IMPLEMENTACIÓN DE UN MODELO DE PREDICCIÓN DE CUPOS DE INTERCAMBIOS PARA LA DIRECCIÓN DE INTERNACIONALIZACIÓN

Alejandro Garcia Flores, Fabian Castellanos, Brayann Quevedo

Contents

1. Entendimiento del problema.

1.1. Contexto.

La dirección de internacionalización de la Universidad de los Andes es la Unidad encargada de Cooperación internacional. En este momento cuenta con seis ejes de gestión: Doctorados, Centro del Japón, Instituto Confucio, Cooperación Académica, Cooperación para la Investigación y Creación y Movilidad académica. Esto significa que se encuentra a cargo de los procesos de movilidad académica de los estudiantes de la universidad y de gestionar y promocionar acuerdos de intercambio, cooperación e investigación con instituciones externas para poder realizar los procesos de movilidad académica.

La movilidad académica es de gran interés para la Universidad ya que en primer lugar es un complemento importante para la formación de los estudiantes y tiene un impacto

significativo en los indicadores de empleabilidad y calidad de vida en el futuro de los estudiantes. De la misma forma es requerida por el CNA (Centro Nacional de Acreditación) para poder certificar los programas que se imparten dentro de la Universidad, y el número de intercambios realizados en un año especifico con relación al número de estudiantes inscritos es un KPI clave en los rankings internacionales de instituciones de educación superior.

Para poder realizar la tarea de gestión de la movilidad la dirección de internacionalización cuenta desde 2018 con una plataforma tipo SaaS llamada *MoveON* desarrollada por la empresa canadiense QS, en donde se almacenan y gestionan los datos de todos los estudiantes que se postulan a intercambios internacionales y nacionales tanto entrantes como salientes. Adicionalmente, esta plataforma es el repositorio oficial de todos los convenios de la Universidad. En promedio en MoveON se gestionan alrededor de 500 intercambios semestrales y la Universidad cuenta con 571 convenios activos en este momento, de los cuales 244 son convenios de intercambio, 14 de doble titulación y 31 son acuerdos de investigación.

1.2. Problemática.

Todos los acuerdos de intercambio que la Universidad firma con instituciones externas impactan la relación que tiene la Universidad con sus socios, el éxito de los acuerdos depende de que los estudiantes utilicen los cupos que se establecen, pero lastimosamente esto no siempre sucede, lo que significa que los convenios cuyos cupos no fueron utilizados no son renovados, y la oportunidad de que los estudiantes viajen a estas instituciones en un futuro se pierde, en su mayor parte de manera permanente.

Una vez por semestre, se realiza una negociación con los socios de la Universidad con los cuales se tienen acuerdos de intercambio en donde los socios examinan los resultados de los intercambios de los últimos semestres y deciden que tantos estudiantes ambas partes están dispuestas a recibir y enviar. La Universidad firma dos tipos de convenios de intercambio: específicos y generales, los específicos son convenios de intercambio que solo estudiantes de una facultad o programa pueden usar, los convenios generales son convenios de intercambio abiertos a estudiantes de todas las carreras.

En este momento, la universidad cuenta con el registro de 35 convenios de intercambio que finalizaron sin ser renovados de los cuales 16 no fueron renovados por falta de utilización por parte de los estudiantes, y 8 convenios de intercambio que fueron cancelados durante su periodo de vigencia por desbalances de intercambios entre las partes.

2. Design thinking.

Para dar solución a la problemática de este proyecto, es importante realizar el proceso de conceptualización y diseño, este último se realiza mediante design thinking que permitirá incluir todos los elementos principales que nos acercarán a la solución de la problemática.

2.1. Arquetipos.

Se han diseñado 3 arquetipos que permitan relacionar y entender desde el punto de vista de 3 estudiantes seleccionados e interesados en formar parte del proceso de intercambio de la universidad. La información puede ser consultada en *Anexos/Design Thinking.pdf* slides 1-3.

2.2. Ideación.

Para este momento se han planteado por parte de los integrantes de este equipo de trabajo, 4 ideas que permitirán ayudar a resolver la problemática planteada en este proyecto.

Es importante resaltar que, para el proceso de selección de ideas, los integrantes del equipo contaron con la posibilidad de votar por cada idea, incluyendo la propia postulada, también con posibilidad de no apoyar una idea, en la medida que se contabilizaron cada una de las preguntas. Fortalecer el esquema finalmente fue la ganadora en este proceso.

2.3. Story board.

La técnica de los story board nos permiten comprender de una mejor manera la solución planteada dada una problemática, por consenso total del equipo de trabajo, orientamos esta solución en función de la interacción entre estudiantes interesados por parte del proceso y la percepción que tendría la selección con los ajustes obtenidos a partir de los resultados de este proyecto o iniciativa. La información puede ser consultada en *Anexos/Design Thinking.pdf* slide 6.

3. Enfoque analítico.

3.1. Preguntas de negocio.

Para poder entender los datos a trabajar primero buscaremos contestar las siguientes preguntas acerca de los datos y el negocio:

- ¿Cuáles son las carreras o facultades que utilizan más cupos en convenios generales?
- ¿Cuáles son las facultades que más postulaciones por convenios específicos tienen y cuál es su tasa de utilización de cupos específicos?
- ¿Cuáles son las 30 universidades que más postulaciones de pregrado reciben?
- ¿Cuál es el porcentaje de estudiantes de pregrado que se postulan a las 30 universidades con más postulaciones?
- ¿Cuáles son las universidades con menos cupos utilizados?
- ¿Qué países realizan más intercambios?
- ¿Pueden las recomendaciones mejorar el ***kpi*** de asignación de cupos en la universidad?

3.2. Análisis predictivos.

Se construirá un modelo de regresión lineal para predecir el número de cupos para la última convocatoria de pregrado y se comprará con los cupos reales utilizados para ver la precisión del modelo construido.

3.3. Objetivo principal.

En el marco de ejecución de este proyecto el propósito principal del proyecto será el siguiente:

- Construir un modelo de regresión lineal que prediga el numero óptimo de cupos para el siguiente semestre para una universidad con un histórico de cupos dados. De tal forma que se utilicen la mayor cantidad de cupos posibles, sin que se desperdicien.

3.4. Objetivos específicos.

Los objetivos específicos que permiten cumplir el objetivo principal del proyecto son los siguientes:

- Entender con profundidad los datos obtenidos para tener claridad sobre el negocio.
- Construir un modelo de regresión lineal que prediga el numero óptimo de cupos base para un convenio nuevo, con una universidad con la cual no se tiene un historial de cupos.
- Responder acertadamente las preguntas de negocio con un enfoque analítico.

3.5. Métricas.

Para lograr responder a las preguntas de negocio es importante lograr como mínimo un nivel de ocupación de los cupos al 80%, kpi base de verificación de la hipótesis.

4. Entendimiento de los datos.

La base de datos de movilidad y convenios de La Dirección de Internacionalización (MoveON) que está a cargo de Alejandro Garcia Flores, fue obtenida con permiso de la directora de internacionalización. La base de datos fue exportada de MoveOn en formato Excel. Esta cuenta con 4 tablas que se utilizaran en el proyecto:

Conjunto	Descripción	Total de registros
Institution	Registro de las instituciones, de las cuales 2008 son externas	4297
Relation Institution	Registro de relación entre las instituciones, caracterización y tipos	2008
Seat	Cupos aperturados semestralmente	2283
Stay Wishes	Postulaciones para intercambios salientes desde 201820 a 202310	8154
Courses	Programas académicos de las instituciones	4557

El detalle del diccionario de datos puede ser consultado en *Anexos/Diccionario Datos.xlsx*

5. Implicaciones éticas.

Los datos usados en el modelo son reales relacionados con los convenios de la universidad. Teniendo en cuenta la ley 1266 de hábeas data, la información es sensible y está relacionada con los datos personales, por lo tanto, debe ser protegido y tratado adecuadamente.

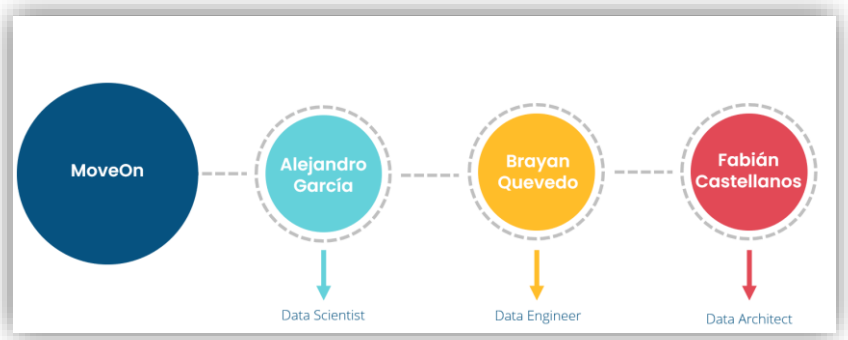
Los datos sensibles de este proyecto corresponden a los datos de identificación de las personas que desean obtener un cupo. Desde el punto de vista de ciencia de datos, los atributos a proteger tampoco son relevantes, por lo tanto, la información tendrá el siguiente tratamiento:

- Alejandro García recibió la información directamente registrada.
- Con el punto anterior, Alejandro realizó el proceso de anonimización (*transformar los datos para protección de la información*) de la información de las postulaciones de los estudiantes.
- Excluir información confidencial de la universidad y los estudiantes
- Excluir la información no relevante para el modelo.

En este punto los integrantes restantes del proyecto, Brayann Quevedo, Fabián Castellanos y las personas que lean el contenido de este proyecto no incurrirán en ninguna implicación ética con la información, teniendo en cuenta que el repositorio es público, el tratamiento de datos cumple a cabalidad la protección que los datos requieren.

6. Organigrama del proyecto.

A continuación, se relacionan los roles que tendrán cada uno de los participantes del equipo:



7. Metodología de trabajo.

Las preguntas de negocio a resolver con la información del proyecto requieren un marco de trabajo para la cual usaremos ASUM-DM. La documentación de este documento corresponde al tercer sprint del proyecto.

8. Modelo.

A continuación, es relacionada la información relevante del modelo de machine learning para MoveOn.

8.1. Código fuente del modelo.

La información puede ser accedida en la siguiente ruta.

8.2. Análisis exploratorio.

Al realizar la exploración de los datos, encontramos que particularmente la tabla de postulaciones (Stay Wishes) y los convenios y oportunidades de intercambio (Relations) tenían problemas grandes de limpieza de datos. Sin embargo, después de depurarlos, imputando y filtrando por los datos que realmente eran pertinentes al enfoque del estudio, ambas tablas se pudieron usar. Como respuesta a las preguntas encontramos lo siguiente:

¿Cuáles son las carreras o facultades que utilizan más cupos en convenios generales?

- Facultades: Ingeniería, Arquitectura y diseño, Economía.
- Programas: Economía, Derecho, Diseño.

¿Cuáles son las universidades que más postulaciones de pregrado reciben?

- Politecnico di milano (227)
- Universitat Ramon Llull (213)
- Universitat Politècnica de Catalunya (175)

¿Cuáles son las universidades con menos cupos utilizados?

- North Dakota State University
- National Chengchi University
- École de Technologie Supérieure (ETS)

¿Cuáles son las facultades que más postulaciones exitosas por convenios específicos tienen y cuál es su tasa de utilización de cupos específicos?

FACULTAD	CUPOS USADOS	CUPOS USADOS/CUPOS DISPONIBLES
SCHOOL OF MANAGEMENT	321	- 2.45%
SCHOOL OF ENGINEERING	312	- 3.71%
SCHOOL OF ARCHITECTURE AND DESIGN	96	- 3.51%

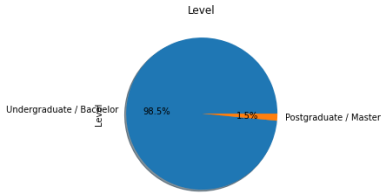
¿Cuál es el porcentaje de estudiantes de pregrado que se postulan a las universidades con más postulaciones?

- Politecnico di milano: 3.24%
- Universitat Ramon Llull: 3.04%
- Universitat Politècnica de Catalunya: 2.49%

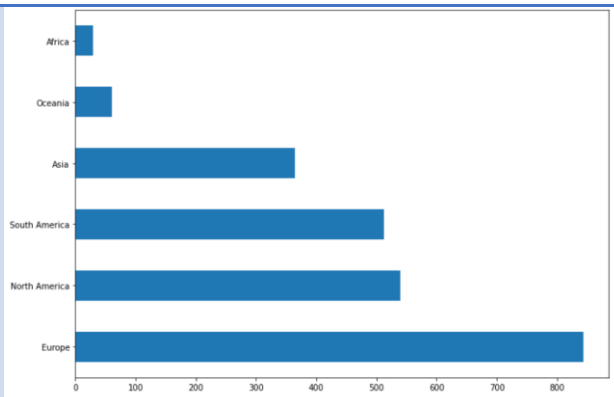
¿Qué países realizan más intercambios?

- France (307)
- Germany (245)
- Spain (190)

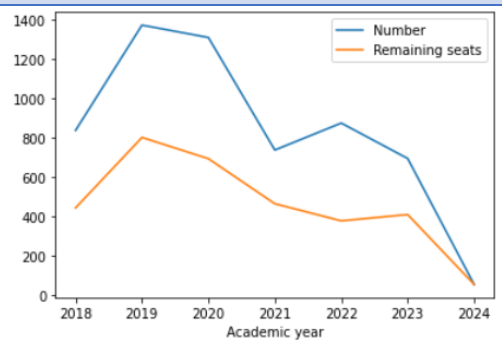
1. La mayoría de las postulaciones para realizar los intercambios son gente de pregrado exactamente el 98.3%
2. Los estudiantes prefieren postularse a España, Francia y Alemania, pues estos países son con los que más cuentan postulaciones
3. Del total de postulaciones tenemos que hay 4977 estudiantes que han sido rechazados desde el proceso de selección
4. Luego del proceso de selección tenemos que una gran proporción quedan completados (es decir, terminando el proceso de intercambio)



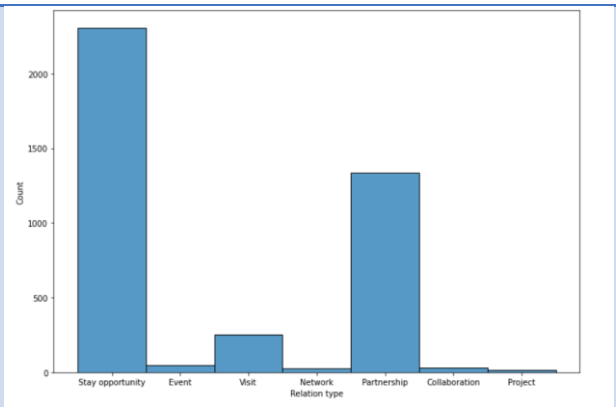
5. Ubicación geográfica de los cupos de las universidades

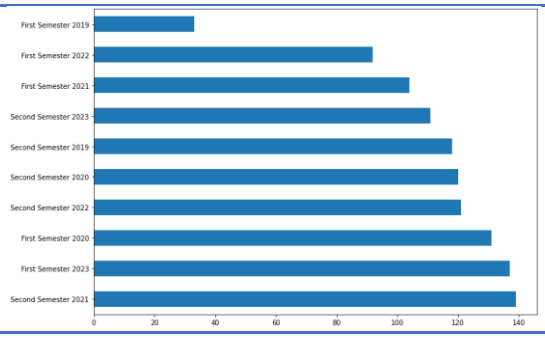


6. Cantidad de cupos a través de los años



7. Conteo sobre las oportunidades que más seleccionan los estudiantes



8. **Cantidad de postulaciones a nivel semestral**	

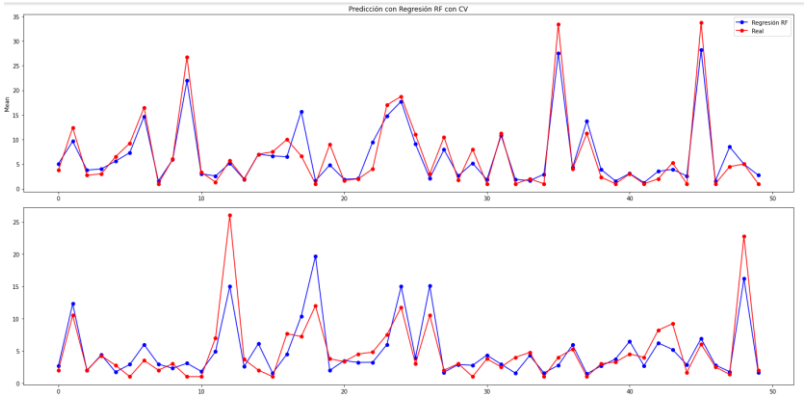
8.3. Construcción del modelo base.

La cardinalidad en la variable a predecir es alta, fue necesario reducir el set de datos para el procesamiento en los diferentes modelos.

8.4. Modelos experimentales.

Se construyeron 3 modelos, polinomal lasso, nearest neighbor y random forest. Para la segunda entrega y Mediante Grid seach fueron identificados los mejores parámetros que permitieran ser usados para construir estos modelos. Random forest mostró un mejor comportamiento en los 3 construidos para ese momento. Siguiendo las recomedaciones dadas se realizaron 4 modelos más: Support Vector Regression, Decision Tree Regresor,XGBoost Regression y por último, un modelo de redes neuronales Multilayer Perceptron

Se presentarán los resultados con los diferentes modelos nuevos, teniendo en cuenta que para la segunda entrega Random Forest tiene los siguientes resultados:




```
------------ Regresión RF con entrenamiento-----------
Residual sum of squares (MSE): 7.04
R2-score: 0.82532
Adj R2-score: 0.72515
------------ Regresión RF con evaluación ------------
Residual sum of squares (MSE): 11.38
R2-score: 0.58950
Adj R2-score: 1.87231
```

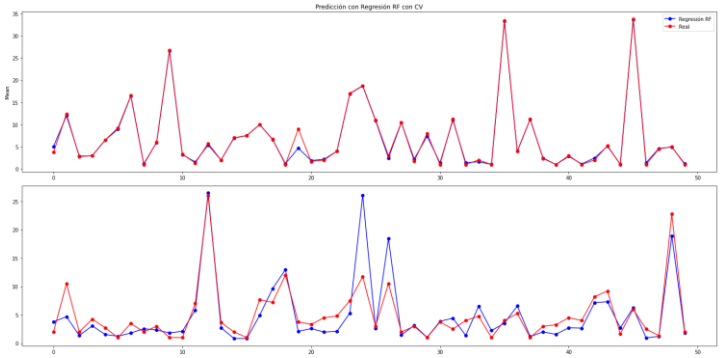
Para Support Vector Regresion se tiene:


```
------------ Regresión RF con entrenamiento-----------
Residual sum of squares (MSE): 15.17
R2-score: 0.62367
Adj R2-score: 0.40789
------------ Regresión RF con evaluación ------------
Residual sum of squares (MSE): 12.94
R2-score: 0.53326
Adj R2-score: 1.99183
```

Para Decision Tree Regresor:


```
------------ Regresión RF con entrenamiento------------
Residual sum of squares (MSE): 5.14
R2-score: 0.87259
Adj R2-score: 0.79954
------------ Regresión RF con evaluación ------------
Residual sum of squares (MSE): 25.56
R2-score: 0.07850
Adj R2-score: 2.95819
```

Para XGBoost Regresion:




```
------------ Regresión RF con entrenamiento------------
Residual sum of squares (MSE): 0.38
R2-score: 0.99046
Adj R2-score: 0.98499
------------ Regresión RF con evaluación ------------
Residual sum of squares (MSE): 12.71
R2-score: 0.54177
Adj R2-score: 1.97374
```

Como se observa en los resultados de los modelos experimentales Random Forest, sigue siendo uno de los mejores modelos para nuestro proyecto, debido a que es el que más se acomoda a los datos.

Ahora bien, el modelo de XGBoost genera un super overfitting y se puede apreciar en los R2, el modelo de Decision Tree, genera overfitting y el modelo de Support Vector Regression obtiene un puntaje bastante bajito, por lo tanto, no tiene overfitting y es por estas razones que Random Forest sigue siendo el mejor Modelo

9. Conclusiones.

- Con la manipulación de los datos obtenidos y el estudio realizado, se lograrían responder las preguntas del enfoque analítico.
- Para mejorar los modelos o los resultados obtenidos es bastante complicado, pues la data que tenemos es suficiente y tendríamos que esperar unas series de tiempo (cada semestre académico) para tener más datos, por lo tanto, se aprecia que es una modelo de predicción bastante real.
- Resolviendo una pregunta del modelo de negocio que resulta de la segunda entrega, la pandemia no afectó el número de cupos debido a que desde el primer semestre de 2020 hasta el segundo semestre de 2022 las postulaciones aumentan.