ECON441B : Intro Machine Learning Lab

Week4, Lecture 4 | Decision Trees

Sam Borghese

Wednesday, January 24th, 2024

1. Decision Trees Conceptually
2. Gini Coefficient
3. Applied Uses of Decision Trees
4. ROC & AUC
5. Classwork Assignment

https://www.anderson.ucla.edu/about/clubs-and-associations/professional/tech-business-association-at-anderson-andertech/unchained

Friday, February 2, 2024

1:30 – 6:00 p.m.

Check-in: 1:00 – 1:30 p.m.

Conference: 1:30 – 4:30 p.m.

Networking with Refreshments: 4:30 – 6:00 p.m.

Conference: Crown Auditorium in Marion Anderson Hall at UCLA Anderson
Networking: The Grand Salon in Marion Anderson Hall at UCLA Anderson

REGISTER NOW

EVENT AT ANDERSON BUSINESS SCHOOL

Decision Trees Conceptually

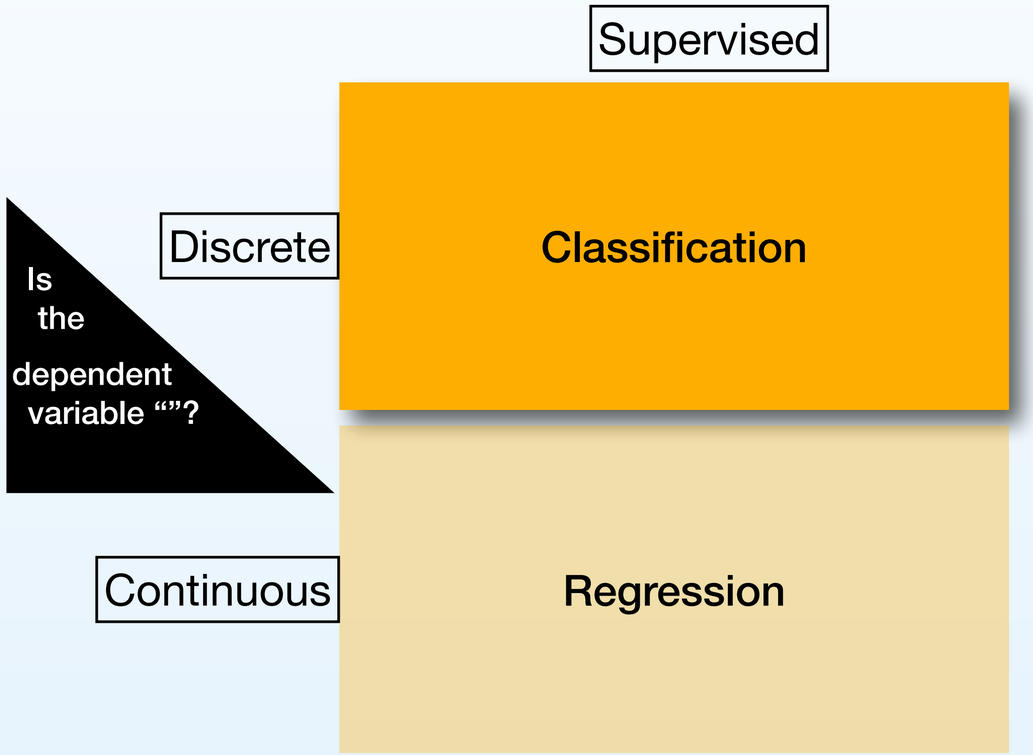
First Decision Tree (Classification Tree)

1966 : Trying to understand the concept of human learning. Created in Psychology

1972 : Paper that split data repeatedly to maximize the homogeneity of the sub classes

Regression vs. Classification

Supervised learning models (Classification)



Models :
Logistic regression
Logistic Lasso
Logistic Ridge
Feed Forward Neural network for classification
Decision Trees
Voting - KNN

Examples :
Email spam detection
Image Recognition
Medical diagnosis
Fraud Detection
Natural Language Processing
Recommended systems

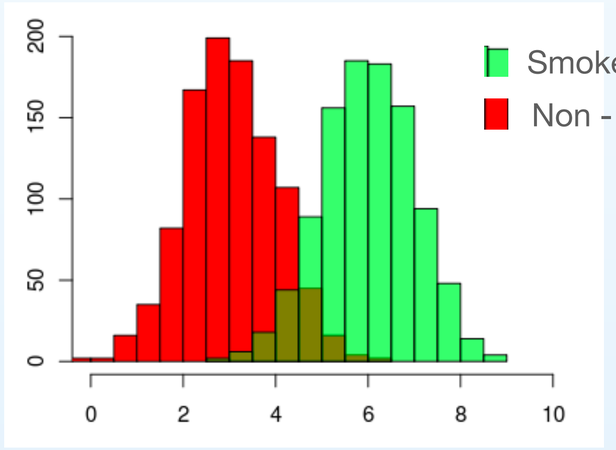
Homogeneity of Subclasses

Read a decision tree

Homogeneity of Subclasses

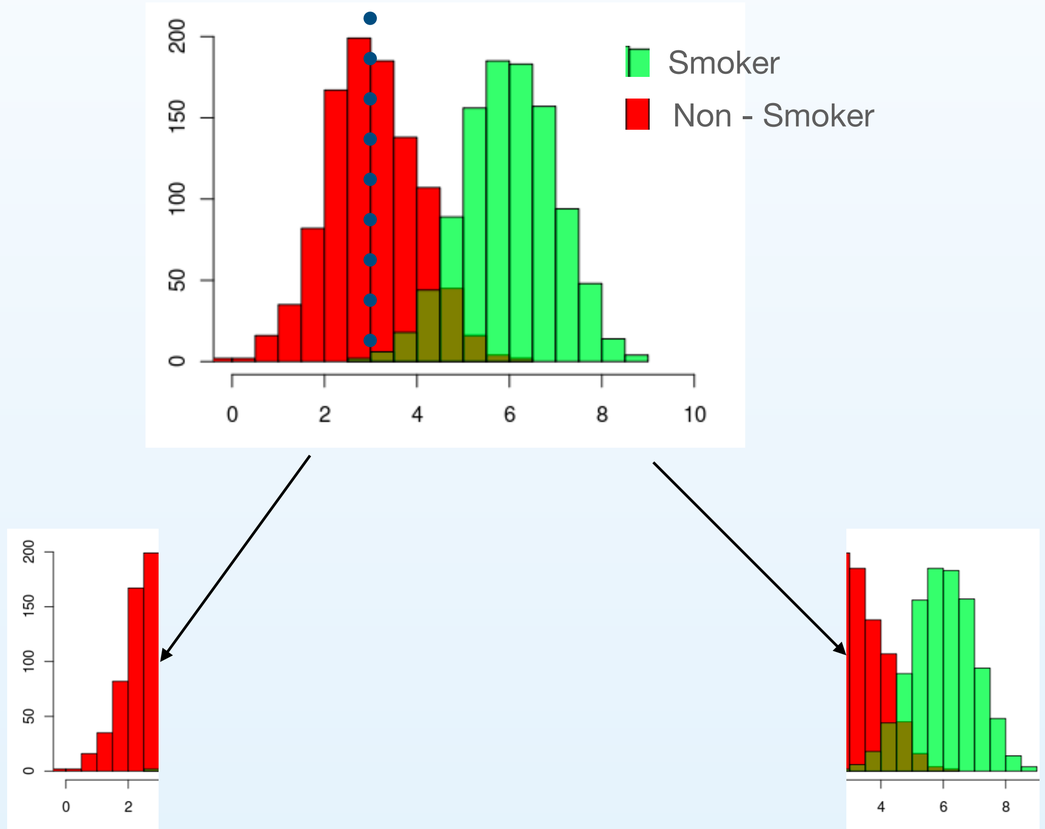
Learn by example : Population of people make a decision tree to predict if someone is a smoker or not

Independent Variables :
Age



Homogeneity of Subclasses

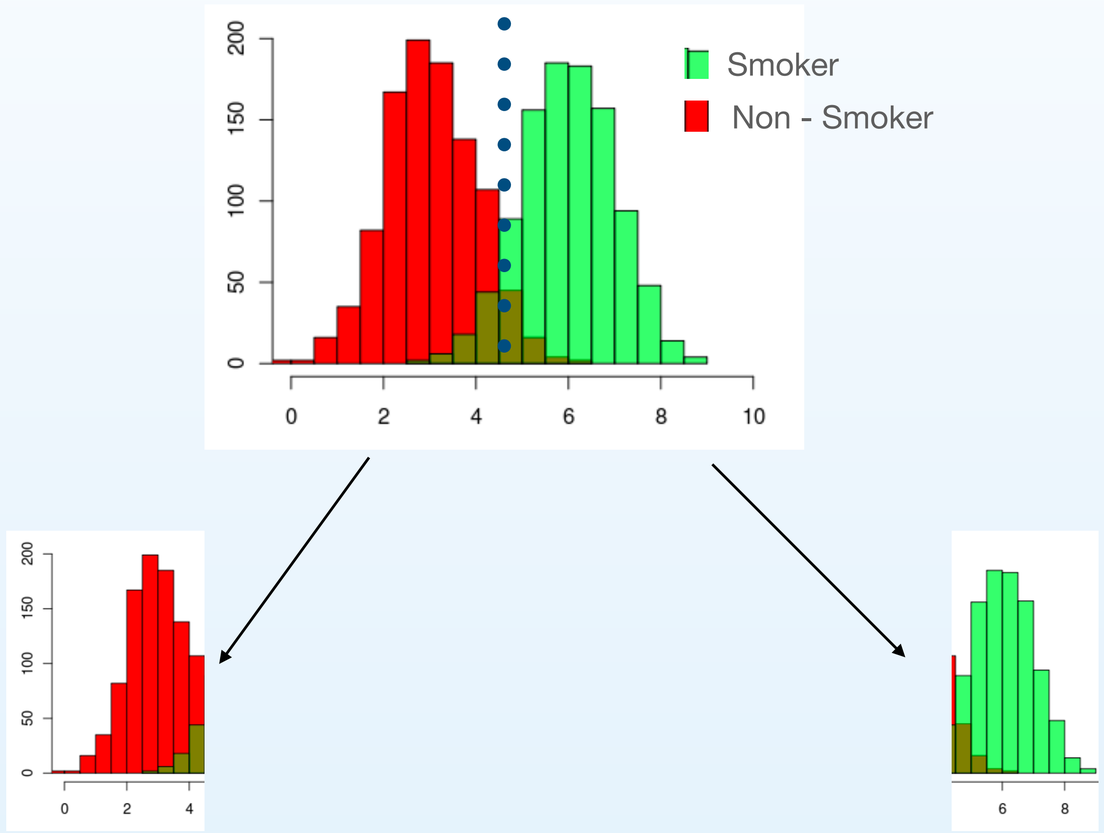
We want to split the population based on independent variables to make the subclasses the most homogenous



Independent Variables :
Age

Homogeneity of Subclasses

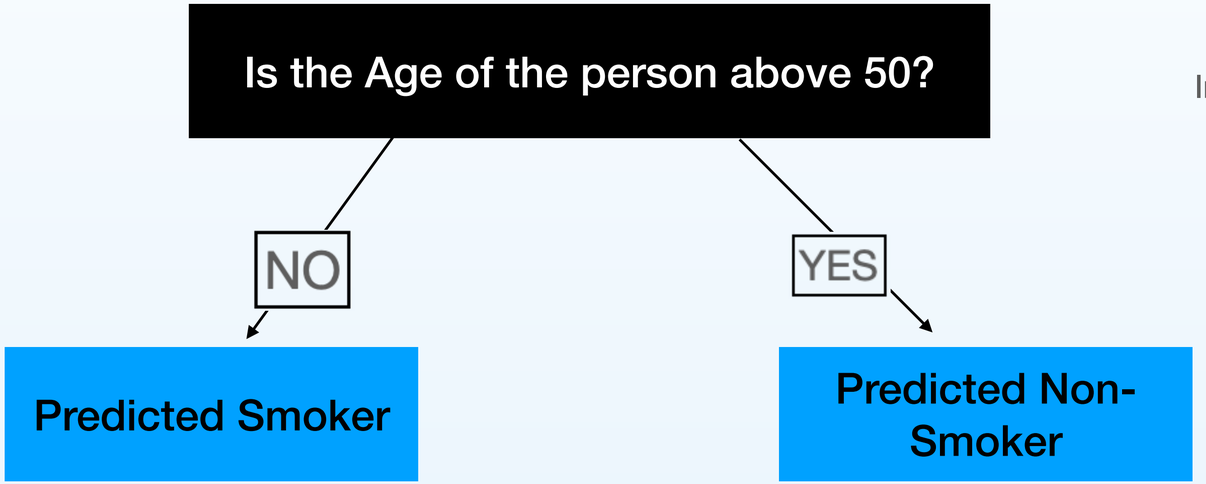
The threshold is fit by the algorithm to maximize homogeneity of all subclasses



Independent Variables :
Age

Predicting

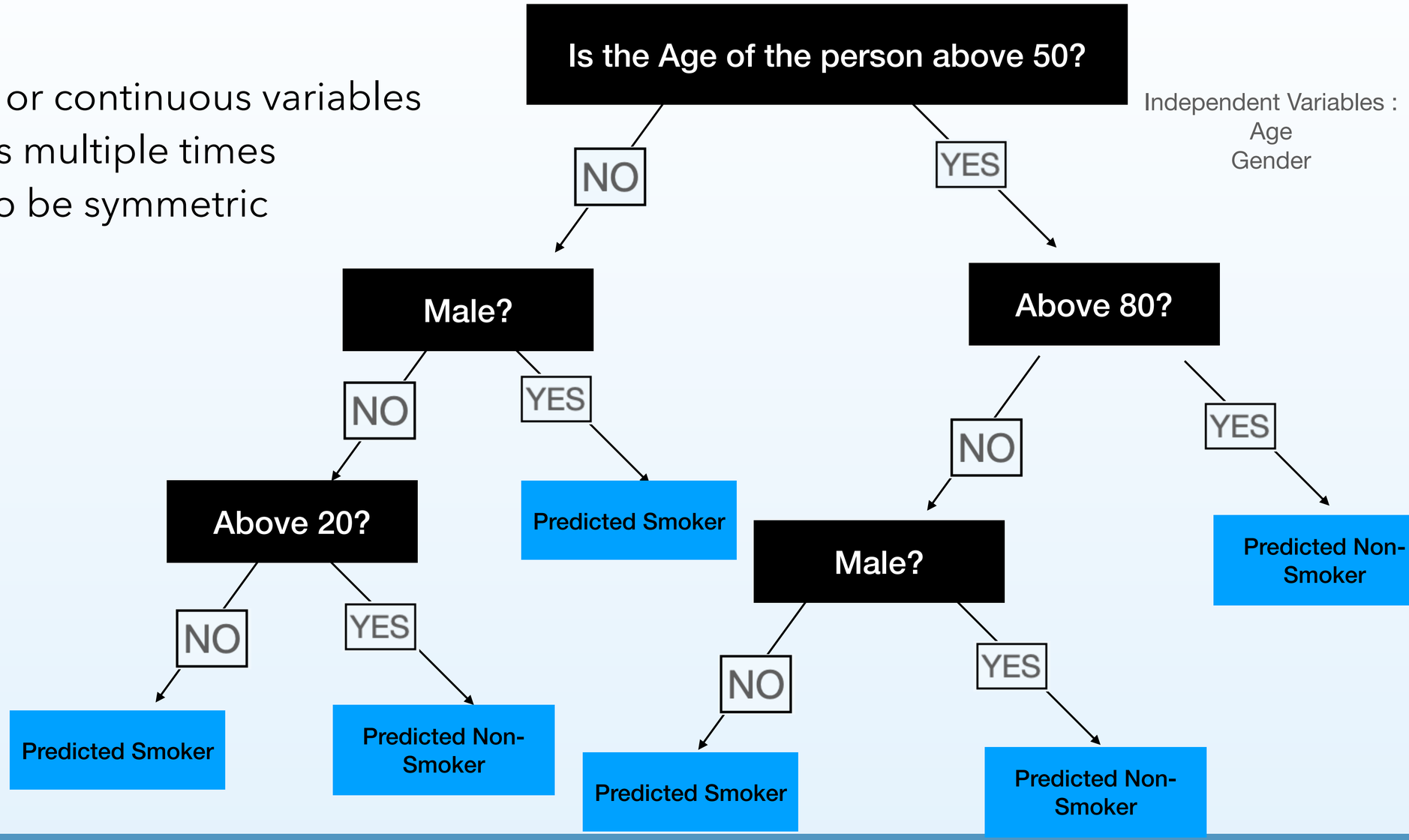
Using historical data a tree can be fit to predict new data points



Independent Variables :
Age

Predicting

1.) Can split on discrete or continuous variables
2.) Can split on variables multiple times
3.) Tree does not have to be symmetric



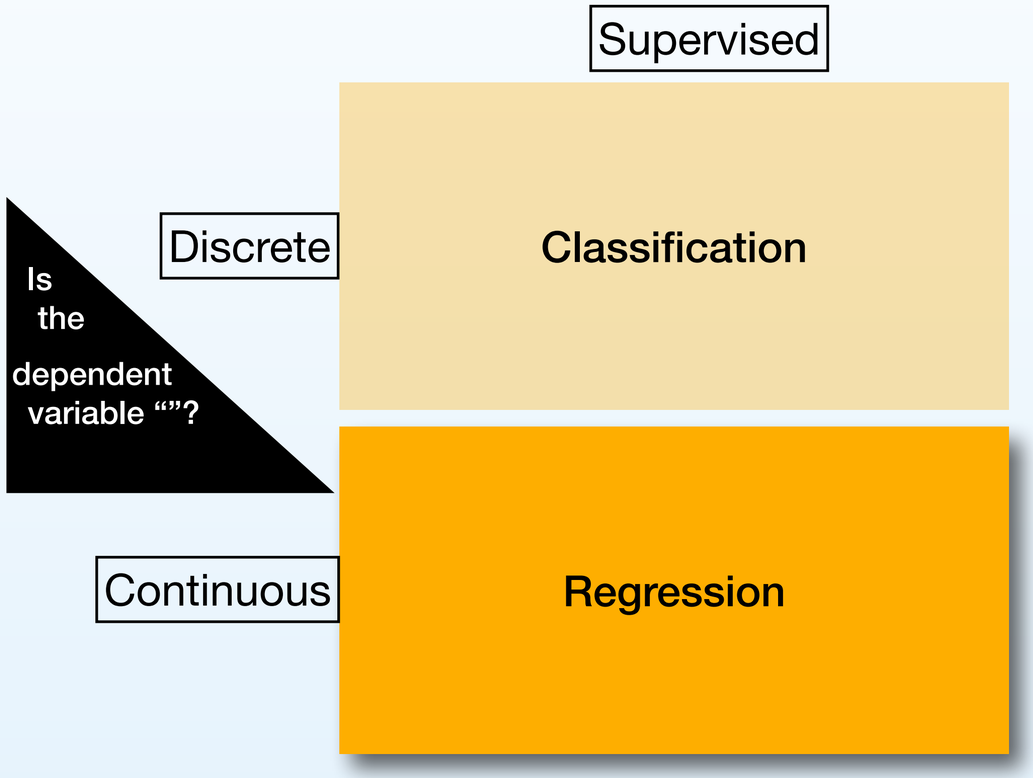
When does the tree stop growing?

1.) Random Forest to Avoid Overfitting

2.) Cross validation (Train test split)

Regression vs. Classification

Supervised learning models (Regression)

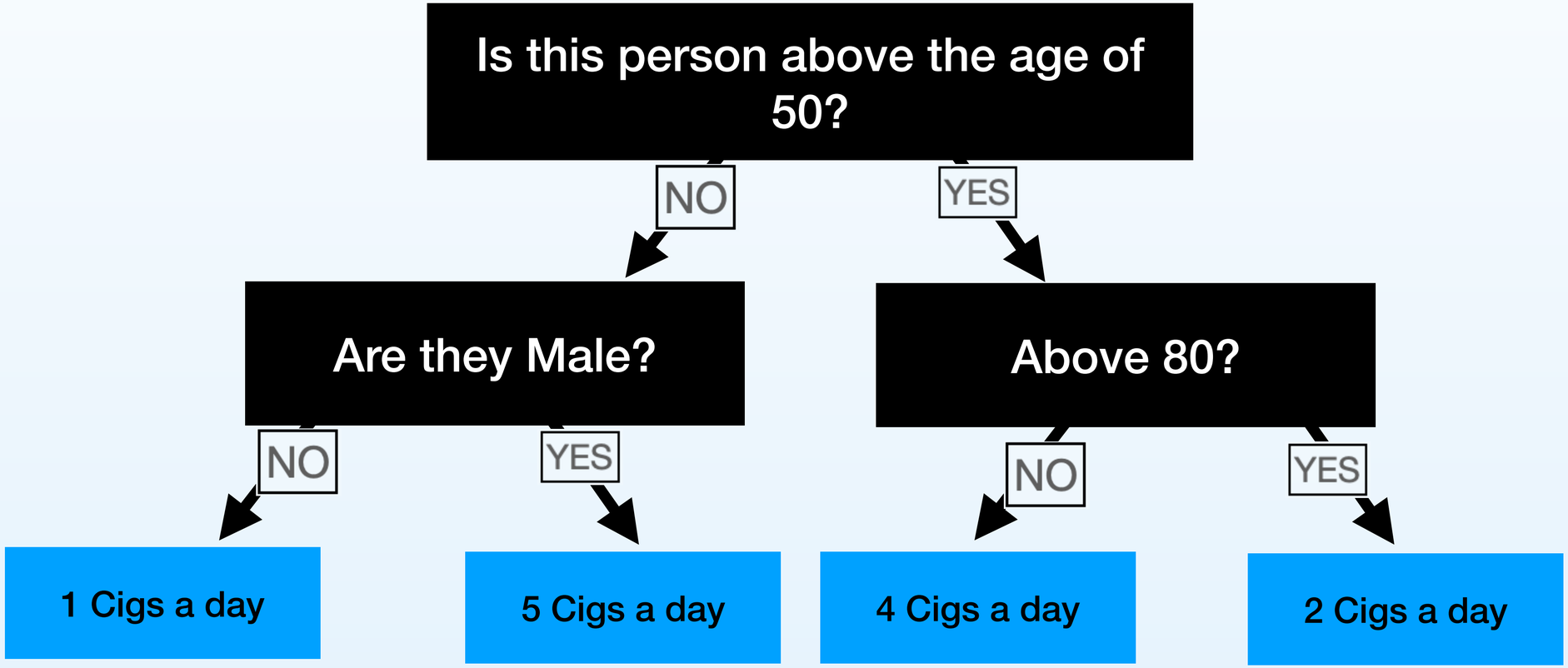


Models :
Linear regression
Lasso
Ridge
Feed Forward Neural network
Regression Trees
Mean - KNN

Examples :
Predict Income
Predict Lifespan
Predict Height
Time for inventory to sell
Amount of rain falling
Value of a house

A regression tree has a continuous output

The question could be : How much do they smoke a day

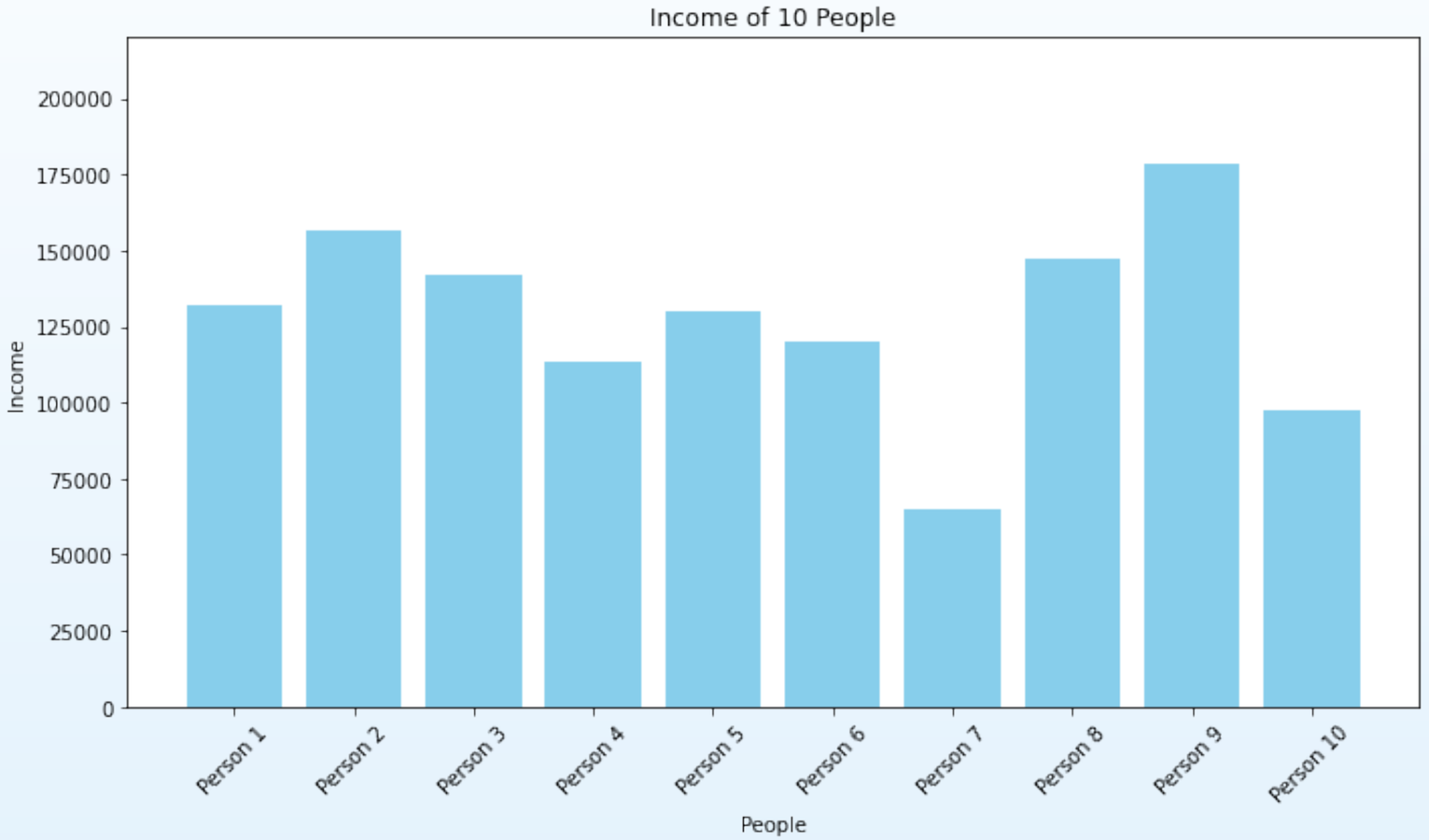


GINI COEFFICIENT

Gini Coefficient

Measure of inequality of a Group : Let us look at income inequality

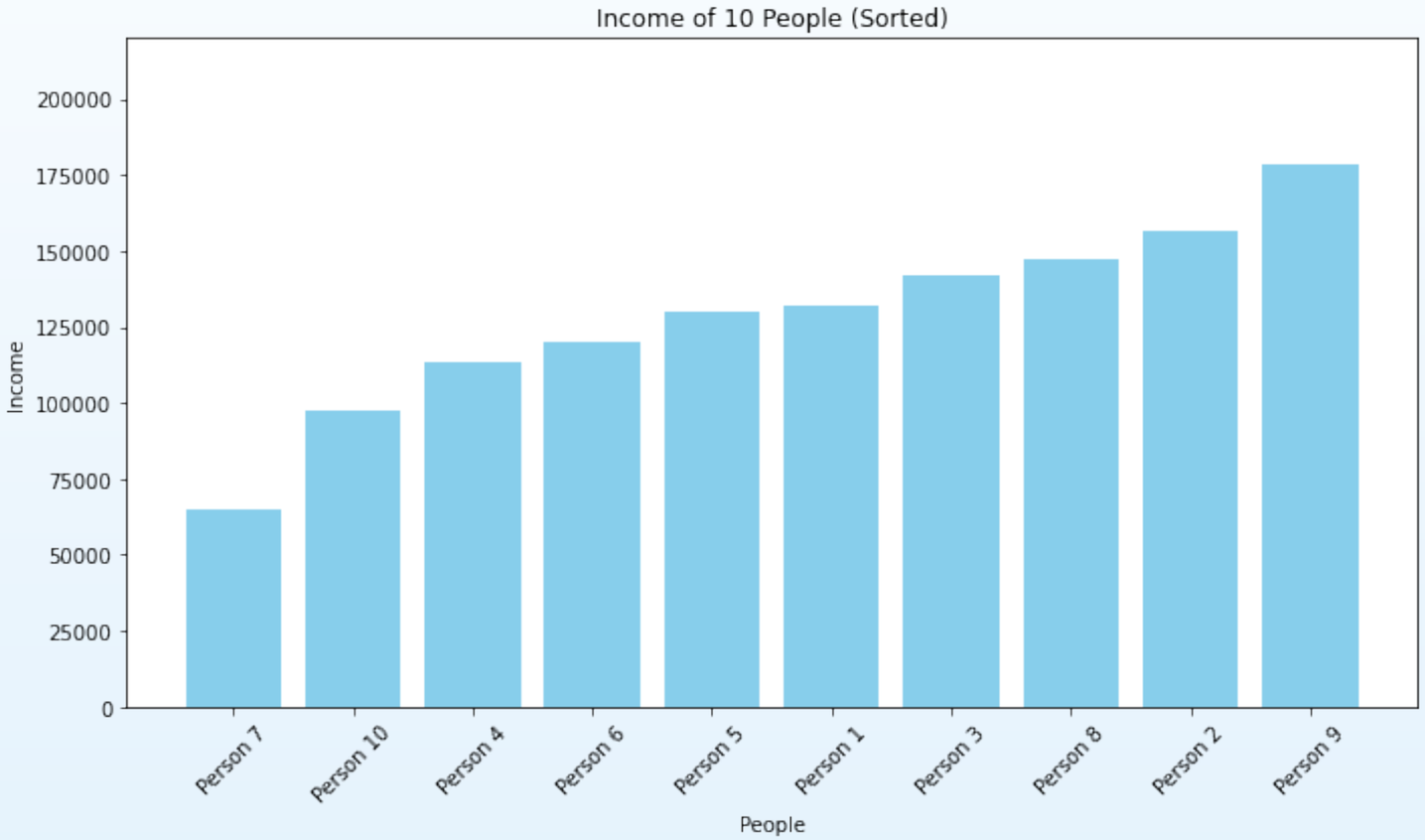
This is a population of people and their respective incomes



Gini Coefficient

Measure of inequality of a Group : Let us look at income inequality

Let us order the people based on their income

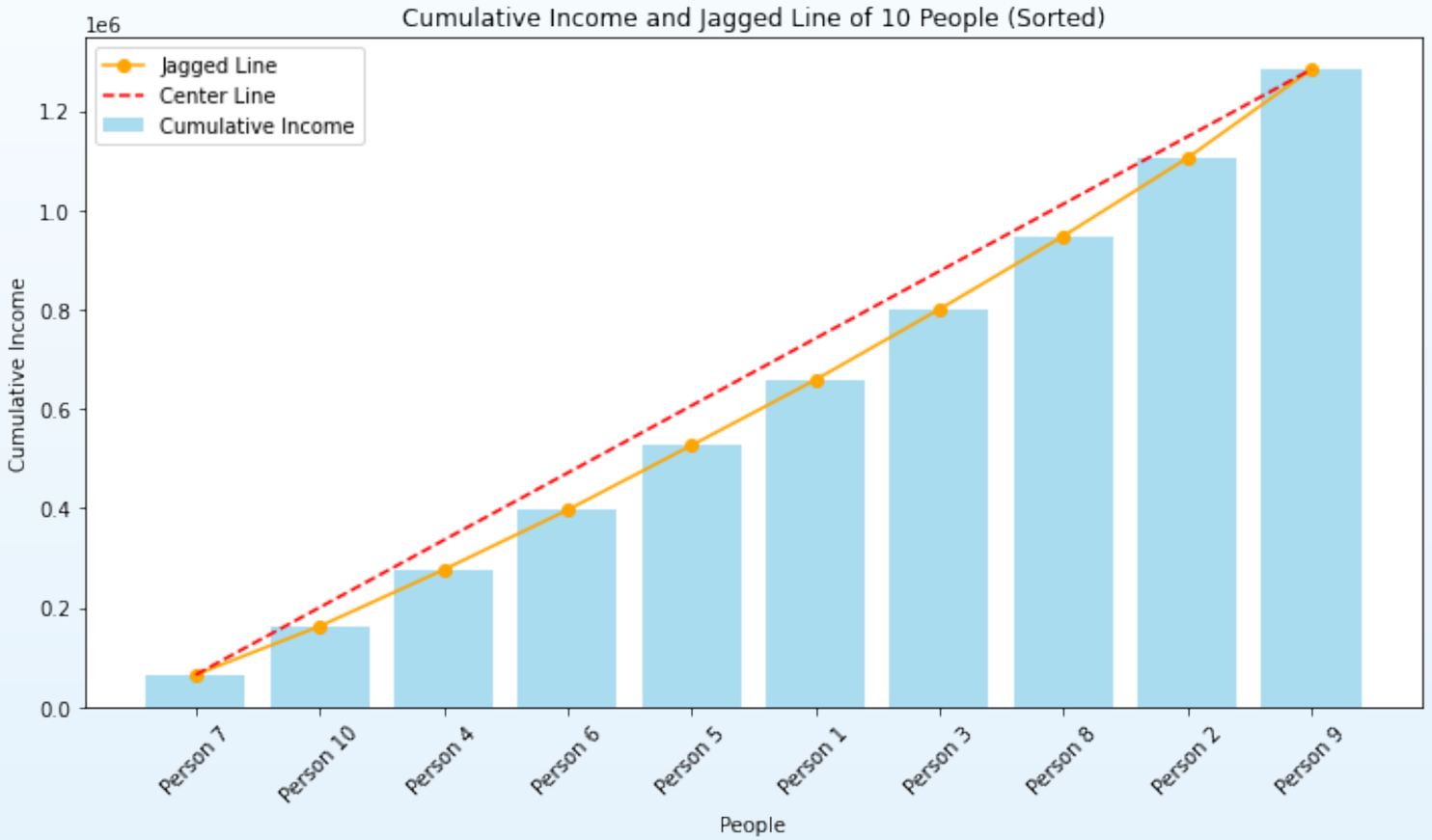


Gini Coefficient

```
Gini Coefficient: 0.12
```

Measure of inequality of a Group : Let us look at income inequality

Then take the cumulative sum of the ordered distribution.
This create two sections



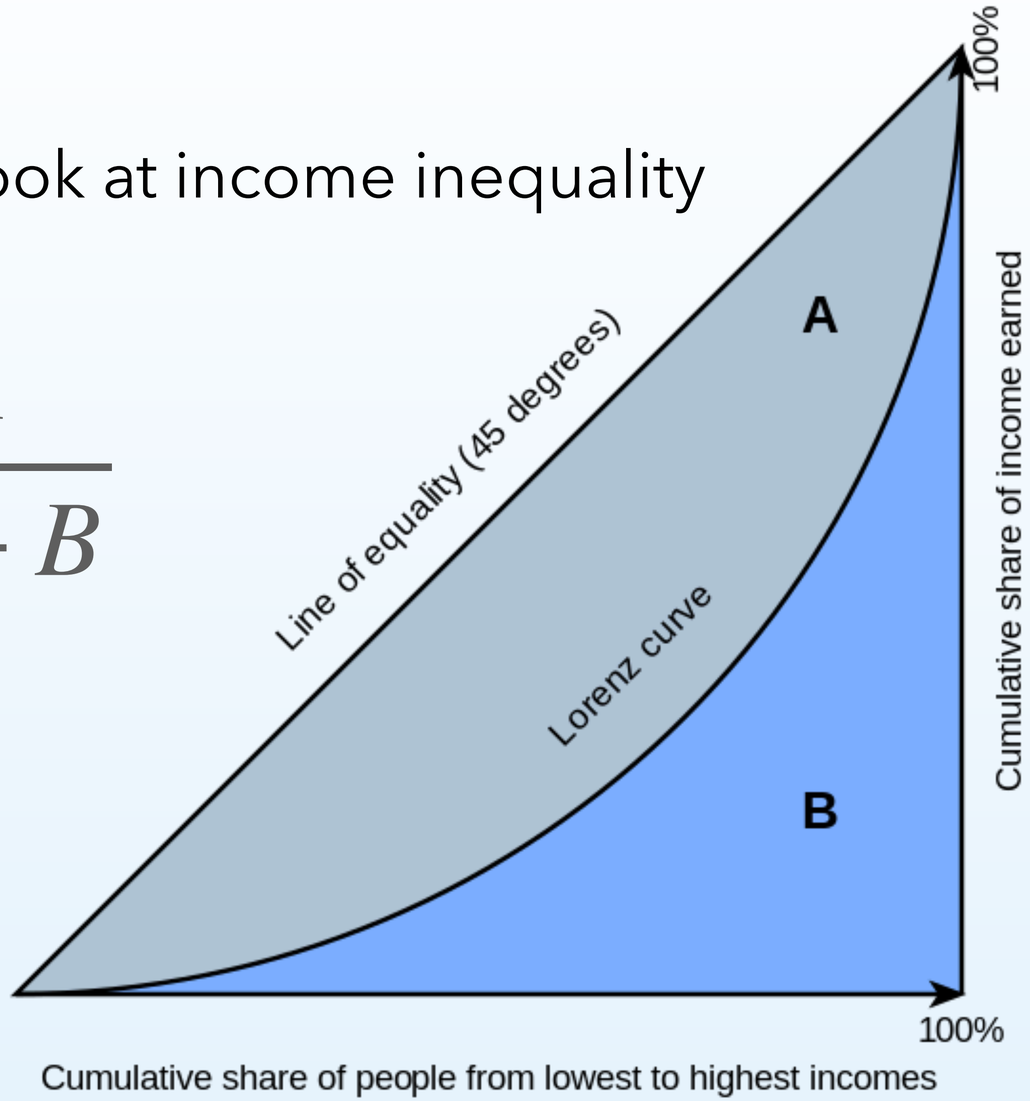
Gini Coefficient

Measure of inequality of a Group : Let us look at income inequality

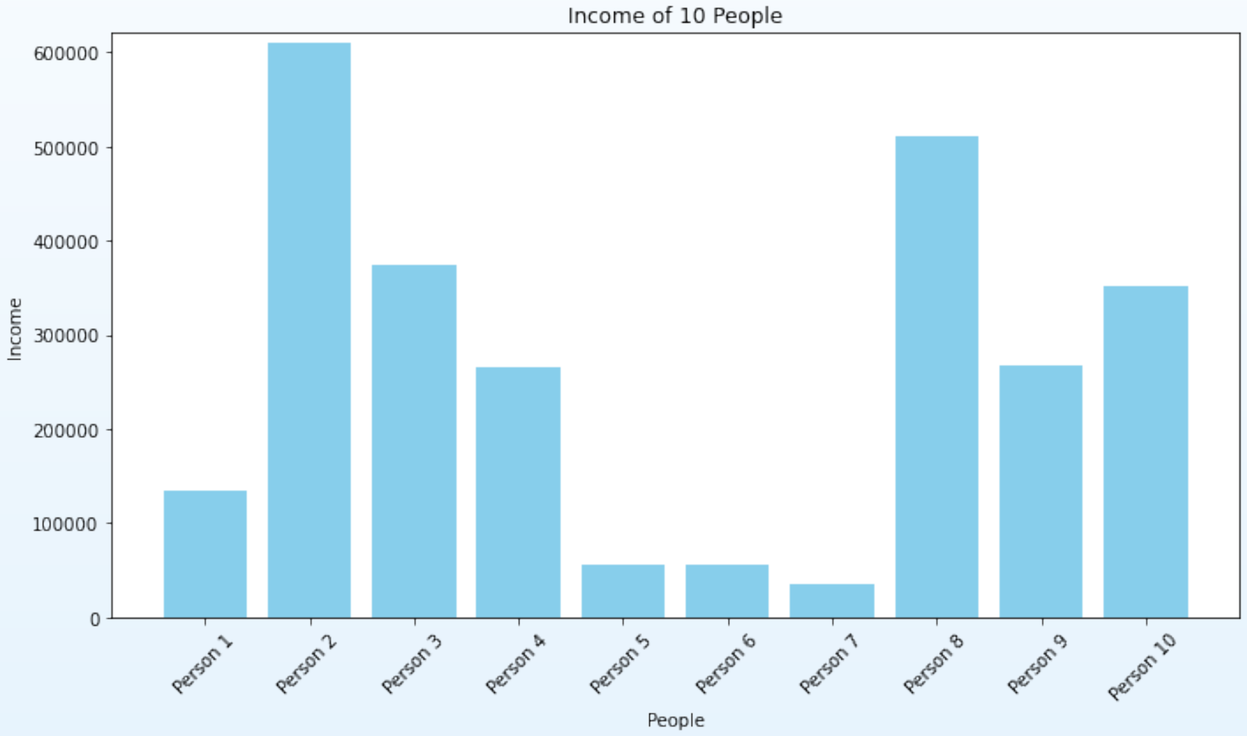
Gini closer to 0 means more equal

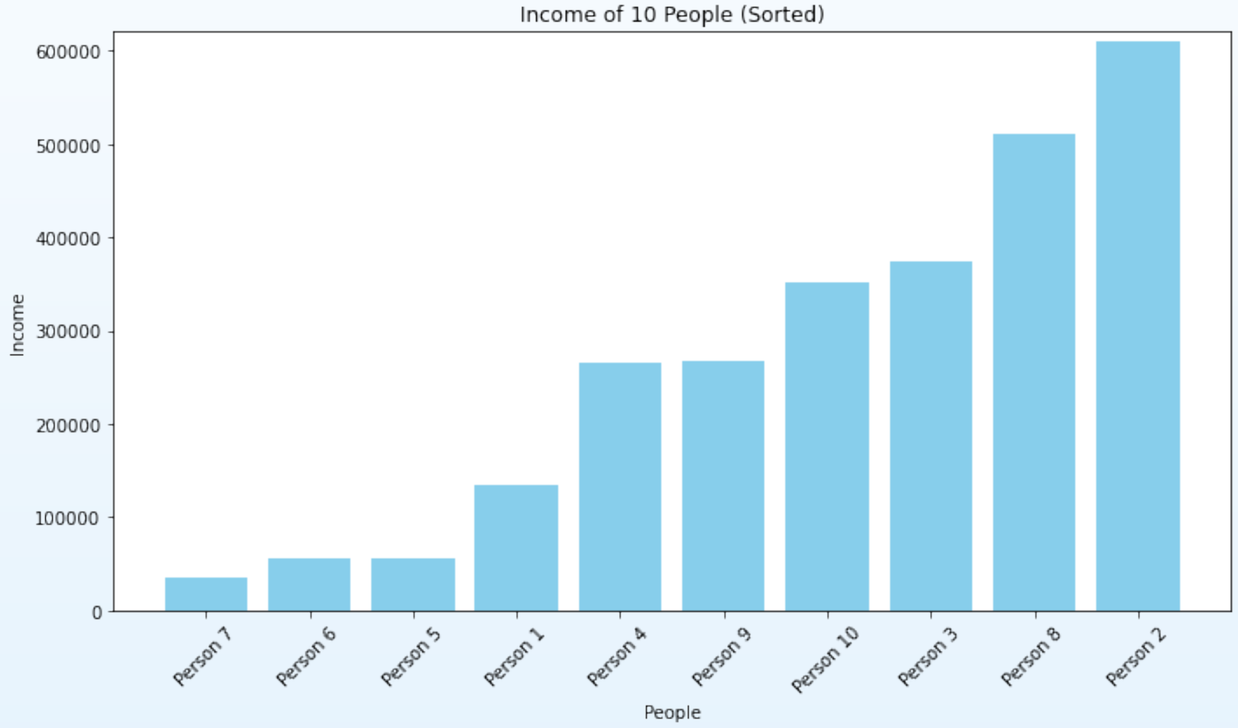
Gini higher is less equal

$$Gini = \frac{A}{A + B}$$



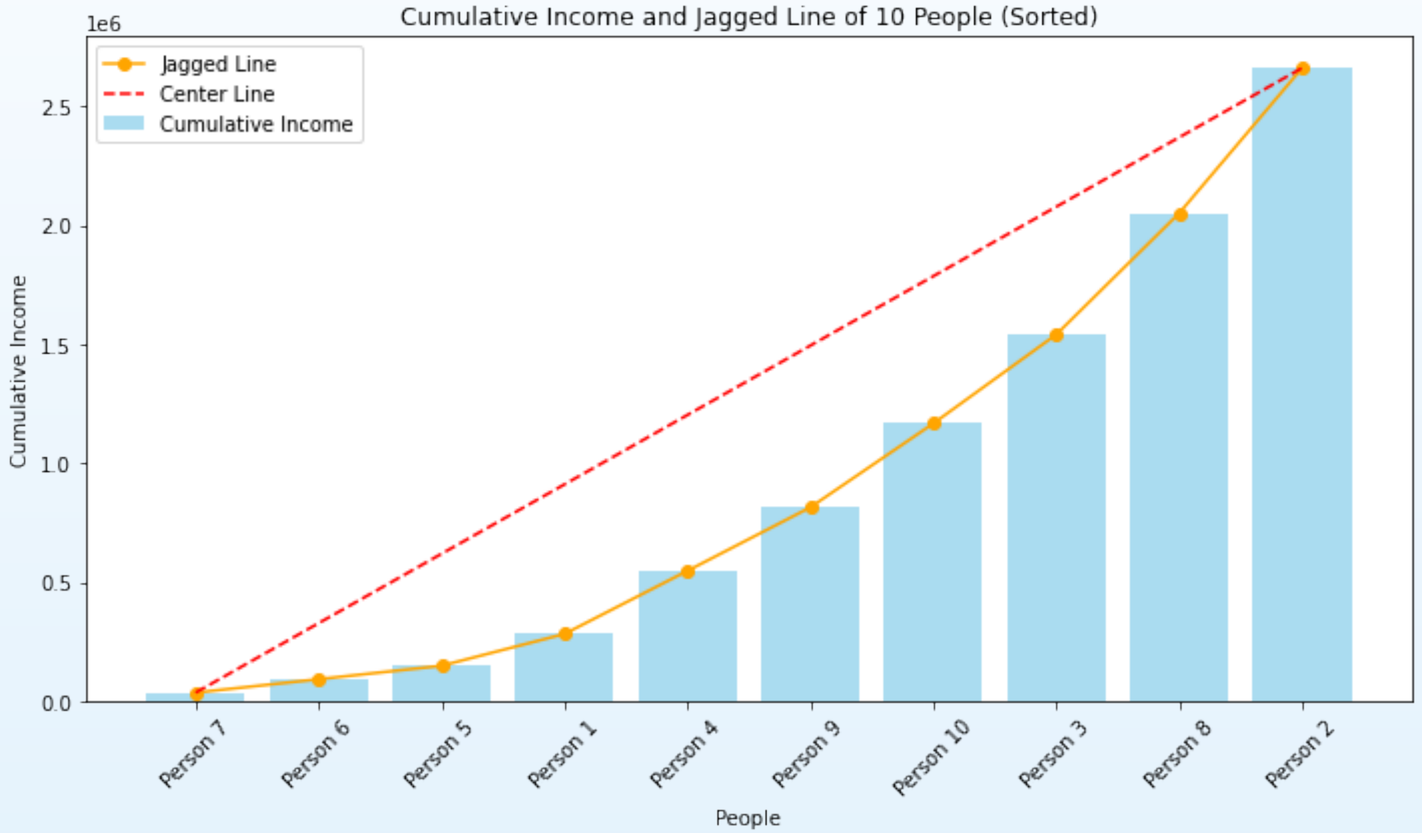
Inequal Population

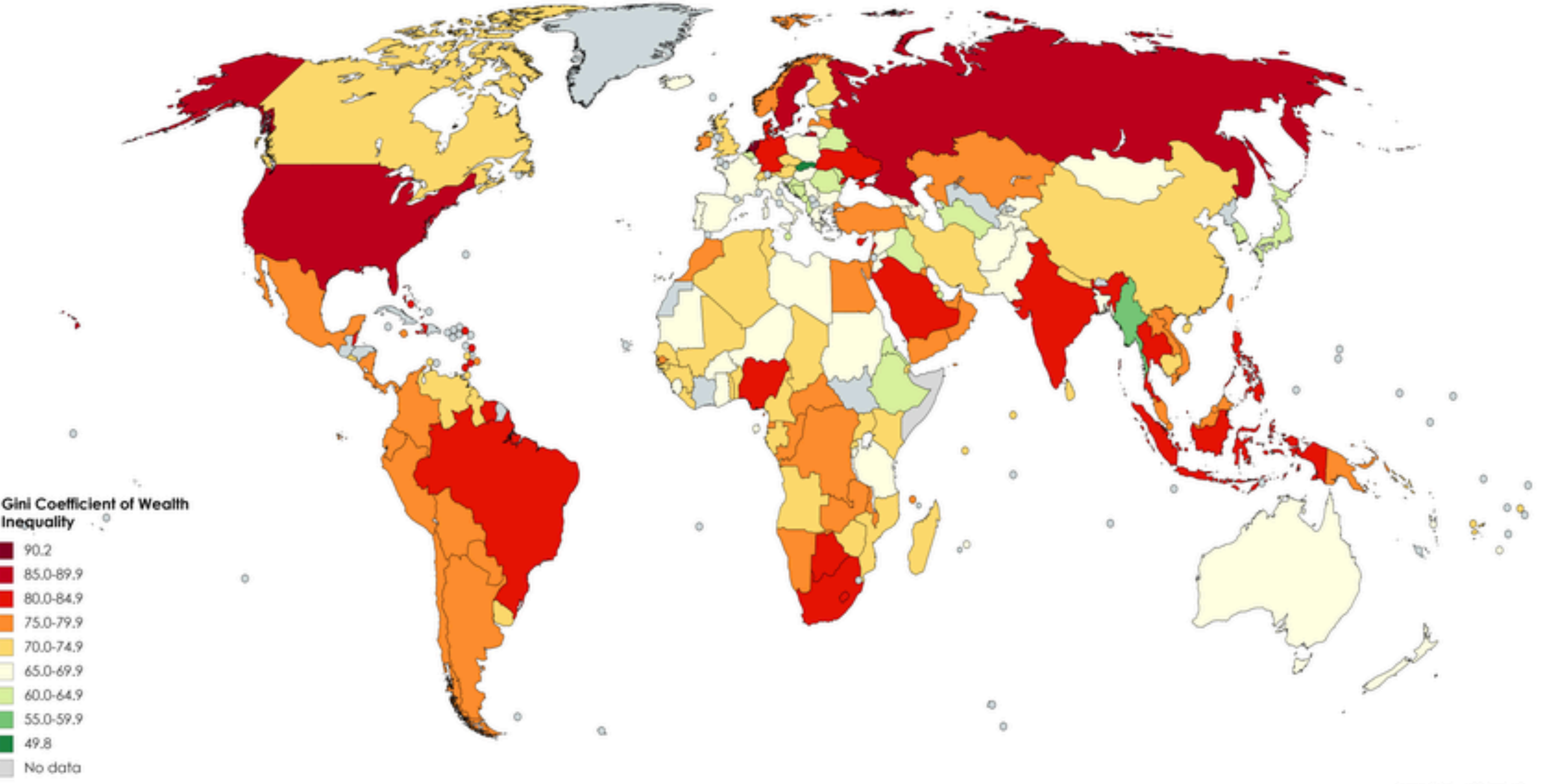




Inequal Population

```
Gini Coefficient: 0.25
```





Source: Global Wealth Databook, Credit Suisse, 2019, Pages 117-120

Created with mapchart.net ©

Applied Use of Decision Trees

Issues with Decision Trees (Weak Learning Models)

It is rare that the performance of a decision tree will beat models like neural networks or even regressions

Large amount of time to train model

Sensitive to outliers and poor data quality

Very hands-on for determining max depth, pruning and bagging/ boosting

Why use decision trees?

Highly Interpretable!

Gives insight into unknown decision making processes

College Admission

Court rulings/sentencing

Loan Decisions

Build a decision Tree

While a college admissions counselor may not be able to say why they decided to admit a student. A model trained on their behavior will be able to.

Putting statistics to “Wholesome admission process”

Example : Decision Tree

Parole Board Interviews

Can we make the decision of giving someone parole based statistically

```
df[ "interview decision"].unique()
```

```
'*', '**********', 'DENIED', 'OPEN DATE', 'GRANTED', 'OR EARLIER'
'NOT GRANTD', 'PAROLED', 'RCND&RELSE;', 'RCND&HOLD;', 'REINSTATE'
```

Example : Parole Decisions

Data Cleaning

X variables have to all be turned to numeric

X

e	race / ethnicity_ASIAN/PACIFIC	race / ethnicity_BLACK	race / ethnicity_HISPANIC
0	0	1	0
0	0	1	0

Y variables do not

y

	interview decision
660	DENIED
661	DENIED
662	DENIED
663	DENIED
664	DENIED

Data Cleaning

How would you convert these dates into useful numeric data

	birth date	parole eligibility date	parole board interview date
	2063-08-24	2009-12-11	*
	1973-08-05	2006-09-20	*
	1980-03-04	2012-04-12	*
	1982-02-01	2008-02-02	*
	1953-11-03	2001-10-20	*
	...	...	...
4	1988-09-01	2017-05-29	2016-06-*
5	1989-10-04	2018-03-08	2016-06-*
6	1948-10-15	1998-09-13	2016-06-*
7	1964-03-16	2009-10-14	2016-06-*
8	1955-01-26	2016-10-29	2016-06-*

rows × 3 columns

Example : Parole Decisions

Data Cleaning

As we have seen before, converting “dates” into something more useful is important

```
df_select_clean.loc[df_select_clean["age"] < datetime.timedelta(0),"age"] = df_select_clean[df_select_clean["age"] < datetime.timedelta(0)]["age"
df_select_clean["time_to_elig"] = df_select_clean["parole board interview date"] - df_select_clean["parole eligibility date"]
```

Model Fitting

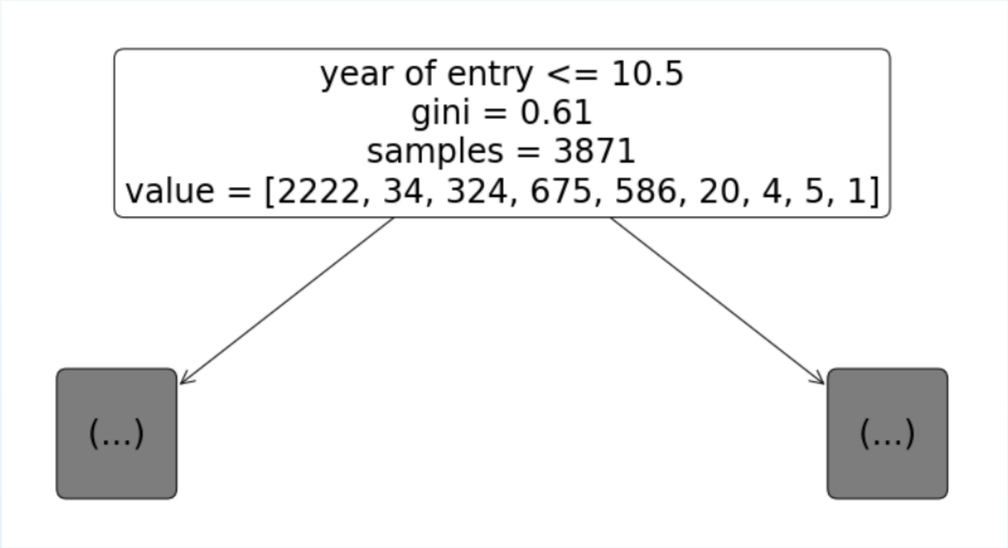
```
X = final.drop("interview decision", axis =1).astype(int)
y = final[["interview decision"]]


clf = tree.DecisionTreeClassifier(max_depth = 2)
clf = clf.fit(X, y)
```

Visualization

Plotting Trees

```
plt.figure(figsize = (20,8))
tree.plot_tree(clf, max_depth = 0, rounded = True, feature_names=X.columns)
```

This max depth will cut the model off

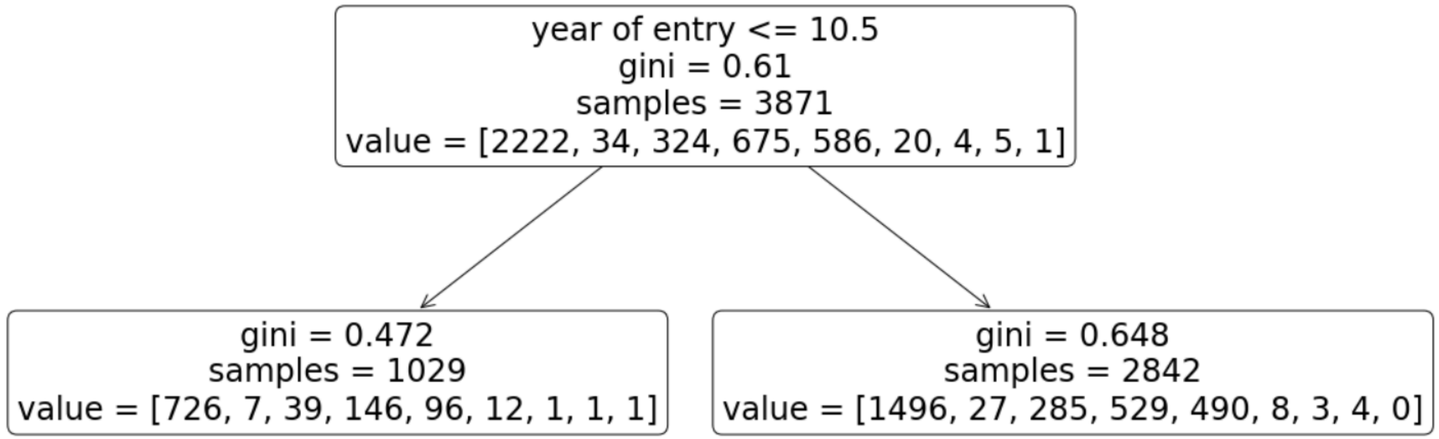


Decision tree

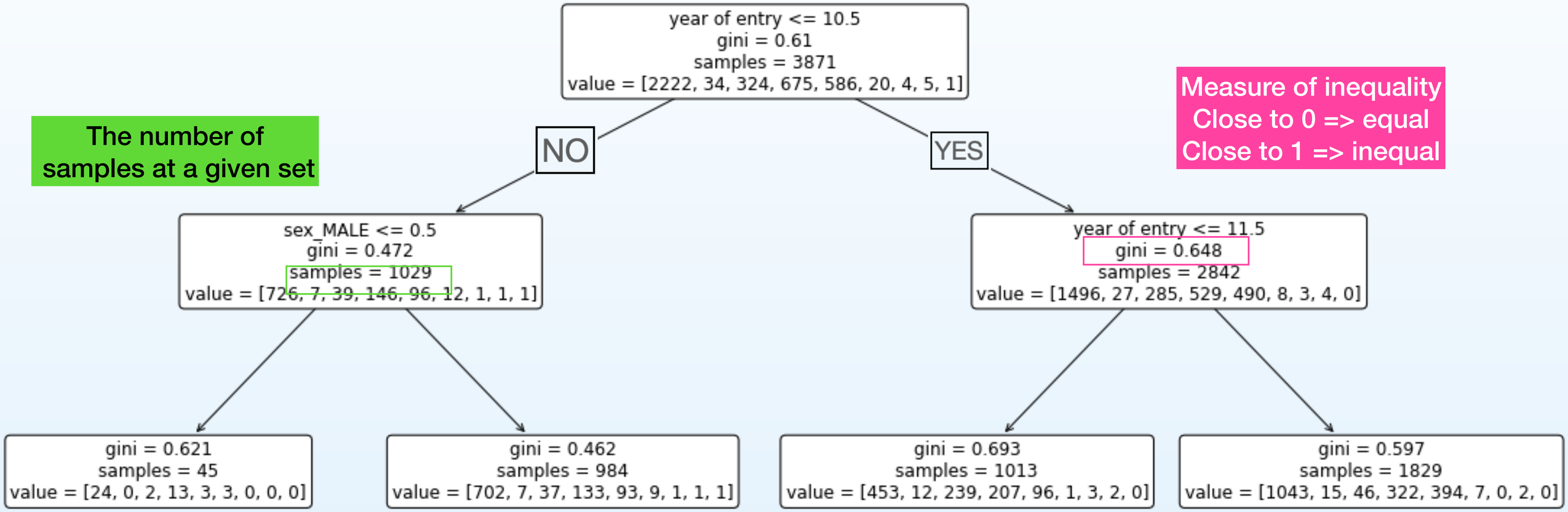
If this max_depth
Is larger than the max depth of
The model

```
plt.figure(figsize = (20,8))
tree.plot_tree(clf, max_depth = 1, rounded = True, feature_names=X.columns)
```

```
[Text(0.5, 0.75, 'year of entry <= 10.5\ngini = 0.61\nsamples = 3871\nvalue =
[2222, 34, 324, 675, 586, 20, 4, 5, 1]'),
 Text(0.25, 0.25, 'gini = 0.472\nsamples = 1029\nvalue = [726, 7, 39, 146, 96,
12, 1, 1, 1]'),
 Text(0.75, 0.25, 'gini = 0.648\nsamples = 2842\nvalue = [1496, 27, 285, 529,
490, 8, 3, 4, 0]')]
```



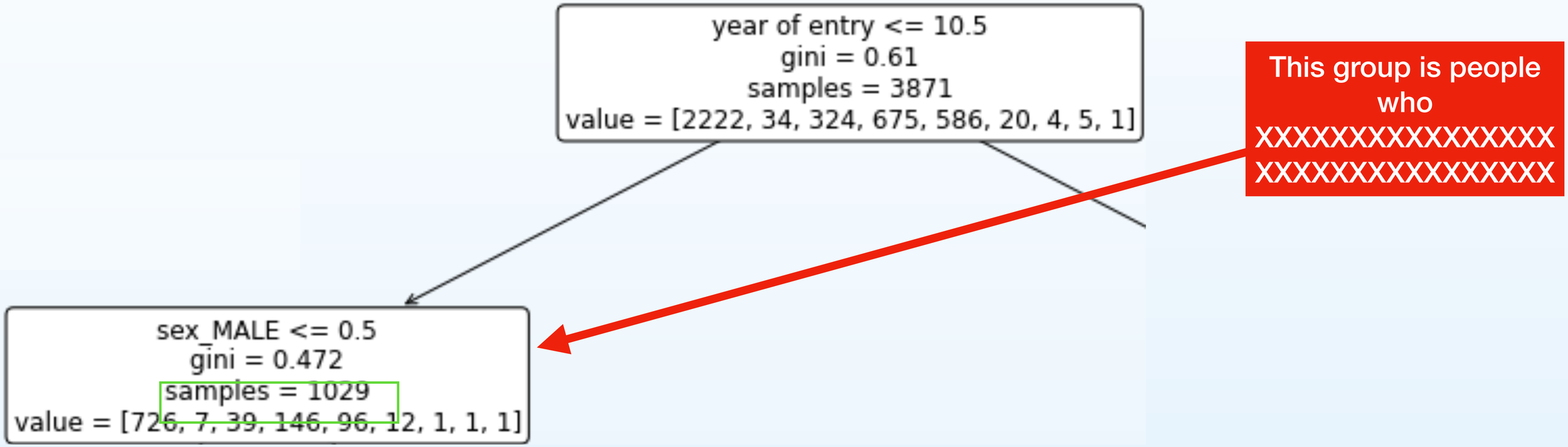
Output



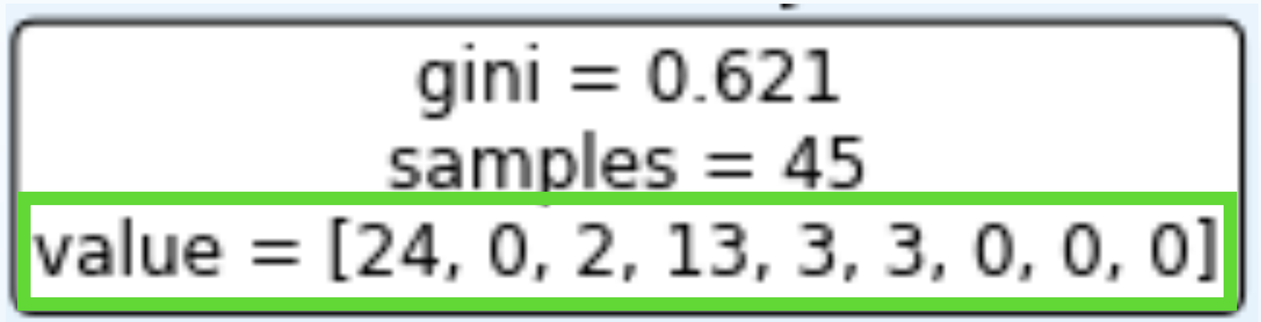
The number of samples at a given set

Measure of inequality
Close to 0 => equal
Close to 1 => inequal

Understanding the Leafs



Value Interpretation



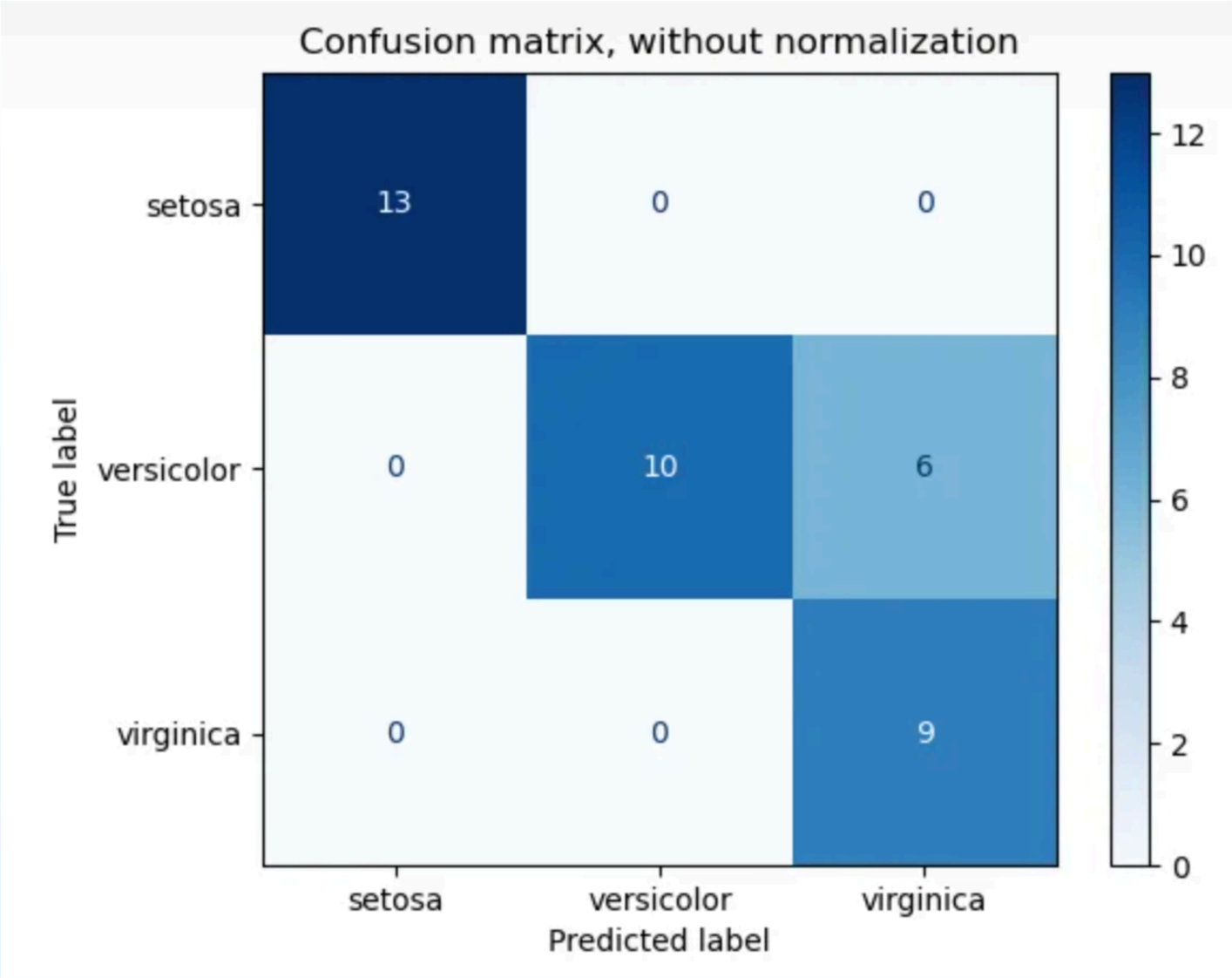

```
'*', '**********', 'DENIED', 'OPEN DATE', 'GRANTED', 'OR EARLIER'
'NOT GRANTD', 'PAROLED', 'RCND&RELSE;', 'RCND&HOLD;', 'REINSTATE'
```

Out of the 45 samples in this group :
24 are DENIED
2 GRANTED
13 OR EARLIER
3 NOT GRANTED
3 PAROLED

ROC, AUC, Precision, Recall and F1

We saw a confusion Matrix

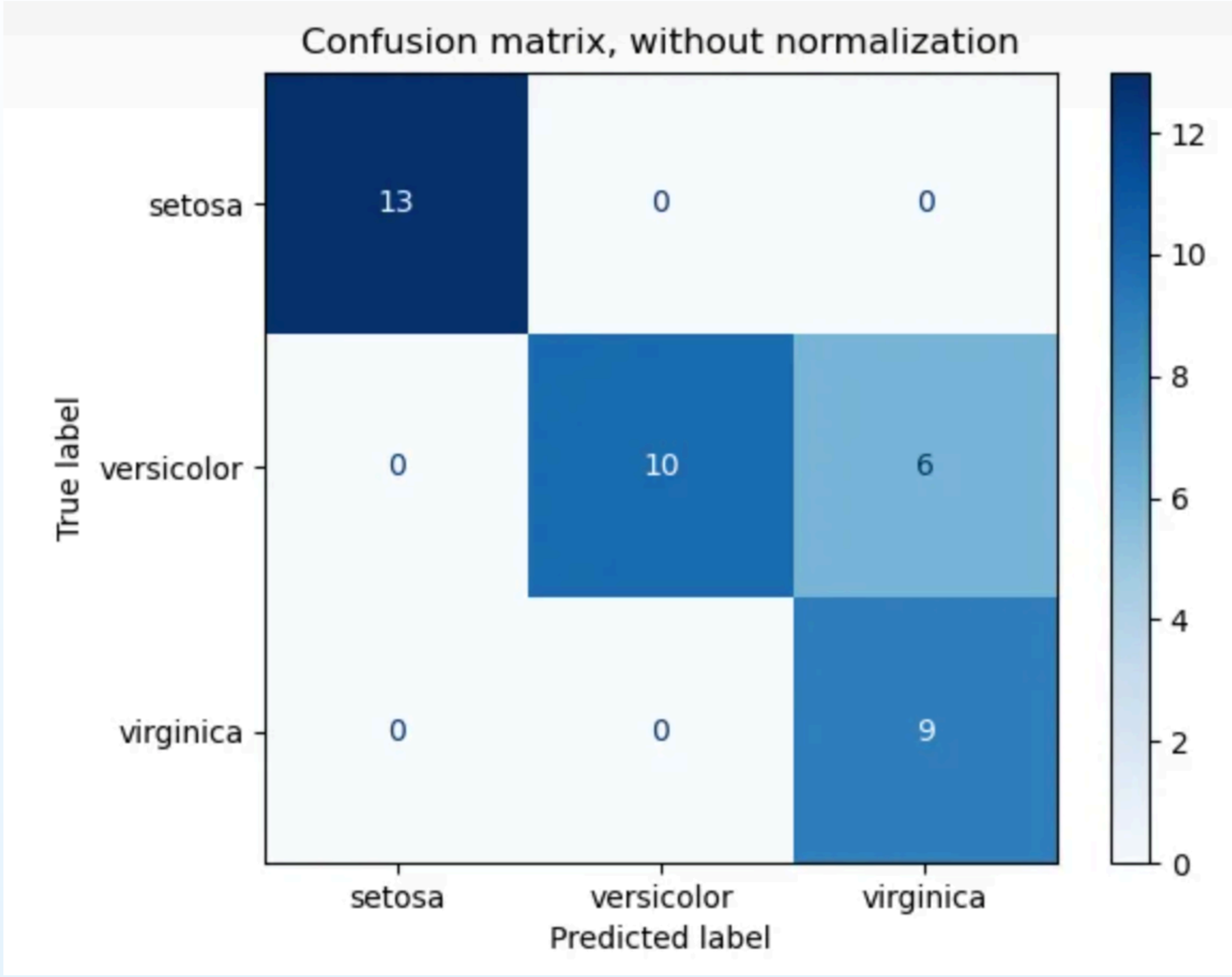
Recall True Positives, True Negatives, False Positives and False Negatives



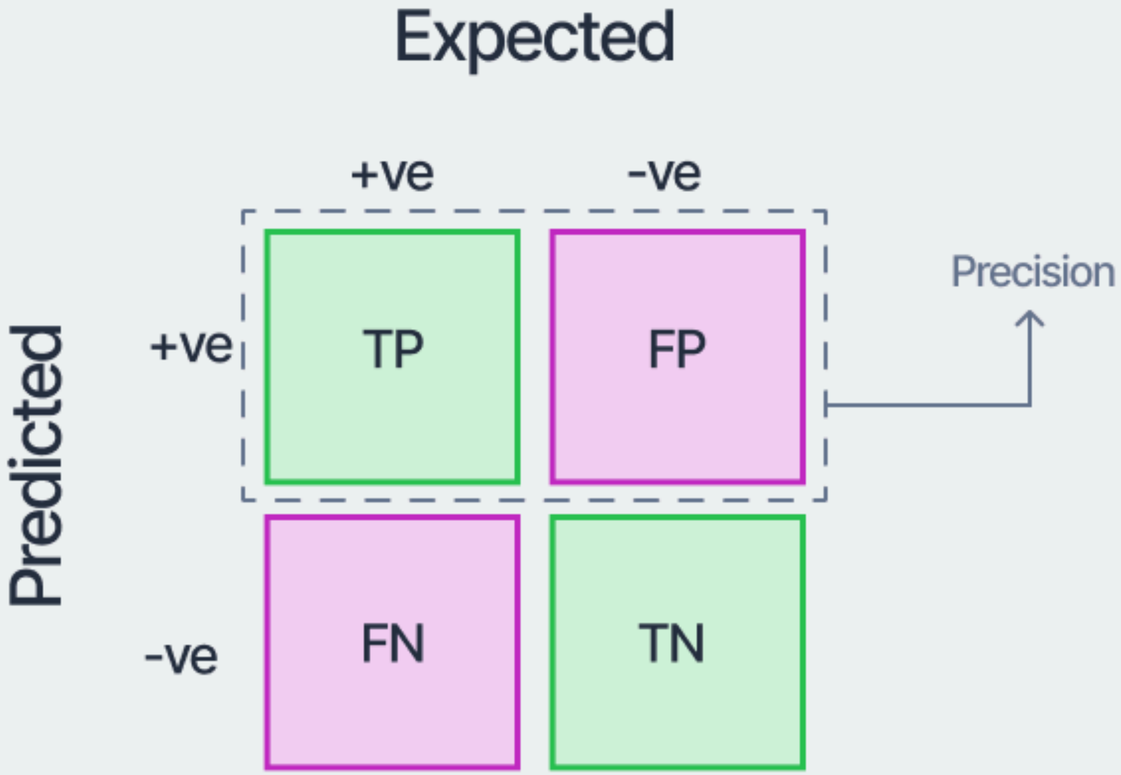
Precision and Recall

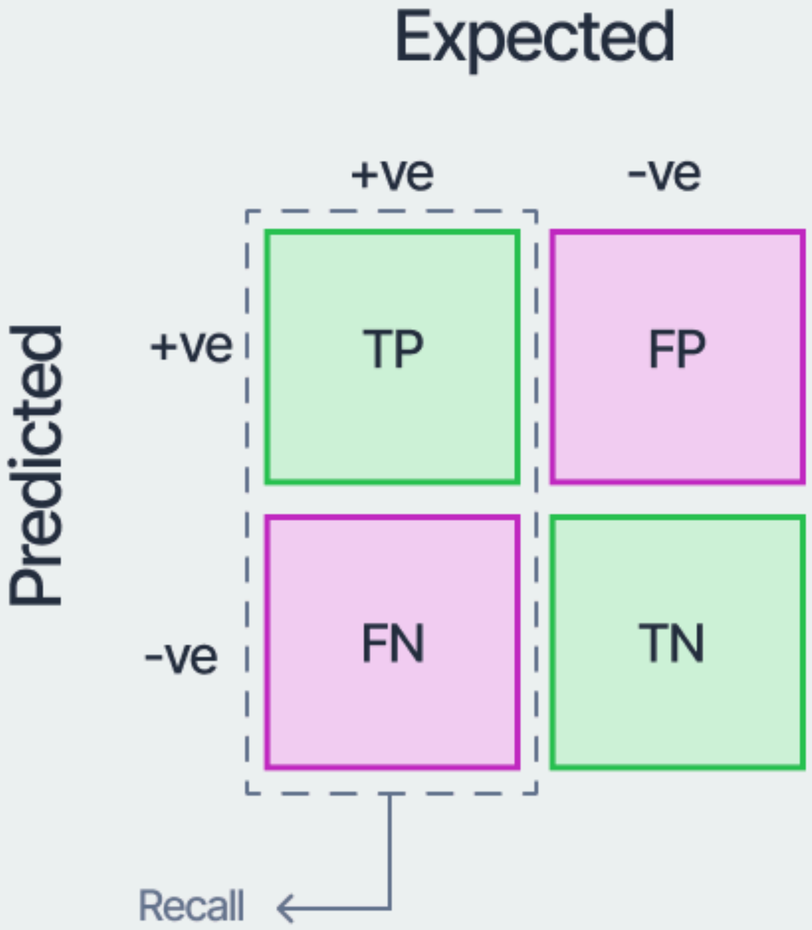
$$Precision = \frac{(true\ positives)}{(true\ positives + false\ positives)}$$

$$Recall = \frac{(true\ positives)}{(true\ positives + false\ negatives)}$$



Visual Intuition





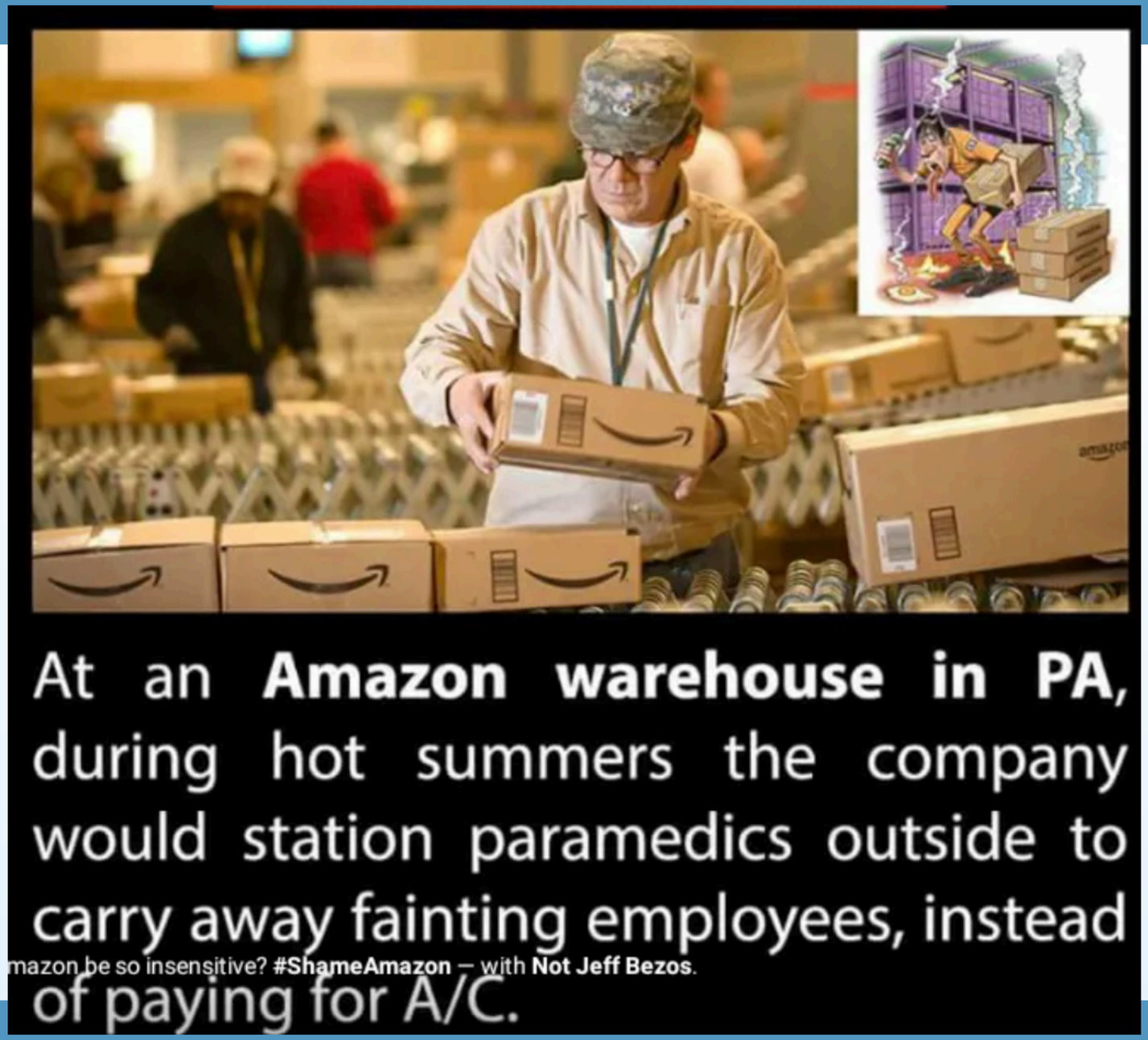
F1-Score

$$F1\ score = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

In-Class Assignment

ONLY NEED TO DO Q5-8.)

5.) Looking at the graphs. what would be your suggestions to try to improve customer retention? What additional information would you need for a better plan. Plot anything you think would assist in your assessment.
6.) Using the Training Data, if they made everyone work overtime. What would have been the expected difference in client retention?
7.) Is it profitable for the company to remove overtime? If so/not by how much?
8.) Use your model and get the expected change in retention for raising and lowering peoples income.



mazon be so insensitive? **#ShameAmazon** – with **Not Jeff Bezos**.

Amazon Responds To Release Of Leaked Documents Showing 150% Annual Employee Turnover