Estudiante	**MARIA CECILIA ZUBIN**
Curso/Comisión	**61615**
Proyecto	**Segunda entrega – Proyecto final**

Temática a Resolver

En base a datos históricos de un trimestre, una empresa de servicios nos consulta con el objeto de predecir e inferir si el título universitario es un elemento fundamental al momento de acceder a puestos jerárquicos percibiendo un salario superior promedio más alto que las personas que ni tiene título universitario.

Adicionalmente, desean saber si la ubicación geográfica, en donde la empresa tiene establecimientos, está vinculada o no a la obtención de un título universitario completo.

Como tercera alternativa resulta interesante, analizar de la población que posee título universitario, que proporción pertenecen al sexo masculino y cuanta al sexo femenino.

La intención de la empresa con esta solicitud, es implementar un sistemas interno a través del departamento de desarrollo de RRHH un esquema de beneficios & compensaciones especialmente direccionado a promover e incentivar al personal a estudiar una carrera universitaria y/o finalizarla en los casos en los que se encuentre incompleta.

Resumen de Hipótesis

Hipótesis Nº1: El poseer un titulo universitario es un factor determinante para tener acceso a puestos jerárquicos, por ende, de mayor ingreso.

Hipótesis Nº2: La ubicación geográfica del empleado influye en la obtención de un título universitario.

Hipótesis Nº3: La población de sexo femenino que posee título universitario es mayor, que la población de sexo masculino,

Detalle de variables contempladas en el dataset

Legajo: numero identificador del empleado, ID univoco que no se repite dentro del dataset.

Estado: hace referencia a si el empleado se encuentra activo al momento del armado del dataset.

Estructura: detalla la estructura de liquidación bajo la cual se encuentra contenido el empleado. Los puestos asociados a staff, analistas, supervisores, responsables, etc., son los asociados al personal fuera de convenio. El personal que realiza tareas operativas detalladas explícitamente en una convención colectiva de trabajo nomenclada y registrada como tal, están asociados al personal dentro de convenio. [Fuera de convenio, Dentro de convenio]

Convenio: detalla los posibles convenios al que puede pertenecer el empleado [FCON,CC-781,ATCC]

ID_Convenio	Convenio del Empleado
33	FCON
22	CC-781
11	ATCC

Categoría: detalla la categoría a la que pertenece el empleado:

ID_Cat	Categoría	ID_Cat	Categoría
FCC	Fuera de Convenio	OP20	20 Hs. CATEGORIA OP. A CCT 781
H01	36,0 Hs. CATEGORIA 3 CCT688/14	OP24	24 Hs. CATEGORIA OP. A CCT 781
H03	35,0 Hs. CATEGORIA 3 CCT688/14	OP25	25 Hs. CATEGORIA OP. A CCT 781
H11	31,0 Hs. CATEGORIA 3 CCT688/14	OP27	27 Hs. CATEGORIA OP. A CCT 781
H13	30,0 Hs. CATEGORIA 3 CCT688/14	OP30	30 Hs. CATEGORIA OP. A CCT 781
H21	26,0 Hs. CATEGORIA 3 CCT688/14	OP35	35 Hs. CATEGORIA OP. A CCT 781
H23	25,0 Hs. CATEGORIA 3 CCT688/14	OP36	36 Hs. CATEGORIA OP. A CCT 781
H25	24,0 Hs. CATEGORIA 3 CCT688/14		
H30	21,5 Hs. CATEGORIA 3 CCT688/14		
H33	20,0 Hs. CATEGORIA 3 CCT688/14		

ID_Cat: código de la categoría del empleado.

Ingreso: fecha de ingreso del empleado a la empresa.

Sexo: sexo del empleado [M: Hombre, F: Mujer]

Provincia: sucursal en la que el empleado se encuentra actualmente prestando servicios [FORM,CBA, CABA2]

Puesto: se detallan las codificaciones y puestos de todo el personal

ID_PuestoBasico	Puesto
ACC	Analista de Calidad & Capacit.
ACE	Analista de Customer Experience
ACH	Analista Gestion Cap. Humano
ACI	Analista de Comunicaciones
ACS	Analista de Control C de Serv
ADP	Analista de Adminis. de Pers.
ADS	Analista de Servidores
AFR	Analista de Fraude
AGE	Analista de Gestion
AGS	Analista de Gestión Senior
AMA	Analista de Mesa de Ayuda
AMG	Analista de Modelo de Gestión
AMT	Analista de Mantenimiento
ANL	Analista
ANR	Analista de Novedades y RRHH
APA	Analista de Aplicaciones
APO	Analista de Planeamiento Oper.
ARF	Analista de Rep. & Facturacion
ARH	Analista de Recursos Humanos
ARL	Analista de Relaciones Labor.
ASE	Analista de Seleccion
AST	Analista de Soporte Tecnico
ASY	Analista de Selección y Desarrollo
CAL	Analista de Calidad
CAP	Coordinador Adm. de Personal
CDT	Operador
CFA	Coordinador de Facturación
CGC	Coordinador de Gestión Comercial
CLW	Coordinador de LAN / WAN
COO	Coordinador
CPA	Analista de Compras
CPO	Coordinador Planea. Operativo
CSE	Coordinador de Selecciòn & Emp
CSS	Coordinador de Servicios IT
CST	Coordinador de Soporte Tecnico
DIE	Director Ejecutivo
DIO	Director de Operaciones
EJU	Ejecutivo Comercial
FOR	Formador
GAF	Gerente de Administración & F
GCG	Gerente Ctrol. Gestion
GCM	Gerente de Comunicaciones y Marketing
GCO	Gerente Comercial
GEO	Gerente Operaciones
GLE	Gerente de Legales
GMI	Gerente Mantenim. e Infraest.
GRH	Gerente de RRHH
GST	Gerente de Sistemas & Tecnolog
JAN	Jefe Administracion de Personal y Novedades
JCR	Jefe de Comunicaciones & RSE
JDC	Jefe Comercial
JDS	Jefe de Desarrollo de Sistemas
JFD	Jefe Formacion & Desarrollo
JGP	Jefe de Contrrol de Gestión y Plan..Operativo
JNS	Jefe de NOC y Seguridad de la Información
JOP	Jefe de Operaciones
JRM	Jefe RRLL & Medic. Laboral
JSH	Jefe Seguridad e Higiene
LDT	Lider de Telefonia
REC	Recepcionista
RSG	Responsable Servicios Grales.
SCA	Supervisor de Calidad
SCC	Supervisor Calidad & Capacit.
SCL	Supervisor de Operaciones
SEO	Analista Soporte Estratégico Operativo
SIM	Supervisor de Infraestructura & Mantenimiento
SPO	Supervisor de Plan. Operativo
SPR	Supervisor de Reporting
SUN	Supervisor de Novedades
SWF	Supervisor de Workforce
TLC	Team Leader Operaciones

Pirámide: detalla el nivel de jerarquía en el organigrama de la empresa, siendo 8 el valor que se ubica en la base de la pirámide y 0 el ultimo valor de la cima de la misma [0,1,2,3,4,5,6,7,8]

Título: hace referencia a la posesión del mismo [SI,NO]

ID_PuestoBasico: código del puesto del empleado.

ID Banco: código del banco donde el empleado percibe sus haberes.

Banco: detalle de banco donde el empleado percibe sus haberes:

ID_Banco	Banco Pago de Rem
7	BANCO DE GALICIA Y BUENOS AIRES S.A.
11	BANCO DE LA NACION ARGENTINA
15	INDUSTRIAL AND COMMERCIAL BANK OF CHINA
17	BBVA BANCO FRANCES S.A.
20	BANCO DE LA PROVINCIA DE CORDOBA S.A.
27	BANCO SUPERVIELLE S.A.
44	BANCO HIPOTECARIO S.A.
72	BANCO SANTANDER RIO S.A.
150	HSBC BANK ARGENTINA S.A.

Análisis Exploratorio de Datos (EDA)

El EDA se centrará en:

• Resumen estadístico de las variables clave.

• Visualización de distribuciones de ingreso, jerarquía de puesto y título universitario.

```
Legajo              object
Estado              object
Estructura          object
Convenio            object
Categoria           object
ID_Cat              object
Ingreso             object
Sexo                object
provincia           object
puesto              object
piramide             int64
Titulo              object
ID_PuestoBasico     object
ID_Banco             int64
Banco Pago de Rem   object
Enero                int64
Febrero              int64
Marzo                int64
dtype: object
```

Se detalla el tipo de cada variable.

Estructura del dataset

```
[150] df.shape

(2205, 18)
```

El dataset obtenido posee 2205 filas y 18 columnas

Análisis de existencia de valores duplicados.

```
df.duplicated().value_counts()
```

	count
False	2205

dtype: int64

El dataset no posee filas repetidas.

Análisis de existencia de posibles valores nulos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2205 entries, 0 to 2204
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Legajo             2205 non-null   object
 1   Estado             2205 non-null   object
 2   Estructura         2205 non-null   object
 3   Convenio           2205 non-null   object
 4   Categoria          2205 non-null   object
 5   ID_Cat             2205 non-null   object
 6   Ingreso            2205 non-null   object
 7   Sexo               2205 non-null   object
 8   provincia          2205 non-null   object
 9   puesto             2205 non-null   object
 10  piramide           2205 non-null   int64
 11  Titulo             2205 non-null   object
 12  ID_PuestoBasico    2205 non-null   object
 13  ID_Banco           2205 non-null   int64
 14  Banco Pago de Rem  2205 non-null   object
 15  Enero              2205 non-null   int64
 16  Febrero            2205 non-null   int64
 17  Marzo              2205 non-null   int64
dtypes: int64(5), object(13)
memory usage: 310.2+ KB
```

Propuesta de Visualizaciones para la primera etapa del trabajo

1. Boxplot

2. Histograma

3. Pie Chart

Primeras conclusiones

El presente trabajo fue realizado bajo el concepto de estudio solicitado inicialmente, realizando una única iteración sobre las variables elegidas, los motivos y la escueta selección de modelos.

En esta primera instancia, las hipótesis se cumplieron.

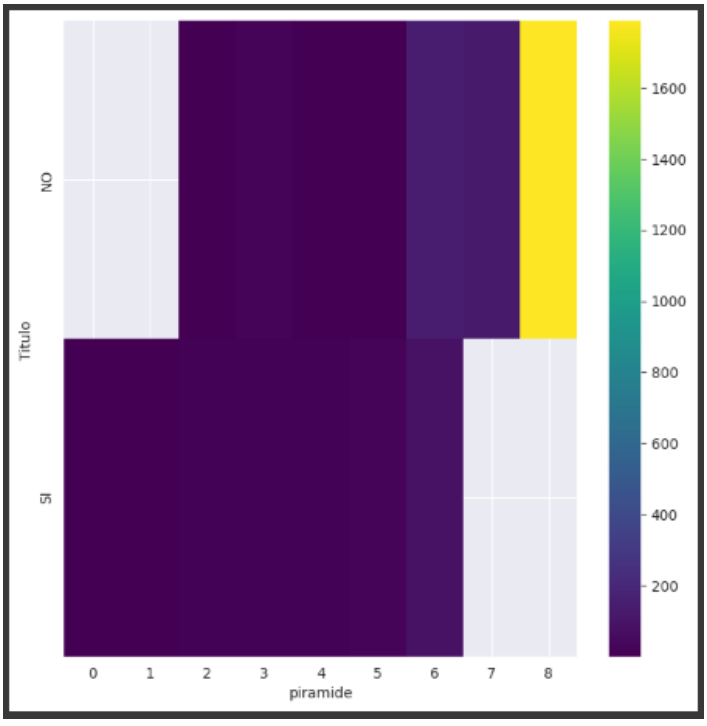
Feature Engineering

En esta segunda interación, se dispuso a encodear las variables categóricas ordinales en enteros representativos y las nominales se binarizaron en 0 y 1, quedando el dataset de la siguiente manera:

	Legajo	Estado	Estructura	Convenio	Categoria	ID_Cat	Ingreso	Sexo	provincia	puesto	piramide	Titulo
0	231	ACTI	Fuera de convenio	FCON	Fuera de Convenio	FCC	8/9/2015	1	FORM	Analista de Mantenimiento	6	0
1	236	ACTI	Dentro de convenio	CC-781	27 Hs. CATEGORIA OP. A CCT 781	OP27	14/9/2015	0	FORM	Operador	8	0
2	1420	ACTI	Fuera de convenio	FCON	Fuera de Convenio	FCC	2/2/2009	0	CBA	Team Leader Operaciones	7	0
3	1444	ACTI	Dentro de convenio	ATCC	36,0 Hs. CATEGORIA 3 CCT688/14	H01	2/2/2009	0	CBA	Operador	8	0

Analisis de las variables piramide vs. titulo

Del grafico surge la relación entre la variable pirámide y la variable titulo, confirmando que quienes poseen titulo universitario, efectivamente acceden a puestos de mayor jerarquía, ubicados en la parte superior de la pirámides.



Baseline

Como primera iteración se entrenará el modelo mediante un árbol de decisión clásico y un Random Forest sin optimizar hiperparámetros ni utilizar validación cruzada.

Test - Train - Validation Split

Se destinará un 80% del dataset para entrenamiento y un 20% para el testeo. El 10% del set de entrenamiento será destinado para el set de validación.

Métricas

El modelo que mejor performó fue el Random Forest con su configuración estándar, obteniéndose un accuracy de 0.99 y un recall de 0.92 para la clase 1.

Primeras conclusiones

El presente trabajo fue realizado bajo el concepto de estudio de factibilidad, realizando una única iteración sobre las variables elegidas, los motivos y la escueta selección de modelos. Aun así, se concluye que la clasificación es factible de ser realizada, habiendo alcanzado valores de métricas altos, susceptibles de ser utilizados en la hipótesis planteada inicialmente.

Siguiente instancia

Habiendo obtenido un baseline en la primera iteración que valida la hipótesis de factibilidad de la clasificación, se realizará una segunda iteración utilizando otros modelos y mejorando los hiperparámetros

El objetivo es analizar si realmente el punto de corte por default para la clasificación binaria, es realmente el punto de corte ideal que maximiza el TPR (recall), que - junto con el accuracy, son las dos principales métricas de negocio a optimizar. Las nuevas métricas y modelos se evaluarán bajo la misma estratégica de encodeo y feature engineering utilizada en la primera iteración, trabajando sobre el dataset balanceado obtenido en la primera entrega

Modelos a implementar

Se probarán y optimizarán hiperparámetros sobre los siguientes modelos:

Random forest: se optimizó la cantidad de estimadores a utilizar en función del estancamiento del accuracy en validación y la caída del mismo en train.

Tratandose de una clasificacion binaria, el modelo elegido para realizar las pruebas es el de Random Forest. Se trata modelo flexible, preciso y fácil de ajustar que es robusto tanto para tareas de clasificación como de regresión. Ofrece una buena capacidad de generalización al reducir el riesgo de sobreajuste que suele afectar a los árboles de decisión individuales.

```
Exactitud del modelo: 0.9931972789115646

Reporte de clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       415
           1       0.96      0.92      0.94        26

    accuracy                           0.99       441
   macro avg       0.98      0.96      0.97       441
weighted avg       0.99      0.99      0.99       441
```

Árbol de decisión: se decidió achicar la profundidad del árbol de 12 a 6, debido a que se encontraba copiando el set de entrenamiento, comportamiento que se comprobó mediante el accuracy y la función de pérdida utilizada, la cuál divergía considerablemente en el set de validación.

Es una estructura similar a un diagrama de flujo, donde cada nodo representa una característica (o una decisión), cada rama representa un resultado de esa característica y cada hoja representa una etiqueta de clase o valor, son modelos interpretables y fáciles de implementar tanto para clasificación como para regresión. Aunque pueden sufrir de sobreajuste, ajustando parámetros como la profundidad máxima o el número mínimo de muestras por nodo, puedes mejorar su rendimiento. Además, puedes visualizar el árbol de decisión, lo que facilita la interpretación de las reglas de decisión del modelo.

```
Exactitud del modelo: 0.9909297052154195

Reporte de clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       415
           1       0.92      0.92      0.92        26

    accuracy                           0.99       441
   macro avg       0.96      0.96      0.96       441
weighted avg       0.99      0.99      0.99       441
```

En esta tabla es posible observar los beneficios de haber optimizado los hiperparámetros sobre el set de validación. Se observa un incremento en la exactitud de todos los modelos utilizados respecto al baseline sobre el set de test, siendo el Random Forest nuevamente el que mejor performa. Se observa que nuevamente el Random Forest es el modelo que maximiza el Recall.

Se mejoraron todas las métricas obtenidas aplicando la optimización manual de hiperparámetros Se comprobó a su vez que tanto una reducción en la dimensionalidad como un cambio en el umbral de cutoff no varían la performance del Random Forest, siendo éste el mejor modelo obtenido hasta el momento.

Conclusión Final

Los valores de ramndom forest, en cuando al accurancy y al recall, reflejan con claridad que es superior al árbol de decisión, por lo tanto se opta por el random forest, debido a la precisión y la robustez del modelo, y no se requiere una gran interpretabilidad.