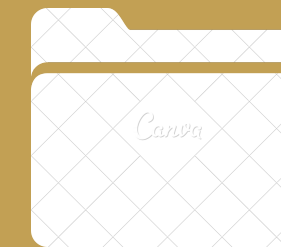
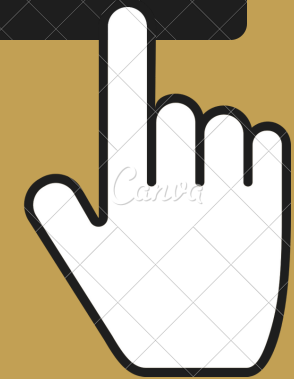
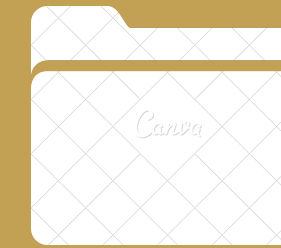


# Using data mining to improve assessment of credit worthiness

Chirat Suwannachote



TUSEDAY

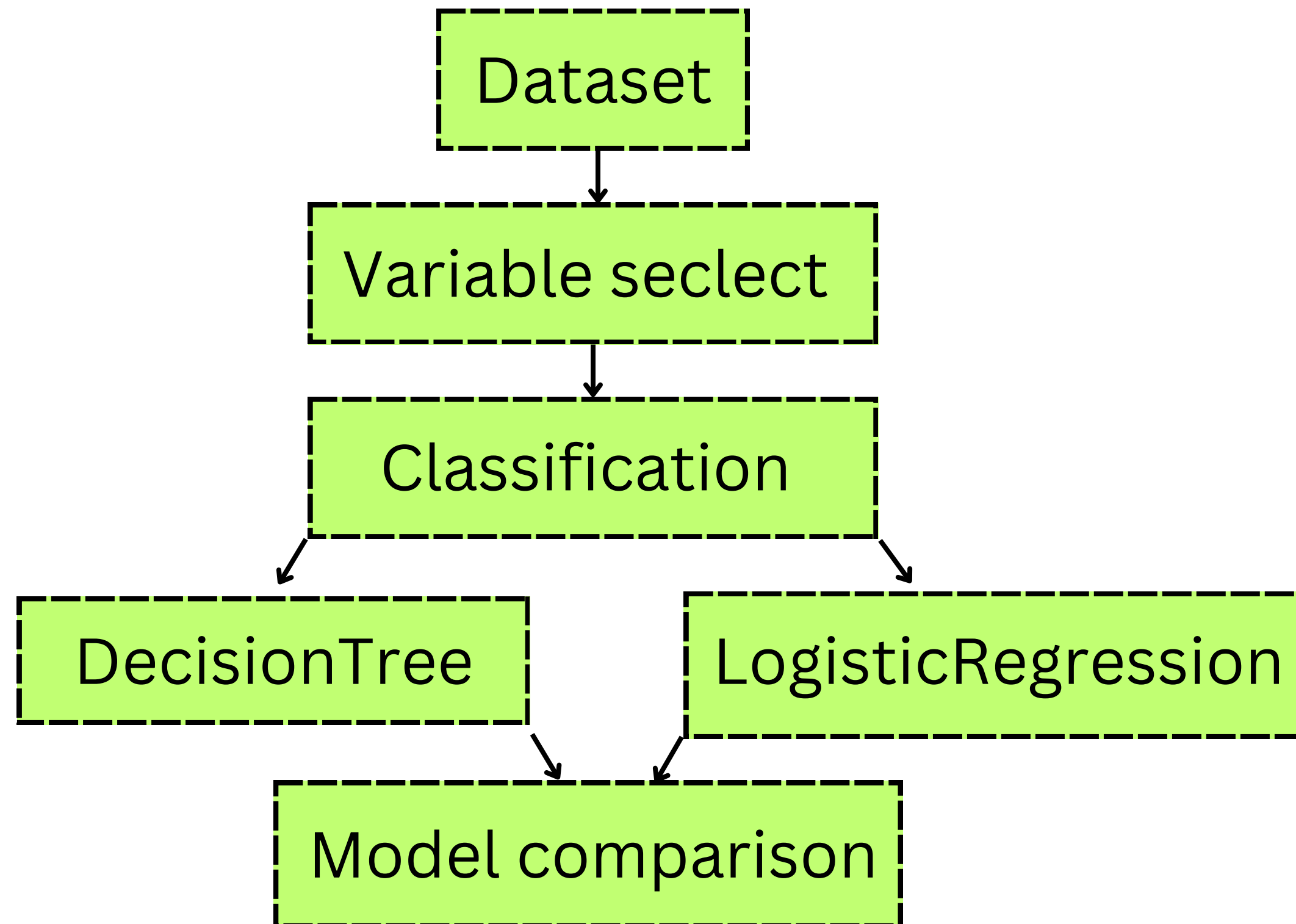
17

OCT 2023

# Background

## ตัวแบบการประเมินสินเชื่อ

เป็นอีกหนึ่งองค์ประกอบที่สำคัญในการประเมินสินเชื่อของลูกค้า เพื่อให้ธนาคารสามารถลดความเสี่ยงของตัวธนาคารได้มากที่สุด ธนาคารจึงพยายามพัฒนาตัวแบบการประเมินสินเชื่อที่มีประสิทธิภาพ ซึ่งตัวแบบนี้นำเทคนิคคัดเลือกตัวแปรอิสระมาใช้ร่วมกับการจำแนกรูปแบบต่างๆที่ใช้อยู่ทั่วไป เพื่อตัดตัวแปรที่ไม่ส่งผลต่อการพยากรณ์ในข้อมูลนั้นๆออก โดยใช้ข้อมูลประวัติการค้าชำระ สถานะการเป็นหนี้ ในช่วยจำแนกตัวแปรที่มีความสำคัญในการประเมินความเสี่ยง



# Data set



```
1 CD.head()
```

	age	income	home_ownership	emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1
1	21	9600	OWN	5.0	EDUCATION	B	1000	11.14	0
2	25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1
3	23	65500	RENT	4.0	MEDICAL	C	35000	15.23	1
4	24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1

from <https://www.kaggle.com/datasets/laotse/credit-risk-dataset>

# Data set

loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
1	0.59	Y	3
0	0.10	N	2
1	0.57	N	3
1	0.53	N	2
1	0.55	Y	4

from <https://www.kaggle.com/datasets/laotse/credit-risk-dataset>

# Data info.



```
1 CD.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  -
 0   age                 1000 non-null   int64
 1   income              1000 non-null   int64
 2   home_ownership      1000 non-null   object
 3   emp_length          970 non-null    float64
 4   loan_intent          1000 non-null   object
 5   loan_grade          1000 non-null   object
 6   loan_amnt           1000 non-null   int64
 7   loan_int_rate       898 non-null    float64
 8   loan_status         1000 non-null   int64
 9   loan_percent_income 1000 non-null   float64
10   cb_person_default_on_file 1000 non-null   object
11   cb_person_cred_hist_length 1000 non-null   int64
dtypes: float64(3), int64(5), object(4)
memory usage: 93.9+ KB
```

# variable define

**1.age = อายุ**

**2.income = รายได้**

**3.home\_ownership = ความเป็นเจ้าของในที่อยู่อาศัย**

**4.emp\_length = ประสบการณ์ทำงาน (ปี)**

**5.loan\_intent = จุดประสงค์การกู้**

**6.loan\_grade = คุณภาพการกู้**

**7.loan\_amnt = จำนวนเงินกู้**

**8.loan\_int\_rate = อัตราดอกเบี้ย**

**9.loan\_status = สถานะการเป็นหนี้**



# variable define

**10.loan\_percent\_income**

**= อัตราส่วนผลตอบแทนของหนี้**

**11.cb\_person\_default\_on\_file**

**= ประวัติการเป็นหนี้(มี หรือ ไม่มี)**

**12.cb\_person\_cred\_hist\_length**

**= ระยะเวลาที่จ่ายตรงหนี้ตรงเวลา (ปี)**





# variable define

**10.loan\_percent\_income**

**= อัตราส่วนผลตอบแทนของหนี้**

**11.cb\_person\_default\_on\_file**

**= ประวัติการเป็นหนี้(มี หรือ ไม่มี)**

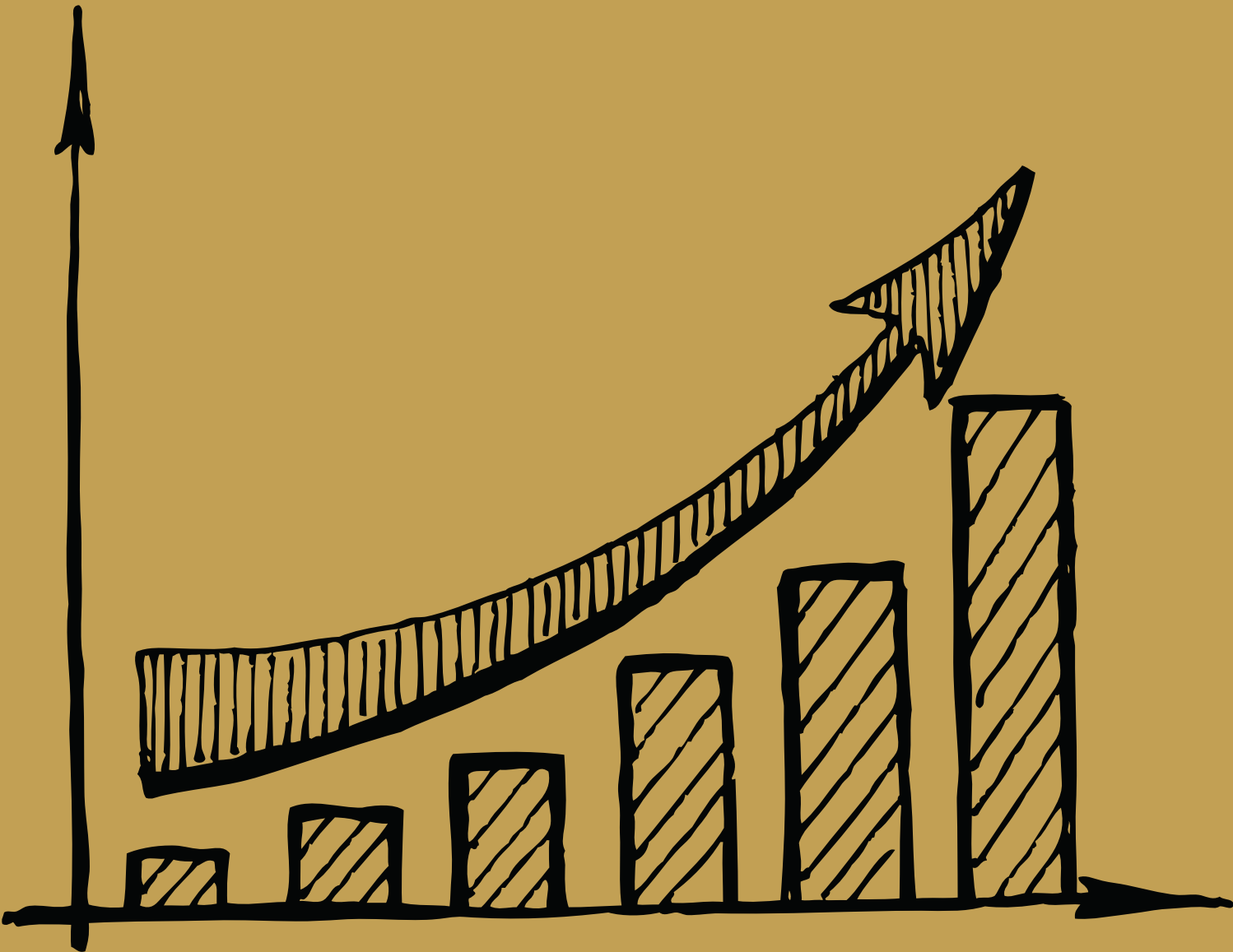
**12.cb\_person\_cred\_hist\_length**

**= ระยะเวลาที่จ่ายตรงหนี้ตรงเวลา (ปี)**





**All  
variable**



# Check Missing data



```
1 CD.isnull().any()
```

age	False
income	False
home_ownership	False
emp_length	True
loan_intent	False
loan_grade	False
loan_amnt	False
loan_int_rate	True
loan_status	False
loan_percent_income	False
cb_person_default_on_file	False
cb_person_cred_hist_length	False
dtype:	bool



```
[79] 1 CD = CD.dropna()  
      2 CD.shape
```

```
(869, 12)
```

# Check Missing data

```
[81] 1 print(CD.home_ownership.unique())
      2 print(CD.loan_intent.unique())
      3 print(CD.loan_grade.unique())
      4 print(CD.cb_person_default_on_file.unique())
      5

['RENT' 'OWN' 'MORTGAGE' 'OTHER']
['PERSONAL' 'EDUCATION' 'MEDICAL' 'VENTURE' 'HOMEIMPROVEMENT'
 'DEBTCONSOLIDATION']
['D' 'B' 'C' 'A' 'E' 'F' 'G']
['Y' 'N']
```

# Check Missing data



```
83] 1 CD.head()
```

	age	income	home_ownership	emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	cb_person
0	22	59000	2	123.0	1	4	35000	16	1	
1	21	9600	1	5.0	2	2	1000	11	0	
2	25	9600	3	1.0	3	3	5500	12	1	
3	23	65500	2	4.0	3	3	35000	15	1	
4	24	54400	2	8.0	3	3	35000	14	1	

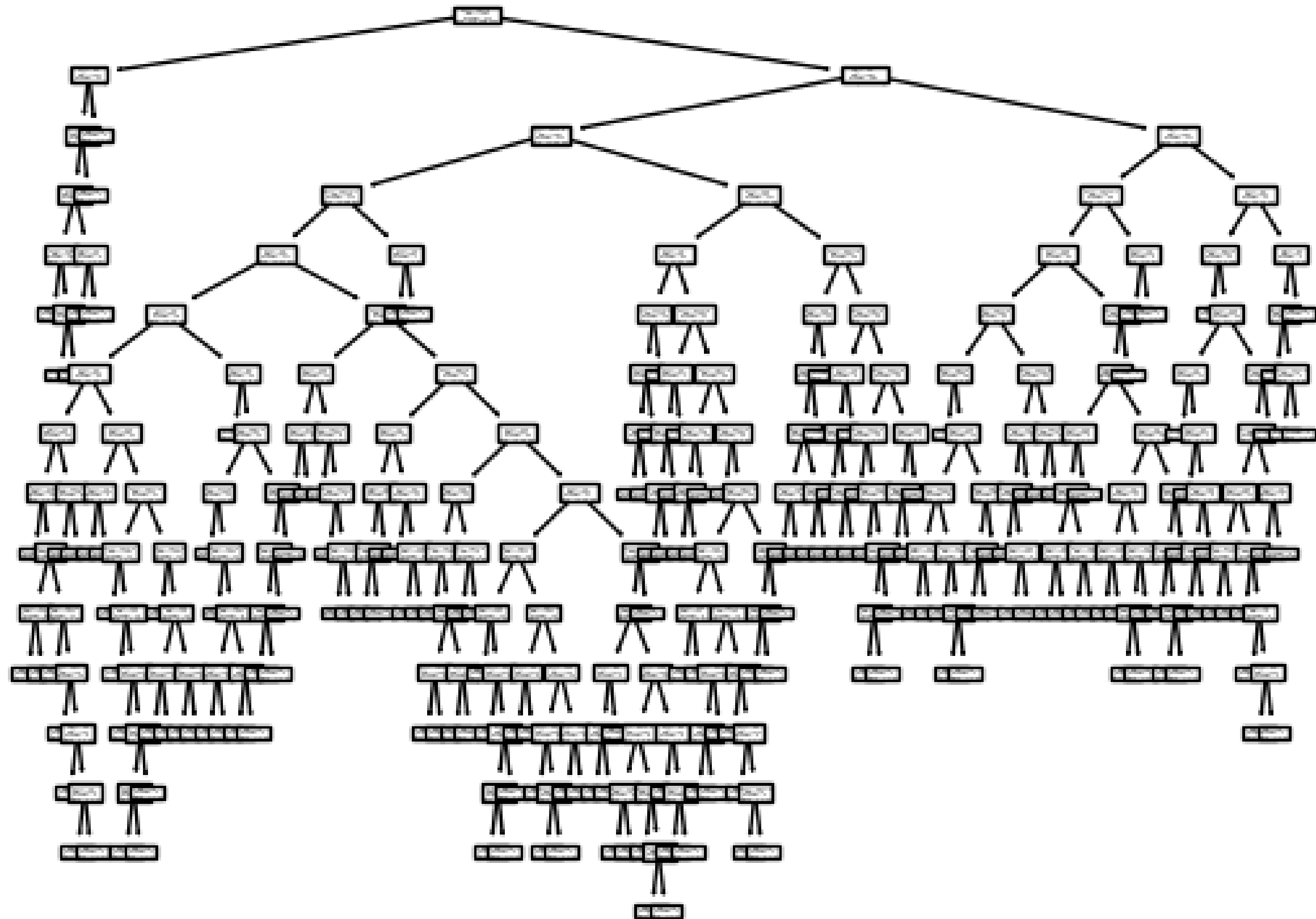
# Set train and test



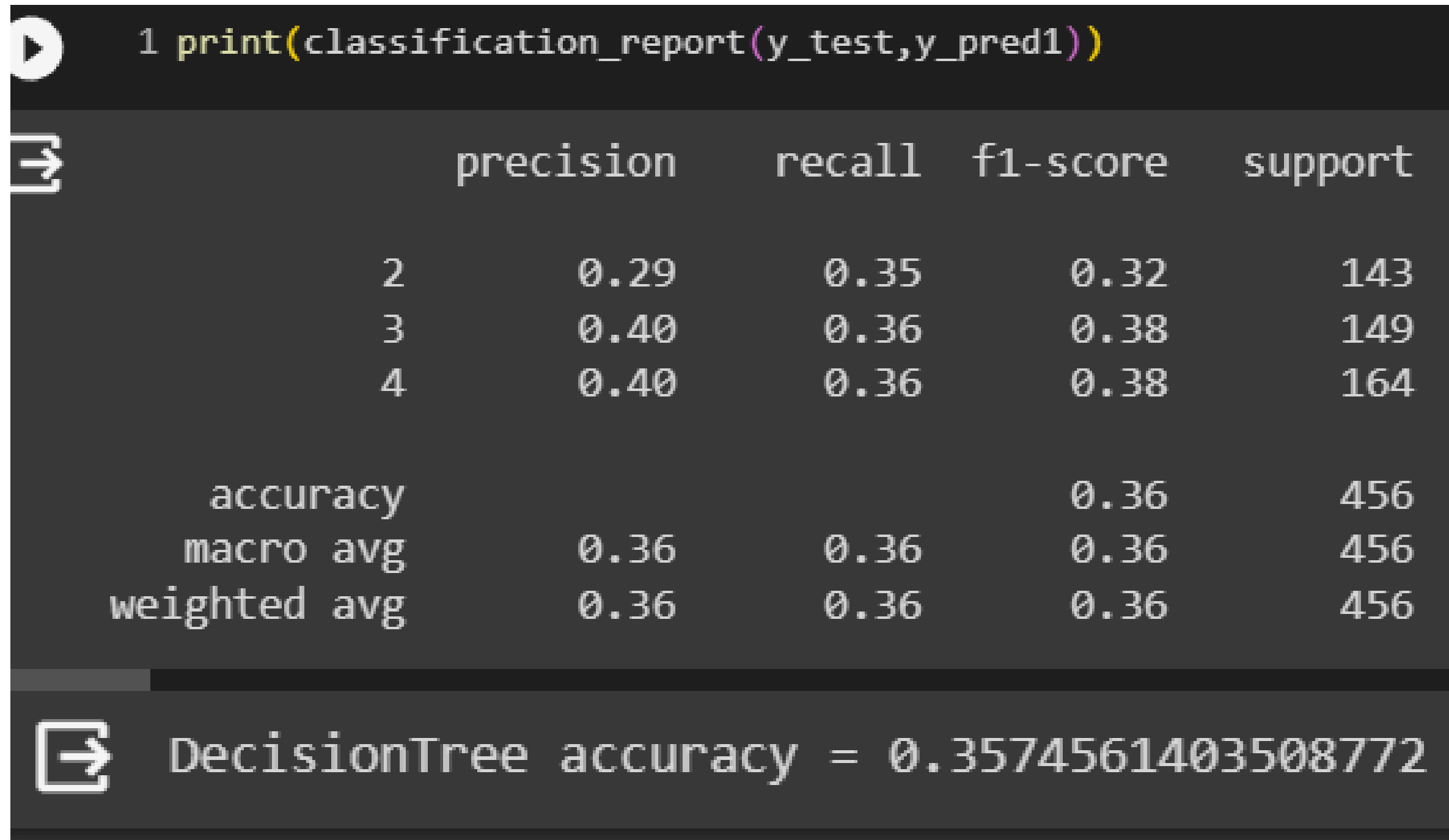
```
1 a = np.random.randint(2, size = len(CD))
2 train = CD[a==1]
3 test = CD[a==0]
4 X_train = train.iloc[:, :-1]
5 y_train = train.iloc[:, -1]
6 X_test = test.iloc[:, :-1]
7 y_test = test.iloc[:, -1]
```



# DecisionTree



# DecisionTree



The image shows a Jupyter Notebook interface with a dark theme. At the top, a code cell contains the command `1 print(classification_report(y_test,y_pred1))`. Below it, the output is displayed as a table with columns for precision, recall, f1-score, and support. The table lists metrics for classes 2, 3, and 4, followed by macro and weighted averages. At the bottom, a text cell displays the overall DecisionTree accuracy as 0.3574561403508772. A vertical scrollbar is visible on the right side of the output area.

```
1 print(classification_report(y_test,y_pred1))
```

	precision	recall	f1-score	support
2	0.29	0.35	0.32	143
3	0.40	0.36	0.38	149
4	0.40	0.36	0.38	164
accuracy			0.36	456
macro avg	0.36	0.36	0.36	456
weighted avg	0.36	0.36	0.36	456

DecisionTree accuracy = 0.3574561403508772



# LogisticRegression

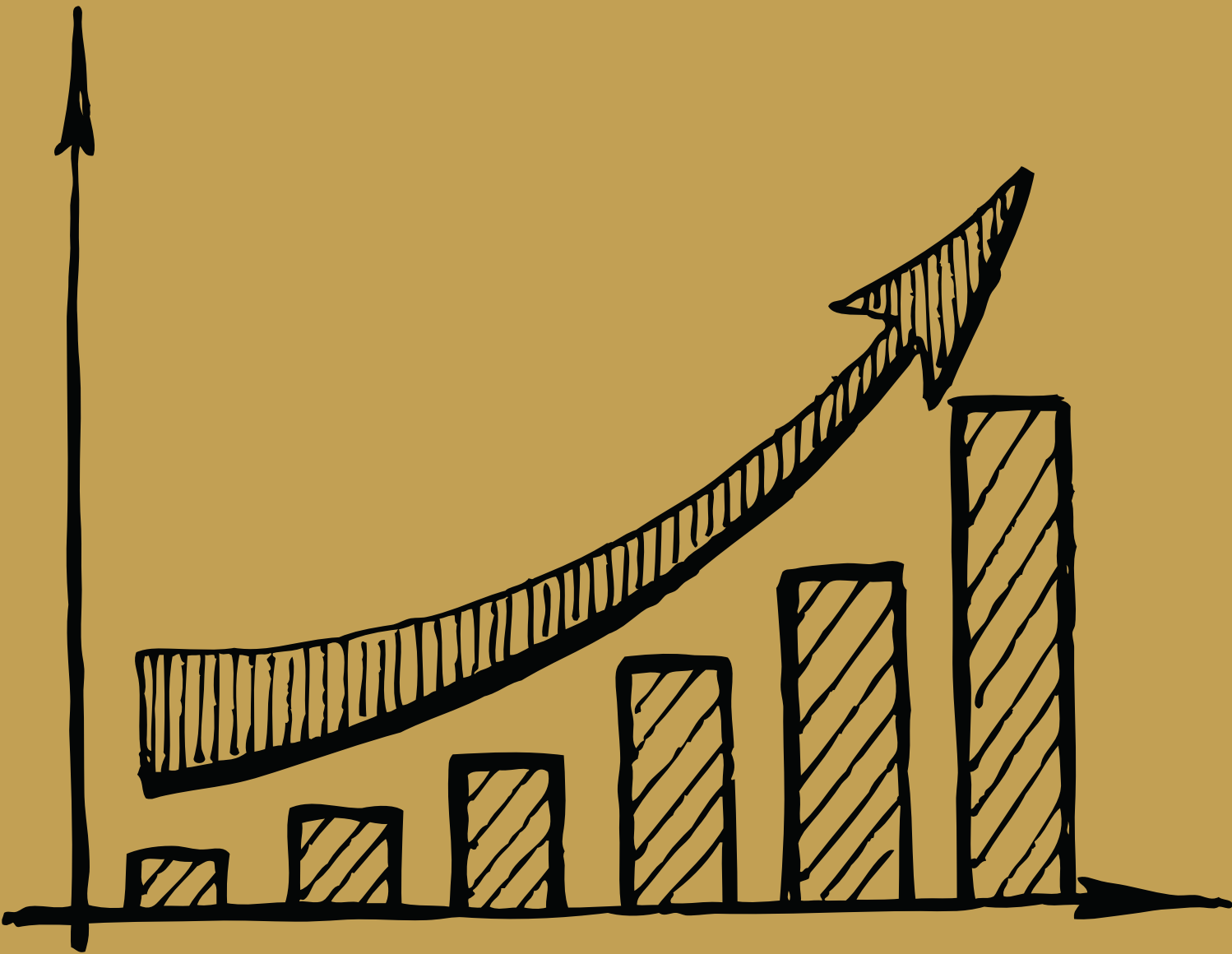
```
[94] 1 print(classification_report(y_test,y_pred2))
```

	precision	recall	f1-score	support
2	0.32	0.74	0.45	143
3	0.36	0.19	0.25	149
4	0.41	0.12	0.19	164
accuracy			0.34	456
macro avg	0.36	0.35	0.29	456
weighted avg	0.37	0.34	0.29	456

```
➞ LogisticRegression accuracy = 0.33771929824561403
```



**select  
variable**





# Defind Event

**event**



×



loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
0.59	1	3
0.10	0	2
0.57	0	3
0.53	0	2
0.55	1	4

# WOE

WOE (Weight of Evidence) เป็นหนึ่งในค่าทางสถิติที่ใช้ในงานการประเมินความเสี่ยงในเชิงการเงินและการวิเคราะห์เครดิต (credit scoring)

WOE วัดความสัมพันธ์ระหว่างความถี่ของคลาสที่ตั้งใจสำรวจ (event) และความถี่ของคลาสที่ไม่ตั้งใจสำรวจ (non-event) ภายในตัวแปรอิสระหรือคุณสมบัติที่นำเข้าคำนวณในโมเดลการประเมินความเสี่ยง

$$WOE = \ln \left( \frac{\text{Event}\%}{\text{Non Event}\%} \right)$$



# WOE



โดยที่:

- non-events % คือ ร้อยละของผู้ผิดชำระในคลาสที่ไม่ตั้งใจสำรวจ
- events % คือ ร้อยละของผู้ผิดชำระในคลาสที่ตั้งใจสำรวจ

ค่า WOE มีความหมายดังนี้:

- $WOE = 0$ : คลาสนี้ไม่มีผลต่อเหตุการณ์หรือไม่มีความสัมพันธ์
- $WOE > 0$ : คลาสนี้เสี่ยงต่อเหตุการณ์
- $WOE < 0$ : คลาสนี้ไม่เสี่ยงต่อเหตุการณ์



# IV

IV (Information Value) ค่าที่แสดงถึงระดับความสัมพันธ์ของตัวแปรกับการทำนาย กล่าวคือ ยิ่งค่า IV มาก หมายความว่าตัวแปรนั้นมีผลต่อการทำนายของโมเดลมาก

$$IV = \sum (\text{Event\%} - \text{Non Event\%}) * (\text{WOE})$$

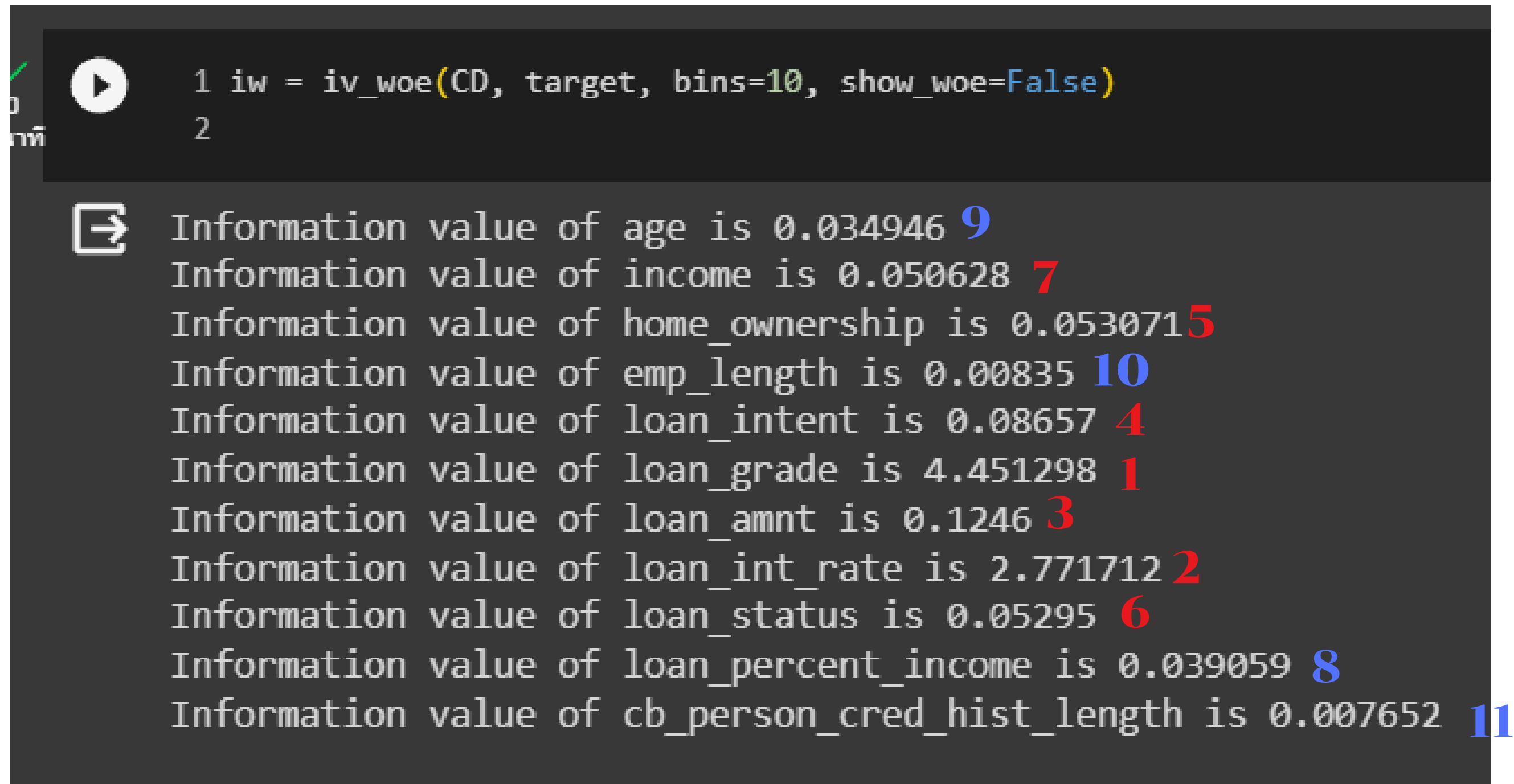
# IV



Information Value (IV)	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	weak predictor
0.1 to 0.3	medium predictor
0.3 to 0.5	strong predictor
> 0.5	suspicious or too good to be true



# IV



```
1 iw = iv_woe(CD, target, bins=10, show_woe=False)
2
```

Information value of age is 0.034946 9  
Information value of income is 0.050628 7  
Information value of home\_ownership is 0.053071 5  
Information value of emp\_length is 0.00835 10  
Information value of loan\_intent is 0.08657 4  
Information value of loan\_grade is 4.451298 1  
Information value of loan\_amnt is 0.1246 3  
Information value of loan\_int\_rate is 2.771712 2  
Information value of loan\_status is 0.05295 6  
Information value of loan\_percent\_income is 0.039059 8  
Information value of cb\_person\_cred\_hist\_length is 0.007652 11



# select variable

```
1 CD = CD[['loan_grade', 'loan_int_rate', 'loan_amnt', 'loan_intent', 'home_ownership', 'loan_status', 'income']]
```

```
1 CD.head()
```

	loan_grade	loan_int_rate	loan_amnt	loan_intent	home_ownership	loan_status	income
0	D	16.02	35000	PERSONAL	RENT	1	59000
1	B	11.14	1000	EDUCATION	OWN	0	9600
2	C	12.87	5500	MEDICAL	MORTGAGE	1	9600
3	C	15.23	35000	MEDICAL	RENT	1	65500
4	C	14.27	35000	MEDICAL	RENT	1	54400

# select variable

```
] 1 CD.head()
```

	age	income	loan_intent	emp_length	loan_grade	loan_amnt	loan_int_rate
0	22	59000	1	123.0	4	35000	16
1	21	9600	2	5.0	2	1000	11
2	25	9600	3	1.0	3	5500	12
3	23	65500	3	4.0	3	35000	15
4	24	54400	3	8.0	3	35000	14

```
1 CD.shape
```

```
(869, 7)
```



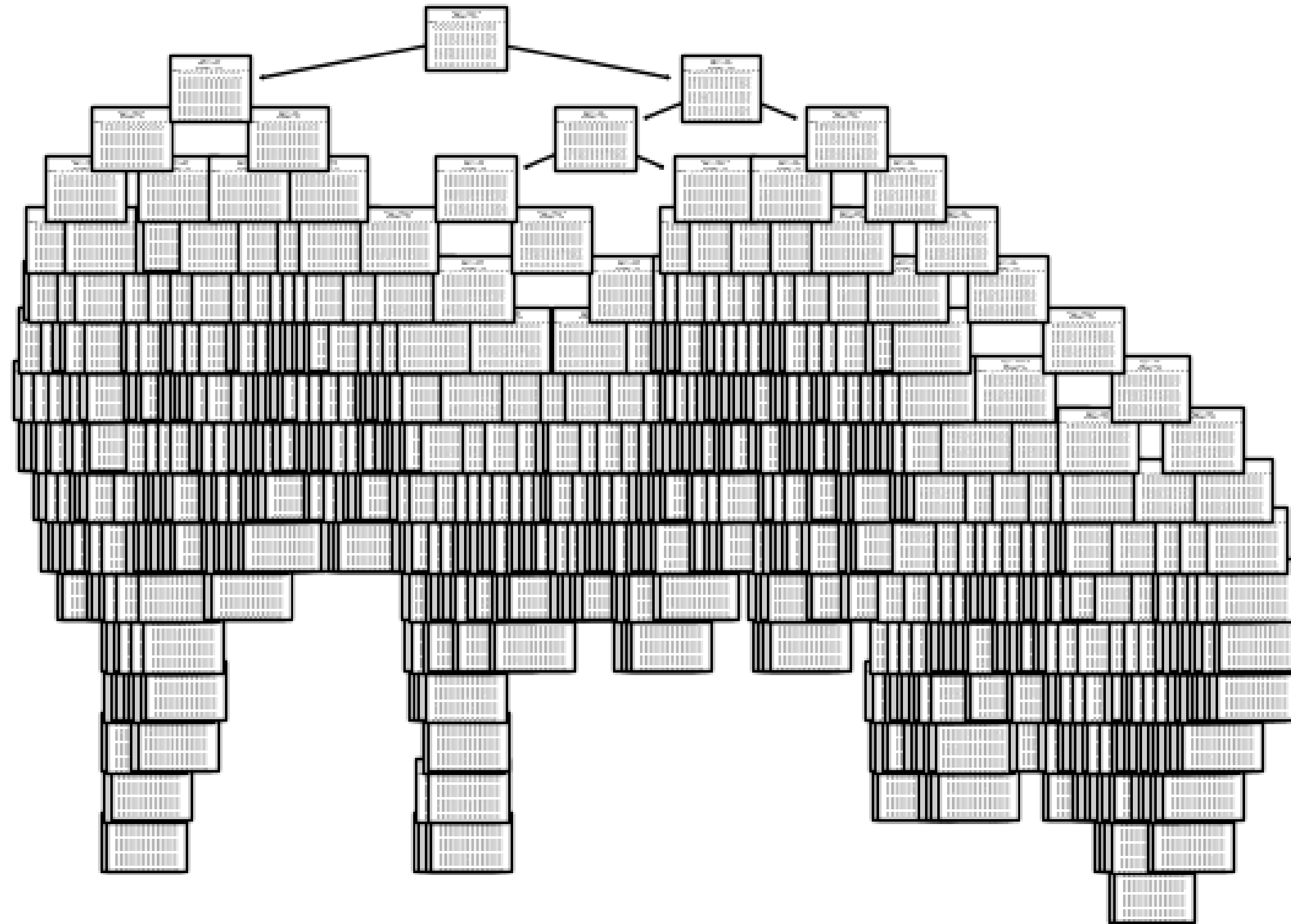
# select variable

```
[71] 1 CD.isnull().any()
```

```
loan_grade      False
loan_int_rate    True
loan_amnt        False
loan_intent      False
home_ownership   False
loan_status      False
income          False
dtype: bool
```

	loan_grade	loan_int_rate	loan_amnt	loan_intent	home_ownership	loan_status	income
0	4	16.02	35000	1	1	1	59000
1	2	11.14	1000	2	2	0	9600
2	3	12.87	5500	3	3	1	9600
3	3	15.23	35000	3	1	1	65500
4	3	14.27	35000	3	1	1	54400

# DecisionTree



# DecisionTree

	precision	recall	f1-score	support
9600	0.00	0.00	0.00	1
9900	0.00	0.00	0.00	1
10000	0.00	0.00	0.00	2
10500	0.00	0.00	0.00	0
10800	0.00	0.00	0.00	0
10980	0.00	0.00	0.00	0
11000	0.00	0.00	0.00	1
11220	0.00	0.00	0.00	1
11389	0.00	0.00	0.00	1
11760	0.00	0.00	0.00	0
11820	0.00	0.00	0.00	1
12000	0.20	0.12	0.15	8
12240	0.00	0.00	0.00	1
12360	0.00	0.00	0.00	1
accuracy				0.02471
macro avg		0.01	0.01	0.01471
weighted avg		0.03	0.02	0.02471

DecisionTree accuracy = 0.021231422505307854

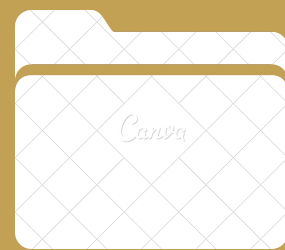
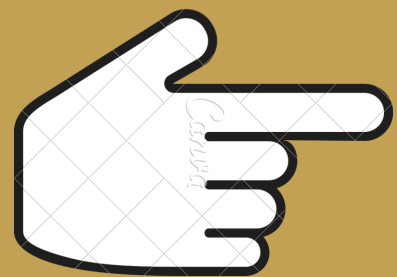
# LogisticRegression



232500	0.00	0.00	0.00	1
234000	0.00	0.00	0.00	3
250000	0.00	0.00	0.00	1
255000	0.00	0.00	0.00	1
275000	0.00	0.00	0.00	1
300000	0.00	0.00	0.00	3
accuracy			0.03	471
macro avg		0.00	0.00	471
weighted avg		0.00	0.03	471

LogisticRegression accuracy = 0.029723991507430998





**Thank  
You**

for listening

