Septiembre 2022

Universidad de Los Andes

Jesús David Barrios 201921887

Sergio Peñuela 201922873

Jhoan Diaz 201819861

Laboratorio 2 Inteligencia de Negocios

1. Descarga y entendimiento de los datos

Se descargaron los datos entregados. Inicialmente se obtuvo información básica de los datos. A continuación, se presentan tablas e imágenes de información básica de los datos.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Unnamed: 0        1569 non-null   int64
 1   Serial No.        1569 non-null   int64
 2   GRE Score         1569 non-null   int64
 3   TOEFL Score       1569 non-null   int64
 4   University Rating 1569 non-null   int64
 5   SOP               1569 non-null   float64
 6   LOR               1569 non-null   float64
 7   CGPA              1569 non-null   float64
 8   Research          1569 non-null   int64
 9   Admission Points  1504 non-null   float64
```

```
Número de filas:  1569
Número de columnas:  10
```

Ejemplo de los datos con sus columnas

	Unnamed: 0	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Admission Points
134	134	119	296	99	2	3.00	3.5	7.28	0	47.0
442	442	465	298	72	2	0.59	3.0	7.21	0	45.0
953	953	463	265	101	4	2.79	3.0	8.25	0	62.0
20	20	83	320	110	5	5.00	4.5	9.22	1	92.0
531	531	65	285	101	3	3.00	3.5	8.70	0	52.0

Dimensiones de calidad

- **Completitud:** Se observa el número de valores nulos en los datos

```
Número de filas con valores nulos:  65
Número de columnas con valores nulos:  1
Lista de columnas con valores núlos y sus tipos:
 Admission Points    True
dtype: bool

Porcentaje de completitud de las columnas: 95.86%
```

Se puede ver que hay 65 filas con valores nulos y que todos se concentran en la misma columna: "Admission Points". Estos valores se tratarán más adelante.

- Unicidad: Se ven el número de valores repetidos en la base

Se observa que no hay filas que se repiten en la base, por lo que no será necesaria hacer la eliminación de filas duplicadas.

```
Número de filas duplicadas:  0
Número de filas con indice duplicado:  0
```

- Consistencia y validez:

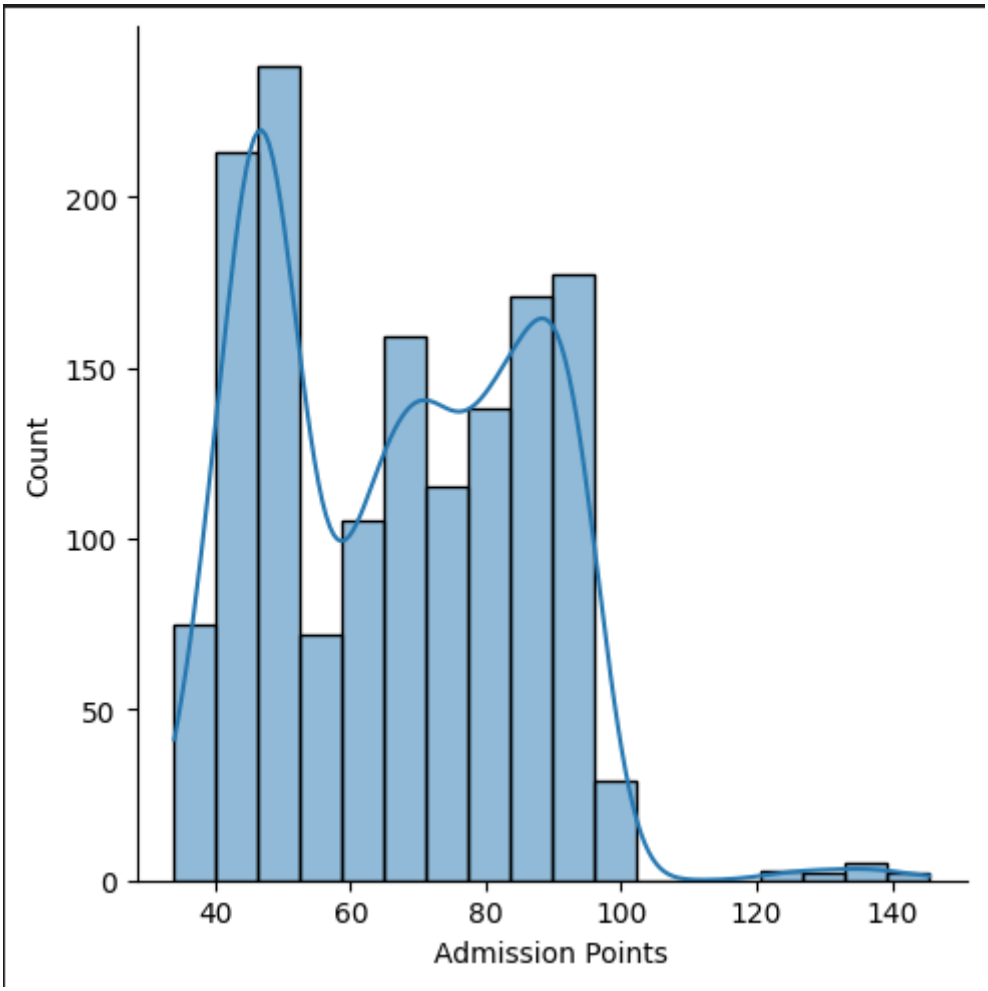
La consistencia hace referencia a la integridad de datos entre fuentes y observaciones. Por su parte, la validez mide si los datos hacen sentido para el contexto específico. En este sentido, se considera que en términos generales se cumple con estas métricas. Todas las columnas dadas son numéricas, algunas de tipo discreto otras de tipo continuo y no se encontraron errores en los valores dados.

2. Preparación de los datos

Todo el procedimiento se puede ver en el notebook titulado "entendimiento_limpieza.ipynb".

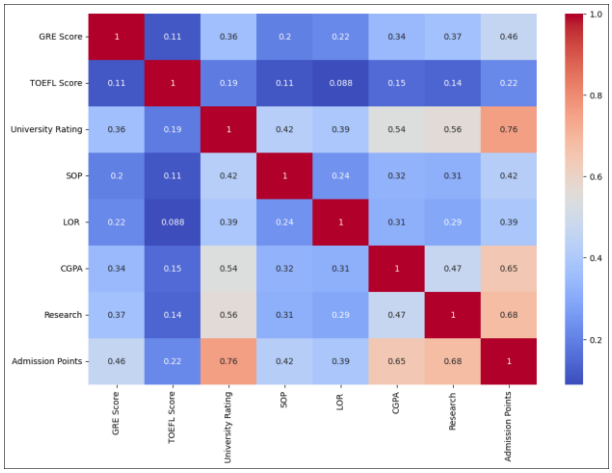
Inicialmente se eliminó la variable "Serial No." pues esta columna no aporta información relevante al problema.

Después de esto se decidió obtener la gráfica para los puntos de admisión para ver como se distribuían los datos.

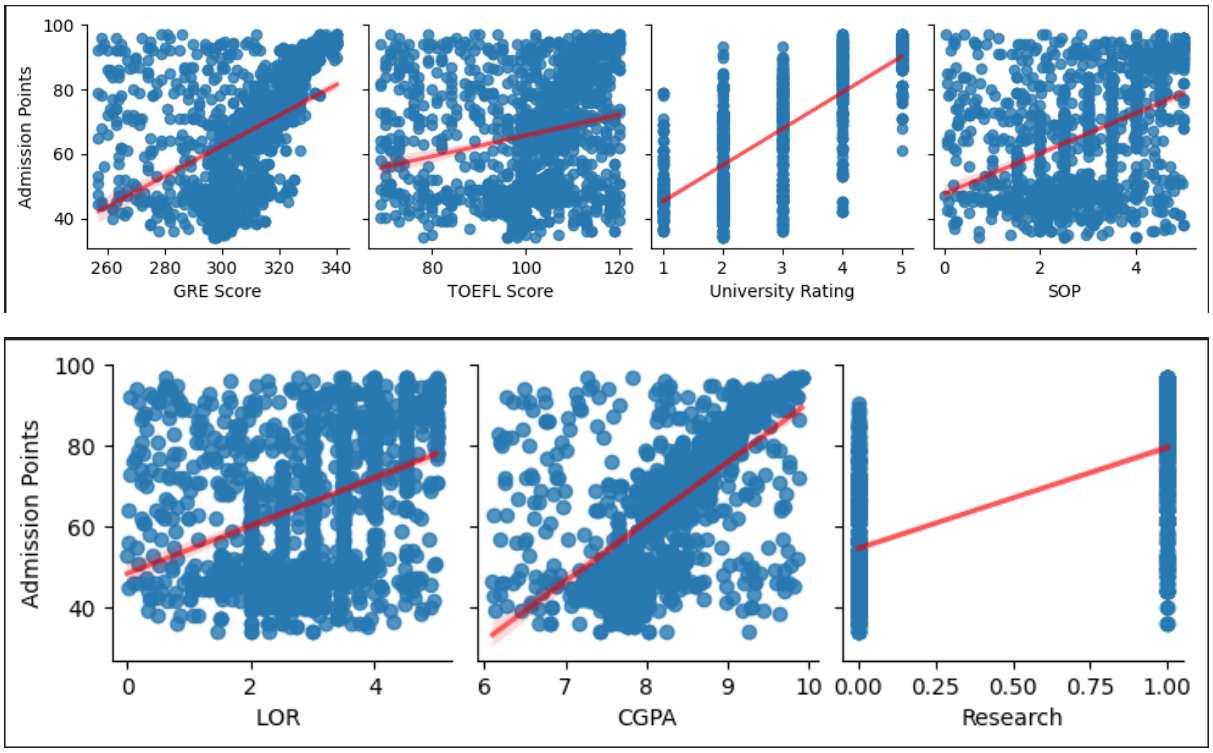


Como se puede ver en la gráfica la mayoría de los valores están en el rango de 40-100, y hay algunos datos atípicos que pueden afectar los resultados del modelo que se encuentran entre 120 y 150. Por ende solo se mantuvo los valores menores a 105.

Posteriormente se buscó la relación que hay entre las diferentes variables y se graficó el resultado.



Posteriormente, se graficó todas las diferentes columnas con respecto a la columna “Admission Points”.



Después de esto solamente se guardó los nuevos datos en un archivo .csv llamado “processed_data.csv”.

3. Modelamiento, validación y visualización

En este caso se hizo uso del algoritmo de regresión lineal. Este algoritmo lo que busca es modelar la relación entre una variable continua y una o más variables independientes. Todo esto se hace mediante el ajuste de una ecuación lineal, en este caso se va a utilizar para predecir el puntaje de admisión de un estudiante dada su información académica. Antes de correr el modelo se estandarizaron los datos para que las diferentes magnitudes entre las columnas no afectarán el resultado del modelo.

	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Admission Points
count	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000
mean	0.610081	0.637846	0.465985	0.592972	0.602602	0.586552	0.465818	0.511277
std	0.214492	0.244508	0.308972	0.247917	0.241887	0.211527	0.498997	0.291625
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.493976	0.529412	0.250000	0.397590	0.409000	0.433901	0.000000	0.222222
50%	0.626506	0.666667	0.500000	0.598394	0.600000	0.583770	0.000000	0.523810
75%	0.774096	0.823529	0.750000	0.799197	0.800000	0.769634	1.000000	0.793651
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Al correr el modelo lo primero que hicimos fue obtener el coeficiente de cada variable y el intercepto de la ecuación lineal. Se obtuvo lo siguiente:

```
Intercepto:  -0.1183142894816871
El coeficiente para GRE Score es 0.14236603022072775
El coeficiente para TOEFL Score es 0.07166836432596357
El coeficiente para University Rating es 0.36864030407185977
El coeficiente para SOP es 0.06557008012763495
El coeficiente para LOR  es 0.051195737619014615
El coeficiente para CGPA es 0.31249092932501377
El coeficiente para Research es 0.16043731376997478
```

El coeficiente nos dice lo siguiente: al mantener todas las otras variables con el mismo valor, si modificamos (para el ejemplo se utilizará el GRE Score) en una unidad el resultado "Y" aumentará en 0.142 unidades. Como se puede ver en la foto la variable que más afecta los "Admission Points" es el University Rating pues una un aumento de decrecimiento de uno (1) en esta variable afectará el resultado del modelo en 0.368 unidades, seguido por el CGPA con 0.312 unidades.

Después de esto para poder evaluar el modelo se hizo uso del coeficiente de determinación (R^2). Este coeficiente toma un valor entre 0 y 1. Entre más cerca este el valor a cero menos ajustado está el modelo y por ende menos fiable para hacer una predicción. Al estar el valor cerca a 1 ocurre lo contrario, el modelo esta mejor ajustado y es más fiable para hacer predicciones. En este caso se obtuvo el siguiente valor:

```
R^2:  0.7490740337295853
```

Este valor nos indica que el modelo tiene un buen ajuste y se puede confiar en este para hacer una predicción de los puntos de admisión que un estudiante va a obtener a partir de sus datos académicos. Por lo que se recomienda el uso de este modelo para la Universidad de los Alpes.