Mini Project

Machine Learning

นาย ธนภูมิ อังอำนวยศิริ 6201011631072

นาย โสภณ สุขสมบูรณ์ 6201011631188

นักศึกษาชั้นปีที่ 4 สาขา วิศวกรรมไฟฟ้า (โทรคมนาคม)

เสนอ

ผศ.ดร. คณบดี ศรีสมบูรณ์ (KDS)

รายงานเล่มนี้เป็นส่วนหนึ่งของวิชา 010113713 Machine Learning

ประจำภาคการศึกษา 1/2565 สาขา วิศวกรรมไฟฟ้า (โทรคมนาคม)

ภาควิชา วิศวกรรมไฟฟ้าและคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

1.การวิเคราะห์ข้อมูล และสาเหตุของการเลือก features ในการนำมา train

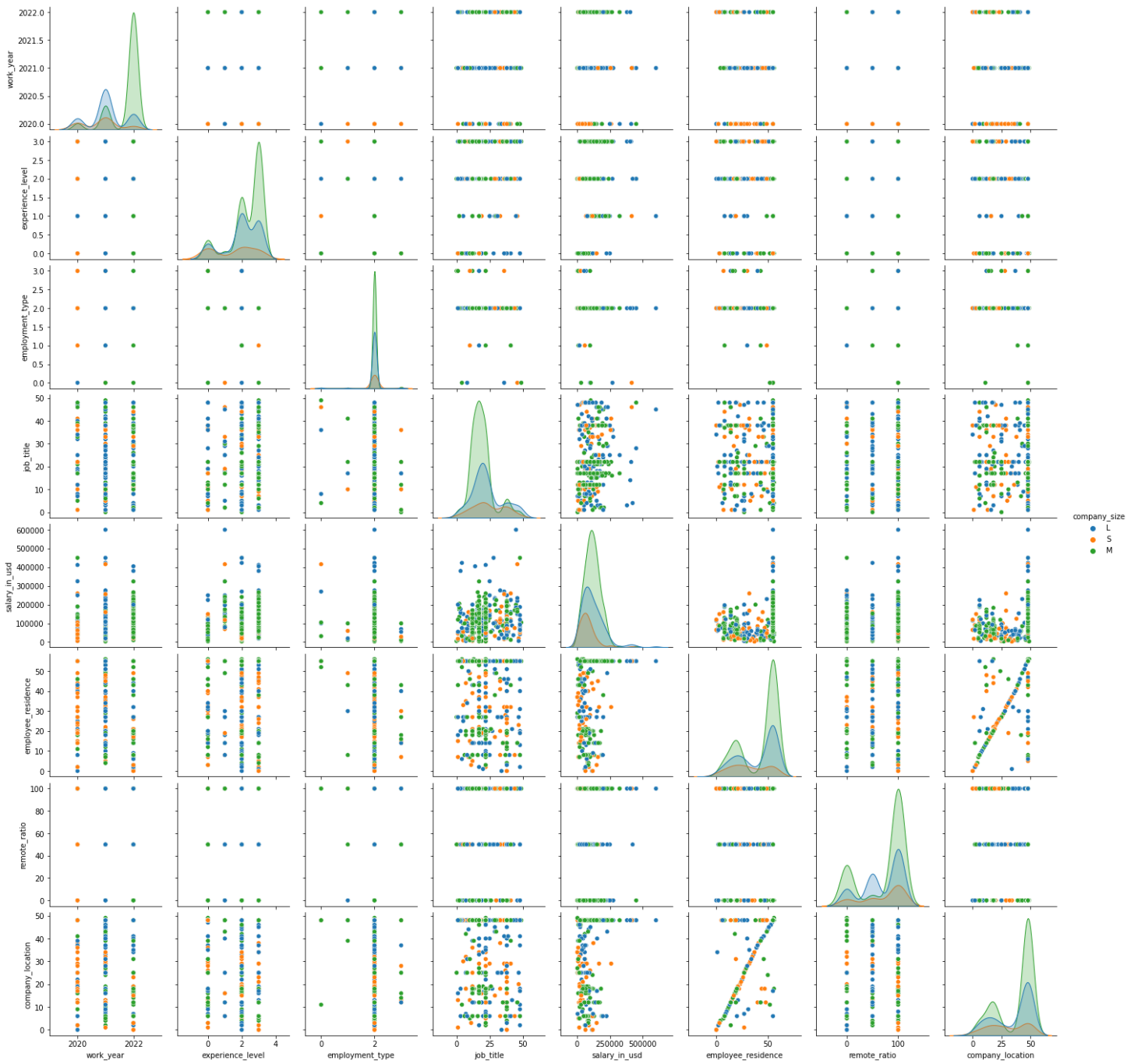
เลือกข้อมูลของ ds_salaries.csv มาวิเคราะห์ชุดข้อมูล

	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

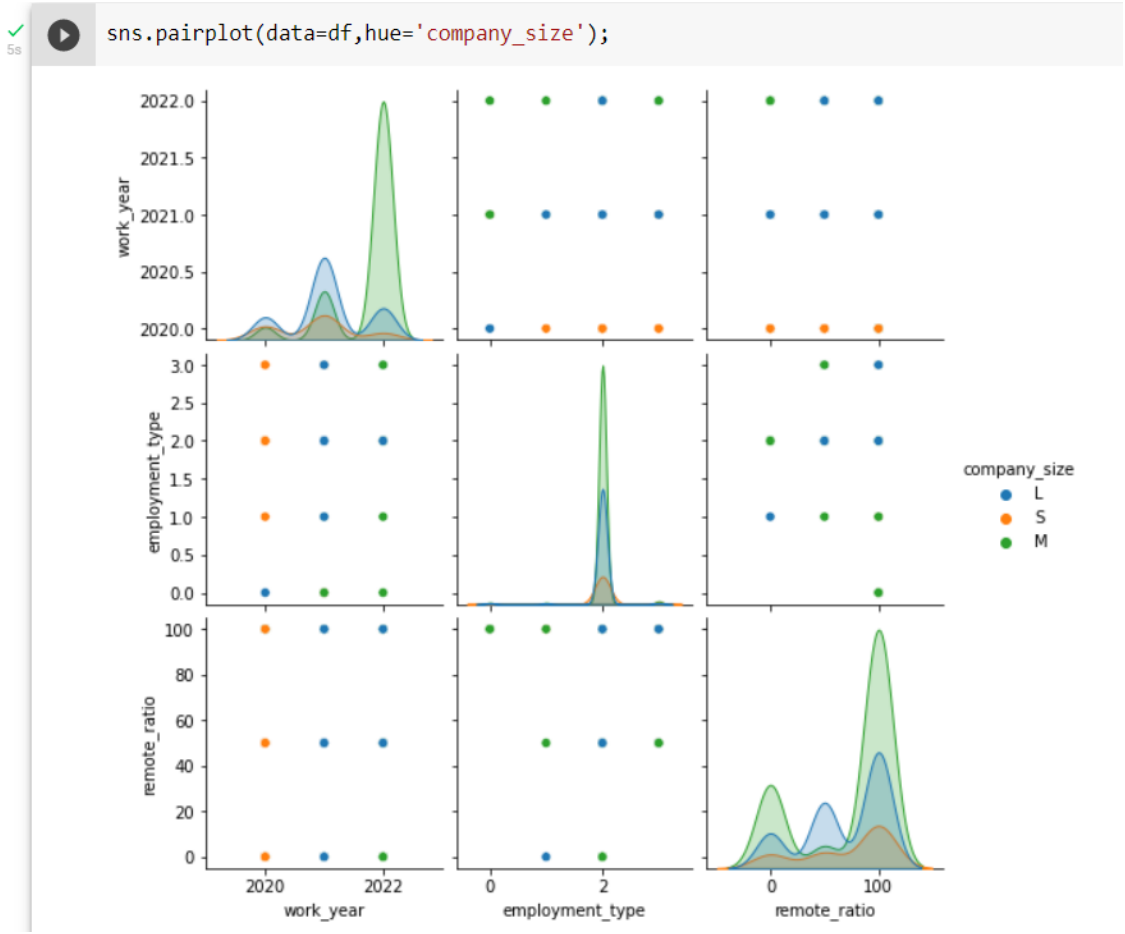
มีทั้งหมด 607 ชุดข้อมูลในการทดสอบเพื่อหาว่า company_size อยู่ในระดับอะไร มี S (small) , M (medium) และ L (large) ในข้อมูลแต่ละชุดจะประกอบไปด้วย work_year, experience_level, employment_type, job_title, salary หน่วย salary_currency, salary_in_usd, employee_residence, remote_ratio และ company_location

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

ทำการนำ feature ทั้ง 8 feature มาแสดงความสัมพันธ์โดยใช้ pairplot ของ seaborn



จากการดูความสัมพันธ์และวิเคราะห์ feature เลือก work_year, employment_type และ remote_ratio ในการนำมาวิเคราะห์



feature ที่เลือก 3 feature มาแสดงความสัมพันธ์โดยใช้ pairplot ของ seaborn

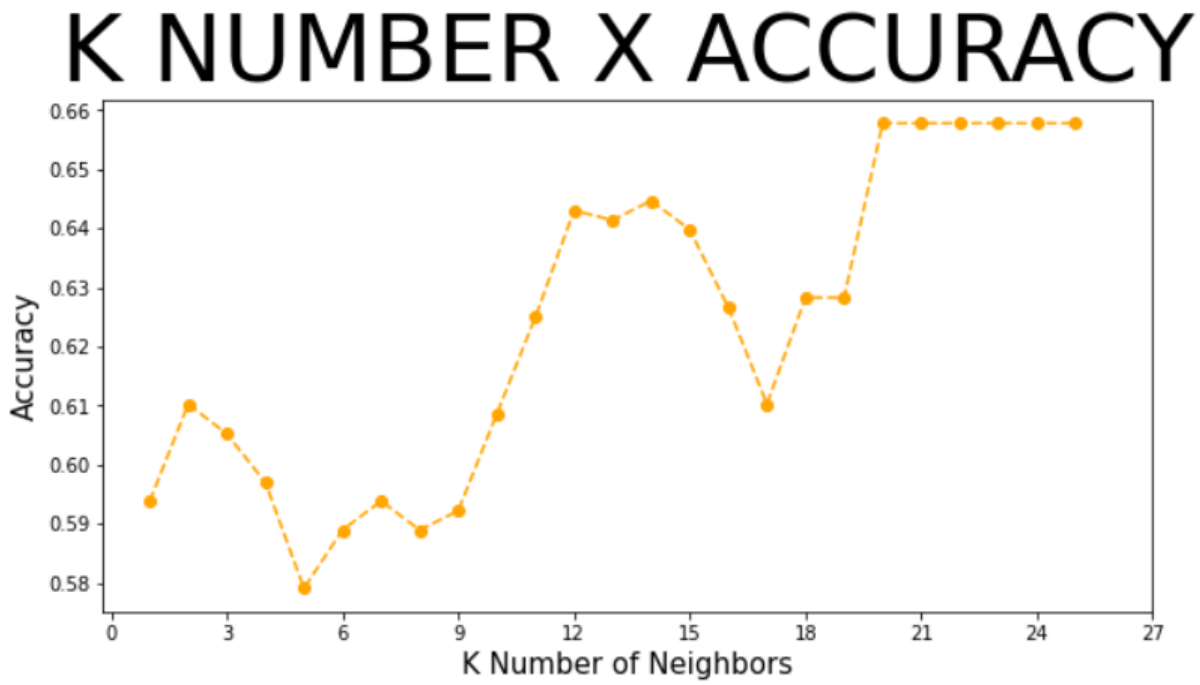
2.สาเหตุที่เลือก learning algorithm มาใช้

เลือกใช้ 3 learning algorithm คือ

1. K-Nearest Neighbors (KNN)

- ใช้หลักการเปรียบเทียบข้อมูลที่สนใจกับข้อมูลอื่นว่ามีความคล้ายคลึงมากน้อยเพียงใด หากข้อมูลที่กำลังสนใจนั้นอยู่ใกล้ข้อมูลใดมากที่สุด ระบบจะให้คำตอบเป็นเหมือนคำตอบของข้อมูลที่อยู่ใกล้ที่สุดนั้น

- เป็น lazy learning คือไม่ต้องทำงาน training ชุดข้อมูล สามารถนำไปใช้ได้เลย ไม่มีการหาค่า optimal parameter แต่ใช้ hyper parameter ที่สุ่มค่าและดู error ที่เกิดขึ้น
- ทำงานง่าย ไม่มีการประมวลผลข้อมูล จึงเหมาะกับข้อมูลที่มี label เยอะๆ
- เหมาะกับงานพวกจำแนกข้อมูล จำแนกใบหน้า หนวด ผม



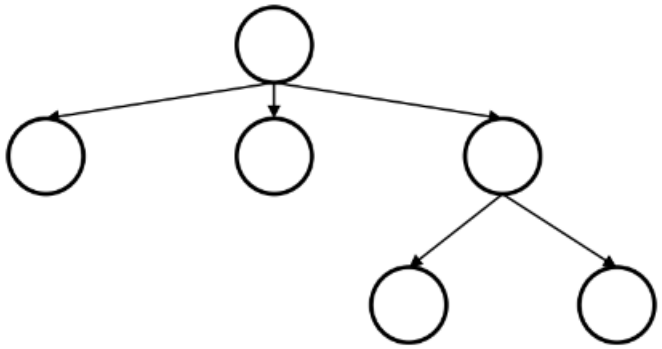

```
The best value of k = {'n_neighbors': 20} with 0.657797046470668 of accuracy.
```

ทดลองหาค่า k ที่เหมาะสมที่สุด และนำมาปรับจูนพารามิเตอร์ของแบบจำลอง เพื่อทดสอบหาประสิทธิภาพของแบบจำลอง สำหรับค่าพารามิเตอร์ต่าง ๆ ร่วมกับการประเมินผลด้วย cross-validation และทดลองพล็อตกราฟค่าประสิทธิภาพออกมา จะได้ค่า K ที่เหมาะสมเท่ากับ 20

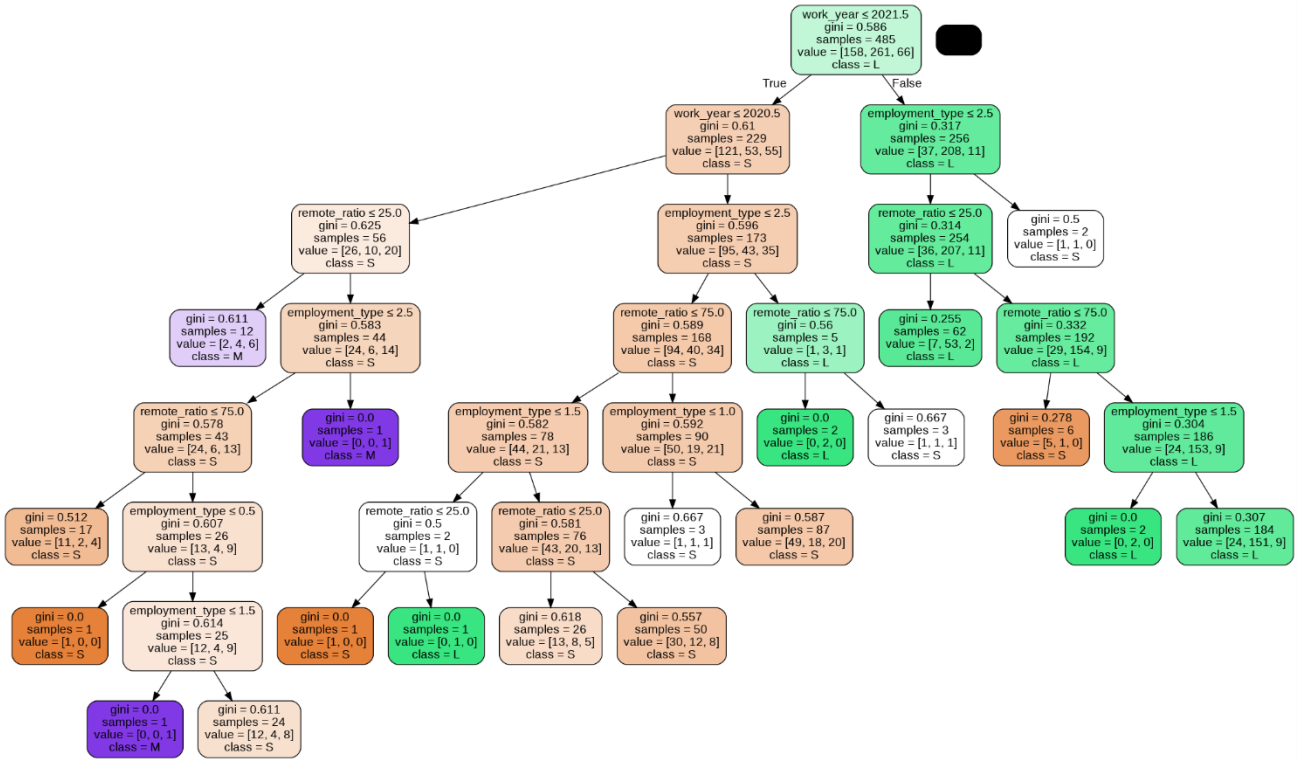
<u>เหตุผลที่เลือก</u> เพราะวิธีนี้น่าจะเหมาะกับชุดข้อมูลที่มีความซับซ้อนมาก และเมื่อดูความสัมพันธ์ระหว่างชุดข้อมูลที่มีอยู่นั้นมีการแบ่งแยกโซนที่พอจะทำงานแบ่งการทำงานได้ดี ว่าโซนนี้เหมาะกับ company_size แบบไหน

2. Decision Tree

- เป็นการจำลองที่ไม่เป็นเชิงเส้น ใช้การทำงานคล้าย if-else ในการนำมาวิเคราะห์ข้อมูล



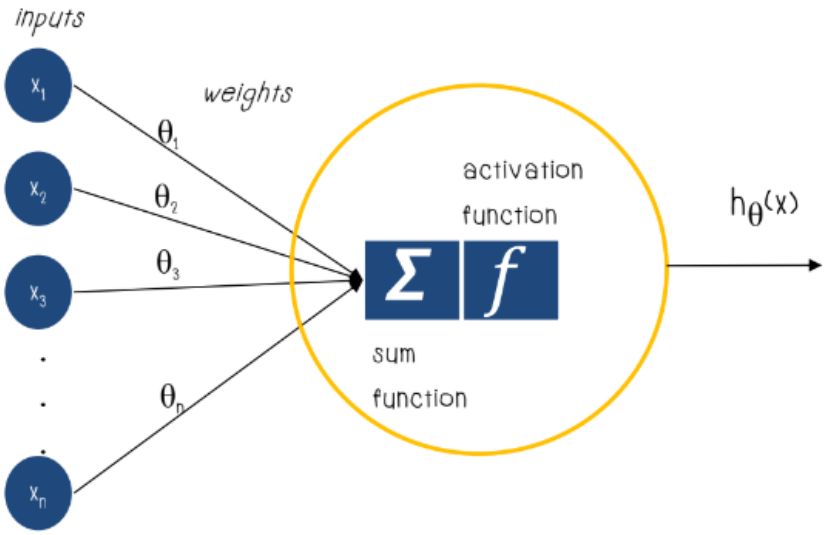
มีการสร้าง Decision tree ด้วยวิธี **gini index** หรือ **Classification and Regression Tree (CART)** ที่ง่ายต่อการสร้าง Decision Tree แต่ไม่ค่อยนิยมเท่ากับ ID3 (Iterative Dichotomiser3)



<u>เหตุผลที่เลือก</u> เพราะชุดข้อมูลที่เลือกในการใช้งานทั้ง 3 feature นั้นมีความสำคัญในการแบ่งข้อมูลไม่เท่ากัน จากรูปจะเห็นว่าการแบ่งกลุ่มจะเริ่มจากการทำงานของ work year เป็น Root node ไล่ไปเรื่อยๆตาม feature เพื่อทำนายว่าข้อมูลในอนาคตที่จะแบ่ง company size เป็นแบบไหน

3. Artificial Neural Network (ANN)

- เป็นโครงข่ายของโหนดการคำนวณแบบง่าย ๆ แต่ละโหนดจะรับค่าตัวแปร คูณกับค่าน้ำหนักรวม และส่งต่อไปยังโหนดถัดไป



- เป็นการที่ค่อยๆเรียนรู้ให้มีการพัฒนาไปเรื่อยๆ ค่าของ weight ต่าง ๆ ในโครงข่ายจะพยายามปรับเพื่อที่จะทำให้การคำนวณเป็นจริง คือถ้า input vector นี้เข้ามา output vector จะต้องเป็นแบบนี้

<u>เหตุผลที่เลือก</u> เพราะชุดข้อมูลมีจำนวน 607 ชุดข้อมูลซึ่งน่าจะมากพอในการเทรนให้ algorithm มีความแม่นยำในการทำนายได้ว่าข้อมูลที่เข้ามาควรจะเป็นแบบใด เพราะวิธีนี้ยิ่งมีชุดข้อมูลที่เยอะ จะยิ่งมีความแม่นยำที่สูงตาม

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. Decision Tree (KNN)

```
from sklearn import metrics
from sklearn.metrics import recall_score
from sklearn.metrics import precision_score
predicted = model.predict(X_test)
print(f'accuracy = {metrics.accuracy_score(y_test, predicted)}')
print(f'precision = {metrics.precision_score(y_test, predicted,average=None)}')
print(f'recall = {metrics.recall_score(y_test, predicted,average=None)}')
```

```
accuracy = 0.6967213114754098
precision = [0.55932203 0.84745763 0.5       ]
recall = [0.825      0.76923077 0.11764706]
```

3. Artificial Neural Network (ANN)

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

```
12/12 [==============================] - 0s 2ms/step - loss: 0.3233 - accuracy: 0.5151
Accuracy: 0.515
```

4.เปรียบเทียบผลของ 3 model พร้อมอธิบายสาเหตุว่า วิธีที่ให้ประสิทธิภาพสูงกว่าเพราะอะไร

จากการทดลองจะพบว่า Accuracy จะเป็น

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

- ที่ทำให้ **Decision Tree** แบบ CART มีค่าสูงที่สุดอาจจะเป็นเพราะมีทำการจัดการกับกรณีที่มีข้อมูลบางตัวหายไป หรือ ข้อมูลที่ไม่สมบูรณ์ โดยที่เราไม่ต้องจัดการเอง รวมถึงข้อมูลผิดปกติอื่นๆ
- ส่วน **K-Nearest Neighbors (KNN)** มีค่า Accuracy ที่น้อยกกว่าแต่ไม่ได้ห่างกันเกินไป แปลว่าวิธีนี้ได้ผลลัพธ์ที่ดีเหมือนกัน อาจเป็นเพราะ feature ที่เลือกมีการจัดกลุ่มที่แบ่งข้อมูลได้ง่าย โซนการแบ่งมีการเกาะกลุ่มกันเยอะ ไม่ปนกันมาก ทำให้ค่า Accuracy สูง
- **Artificial Neural Network (ANN)** ที่น้อยอาจจะเป็นเพราะมีการทำงานของ algorithm ไม่ดีพอ หรือชุดข้อมูลไม่มากพอให้มีการเรียนรู้ในการทำงานได้ดีสำหรับการทำงานกับข้อมูลชุดนี้

5.สรุปผล

ชุดข้อมูลที่เลือกมานั้นมี feature ที่ใช้งานอยู่ 3 อย่างคือ work_year = ปีที่มีการทำงาน employment_type = สายงาน (PT (Part-time), FT (Full-time) ,CT (Contract) ,FL (Freelance)) remote_ratio = รูปแบบการทำงาน (0 = No remote work (less than 20%) 50 = Partially remote 100 = Fully remote (more than 80%)) ทั้ง 3 อย่างนี้มีผลในการตัดสินใจในการทำนายได้ว่า company size ที่เลือกนั้นควรเป็นระดับไหน มีทั้ง S (small) , M (medium) และ L (large)

model ที่ใช้ในการพิจารณาในการนำมาสร้าง และมีค่า Accuracy ที่สูงมากพอในการทำนาย คือ Decision Tree และ K-Nearest Neighbors (KNN) เป็นเพราะด้วยการทำงานของ Decision Tree คือเป็นการเลือกเหมือน if-else ตัดสินใจว่า feature ที่ได้มานั้นผ่านเงื่อนไขไหม ถ้าผ่านจะไปต่อเรื่อยๆ จนกว่าจะถึงผลลัพธ์สุดท้ายที่จะบอกว่าเป็น company size แบบไหน ซึ่งเหมาะมากสำหรับข้อมูลที่เลือกมา

ยังมี K-Nearest Neighbors (KNN) ที่เหมาะจะนำมาทำงานด้วยเหมือนกัน เพราะมีการเกาะกลุ่มของ feature อยู่ทำให้ง่ายต่อการตัดสินใจกับข้อมูลชุดนี้ได้ว่าควรเป็น company size อะไรเมื่อมีข้อมูลชุดใหม่เข้ามา เพราะการเลือก company size จะมีเกณฑ์ในการเลือกอยู่