WEEK 8: CLASSIFICATION METHODS

Dr. Kai Li
School of Information Sciences
University of Tennessee, Knoxville
Spring 2025

Feedback from the survey

- 1. Write code to run the analysis.
  - *I will give you more explicit instructions in the new activities.*
- 2. Group lab
  - *We will do the first two things from this week!*
- 3. Statistics concepts
  - *A few points: (1) this is not a statistical class: statistics is just an (important) application, (2) we don't have the time to talk about all concepts, so I am trying to cover the most basic and classic concepts and models; (3) I was trying to focus on why and how, but certainly not mathematics behind them.*
  - *If you want to learn more about statistics, you certainly need to spend much more time beyond the class.*

Feedback from the survey

- 4. R code and skills
  - *I have some similar points here, that this is not a programming language class, so it's simply impossible teach all the programming skills (as simple as R).*
  - *But teaching/taking this class without any previous technical class is a major challenge for both you and me!*
    - From my side, I am still trying to figure out a better way to deliver the code.

Review

- Regression methods
  - *Regression quantifies as one variable changes by one unit, how the other variables will change accordingly.*
    - For example, coefficient = .8 between variables A and B suggests that as A (IV) increases by one unit, B (DV) will increase by 0.8 unit.
    - So it is the foundation of machine learning (i.e. predicting values).
  - *Correlation measures the strength of a relationship between two variables*
    - For example, coefficient = .8 suggests that variables A and B are highly positively correlated.

Announcements

- I will try the group activity approach from this week.
  - *Feel free to do it yourself if you don't need a group.*
  - *A group (or yourself) will need to answer the questions in the code file.*
  - *Everyone will need to post your answer.*

Announcements

- The plan for the next few weeks:
  - *No class next week (Spring break) and April 17 (Spring recess)*
  - *We have three more weeks:*
    - I am going to use 1-1.5 weeks talking more about statistics and then 1.5-2 weeks talking about visualization.
    - We will talk about unsupervised learning methods and then some generally visualization techniques and concepts.
  - *We will use the last two weeks (April 25 and May 1) for the final project presentations.*

Assignment 2

- This is a semi-structured report for you to answer a series of questions.
- You probably need to save all the four dataset to your folder before doing the assignment.
- Deadline: April 10

Review / Overview

- Regression
- Classification

Numeric vs. Categorical IV coefficients

- The coefficient of any numeric IV shows as the IV increases by one unit (from 0 to 1 dollar of income), the change of the DV.
- The coefficient of any categorical IV shows that comparing to the baseline category (which is not shown in the table), what impact of the shown category has on the outcome.
  - *X1 has 7.6 points lower than X0, on average.*

```
Call:
lm(formula = happiness ~ income, data = income.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.02479 -0.48526  0.04078  0.45898  2.37805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20427    0.08884   2.299   0.0219 *
income       0.71383    0.01854  38.505   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	51.678	.982		52.619	.000
	X1	-7.597	1.989	-.261	-3.820	.000
	X2	3.945	2.823	.095	1.398	.164
	X3	-5.855	2.153	-.186	-2.720	.007

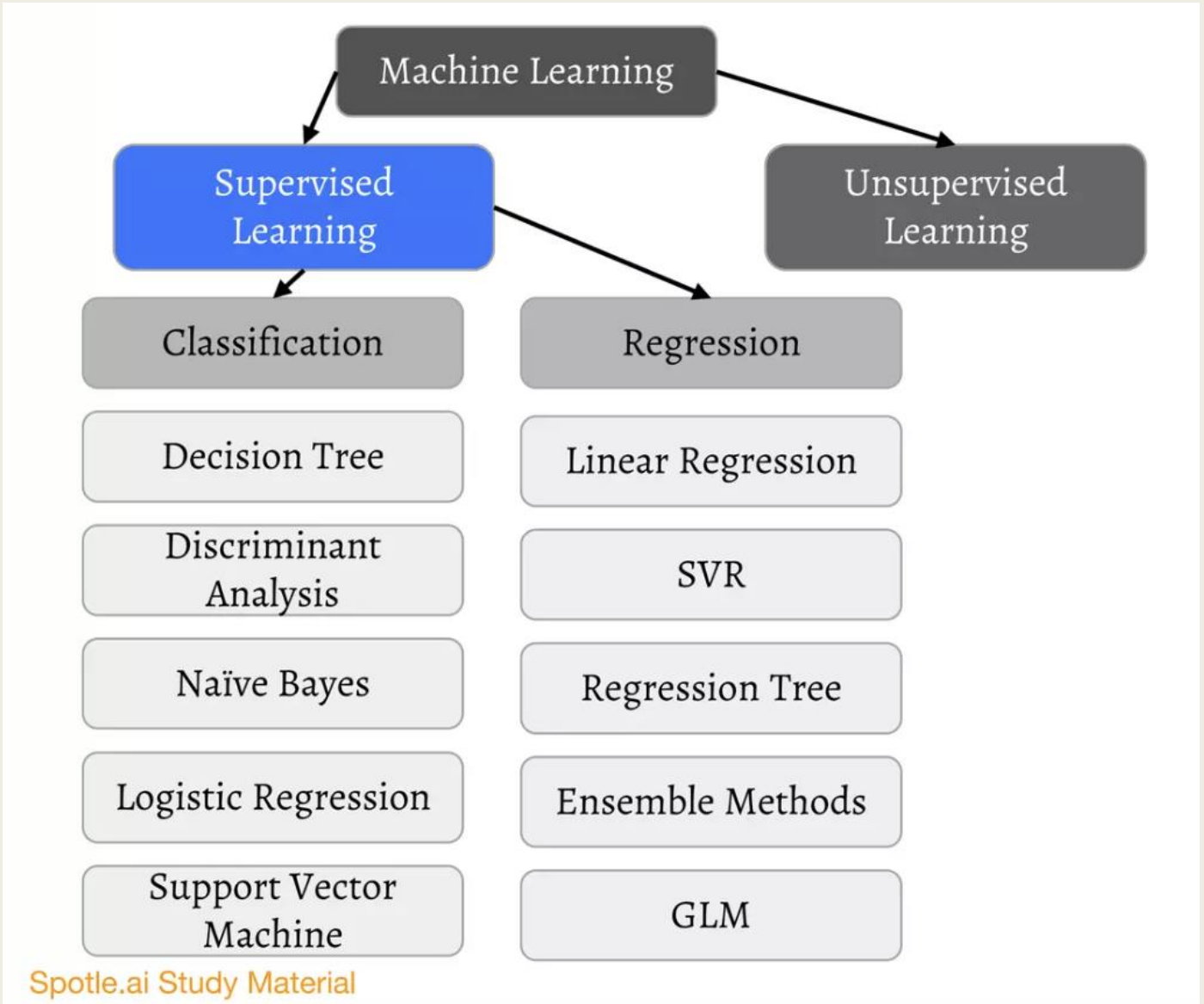
a. Dependent Variable: WRITE

Accuracy evaluation

- Some measurements to evaluate the performance of regression models.
  - *R^2 is the direct evaluation of goodness-of-fit of the model, by showing the share of variance in the data that can be explained by the model. The value is between 0-1 and higher is better (> 0.5 as a rule of thumb but it really depends).*
  - *RMSE (root-mean-squared-error) shows the mean difference between predicted value and the actual value.*
    - In our demonstration, our predicted values is 11.3 points different from the actual values (remember that the value range is 0-100).
  - *MAE (mean absolute error) is another way to calculate the difference between predicted and actual values.*

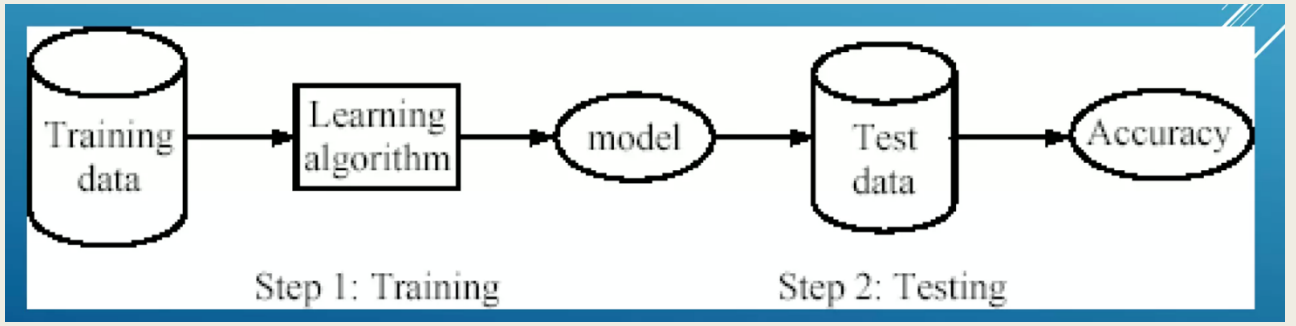
What is machine learning?

- Machine learning (ML) is to program an algorithm to learn from data, so that the algorithm will solve the problems.
  - *For example, in regression methods:*
    - We first use the training set to train a model, which is the process for the algorithm to learn about the relationship between DV and IVs.
    - Then, we can ask the algorithm to use what it learnt to predict values in the testing set.
  - *Classification methods work in the same way but deal with a categorical DV.*



https://www.slideshare.net/SpotleAI/supervised-and-unsupervised-machine-learning

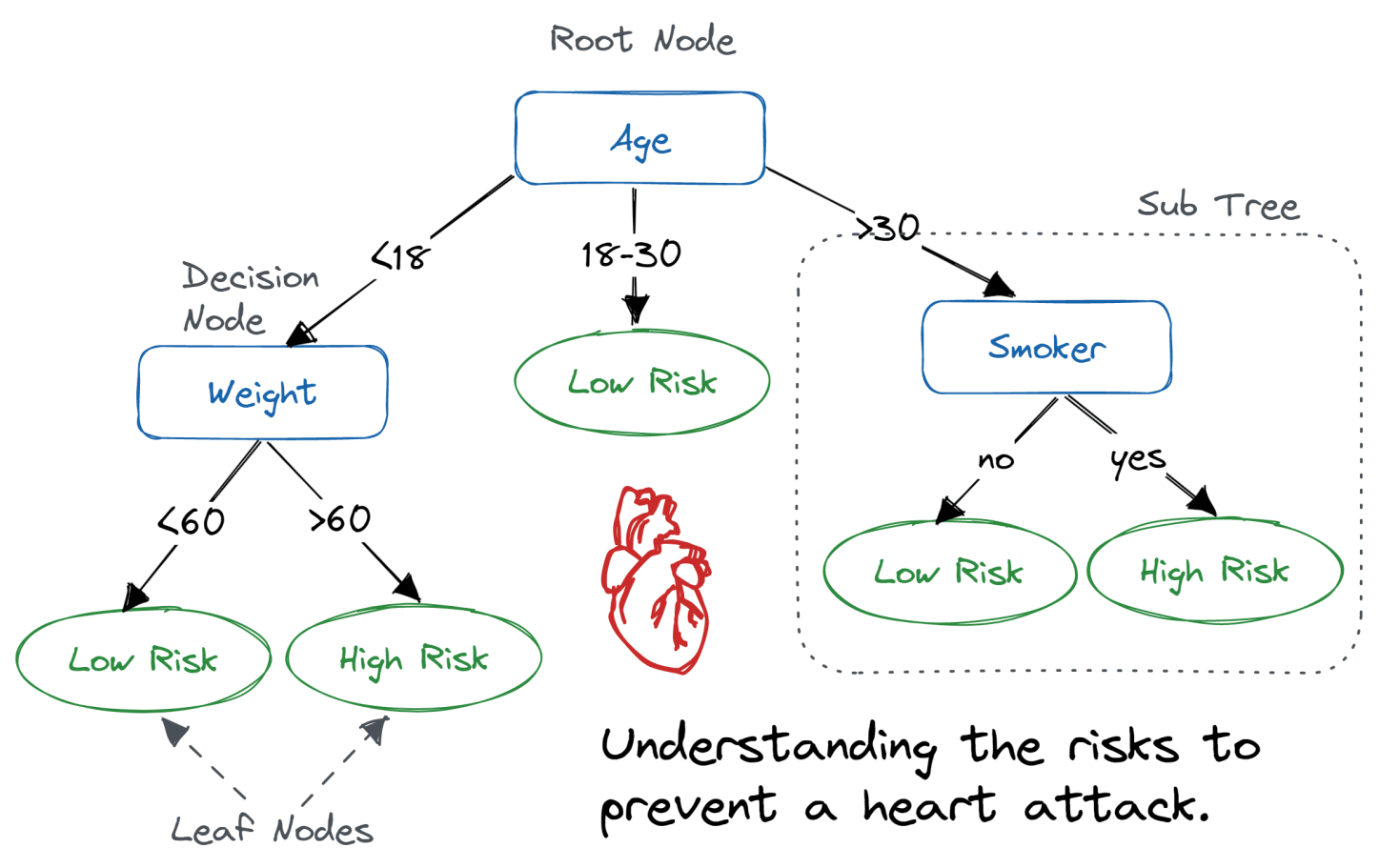
Classification



- Classification methods deal with a categorical DV.
  - *Remember that we can transform a numeric variable into a categorical variable if we want to use a classification method.*
    - For example, we can classify the size of flower into large and small categories.
- By using classification methods, we can still get insights from the data (inference) and/or predict values (prediction).

Classification methods

- Some general methods:
  - *Decision tree*
    - Random forest
  - *Logistic regression*
- These methods are based on different logics. In practice, they don’t have much difference in terms of applicable scenarios.
- We may want to try different models and use the best one for a dataset.

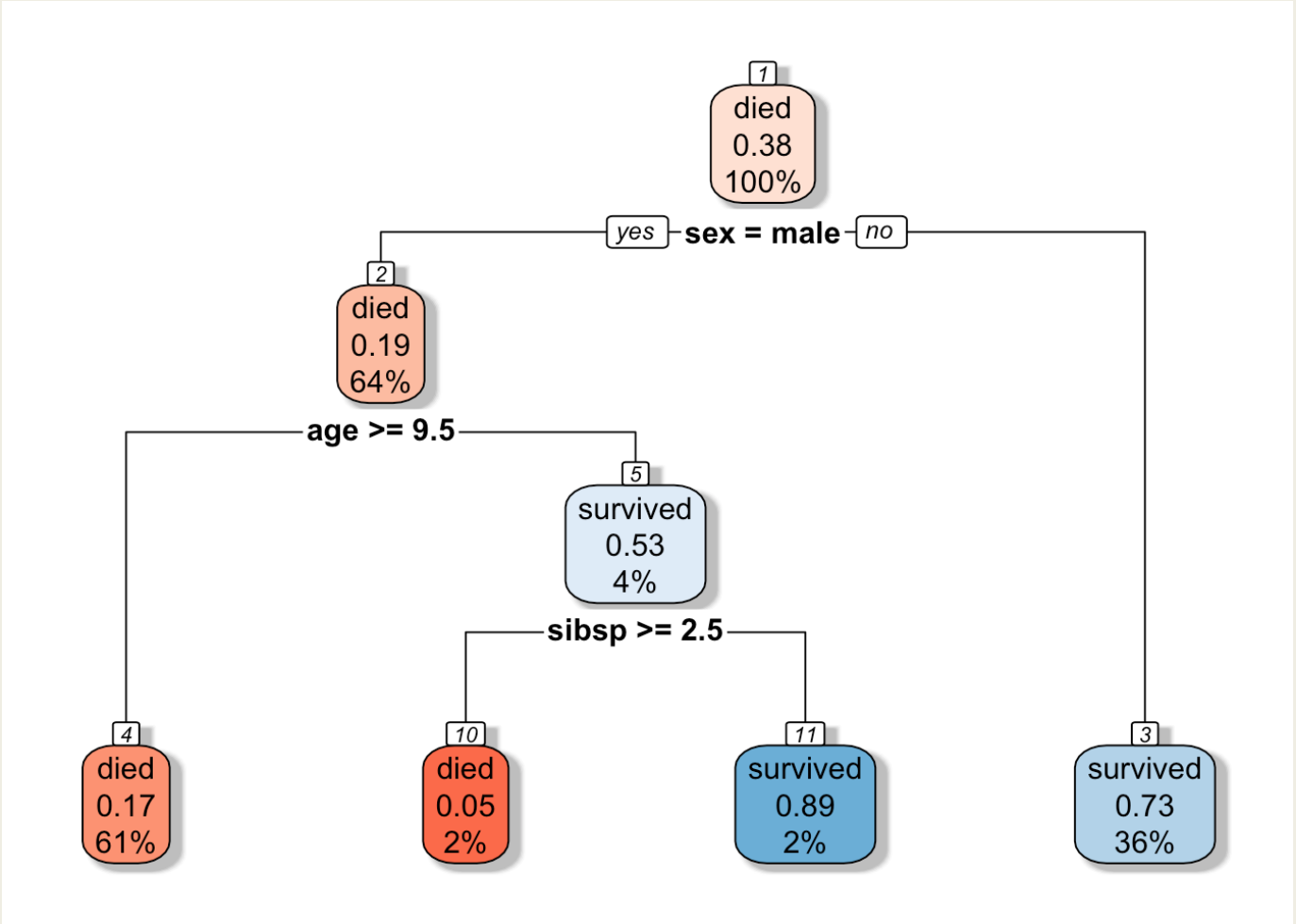


Decision tree

- Decision tree uses a hierarchical tree model to classify all data points into categories.
- It is still a supervised learning method that uses training data to construct the hierarchical model based on variables in the dataset.

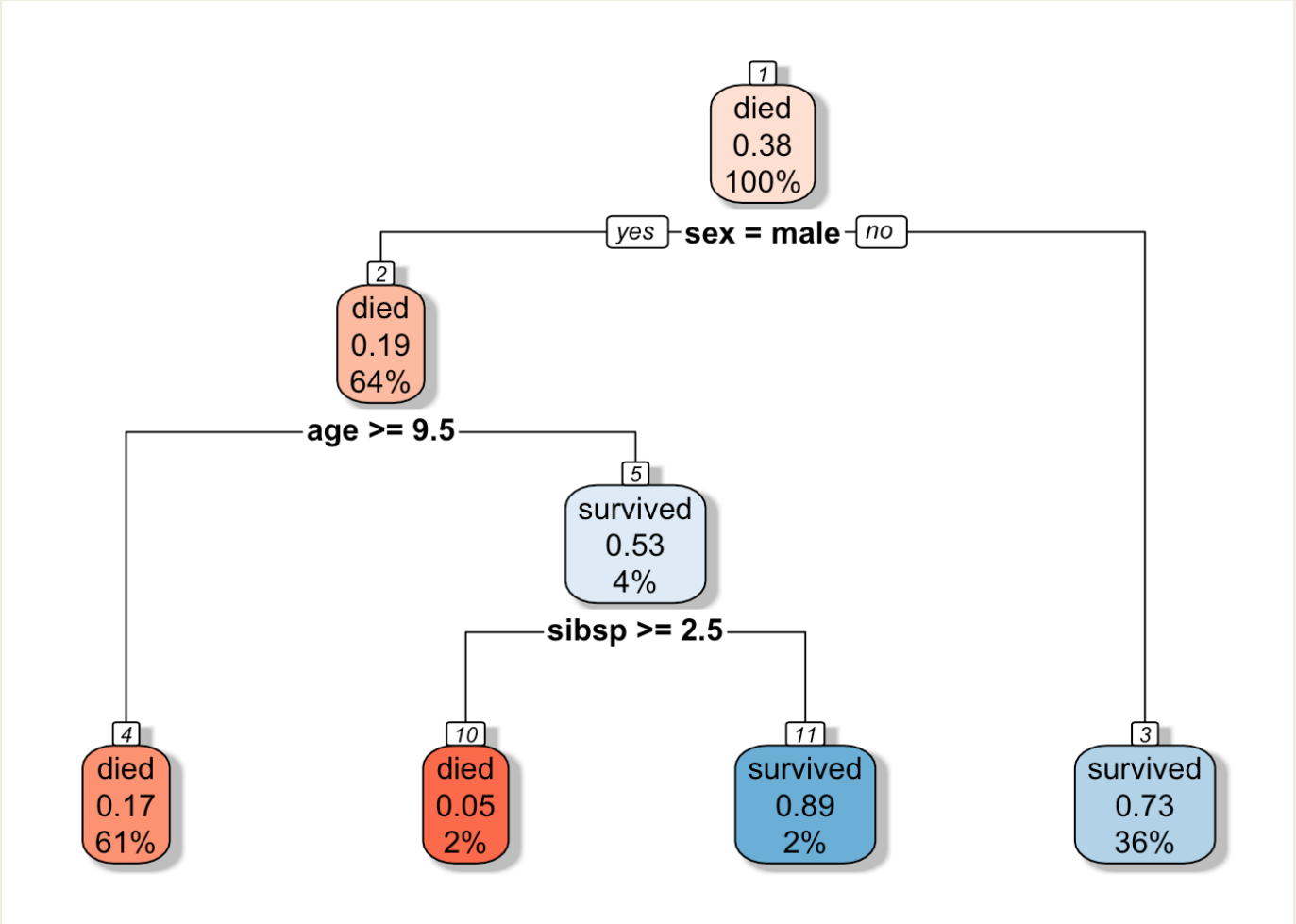
Results of decision tree

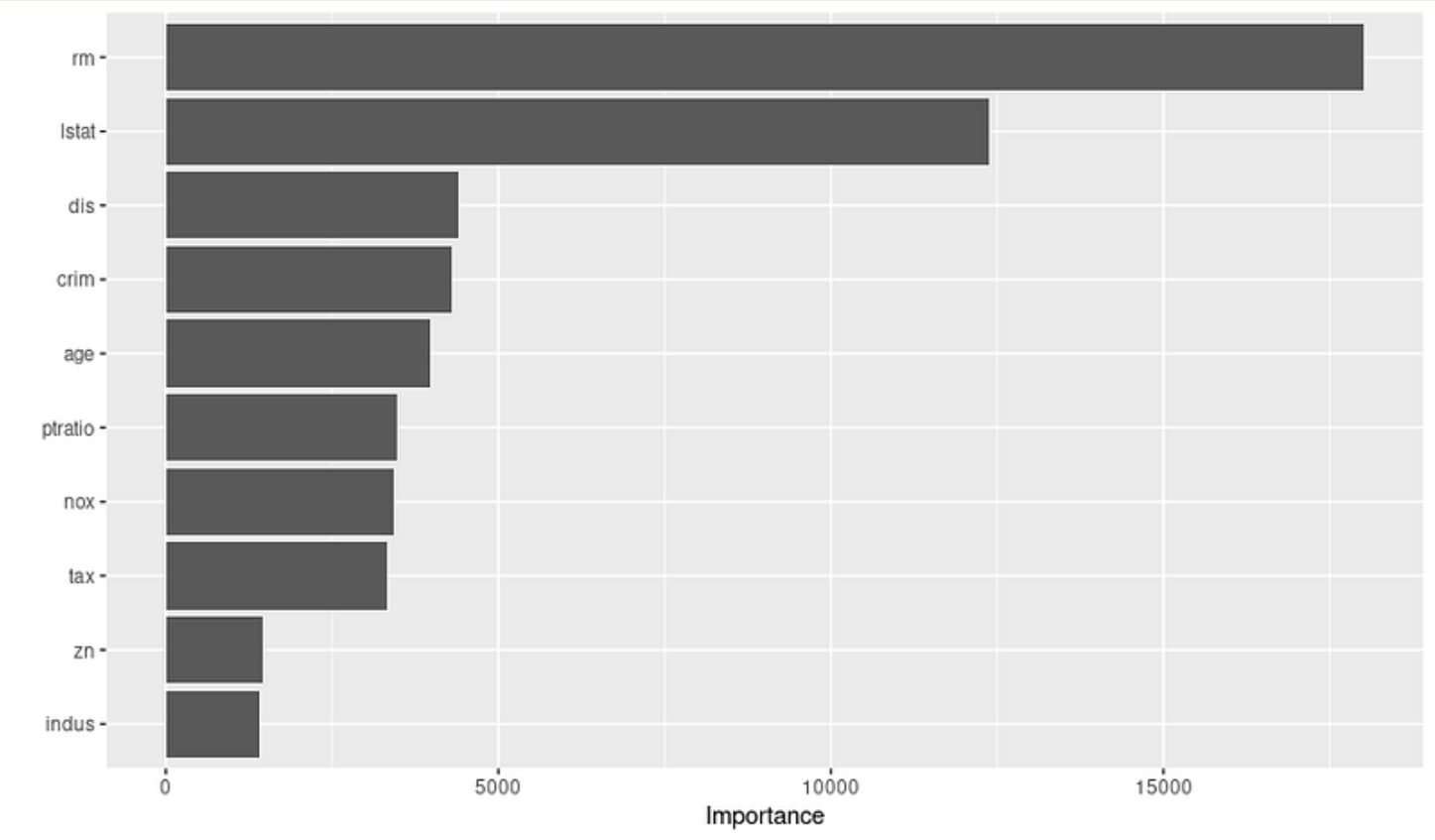
- The results of decision tree models can be visualized into this diagram, showing:
  - *Decisions*
  - *Group size*
  - *Final outcome rate*



Results of decision tree

- For example, 36% of the population is female and they have a 73% of survival rate.
- For all male population, their overall survival rate is 19%.
- But those boys whose age is lower than 9.5 ages, their survival rate is 53%.





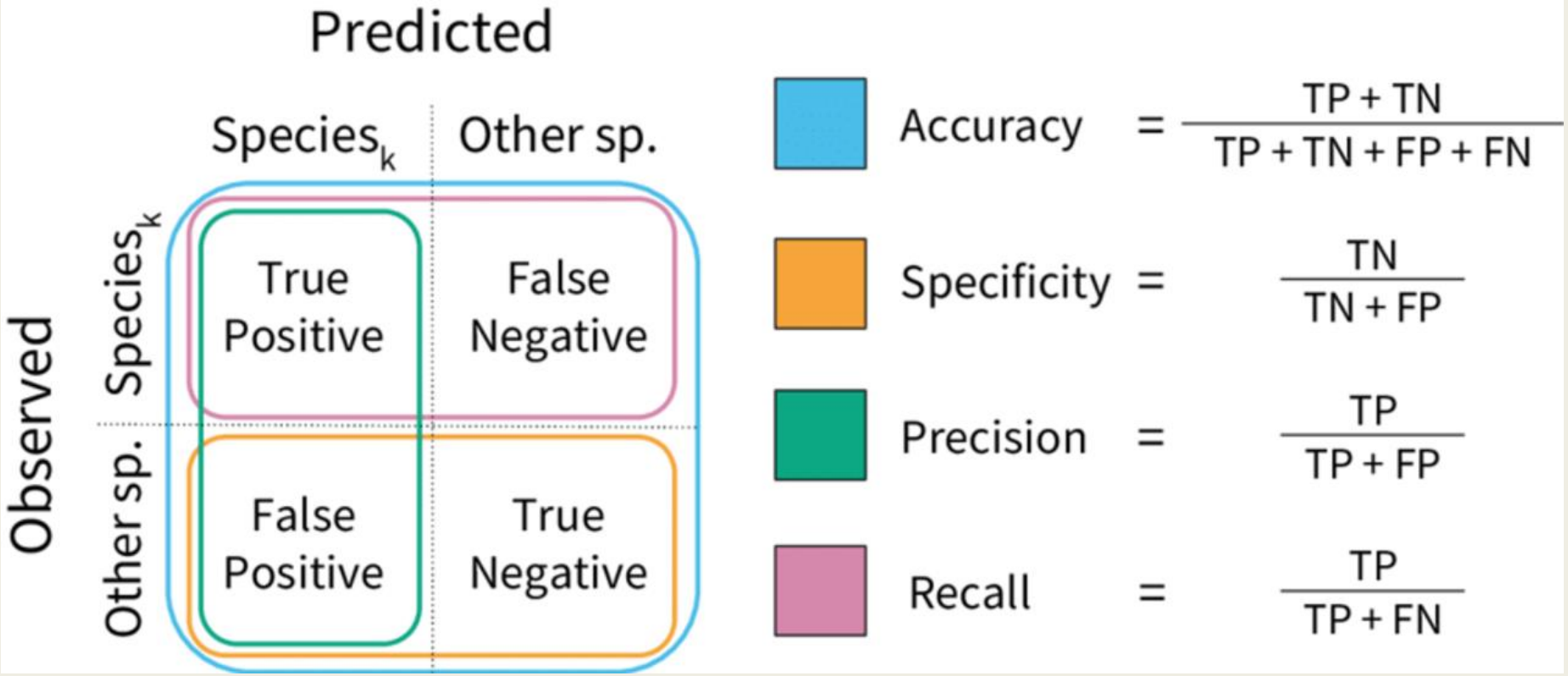
Results of decision tree

- We can also understand how each independent variable contributes to the final outcome.

Prediction using decision tree

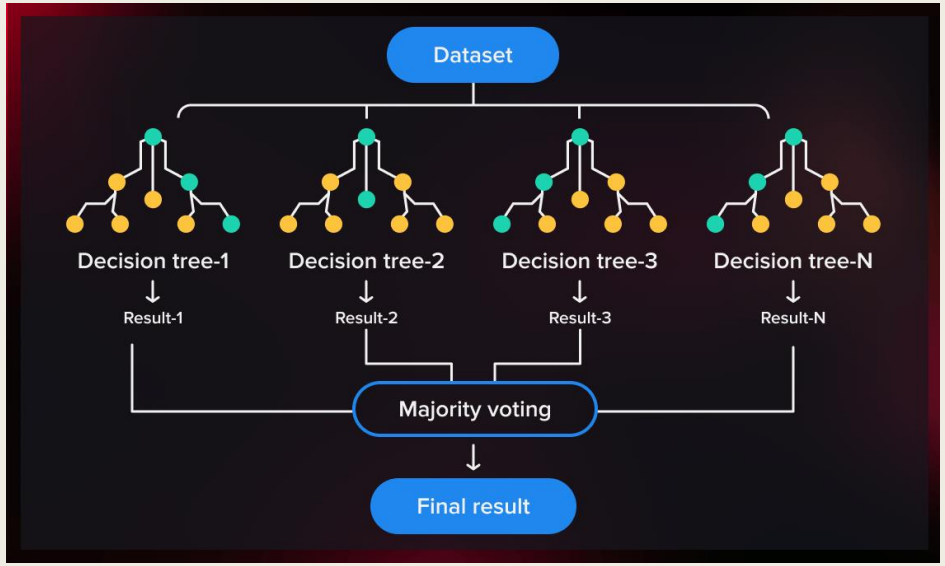
- Similar with regression models, we can use decision tree models to predict the classification of all data points (into categories, of course).
- And then, we can compare our results with baseline results to evaluate the performance of the model.
- We need to rely on very different measurements from regression models, such as precision, recall, and accuracy…

Precision, recall, and F1



Random forest

- Random forest is an expansion of decision tree model, in that it constructs multiple tree models at the same time and draws the most effective variables from all models.
- Feel free to learn about doing it yourself: https://www.r-bloggers.com/2021/04/random-forest-in-r/.



Logistic (logit) regression

- Despite the name, logistic regression is a classification model, butnot regression model.
  - *A typical question: is gender related to one's survival (True or False)?*
- The idea is pretty similar to linear regression models: how IVs could affect the DVs.
- But the coefficient means the log (odds ratio) of something happens (such as survived).
  - *log(1) = 0: so any negative coefficient means IV has a negative impact on the outcome.*
- This chapter gives a pretty good example of interpretation and reporting: https://www.bookdown.org/rwnahhas/RMPH/blr-orlr.html.

```
Call:
glm(formula = Survived ~ Sex, family = binomial, data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6462  -0.6471  -0.6471   0.7725   1.8256

Coefficients:
            Estimate Std. Error z value             Pr(>|z|)
(Intercept)   1.0566     0.1290   8.191 0.0000000000000258 ***
Sexmale      -2.5137     0.1672 -15.036 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.7  on 890  degrees of freedom
Residual deviance:  917.8  on 889  degrees of freedom
AIC: 921.8

Number of Fisher Scoring iterations: 4
```

The estimate for sex_male is the log-odd for male comparing to female. In this case, it says that the survival rate of male is at 8.1% (i.e., exp(-2.5137)) of that of female.

Any negative coefficient means DV is less likely to happen.

```
> exp(coefficients(model))
(Intercept)     Sexmale
 2.87654321  0.08096732
> exp(confint(model))
Waiting for profiling to be done...
                   2.5 %    97.5 %
(Intercept) 2.24473635 3.7245050
Sexmale     0.05804709 0.1118353
```

```
Call:
glm(formula = Survived ~ Age, family = binomial, data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1488  -1.0361  -0.9544   1.3159   1.5908

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05672    0.17358  -0.327   0.7438
Age         -0.01096    0.00533  -2.057   0.0397 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 960.23  on 712  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 964.23

Number of Fisher Scoring iterations: 4
```

WE can also use a numeric IV, in this case, the coefficient shows when the IV changes by one unit, what is the changed log-odd of the DV.

In this case, it says the increase of one year in the age will negatively impact the survival rate by about 1.1%.

```
> exp(coefficients(model))
(Intercept)         Age
  0.9448552   0.9890964
> exp(confint(model))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 0.6722345 1.328528
Age         0.9787246 0.999417
```

Evaluation of logit regression models

- AIC (Akaike information criteria) of the model
  - *This value examines the goodness-of-fit of the model. And the lower the value is, the better the model is.*
  - *It is a useful measurement to compare different models, especially in the model selection method.*
- Predicted value: precision, recall, accuracy…

Demo / group project