G., Witten, D, Hastie, T., & Tibshirani, R. (2014) *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated