

**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**

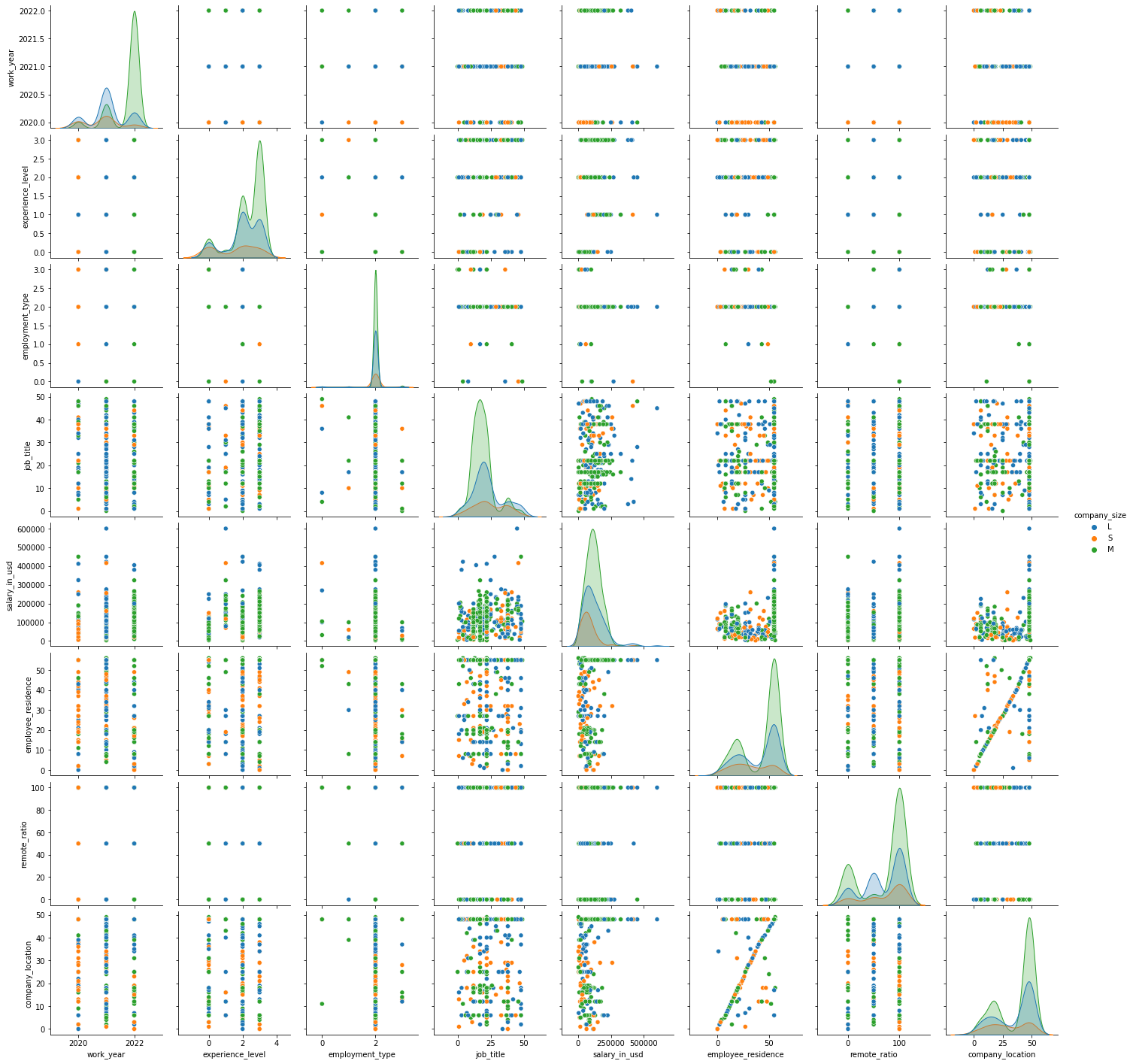
**คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ**

**1.การวิเคราะห์ข้อมูล เเละสาเหตุของการเลือก features ในการนำมา train**

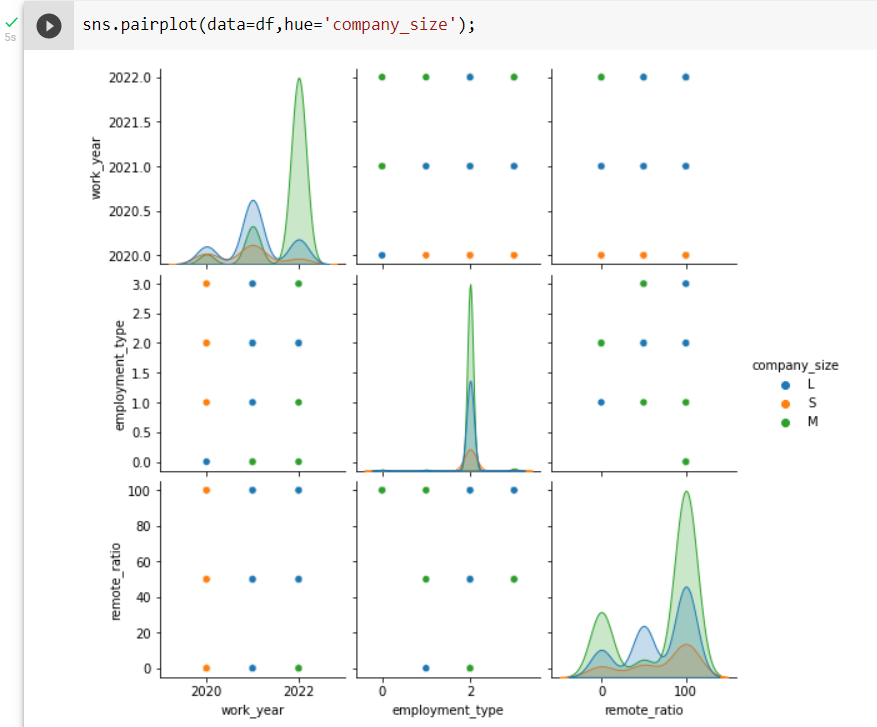
เลือกข้อมูลของ ds\_salaries.csv มาวิเคราะห์ชุดข้อมูล

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

****

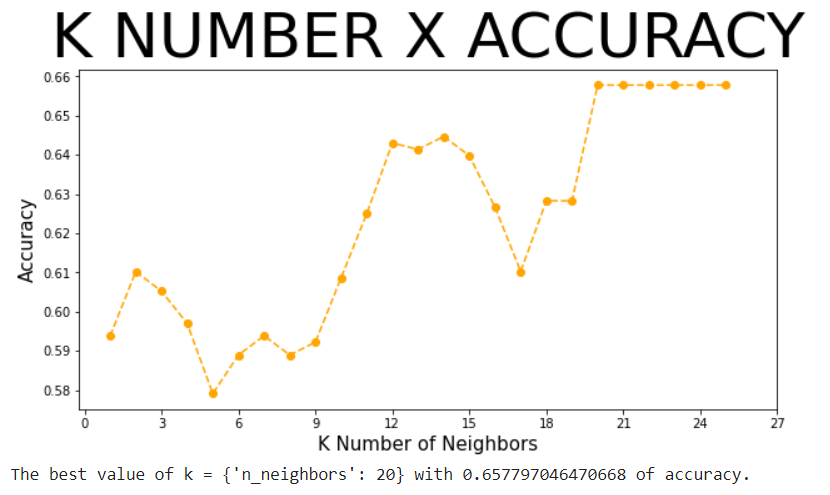
feature ที่เลือก 3 feature มาแสดงความสัมพันธ์โดยใช้ pairplot ของ seaborn

**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

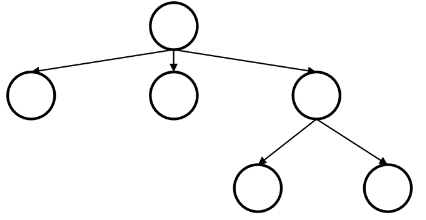


ทดลองหาค่า k ที่เหมาะสมที่สุด และนำมาปรับจูนพารามิเตอร์ของแบบจำลอง เพื่อทดสอบหาประสิทธิภาพของแบบจำลอง สำหรับค่าพารามิเตอร์ต่าง ๆ ร่วมกับการประเมินผลด้วย cross-validation และทดลองพล็อตกราฟค่าประสิทธิภาพออกมา จะได้ค่า K ที่เหมาะสมเท่ากับ 20

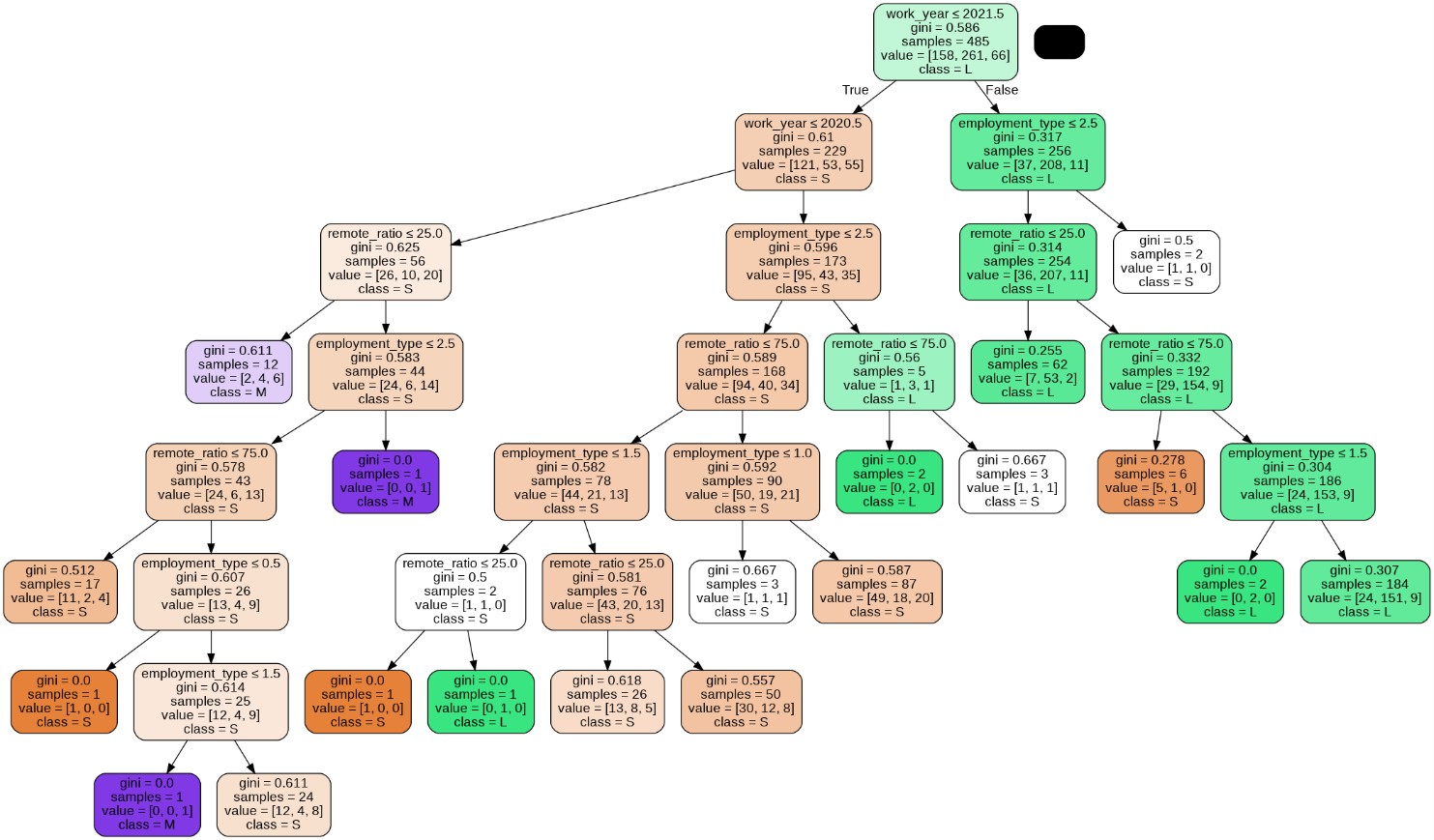
เหตุผลที่เลือก เพราะวิธีนี้น่าจะเหมาะกับชุดข้อมูลที่มีความซับซ้อนมาก และเมื่อดูความสัมพันธ์ระหว่างชุดข้อมูลที่มีอยู่นั้นมีการแบ่งแยกโซนที่พอจะทำงานแบ่งการทำงานได้ดี ว่าโซนนี้เหมาะกับ company\_size แบบไหน

1. Decision Tree

* เป็นการจำลองที่ไม่เป็นเชิงเส้น ใช้การทำงานคล้าย if-else ในการนำมาวิเคราะห์ข้อมูล



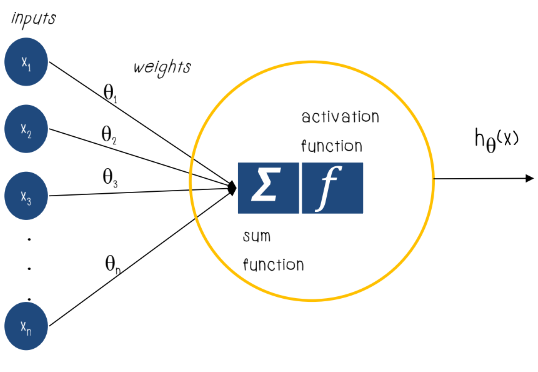
มีการสร้าง Decision tree ด้วยวิธี **gini index** หรือ **Classification and Regression Tree (CART)** ที่ง่ายต่อการสร้าง Decision Tree แต่ไม่ค่อยนิยมเท่ากับ ID3 (Iterative Dichotomiser3)



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

1. Artificial Neural Network (ANN)

* เป็นโครงข่ายของโหนดการคำนวณแบบง่าย ๆ แต่ละโหนดจะรับค่าตัวแปร คูณกับค่าน้ำหนัก รวม และส่งต่อไปยังโหนดถัดไป

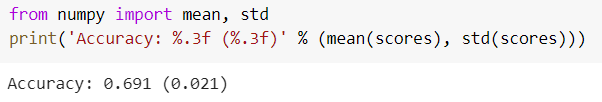


* เป็นการที่ค่อยๆเรียนรู้ให้มีการพัฒนาไปเรื่อยๆ ค่าของ weight ต่าง ๆ ในโครงข่ายจะพยายามปรับเพื่อที่จะทำให้การคำนวณเป็นจริง คือถ้า input vector นี้เข้ามา output vector จะต้องเป็นแบบนี้

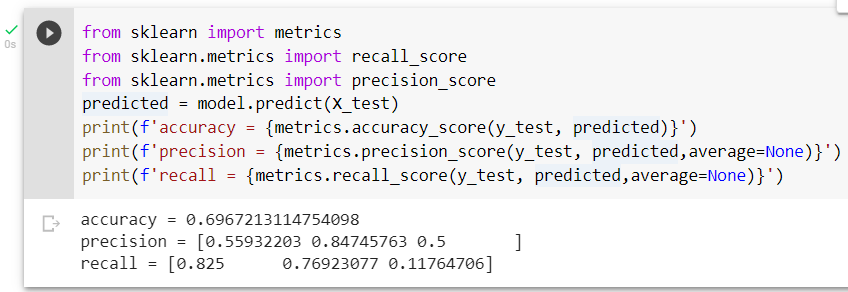
เหตุผลที่เลือก เพราะชุดข้อมูลมีจำนวน 607 ชุดข้อมูลซึ่งน่าจะมากพอในการเทรนให้ algorithm มีความแม่นยำในการทำนายได้ว่าข้อมูลที่เข้ามาควรจะเป็นแบบใด เพราะวิธีนี้ยิ่งมีชุดข้อมูลที่เยอะ จะยิ่งมีความแม่นยำที่สูงตาม

**3.ผลการ testing model**

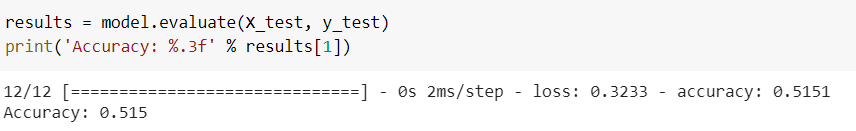
1. K-Nearest Neighbors



1. Decision Tree (KNN)



1. Artificial Neural Network (ANN)

****

**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 ที่สูงมากพอในการทำนายคือ 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