1.การวิเคราะห์ข้อมูล และสาเหตุของการเลือก features ในการนำมา train

เลือกข้อมูลของ ds_salaries.csv มาวิเคราะห์ชุดข้อมูล

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L
1	2020	SE	FT	Machine Learning S	260000	USD	260000	JP	0	JP	S
2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M
3	2020	MI	FT	Product Data Analys	20000	USD	20000	HN	0	HN	S
4	2020	SE	FT	Machine Learning E	150000	USD	150000	US	50	US	L
5	2020	EN	FT	Data Analyst	72000	USD	72000	US	100	US	L
6	2020	SE	FT	Lead Data Scientist	190000	USD	190000	US	100	US	S
7	2020	MI	FT	Data Scientist	11000000	HUF	35735	HU	50	HU	L
8	2020	MI	FT	Business Data Analy	135000	USD	135000	US	100	US	L
9	2020	SE	FT	Lead Data Engineer	125000	USD	125000	NZ	50	NZ	S
10	2020	EN	FT	Data Scientist	45000	EUR	51321	FR	0	FR	S
11	2020	MI	FT	Data Scientist	3000000	INR	40481	IN	0	IN	L
12	2020	EN	FT	Data Scientist	35000	EUR	39916	FR	0	FR	M
13	2020	MI	FT	Lead Data Analyst	87000	USD	87000	US	100	US	L
14	2020	MI	FT	Data Analyst	85000	USD	85000	US	100	US	L
15	2020	MI	FT	Data Analyst	8000	USD	8000	PK	50	PK	L
16	2020	EN	FT	Data Engineer	4450000	JPY	41689	JP	100	JP	S
17	2020	SE	FT	Big Data Engineer	100000	EUR	114047	PL	100	GB	S
18	2020	EN	FT	Data Science Consu	423000	INR	5707	IN	50	IN	M
19	2020	MI	FT	Lead Data Engineer	56000	USD	56000	PT	100	US	M
20	2020	MI	FT	Machine Learning E	299000	CNY	43331	CN	0	CN	M
21	2020	MI	FT	Product Data Analys	450000	INR	6072	IN	100	IN	L
22	2020	SE	FT	Data Engineer	42000	EUR	47899	GR	50	GR	L
23	2020	MI	FT	BI Data Analyst	98000	USD	98000	US	0	US	M
24	2020	MI	FT	Lead Data Scientist	115000	USD	115000	AE	0	AE	L
25	2020	EX	FT	Director of Data Scie	325000	USD	325000	US	100	US	L
26	2020	EN	FT	Research Scientist	42000	USD	42000	NL	50	NL	L
27	2020	SE	FT	Data Engineer	720000	MXN	33511	MX	0	MX	S
28	2020	EN	CT	Business Data Analy	100000	USD	100000	US	100	US	L

มีทั้งหมด 607 ชุดข้อมูลในการทดสอบเพื่อหาว่า company_size อยู่ในระดับอะไร มี S (small) , M (medium) และ L (large) ในข้อมูลแต่ละชุดจะประกอบไปด้วย work_year, experience_level, employment_type, job_title, salary หน่วย salary_currency, salary_in_usd, employee_residence, remote_ratio และ company_location

ไม่เลือกส่วนของ salary เป็น feature เพราะหน่วยที่ไม่เหมือนกันในแต่ละชุดข้อมูลตามค่าเงินนั้นๆ แต่มี feature ที่ทำงานเหมือนกันคือ salary_in_usd ที่ทำการแปลงมาจาก salary ให้เป็นหน่วยของ USD ทั้งหมด

2.สาเหตุที่เลือก learning algorithm มาใช้

เลือกใช้ 2 learning algorithm คือ

1. K-Nearest Neighbors (KNN)

- optimal parameter แต่ใช้ hyper parameter ที่สุ่มค่าและดู error ที่เกิดขึ้น
- ทำงานง่าย ไม่มีการประมวลผลข้อมูล จึงเหมาะกับข้อมูลที่มี label เยอะๆ
- เหมาะกับงานพวกจำแนกข้อมูล จำแนกใบหน้า หนวด ผม

2. Artificial Neural Network (ANN)

- เป็นโครงข่ายของโหนดการคำนวณแบบง่าย ๆ แต่ละโหนดจะรับค่าตัวแปร คูณกับค่าน้ำหนักรวม และส่งต่อไปยังโหนดถัดไป

3.ผลการ testing model

1. K-Nearest Neighbors

```
from numpy import mean, std
print('Accuracy: %.3f (%.3f)' % (mean(scores), std(scores)))
```

```
Accuracy: 0.691 (0.021)
```

2. Artificial Neural Network (ANN)

```
results = model.evaluate(X_test, y_test)
print('Accuracy: %.3f' % results[1])
```

```
13/13 [==============================] - 0s 2ms/step - loss: 0.3233 - accuracy: 0.5151
Accuracy: 0.515
```

4.เปรียบเทียบผลของ 2 model พร้อมอธิบายสาเหตุว่า วิธีที่ให้ประสิทธิภาพสูงกว่าเพราะอะไร

จากการทดลองจะพบว่า Accuracy จะเป็น

K-Nearest Neighbors (KNN) > Artificial Neural Network (ANN)

- ส่วน **K-Nearest Neighbors (KNN)** มีค่า Accuracy ที่น้อยกกว่าแต่ไม่ได้ห่างกันเกินไป แปลว่าวิธีนี้ได้ผลลัพธ์ที่ดีเหมือนกัน
- **Artificial Neural Network (ANN)** ที่น้อยอาจจะเป็นเพราะมีการทำงานของ algorithm ไม่ดีพอ หรือชุดข้อมูลไม่มากพอ

5.สรุปผล

ชุดข้อมูลที่เลือกมานั้นมี feature ที่ใช้งานอยู่ 2 อย่างคือ work_year = ปีที่มีการทำงาน employment_type = สายงาน ทั้ง 2 อย่างนี้มีผลในการตัดสินใจในการทำนายได้ว่า company size ที่เลือกนั้นควรเป็นระดับไหน มีทั้ง S (small) , M (medium) และ L (large)

model ที่ใช้ในการพิจารณาในการนำมาสร้าง และมีค่า Accuracy ที่สูงมากพอในการทำนาย คือ K-Nearest Neighbors (KNN) K-Nearest Neighbors (KNN) ที่เหมาะจะนำมาทำงานด้วยเหมือนกัน