**ML1 Report**

1. **Data Preparation**

* Data จะขึ้นต้นด้วยเว้นวรรคทุกอัน จึงต้องทำการตัดตัวเว้นวรรคออกก่อน
* ค่าที่ไม่รู้จะใส่เป็น ‘?’ จึงทำการเปลี่ยนเป็น NaN ก่อนจะได้เอาไปกำจัด null value ได้ง่ายๆ
* เปลี่ยนค่าใน ‘sex’ จาก ‘Male’ เป็น 0 และ ‘Female’ เป็น 1
* เปลี่ยนค่า target จาก >50K เป็น 0 และ <=50K เป็น 1
* ทำการสร้าง dummies ให้กับ workclass, education, maritial-status, occupation, relationship, race
* ทำการ drop ค่า native-country เพราะว่ามีค่าหลากหลายเกินไป
* กำจัด null values
* สุดท้ายจะได้ข้อมูล 32561 รายการ และมีทั้งหมด 57 features (ไม่รวม target value)

1. **Divide data into train/test data**

* ใช้คำสั่ง train\_test\_split โดยให้ train:test = 70:30 และ random\_state=1234

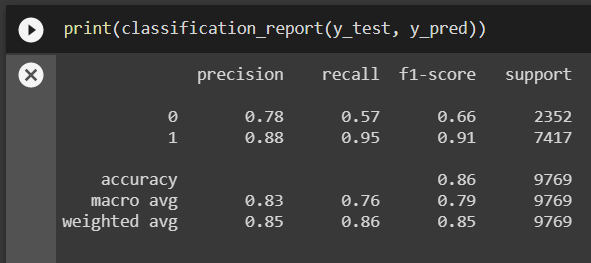
1. **Choosing best model and parameter**

* เลือกทั้งหมด 3 algorithms คือ Random Forest, kNN, Neural Network

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Random Forest | kNN | Neural Network |
| RandomizedSearchCV | Best Parameter | {'criterion': 'entropy', 'max\_depth': None,'min\_samples\_leaf': 10,'min\_samples\_split': 8} | {'n\_neighbors':[3,5,10], 'weights':['uniform','distance']} | {'activation': 'identity', 'solver': 'adam'} |
| Accuracy | 0.8614863520661915 | 0.786591787976875 | 0.8105918491898093 |
| GridSearchCV | Best Parameter | - | {'n\_neighbors': 10, 'weights': 'uniform'} | {'activation': 'identity', 'solver': 'adam'} |
| Accuracy | - | 0.786591787976875 | 0.8105918491898093 |
| **Summary** | | **0.8614863520661915** | **0.786591787976875** | **0.8105918491898093** |

* หมายเหตุ: Random Forest ไม่ได้ทำ GridSearchCV เพราะว่าลองทำแล้วใช้เวลานานมากๆๆๆๆๆ และเท่าที่ดูจาก algorithm อื่นได้ผลคล้ายๆ กันเลยใช้แค่ Randomized
* สรุปแล้วจึงเลือกใช้ **Random Forest** ด้วย parameter = {'criterion': 'entropy', 'max\_depth': None,'min\_samples\_leaf': 10,'min\_samples\_split': 8}

1. **Train Model**



* **สรุป ได้ macro F1-score = 0.79**