

Aprenentatge Computacional
Práctica 1: Regresión
Predicción del salario en jóvenes recién graduados



Bartomeu Vargas Alarcón 1568113
Harold Gonzales Contreras 1571485
Martí Simón Rojas 1568180

Índice

Introducción	2
Librerías utilizadas	2
Apartado (C): Analizando datos	3
Preguntas	4
Apartado (B): Primeras regresiones	9
Preguntas	11
Conclusiones	13
Problemas surgidos	13

Introducción

En esta memoria se mostrará toda la información con la que se ha trabajado. Se ha estudiado una base de datos, analizado sus atributos y valores, seleccionado los datos interesantes, tratado los datos, analizado los datos, y finalmente sacado conclusiones.

La base de datos que se nos ha entregado ha recogido diferentes datos de miles de estudiantes en India durante sus estudios. Concretamente: datos personales: fecha de nacimiento o género; datos académicos: notas del grado 10 y 12, notas de las asignaturas y aptitudes evaluadas en el AMCAT y estudios universitarios y sus notas; y finalmente datos financieros.

Para entender en su plenitud la base de datos entregada, vemos menester hacer una pequeña explicación del sistema educativo indio. Los cursos académicos allí se consideran grados, y el grado 10 y 12 son los previos a los estudios universitarios, encontramos su equivalente aquí con el Bachillerato. Luego se debe hacer un examen de acceso a la universidad llamado AMCAT (la Selectividad de India) donde se evalúa asignaturas obligatorias (Inglés, Matemáticas y Lógica), aptitudes (escrupulosidad, empatía, extrovertismo, neuroticismo y creatividad) y optativas. Finalmente se accede a la universidad con el mismo funcionamiento que aquí.

Para acabar, se han analizado los datos y hecho regresiones que se han creído necesarias para poder hacer conclusiones a partir de los datos entregados. En esta memoria se ha querido reflejar la metodología de trabajo que se ha seguido con los datos, por ello, no dedicamos un apartado en específico para hablar del tratamiento de datos sino que lo comentamos los tratamientos específicos hechos antes del correspondiente apartado.

Repositorio: <https://github.com/haroldgonzales98/practica1-apc>

Librerías utilizadas

Para la realización de la práctica con una mayor eficiencia y comodidad, hemos hecho uso de librerías de Python, algunas necesarias como otras de ayuda para la realización de los códigos. Las principales librerías utilizadas son las siguientes:

- **NUMPY**: incorpora algunas funciones matemáticas de alto rendimiento con las que podemos hacer operaciones de manera más rápida y eficiente.
- **PANDAS**: incorpora estructura de datos que nos facilitan el almacenamiento de tablas y su correcta manipulación. Además, incluye funciones de análisis matemáticos que nos ayudarán a trabajar mucho más cómodos con los datos de la BBDD.
- **MATPLOTLIB Y SEABORN**: son dos librerías que están especializadas en gráficos y análisis de variables.

- **SKLEARN:** librería de aprendizaje computacional que incorpora todos los algoritmos que hemos utilizado (regresiones, normalizaciones...).
- **DATETIME:** librería utilizada para poder analizar las fechas con formato YYYY-MM-DD y poder luego actualizarlo a un valor más manejable.

Apartado (C): Analizando datos

Tratamiento de datos

Varios datos de tipo string daban problemas con funciones de representación de valores, en concreto la función **pairplot()** para poder visualizar qué atributos generaban una campana de Gauss. Para solucionarlo, pasamos dichos strings a valores de discretos:

- Del atributo **DOB**, date of birth, está representado como un string de formato yyyy-mm-dd, se ha pasado a tipo entero de formato yy.yym.mdd (1990-10-23 => 19.901.013) y luego se ha normalizado.
- Del atributo **Gender**, con valores m o f, son sustituidos por 0 y 1 respectivamente.
- Del atributo **Degree**, con strings por nombre de grado, la función scatter muestra los 4 valores que más se han repetido, donde **B.Tech/B.E.** tiene el 92%, **MCA** el 7% y **M.Tech./M.E.** y **M.Sc. (Tech.)** con +1%, por tanto nos hemos decidido por quedarnos solo con estos valores representados por 1, 2, 3 y 4 respectivamente y con 0 los **otros** que no tendrán prácticamente representación en la nube de puntos.
- Del atributo **Specialization**, para evitar una segregación muy alta de datos escogeremos las 5 primeras categorías más repetidas y el resto las juntaremos, por tanto queda: 1 para **electronics and communication engineering**, 2 para **computer science & engineering**, 3 para **information technology**, 4 para **computer engineering**, 5 para **computer application** y 0 para **otros**.
- Del atributo **CollegeState** y se ha categorizado como: 1 para **Uttar Pradesh**, 2 para **Karnataka**, 3 para **Tamil Nadu**, 4 para **Telangana**, 5 para **Maharashtra** y 0 para **otros**.

Los atributos **10board** y **12board** contienen hasta 1093 valores de tipo string lo que complica mucho su clasificación y posterior distribución con los valores de los otros atributos. Por tanto, hemos decidido eliminarlo del dataset ya que, como mencionaremos en el apartado B, no encontramos mucha relación con el atributo objetivo.

En el atributo **GraduationYear** el rango se sitúa entre [2000,2017] pero hay un dato evidentemente erróneo con GraduationYear=1420, para tratar el valor hemos decidido no eliminar toda su fila de datos del estudiante y simplemente asumir que se graduó en el año con más graduaciones (se ha hecho un replace), es decir, 2010 y, aunque probablemente sea erróneo no afectará a la visualización de los datos de la gráfica.

Preguntas

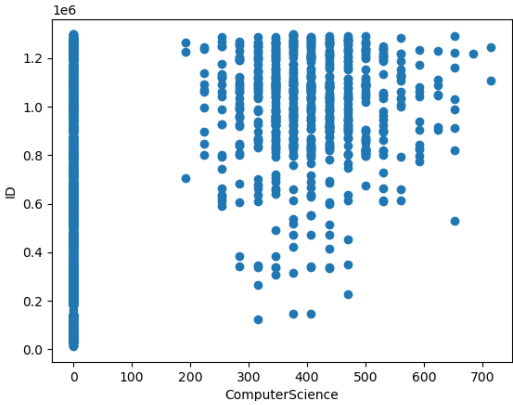
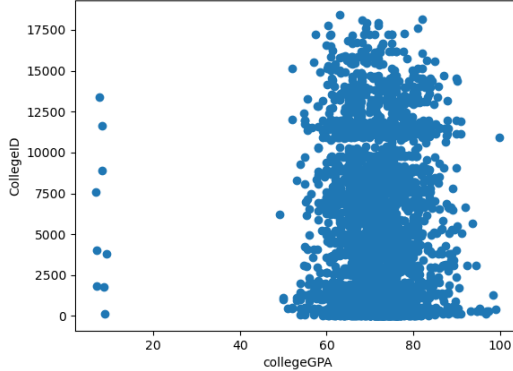
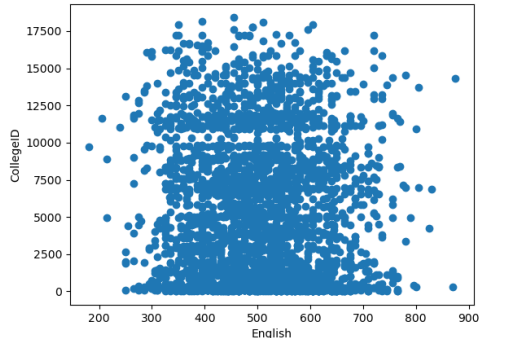
1. ¿Cuál es el tipo de cada atributo?


Trabajamos con un total de 34 atributos que se nos han entregado para poder hacer el estudio y posterior predicción de uno de esos atributos. Hay todo tipo de valores de los atributos que, para poder trabajar con ellos, se han modificado o transformado; pero en la siguiente tabla se indicará el tipo de los datos sin ninguno de estos tratamientos para poder entender de qué base partimos. Las modificaciones de los datos se explican en el momento adecuado en su apartado correspondiente de **Tratamiento de datos**. En la tabla también les acompañará una pequeña descripción para poder entenderlos aunque luego se decida no usarlos.

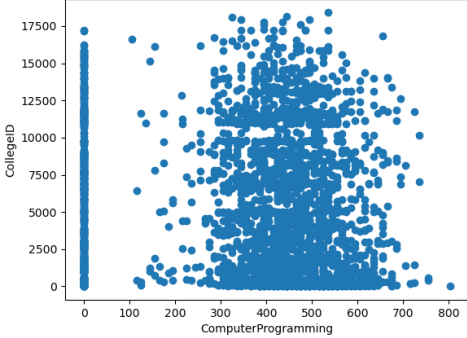
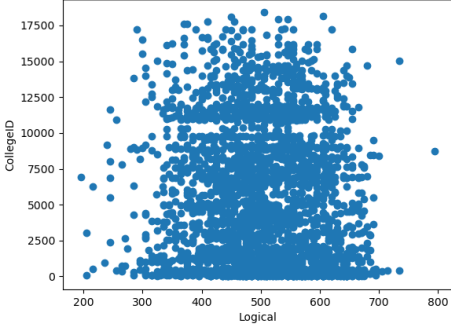
Atributo	Tipo	Descripción	Atributo	Tipo	Descripción
ID	Discreta	Muestra el ID de cada estudiante participe de la muestra.	Gender	Discreta	Muestra el género del estudiante.
DOB	Discreta	Fecha de nacimiento del estudiante.	10percentatge	Continua	Media de las notas de los exámenes de grado 10.
10board	Discreta	Plan de estudios escogido en el grado 10.	12graduation	Discreta	Año de graduación del grado 12.
12percentage	Continua	Media de las notas de los exámenes de grado 12.	12board	Discreta	Plan de estudios escogido en el grado 12.
CollegeID	Discreta	ID de la universidad del estudiante.	CollegeTier	Discreta	Grado de excelencia de la universidad en función de las notas.
Degree	Discreta	Grado estudiado.	Specialization	Discreta	Mención escogida dentro del grado.
CollegeGPA	Continua	Valor de GPA, valor acumulativo en función de las notas.	CollegeCityID	Discreta	ID de la ciudad en la que se sitúa la universidad.
CollegeCityTi	Discreta	Grado de	CollegeState	Discreta	Estado en India

er		excelencia de la ciudad en función de la población			en el que se localiza la universidad.
GraduationYear	Discreta	Fecha de graduación.	English	Continua	Notas en la asignatura de Inglés.
Logical	Continua	Nota en la habilidad de Lógica.	Quant	Continua	Nota en habilidades en cálculo.
Domain	Continua	Nota en las asignaturas opcionales.	ComputerProgramming	Continua	Notas en la asignatura de Computer Programming.
ElectronicsAndSemicon	Continua	Notas en la asignatura de Electronics & Semiconductor Engineering.	ComputerScience	Continua	Notas en la asignatura de Computer Science.
MechanicalEngg	Continua	Notas en la asignatura de Mechanical Engineering.	ElectricalEngg	Continua	Notas en la asignatura de Electrical Engineering.
TelecomEngg	Continua	Notas en la asignatura de Telecommunication Engineering.	CivilEngg	Continua	Notas en la asignatura de Civil Engineering.
conscientiousness	Continua	Nivel de escrupulosidad (hacer las tareas correctamente).	agreeableness	Continua	Nivel de empatía (anteponer las necesidades de los demás).
extraversion	Continua	Nivel de extrovertismo.	neuroticism	Continua	Nivel de neuroticismo (grado de respuesta negativa ante situaciones negativas).
openness to experience	Continua	Nivel de creatividad o imaginación.	Salary	Continua	Salario anual en rupias indias.

2. ¿Qué atributos tienen una distribución Gaussiana?

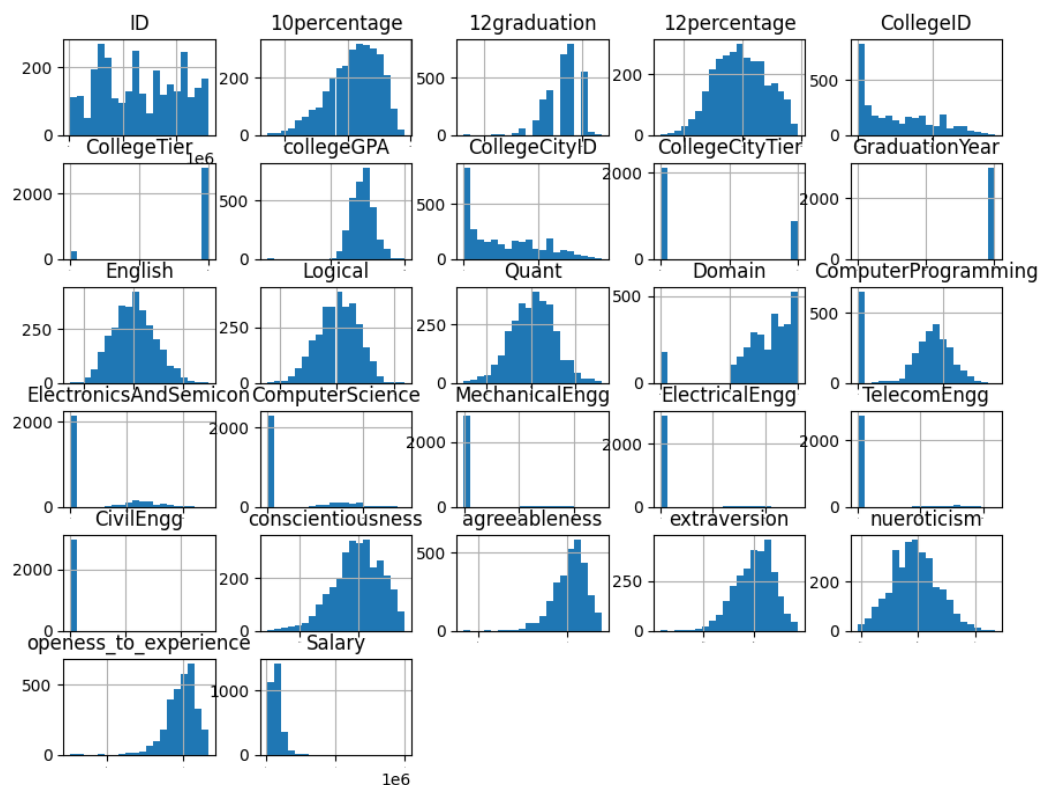
	<p>Observamos que en la asignatura de Computer Science las notas están distribuidas alrededor del aprobado en una campana de Gauss.</p>
	<p>Si recordamos, GPA es un valor acumulativo en función de las notas de un estudiante, por tanto, cuanto más alto mejores cualificaciones ha sacado el estudiante. En la nube de puntos podemos observar que las notas se sitúan por encima del aprobado alrededor de 70.</p>
	<p>En esta nube de puntos podemos observar que las notas de la asignatura de inglés (English) están distribuidas en forma de campana de Gauss para todas las universidades de las cuales se ha tomado muestra.</p>

	<p>Similar a la anterior, en esta ocasión nos muestra que los estudiantes se encuentran</p>
---	---

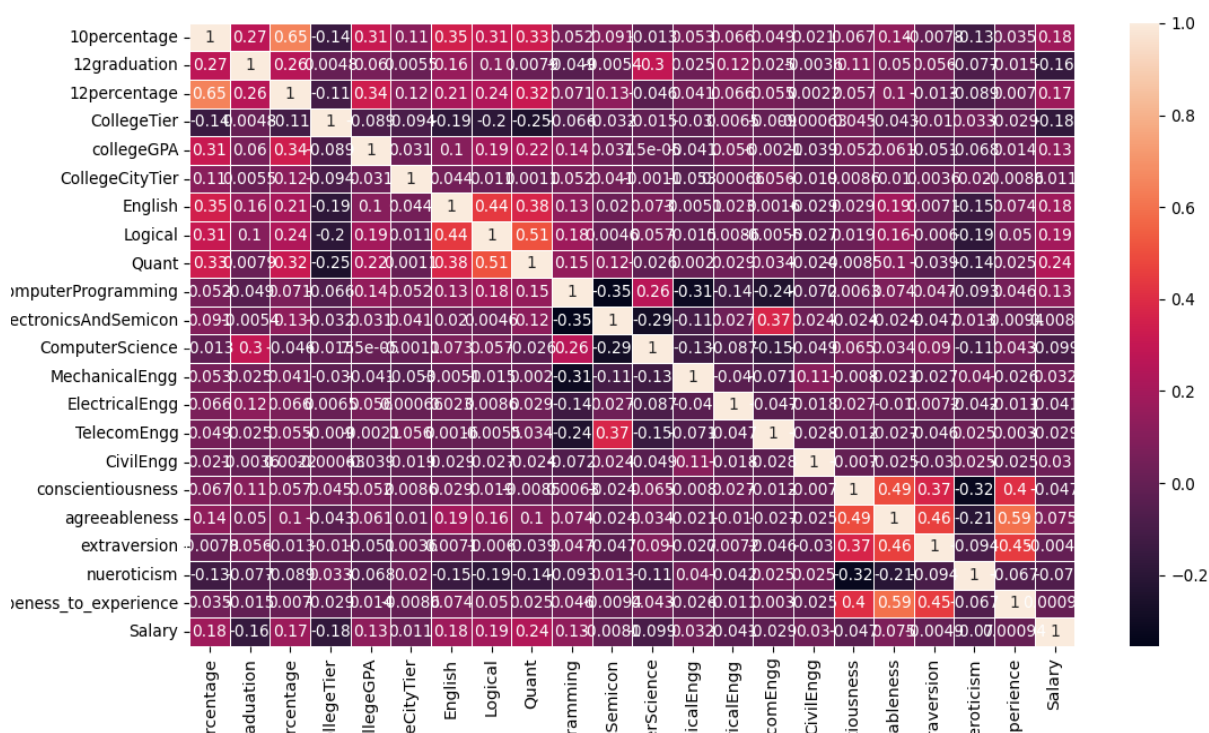
	<p>alrededor del aprobado en su habilidad cuantitativa.</p>
	<p>Nos encontramos una forma Gaussiana alrededor por debajo del aprobado, por tanto lleva a pensar qué Computer Programming es una asignatura con la que los estudiantes encuentran dificultades. Podemos observar también una gran cantidad de 0, que probablemente sean no presentados.</p>
	<p>Esta nube de puntos muestra una forma Gaussiana alrededor del aprobado para la habilidad de Lógica de los estudiantes.</p>

Además, con la función **hist()** de la librería **pandas** podemos ver un histograma de cada atributo para ver desde otra perspectiva los atributos que generan una distribución de Gauss.

También podemos ver que hay atributos que parecen tener una distribución gaussiana pero también una barra muy grande cerca del cero, eso es debido a que tienen valores en la base de datos marcados como -1 equivalentes a NULL.



Haremos un heatmap con la librería **seaborn** para observar de manera visual la correlación entre los distintos atributos.



Podemos ver que las columnas con más correlación son:

- Logical - Quant
- Conscientiousness - agreeableness
- extraversion - open_to_experience
- Logical - English
- Conscientiousness - open_to_experience

Los atributos que más correlación con el atributo objetivo, aunque no tienen mucha, son:

- Quant
- Logical

3. ¿Cuál es el atributo objetivo? ¿Por qué?

El muestreo de datos de la base de datos entregada tiene el objetivo concreto de poder predecir el salario de los jóvenes indios después de graduarse e intentar encontrar qué atributos tienen una implicación en el salario que cobrarán estos jóvenes. Por tanto, hemos escogido **Salary** como el atributo a predecir mediante regresión. Pero los análisis de este primer apartado nos llevan a ciertas conclusiones.

Hasta ahora ya podemos observar ciertos comportamientos extraños en las gráficas del anterior apartado. Todas las campanas de Gauss encontradas muestran relaciones entre las notas de las diferentes asignaturas y aptitudes con los estudiantes de las distintas universidades y ciudades, mostrándonos así dónde se sitúan las notas medias de estas categorías. Luego, en el **heat map** observamos que nuestro atributo objetivo no tiene prácticamente relación con los otros atributos.

Apartado (B): Primeras regresiones

En este apartado calcularemos el error cuadrático medio del regresor para cada atributo de la base de datos. Gracias a este paso, podremos observar qué atributo nos muestra un error cuadrático menor, lo cual nos querrá decir que difiere menos entre la predicción y la realidad.

Para la realización de este paso, nos hemos ayudado de las funciones dadas del Notebook junto a algunas modificaciones de los datos obtenidos (eliminaciones[drops] de atributos, cambio de valores de string a int...). Hemos calculado el error con los datos con y sin normalizar para poder comprobar y observar las diferencias dadas dos maneras de hacerlo, y poder confirmar si la normalización nos ayuda para conseguir un rango de datos por igual.

Podemos conseguir los siguientes errores cuadráticos:

<i>MSE</i>	<i>SIN NORMALIZAR</i>	<i>NORMALIZADO</i>
Atributo 1	1.5512	0.7963
Atributo 2	1.9261	0.8336
Atributo 3	1.7712	0.8274
Atributo 4	1.6313	0.8062
Atributo 5	1.7281	0.8211
Atributo 6	1.6709	0.8128
Atributo 7	1.8684	0.8349
Atributo 8	1.7726	0.8349
Atributo 9	2.1413	0.8346
Atributo 10	1.6776	0.8161
Atributo 11	1.6725	0.8140
Atributo 12	1.8684	0.8349
Atributo 13	2.0047	0.8065
Atributo 14	1.8861	0.8329
Atributo 15	1.6177	0.8065
Atributo 16	1.6639	0.8123
Atributo 17	1.6801	0.8166
Atributo 18	1.5649	0.7978
Atributo 19	1.7581	0.8237
Atributo 20	1.8002	0.8302
Atributo 21	2.1077	0.8347
Atributo 22	1.9322	0.8395
Atributo 23	1.9345	0.8335
Atributo 24	1.9719	0.8352
Atributo 25	1.9301	0.8335
Atributo 26	1.8990	0.8329
Atributo 27	1.9103	0.8330
Atributo 28	1.8870	0.8325

Atributo 29	2.0636	0.8345
Atributo 30	1.9284	0.8357
Atributo 31	2.0920	0.8358

Podemos comprobar que hay dos atributos que comparten un número similar de error cuadrático, que son el atributo 1 (ID) y el atributo 18 (Quant). Podemos descartar el atributo 1, ya que no es un atributo realmente fiable como para conseguir una buena predicción del salario. También podemos comprobar que los datos sin normalizar son más grandes que los normalizados, que según la teoría estaría correctamente realizado.

Ahora, comprobaremos las desviaciones de los atributos

```

Atributo ID : 364834.21522381395
Atributo Gender : 0.4263661350191836
Atributo DOB : 17467.259316938595
Atributo 10percentage : 10.001116331265484
Atributo 12graduation : 1.631542164480328
Atributo 12percentage : 11.118444304164754
Atributo CollegeID : 4775.813178150748
Atributo CollegeTier : 0.26400928688192105
Atributo Degree : 0.3378925865015805
Atributo Specialization : 1.400392696990963
Atributo collegeGPA : 8.121107182358415
Atributo CollegeCityID : 4775.813178150748
Atributo CollegeCityTier : 0.45657915712180364
Atributo CollegeState : 1.590999064570207
Atributo GraduationYear : 1.3079725899431502
Atributo English : 105.2869483425227
Atributo Logical : 87.28528880555089
Atributo Quant : 122.17457367587159
Atributo Domain : 0.4632959139420037
Atributo ComputerProgramming : 204.4921141671563
Atributo ElectronicsAndSemicon : 158.71127084420587
Atributo ComputerScience : 177.75488340868392
Atributo MechanicalEngg : 99.76849499611873
Atributo ElectricalEngg : 86.04038551005691
Atributo TelecomEngg : 103.53569152899614
Atributo CivilEngg : 32.23612312632463
Atributo conscientiousness : 1.02480300127156
Atributo agreeableness : 0.9556715205055258
Atributo extraversion : 0.9625346577114087
Atributo neuroticism : 1.0127322358347157
Atributo openness_to_experience : 1.006966053098755
Atributo Salary : 212295.77905147275

```

Para poder escoger los atributos con más dispersión, hemos decidido seleccionar aquellos que tuviesen un valor de dispersión por encima de los 4000. Estos atributos son los siguientes: 1 (ID), 3 (DOB), 7 (CollegeID), 12 (CollegeCityID).

Gracias a todos los datos obtenidos, podemos responder a las preguntas de este apartado.

Preguntas

- ¿Cuáles son los atributos más importantes para hacer una buena predicción?

Debemos fijarnos en aquellos atributos que tengan una relación entre los resultados obtenidos. Estos pueden ser los atributos 1, 3, 4, 7, 6, 12, 16, 17 y 18.

- ¿Con qué atributo se consigue un MSE menor?

Como hemos podido comprobar en los resultados anteriores, nos da únicamente el atributo 18 (Quant), aunque el valor que obtenemos de este no es tan próximo a 0 como quisiéramos.

- ¿Qué correlación hay entre los atributos de nuestra base de datos?

Según el heatmap sacado en el apartado C, podemos comprobar que aquellos datos que más o menos correlación tienen depende del color que representen. Cuanto más tono rojizo tiene, más correlación presentan los datos. Con respecto al salario (atributo objetivo), aquellos atributos que más relación tienen son (aceptando sólo un valor por encima del 0.17): 10percentage, 12percentage, English, Logical, Quant.

- ¿Cómo influye la normalización en la regresión?

Dados los resultados obtenidos en la tabla anterior, podemos observar que los resultados normalizados son inferiores a los sin normalizar, algo que nos indica que se hizo correctamente. Los datos están en un intervalo más o menos normal, por eso es que no varía mucho los resultados con o sin normalizar. Por tanto en este caso concreto, no hace falta normalizar los datos para conseguir buenos resultados.

- ¿Cómo mejora la regresión cuando se filtran aquellos atributos de las muestras que no muestran información?

Haciendo unas comprobaciones, pudimos saber que hay ciertos atributos que presentan -1 como valor. Estos atributos son:

ATRIBUTO	Nº VECES CON VALOR -1
Domain	179
ComputerProgramming	650
ElectronicsAndSemicon	2133
ComputerScience	2298
MechanicalEngg	2811
ElectricalEngg	2876
TelecomEngg	2724
CivilEngg	2972

Al hacer las filtraciones de estos atributos, los resultados no varían demasiado con respecto a los resultados obtenidos. Por tanto, podemos afirmar que no hay apenas mejoría en la regresión con estos cambios.

- Si se aplica un PCA, ¿a cuántos componentes se reduce el espacio? ¿Por qué?

Para poder visualizar los 34 atributos que tiene nuestra base de datos en un espacio visible, se podría aplicar un PCA (Principal Component Analysis) para reducir la dimensión del espacio a uno observable (como 2D o 3D). Si quisiéramos aplicarlo a nuestra BBDD, nos quedaría en un espacio de 2 dimensiones, porque se puede visualizar mejor, y podemos ver de forma más directa la relación que existe entre los 40 atributos de entrada de nuestra BBDD.

Conclusiones

En la vista de los datos que se han podido presentar en esta memoria podemos concluir que: no hay ningún o varios atributos en específico que permitan hacer una regresión para poder predecir el salario que cobrará el estudiante una vez graduado en función de sus aptitudes en el **AMCAT**, el grado o la especialización seleccionado o las notas que haya sacado.

En el apartado C podemos observar que no se ha encontrado ningún atributo que genere una campana de Gauss con el atributo salario que permita predecir este último en función de la evolución del primero, todo son campanas de Gauss que muestran dónde se sitúa la nota media de ciertas asignaturas y aptitudes para las distintas universidades. Mientras que en el **heat map** no muestra ninguna relación fuerte con ningún atributo en concreto, otra vez, si lo hace con las notas.

En el apartado B, al calcular la función de coste con los distintos atributos nos encontramos que nos salen errores prácticamente iguales para todos los atributos mostrando que ninguno tiene prácticamente relación con el salario. Por otra parte, los errores son muy grandes (cerca de 1) lo que implicaría así que la recta sacada de la regresión lineal no daría una correcta representación de los atributos.

Este tipo de resultados nos lleva a concluir que, después de que los estudiantes se gradúen, las empresas no tienen en consideración los esfuerzos hechos durante los estudios de los estudiantes, ni (menos mal) el lugar de procedencia o género; entienden que son principiantes y los contratan para formarlos en un entorno laboral y más acorde a lo que pide el mercado.

Problemas surgidos

Durante la realización de la práctica, hemos tenido ciertas dificultades que han ido apareciendo pero que, gracias a la ayuda de información del foro del campus e Internet, hemos podido solucionar gran parte de ellos. Pasamos a nombrar algunos de los más destacables:

- Tenemos una BBDD que combina valores tipo int, float y string, y no se puede hacer cálculos de regresión a variables que son de tipo string. Para solucionar este problema, hemos accedido a aquellos atributos que son de tipo string o valores raros y los hemos modificado según este criterio:

ATRIBUTO	CAMBIOS HECHOS
Gender	<ul style="list-style-type: none"> • Si el valor es 'm': pasa a valor 0 • Si el valor es 'f': pasa a valor 1
Degree	<ul style="list-style-type: none"> • Si el valor es 'B. Tech/B.E.': pasa a valor 1 • Si el valor es 'MCA': pasa a valor 2 • Si el valor es 'M.Tech/M.E.': pasa a valor 3 • Si el valor es 'M.Sc. (Tech)': pasa a valor 4
Specialization	<ul style="list-style-type: none"> • Si el valor es "electronics and communication engineering": pasa a valor 1 • Si el valor es "computer science & engineering": pasa a valor 2 • Si el valor es "information technology": pasa a valor 3 • Si el valor es "computer engineering": pasa a valor 4 • Si el valor es "computer application": pasa a valor 5 • Para cualquier otro valor: pasa a valor 0
CollegeState	<ul style="list-style-type: none"> • Si el valor es "Uttar Pradesh": pasa a valor 1 • Si el valor es "Karnataka": pasa a valor 2 • Si el valor es "Tamil Nadu": pasa a valor 3 • Si el valor es "Telangana": pasa a valor 4 • Si el valor es "Maharashtra": pasa a valor 5 • Para cualquier otro valor: pasa a valor 0
GraduationYear	<ul style="list-style-type: none"> • Todos los valores menores de 2000: pasa a valor 2010

Tanto en Specialization y CollegeState hemos decidido coger 4 o 5 valores base, y los demás valores los consideramos como "otros". Además, el atributo de GraduationYear presenta algunos años que son muy inferiores (como 1500), y los ajustamos porque no tienen sentido.

- Al hacer estos cambios, no hemos tenido en cuenta que al cambiar un tipo string a un tipo int no se reflejaba bien en el numpy array, pues en vez de considerarlos de tipo "integer" lo consideraba de tipo "object", y con este tipo de variable no podemos trabajar para realizar las regresiones. Decidimos hacer el cambio con una función de pandas llamada "replace".
- Nos ha dado problemas guardar los datos resultantes en un archivo de texto para guardarlo en el proyecto (datos cómo MSE, dispersión, etc). Simplemente decidimos printar los resultados en PyCharm o debugando pudimos acceder a los resultados.