Segunda entrega de proyecto

POR:

Claudia Yaneth Giraldo Vergara

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raúl Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA MEDELLÍN

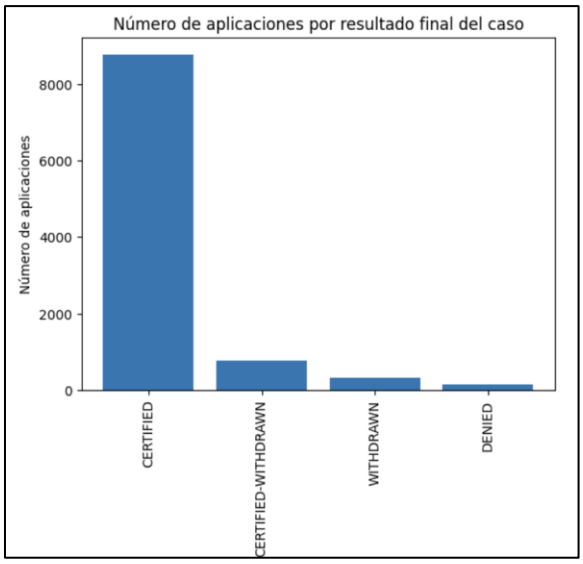
2023

PROYECTO: PREDICCIÓN DE VISAS

1. Exploración de datos

Vistas de los casos de forma general

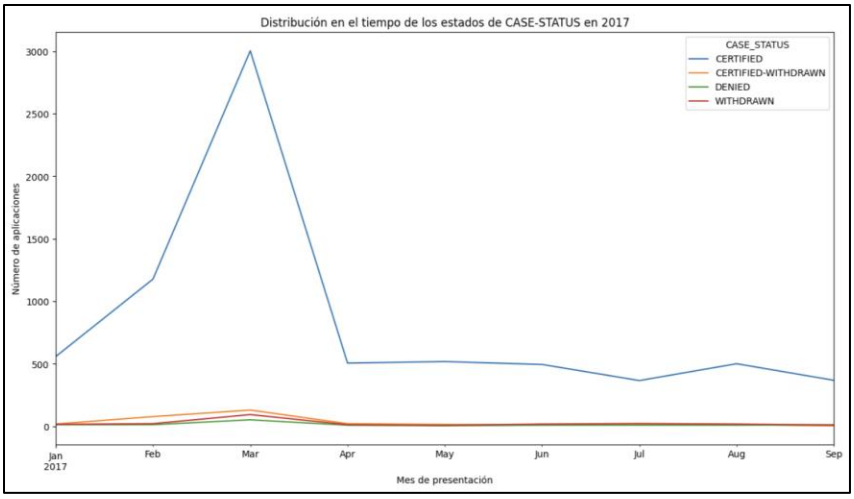
En primer lugar, se quiere observar cómo se distribuyen las decisiones finales a las solicitudes que hacen las personas al gobierno de Estados Unidos para obtener una visa laboral.



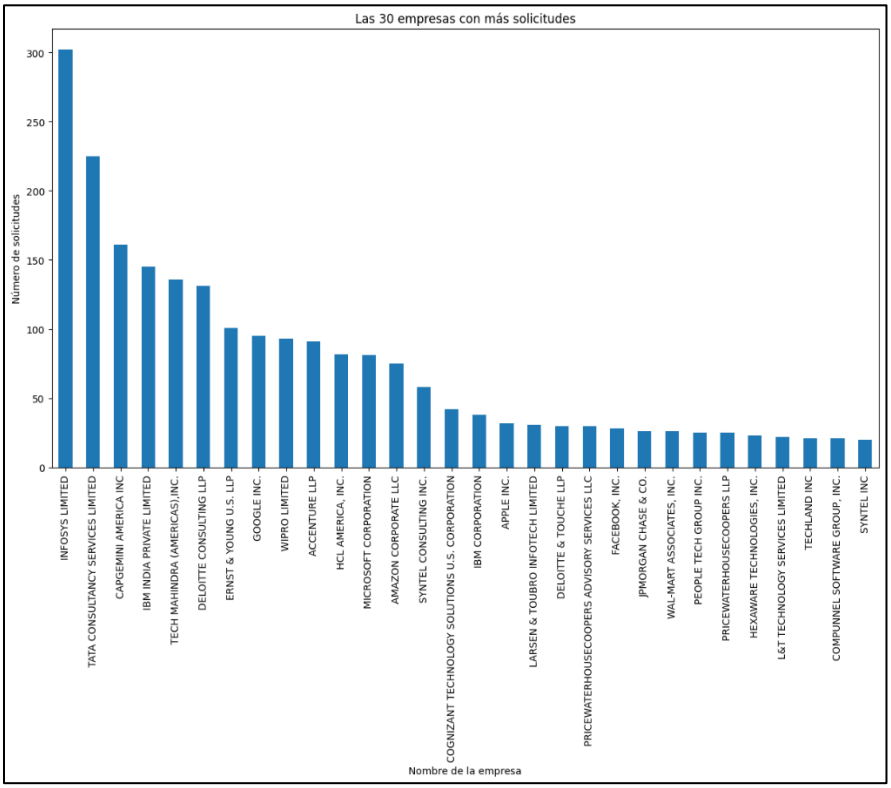
La mayoría de las solicitudes se encuentran en estado CERTIFICADO (con 8776), le siguen CERTIFICADO-RETIRADO (con 767), RETIRADO (con 314) y DENEGADO (con 143).

De los 10 mil datos, solo un 1.43% de los casos totales son denegados y más de un 87% son certificados, eso significa que hay un gran porcentaje de aceptación para este tipo de visa.

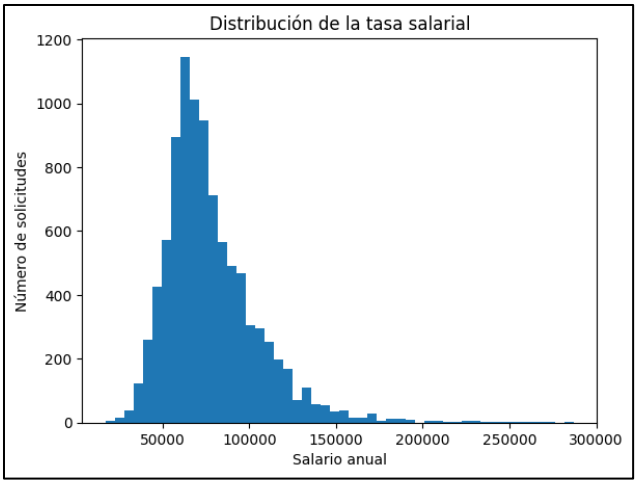
Vista de los casos por distribución en el tiempo



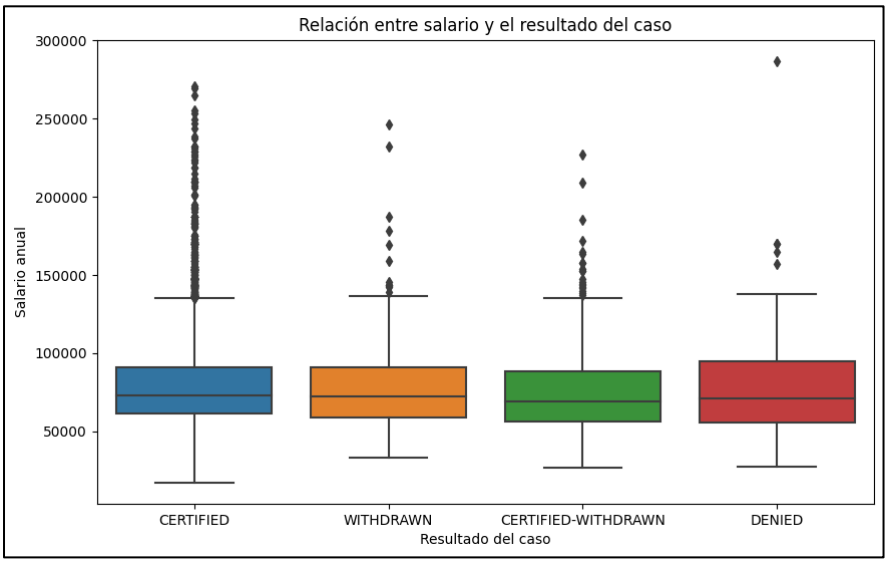
Se observa un aumento considerable de las aplicaciones entre los meses febrero y marzo, la gran mayoría de estas solicitudes obtienen el valor de 'CERTIFICADO'.



Observando este gráfico se encuentra algo interesante, y es que 4 de las 5 empresas con más solicitudes de visa laboral son empresas de India (Infosys, Tata Consultancy, IBM India y Tech Mahindra).



La distribución de la tasa salarial sigue una distribución normal, con una media ligeramente por encima de los 50 mil dólares anuales. Según la Oficina de Estadísticas Laborales (BLS, por sus siglas en inglés), el salario medio de los trabajadores en los Estados Unidos es de 54 mil dólares anuales por lo que se puede asegurar que este promedio aplica también para los extranjeros.



Mediante este boxplot se puede confirmar que el salario influye en el resultado del caso, se observa que, aunque para los 4 posibles resultados la media ronda entre los 50 mil y 100 mil dólares, los valores atípicos del resultado 'CERTITIFICADO' van dese 140 mil (aproximadamente) hasta más de 250 mil dólares anuales, es decir, entre más alto sea el salario, más probable es conseguir una visa laboral en Estados Unidos.

2. Preprocesado de datos

Se convierten los valores o etiquetas de forma numérica para que sean legibles para el algoritmo. Las variables 'CASE_STATUS', 'FULL_TIME_POSITION', 'PW_WAGE_LEVEL', 'H1B_DEPENDENT', 'WILLFUL_VIOLATOR', 'AGENT_REPRESENTING_EMPLOYER' son fáciles de convertir, mientras que 'SOC_NAME' y 'EMPLOYER_NAME' son más complejas.

	CASE_STATUS	AGENT_REPRESENTING_EMPLOYER	FULL_TIME_POSITION	PW_WAGE_LEVEL	H1B_DEPENDENT	WILLFUL_VIOLATOR	SOC_N	EMP_N
0	0	0	1	1	0	0	11	12
1	0	1	1	1	0	0	11	12
2	0	1	1	3	1	0	11	14
3	0	1	1	3	0	0	11	12
4	0	0	1	2	1	0	11	10

3. Modelo de prueba

El primer algoritmo ha utilizar será la regresión logística, el desempeño con este primer modelo de prueba es el siguiente:

	precision	recall	f1-score	support
0	0.88	1.00	0.93	4379
1	0.00	0.00	0.00	388
2	0.00	0.00	0.00	159
3	0.00	0.00	0.00	74
accuracy			0.88	5000
macro avg	0.22	0.25	0.23	5000
weighted avg	0.77	0.88	0.82	5000

Se encuentra que el algoritmo es capaz de predecir con buena precisión el valor 0 ‘CERTIFIED’ mientras que no es capaz de predecir lo demás valores, esto significa que se debe mejorar el modelo o cambiar el algoritmo.