Autumn School in Data Science ’19

UNIVERSITY OF CAMBRIDGE

Introduction to Machine Learning for Biomedical Data Analysis in Python.

Adriano Barbosa da Silva, Ph.D.

UKRI Rutherford Fund Innovation Fellow

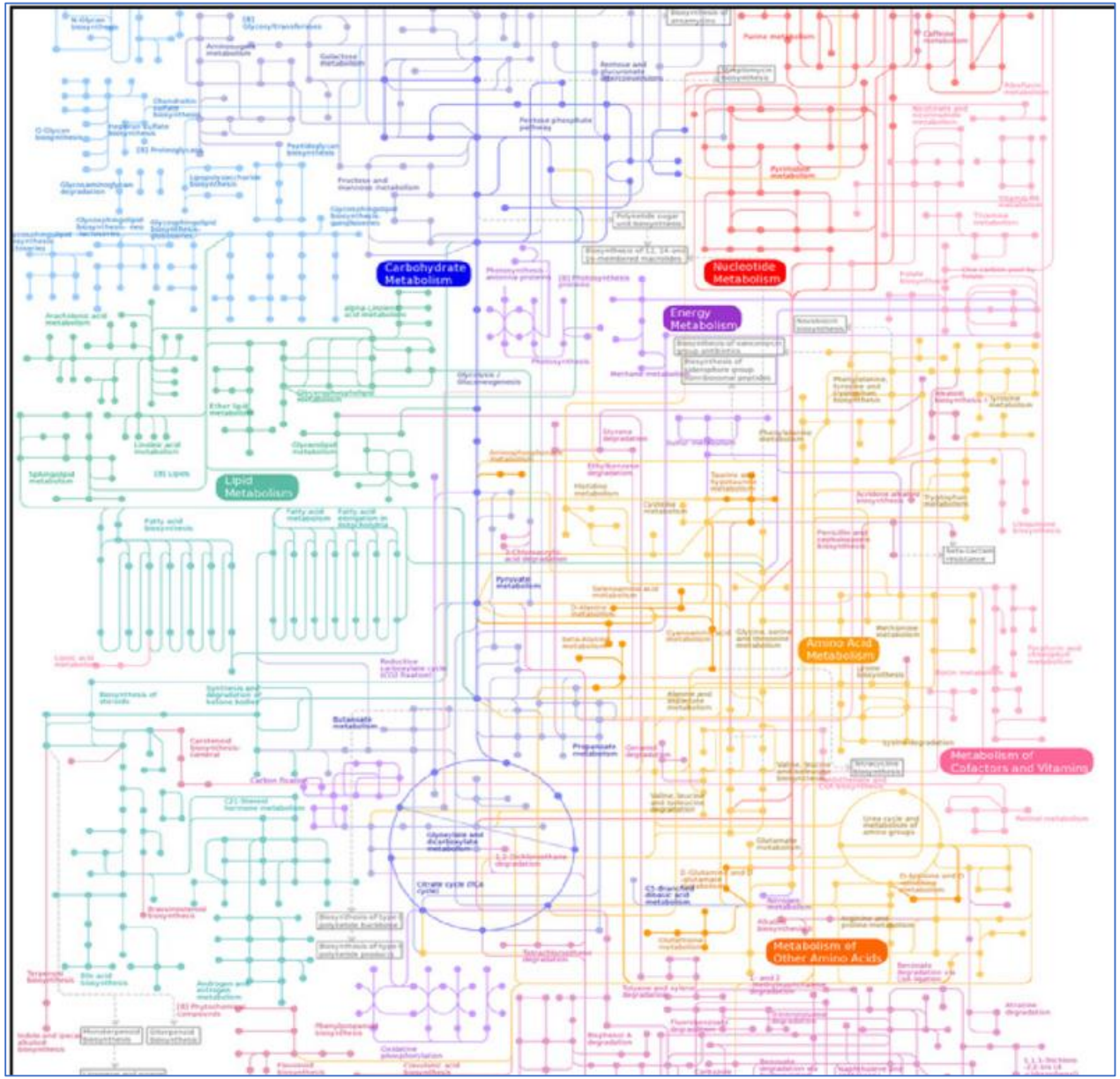
William Harvey Research Institute, Centre for Translational Bioinformatics

https://www.qmul.ac.uk/c4tb/

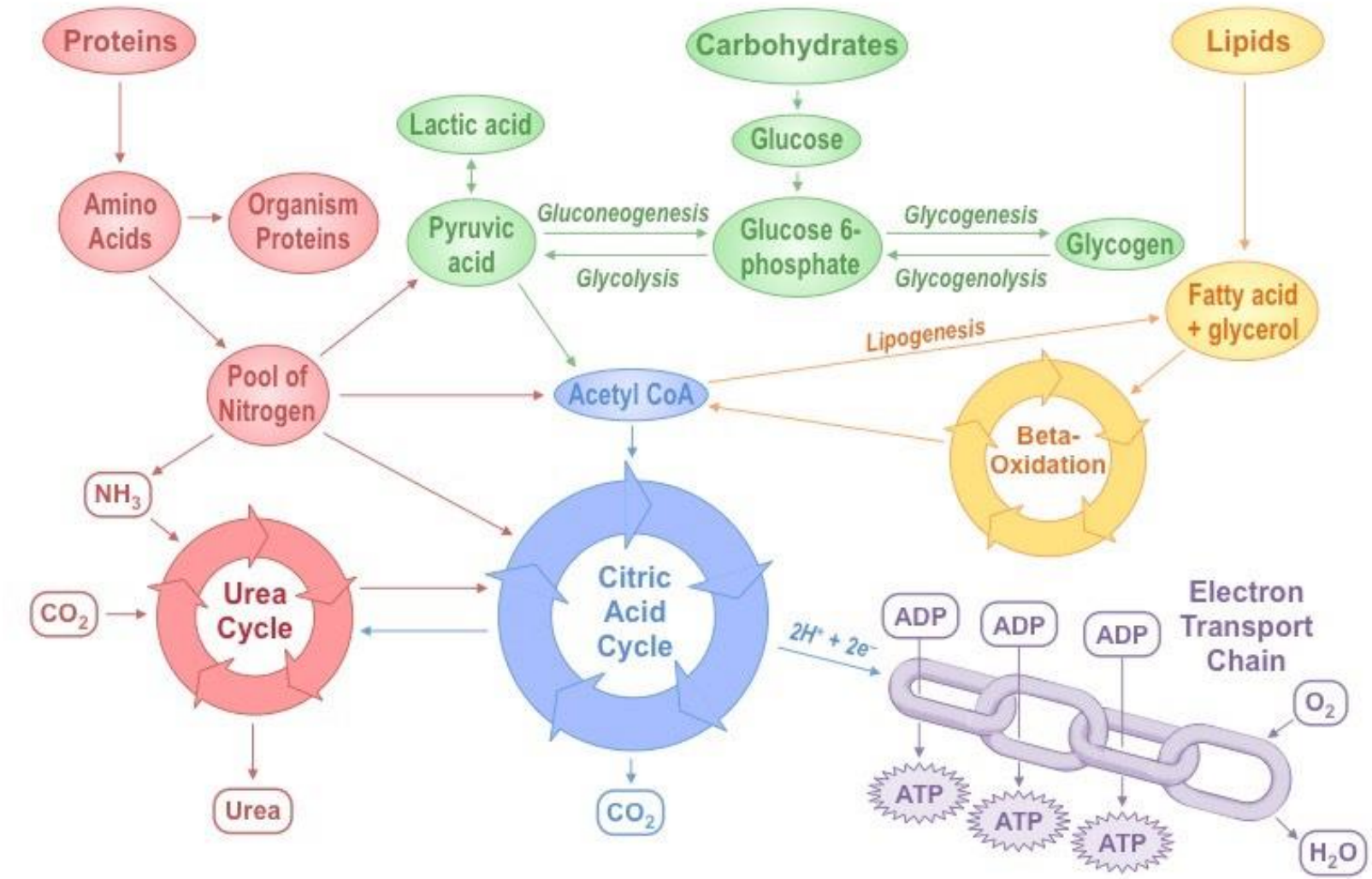
Queen Mary University of London

KEGG

UNIVERSITY OF CAMBRIDGE

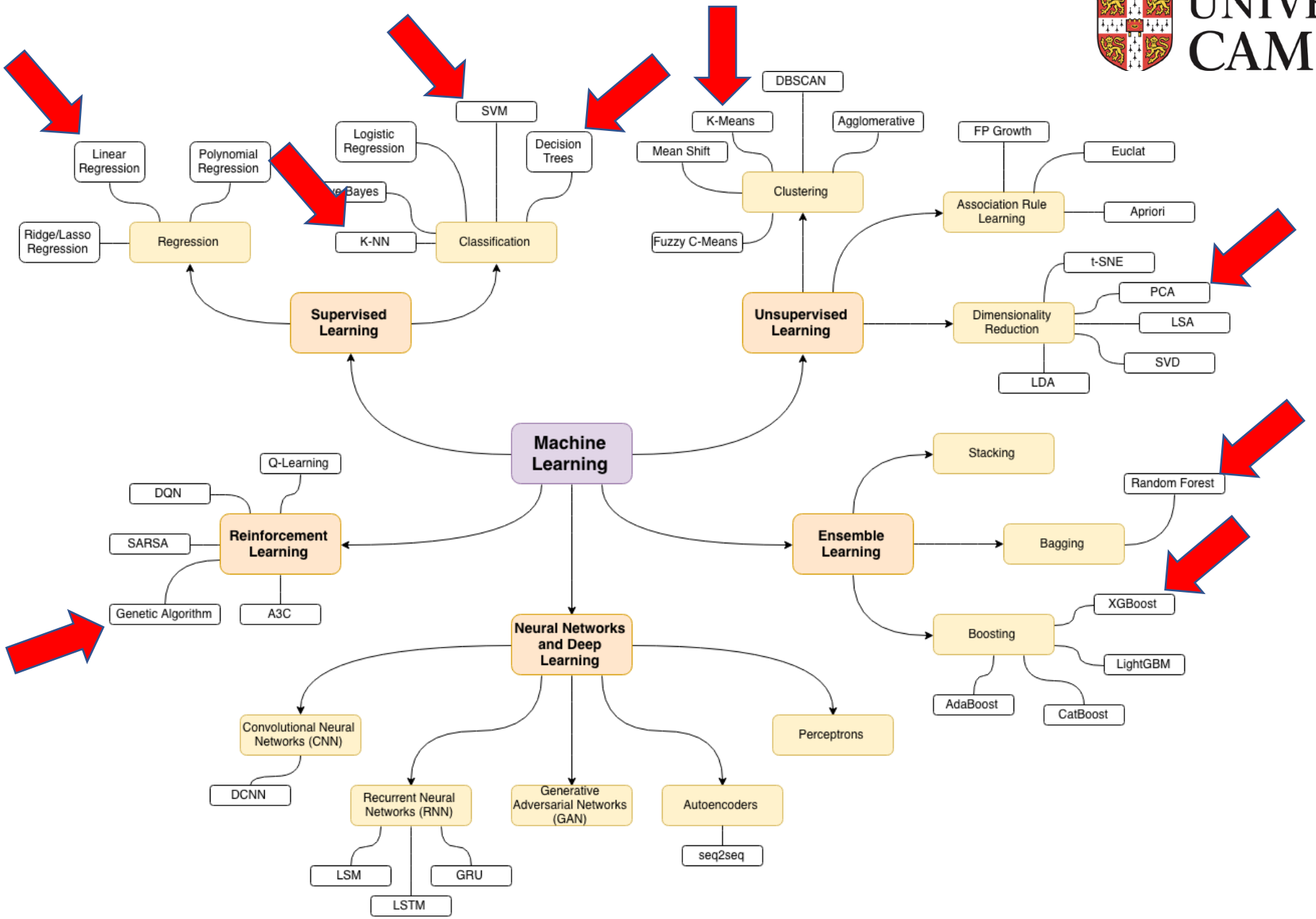


Kegg



https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/21-molecules-to-metabolism/metabolism.html

UNIVERSITY OF CAMBRIDGE



https://medium.com/datadriveninvestor/homemade-machine-learning-in-python-ed77c4d6e25b

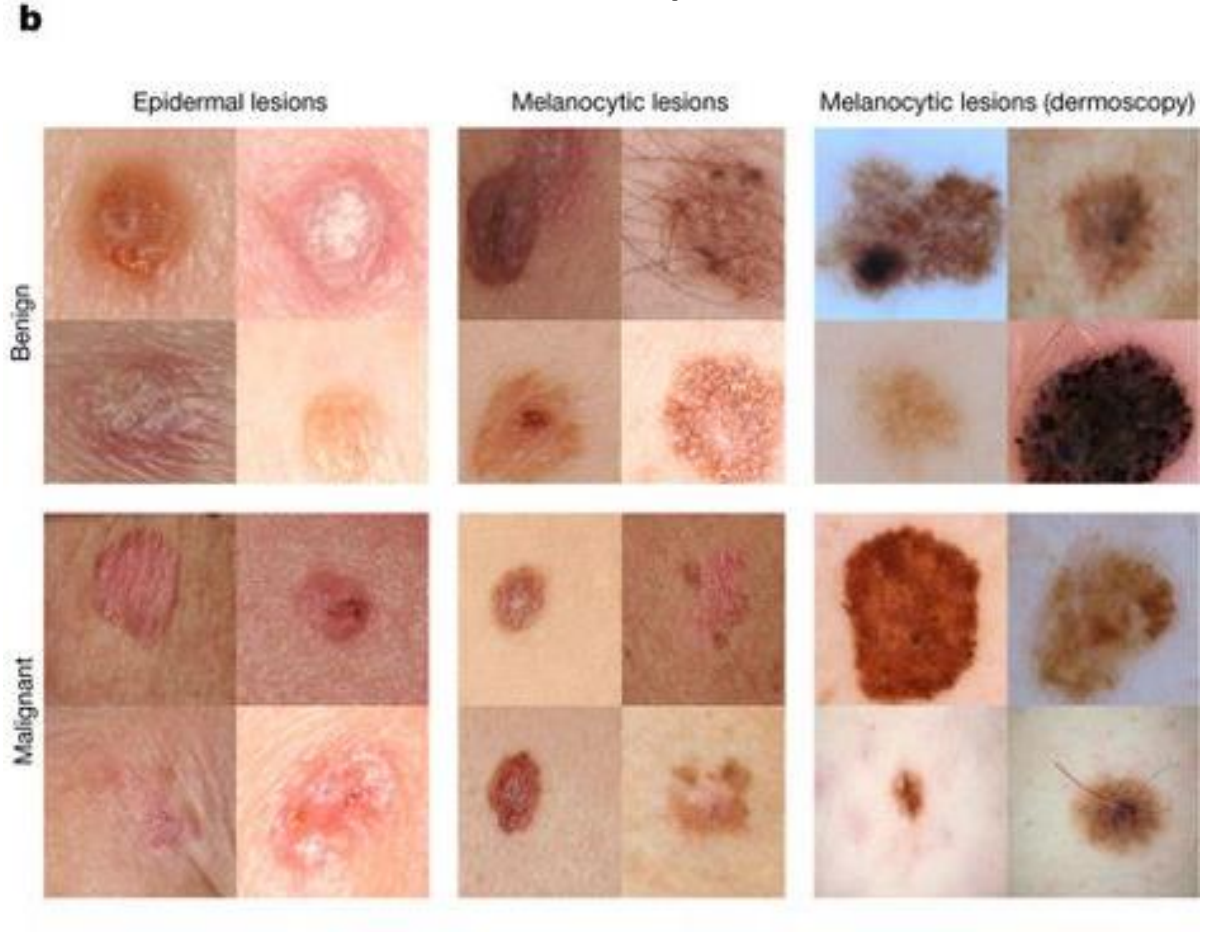
AI Definition

- Machines performing complex tasks;
- Capacity to solve problems;
- Intelligent systems;

Accelerating Artificial Intelligence in health and care: results from a state of the nation survey
https://www.kssahsn.net/what-we-do/our-news/news/Documents/AI-Strategy.pdf

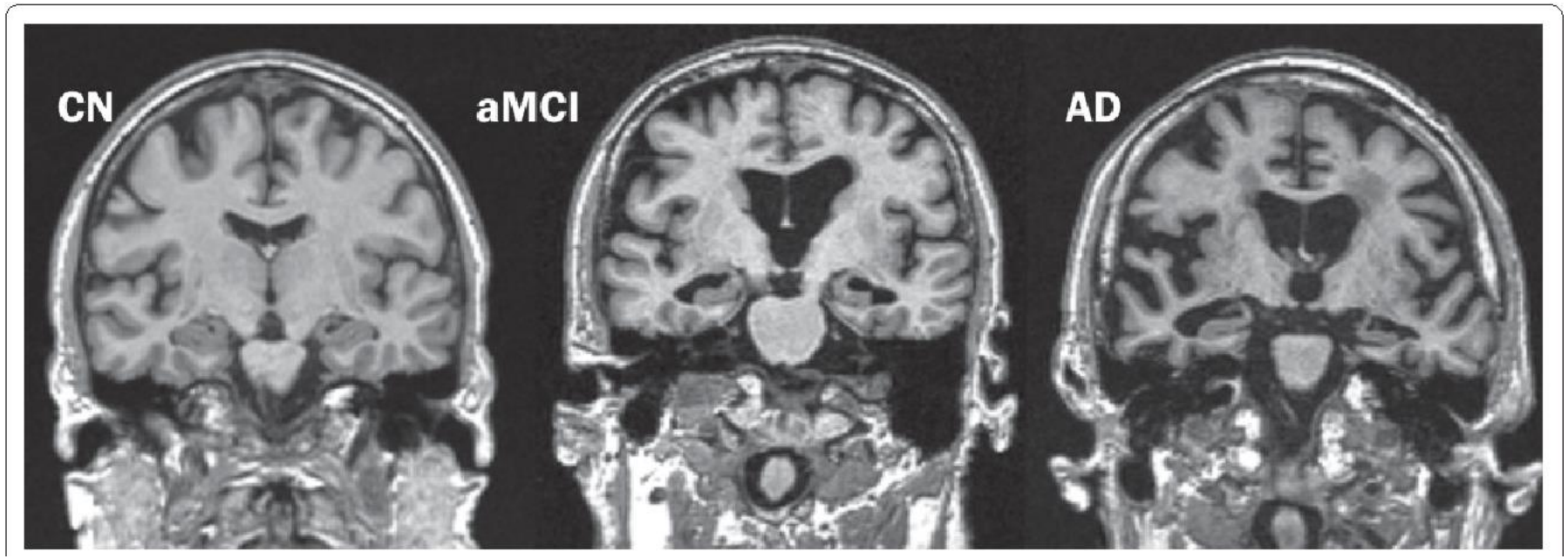
https://www.telegraph.co.uk/news/newstopics/howaboutthat/12194623/Chihuahua-or-muffin-Internet-goes-crazy-for-animals-v-food-trend.html

Dermatologist-level classification of skin cancer with deep neural networks



https://www.nature.com/articles/nature21056/figures/2

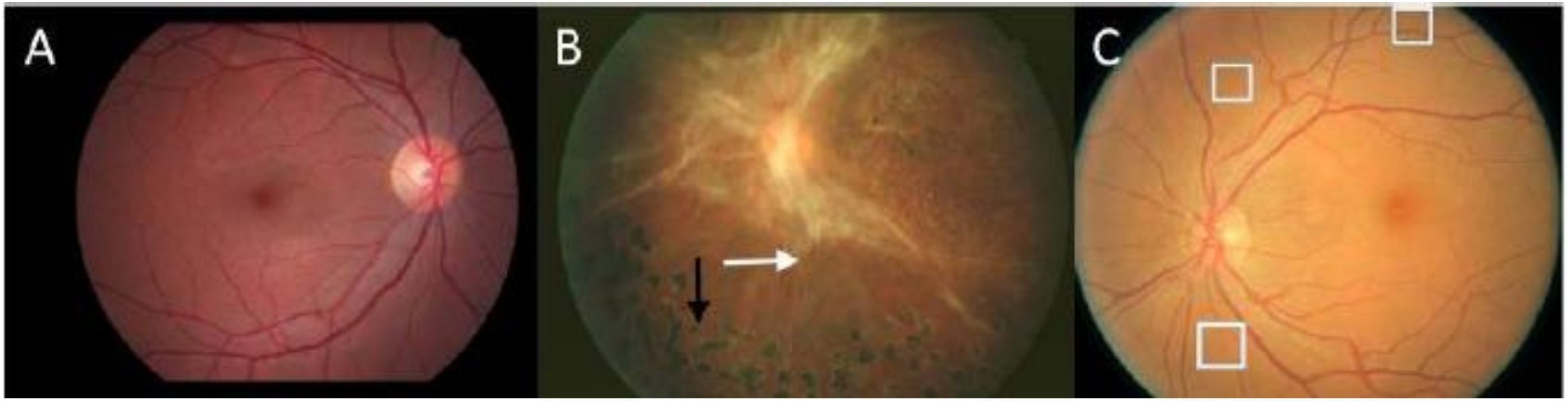
Role of structural MRI in Alzheimer's disease



Progressive atrophy (medial temporal lobes) in an older cognitively normal (CN) subject, an amnestic mild cognitive impairment (aMCI) subject, and an Alzheimer's disease (AD) subject.

Prashanthi Vemuri, Clifford R. Jack
Published in Alzheimer's Research & Therapy 2010
DOI:10.1186/alzrt47

Automated Detection of Diabetic Retinopathy using Deep Learning



Representative retinal images of DR at various stages of the disease, as labeled: A- normal, B- end stage, C- early stage. Arrows in B point to pathological indications. White boxes in C enclose very small lesions that the CNNs have difficulty discerning.

Lam C, et al. *AMIA Jt Summits Transl Sci Proc*. 2018

Definition

- We usually make decisions based on information.
- As information has become more readily **available**, our desire to use this information to help us make decisions has **intensified**.
- To aid in our decision-making processes, we now turn to web-tools to search through data looking for **patterns** that are relevant to return **answers**.

The ‘A’ of AI generally refers to one of the following:

- **ARTIFICIAL (INTELLIGENCE)**
- **AUGMENTED (INTELLIGENCE)**
- **AMBIENT (INTELLIGENCE)**

Accelerating Artificial Intelligence in health and care: results from a state of the nation survey
https://www.kssahsn.net/what-we-do/our-news/news/Documents/AI-Strategy.pdf

Fields

UNIVERSITY OF CAMBRIDGE

- Machine Learning
- Artificial Intelligence
- Pattern Recognition
- Data Mining
- Predictive Analytics
- Knowledge discovery

To make an accurate prediction.

Some applications of AI

- Application in health and social care;
- Personalized and proactive care models;
- Identifying diseases and conditions early;
- Right treatments to patients;

Accelerating Artificial Intelligence in health and care: results from a state of the nation survey
https://www.kssahsn.net/what-we-do/our-news/news/Documents/AI-Strategy.pdf

Some applications of AI

UNIVERSITY OF CAMBRIDGE

Prognosis

Diagnosis

https://www.buttersafe.com/2010/02/25/the-prognosis/

Limitations

- Inadequate pre-processing of the data,
- Inadequate model validation,
- Unjustified extrapolation
- Over-fitting the model to the existing data

Prediction versus Interpretation

- The interest is to accurately predict rather than understand;
- An example could be a physician encounter when contemplating changing treatment therapies;
- The critical question for the doctor and patient is a prediction of how the patient will react to a change in therapy.

Importance of subject-specific knowledge

- If a predictive signal exists in a set of data, many models will find some degree of that signal regardless of the technique or care placed in developing the model;

- Most predictive models are fundamentally influenced by a modeler with expert knowledge and context of the problem;

- Irrelevant information can drive down predictive performance of many models

Expert opinion and data-driven models, together

UNIVERSITY OF CAMBRIDGE

- *"In the end, [predictive modeling] is not a substitute for intuition, but rather a complement"*

- *"Traditional experts make better decisions when they are provided with the results of statistical prediction. Those who cling to the authority of traditional experts tend to embrace the idea of combining the two forms of 'knowledge' by giving the experts 'statistical support' ...* ***Humans usually make better predictions when they are provided with the results of statistical prediction."***

Ayres I (2007). Super Crunchers: Why Thinking–By–Numbers Is The New Way To Be Smart. Bantam.

How to build an effective predictive model

- Intuition and deep knowledge of the problem context;
- Relevant data
- Versatile computational toolbox

UNIVERSITY OF CAMBRIDGE

Terminology

Terminology

- The terms **sample**, **data point**, **observation**, or **instance** refer to a single, independent **unit of data**, such as a patient. The term sample can also refer to a subset of data points, such as the training set sample.
- The **training set** consists of the data used to develop models while the test or validation sets are used solely for evaluating the performance of a final set of candidate models.

Terminology

- The **predictors**, **independent variables**, **attributes**, or **descriptors** are the data used as input for the prediction equation.
- **Outcome**, dependent variable, target, class, or response refer to the outcome event or quantity that is being predicted.
- **Continuous** data have natural, numeric scales, such as Blood pressure.
- **Categorical** data, otherwise known as **nominal**, **attribute**, or **discrete data**, take on specific values that have no scale. Such as nationality.
- **Model building**, **model training**, and **parameter estimation** all refer to the process of using data to determine values of model equations.

UNIVERSITY OF CAMBRIDGE

Data Pre-Processing

Data preparation

- Different models have different sensitivities to the **type** of predictors in the model;
- **Transformations** of the data to reduce the impact of data skewness or outliers can lead to significant improvements in performance;
- How the predictors are encoded, called **feature engineering**, can have a significant impact on model performance.

Data Representation

UNIVERSITY OF CAMBRIDGE

- The “correct” feature engineering depends on several factors:
  - it depends mainly on the model being used and the true relationship with the outcome;

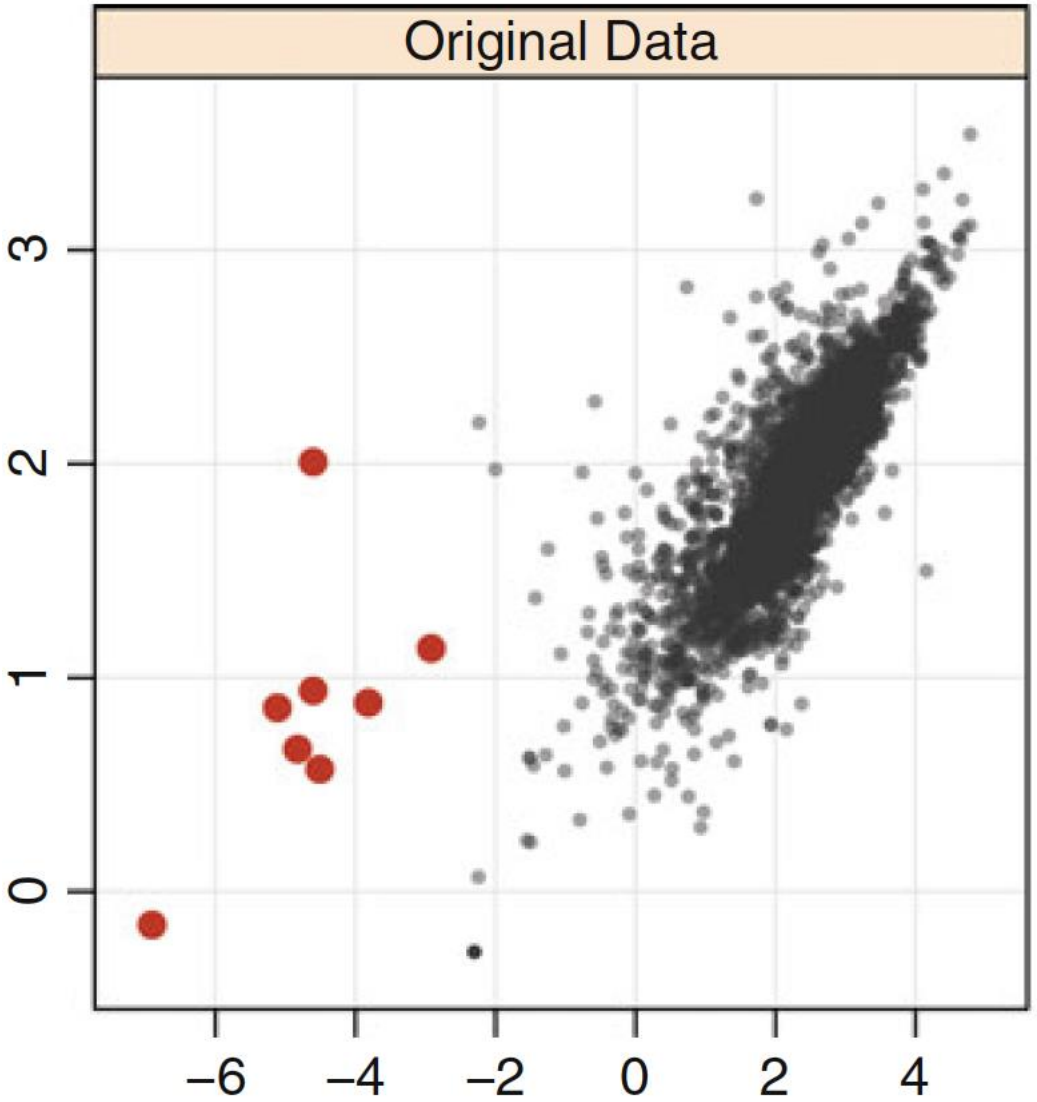
- Some modeling techniques may have strict **requirements**, such as the predictors having a common scale.

Centering and Scaling

- To **center** a predictor variable, the average predictor value is subtracted from all the values. As a result of centering, the predictor has a **zero mean**.

Resolve Outliers

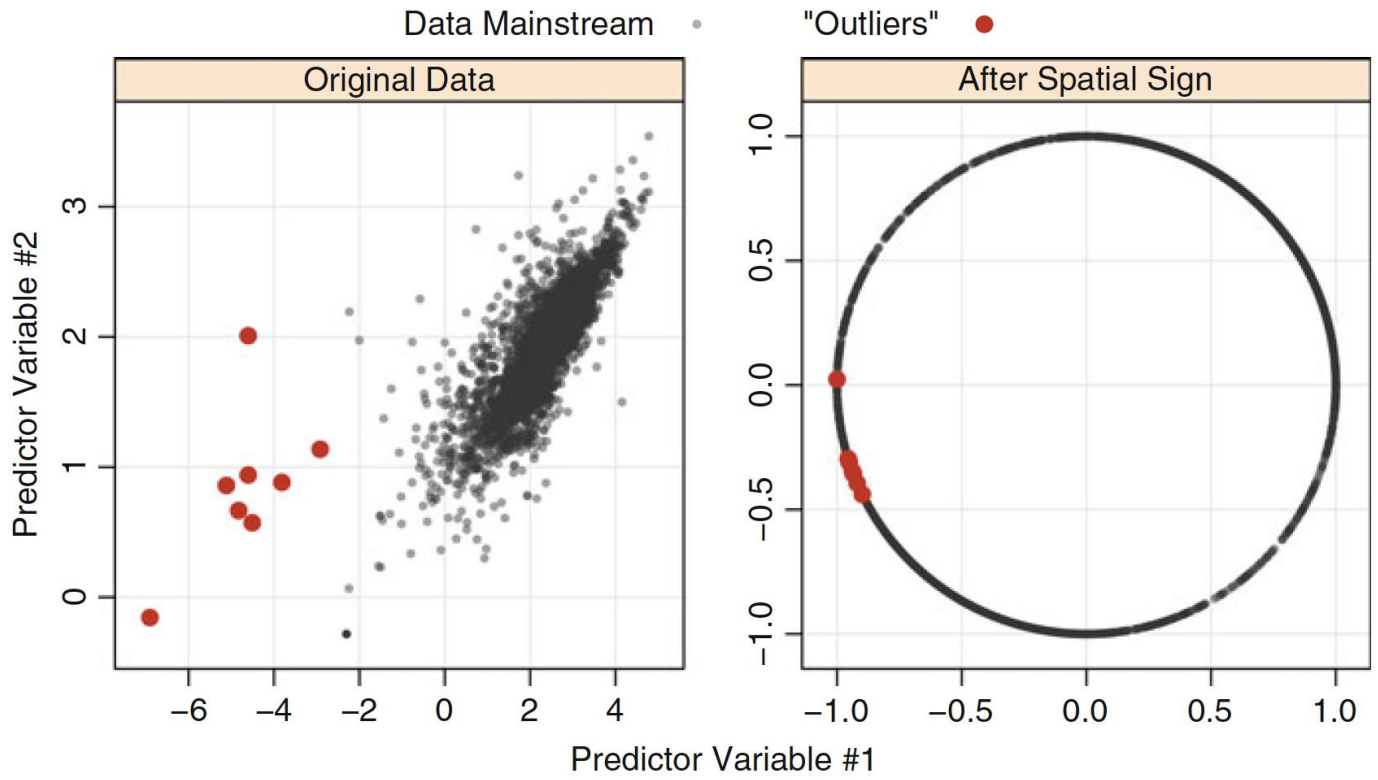
- Samples that are exceptionally **far** from the mainstream of the data;
- When one or more samples are suspected to be outliers, the first step is to make sure that the values are scientifically **valid** (e.g., positive blood pressure) and that no data recording errors have occurred.
- Also, the outlying data may be an indication of a **special part** of the population under study that is just starting to be sampled.



Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Resolve Outliers

- The spatial sign transformation is shown on the right-hand panel where all the data points are projected to be a **common distance** away from the origin.



Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Data Reduction and Feature Extraction

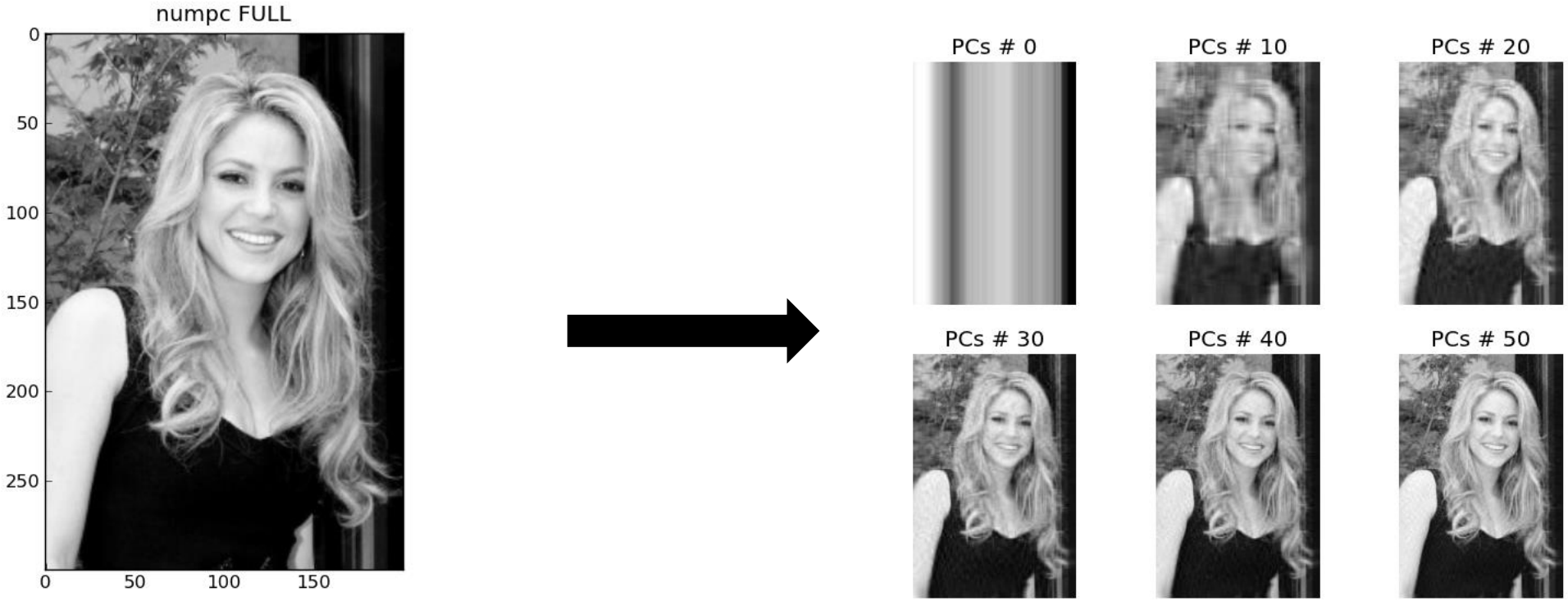
UNIVERSITY OF CAMBRIDGE

- These methods reduce the data by generating a **smaller** set of predictors that seek to capture a **majority** of the information in the original variables.

- The new predictors are functions of the original predictors; therefore, all the original predictors are **still needed** to create the surrogate variables.

- PCA is a commonly used **data reduction** technique

Data Reduction and Feature Extraction



https://glowingpython.blogspot.com/2011/07/pca-and-image-compression-with-numpy.html

Data Reduction and Feature Extraction

- PCA method seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance.

- Mathematically the *j*th PC can be written as:

$$\text{PC}_j = (a_{j1} \times \text{Predictor } 1) + (a_{j2} \times \text{Predictor } 2) + \cdots + (a_{jP} \times \text{Predictor } P).$$

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Data Reduction and Feature Extraction

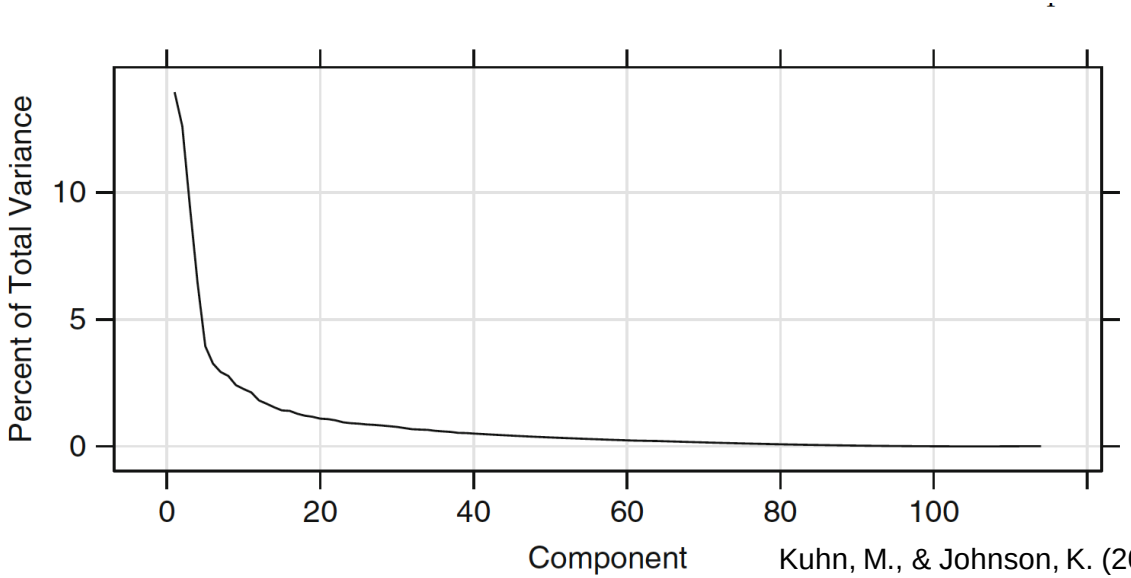
- PCA seeks predictor-set variation without regard to any further understanding of the predictors or to knowledge of the modeling objectives

- PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective

- PCA will be focusing its efforts on identifying the data structure based on measurement scales rather than based on the important relationships within the data for the current problem.

Data Reduction and Feature Extraction

- For data sets with many predictor variables, we must decide how many components to retain.
- For most data sets, the first few PCs will summarize a majority of the variability, and the plot will show a steep descent; variation will then taper off for the remaining components.



Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Missing Values

- In many cases, some predictors have no values for a given sample;
- In other cases, the value cannot or was not determined at the time of model building.
- It is important to know if the pattern of missing data is related to the outcome.

Missing Values

- Missing data should not be confused with **censored data;**
- Censored data can be common when using laboratory measurements.
- For predictive models, it is more common to treat these data as simple missing data or use the censored value as the observed value.

Imputation

- There are cases where the missing values might be concentrated in specific samples.
- Missing data can be imputed.
- In essence: to estimate the values of other (missing) predictors. This amounts to a predictive model within a predictive model.

Removing Predictors

- Fewer predictors means decreased computational time and complexity;
- If two predictors are highly correlated, this implies that they are measuring the same underlying information.
- Some models can be crippled by predictors with degenerate distributions.

Removing Predictors

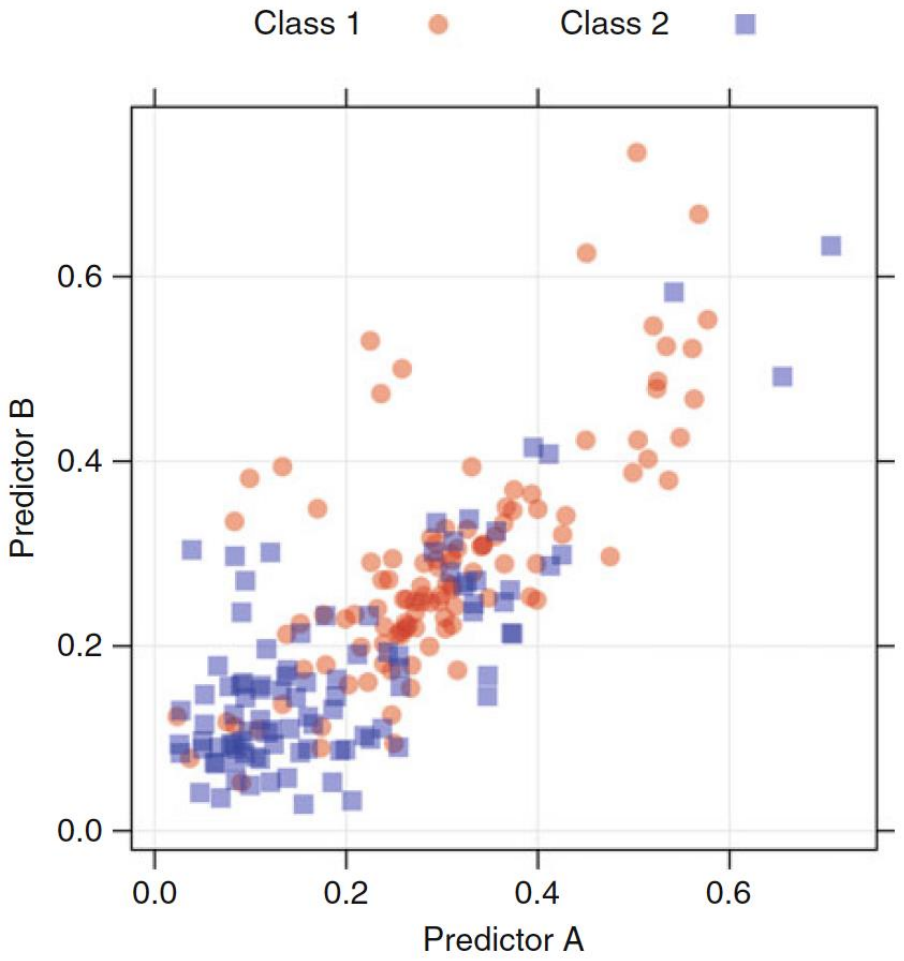
- The number of unique points in the data must be small relative to the number of samples:
  - The fraction of unique values over the sample size is low (say 10%).
  - The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20).
- Collinearity

Overfitting and Tuning

- Models can overemphasize data patterns that are not reproducible;
- Models should predict new samples with a similar degree of accuracy on the set of data for which the model was evaluated.
- Model building efforts are constrained by the existing data.
- While there are ways to build predictive models on small data sets, we generally assume that data quality is sufficient and that it is representative of the entire sample population.

Overfitting

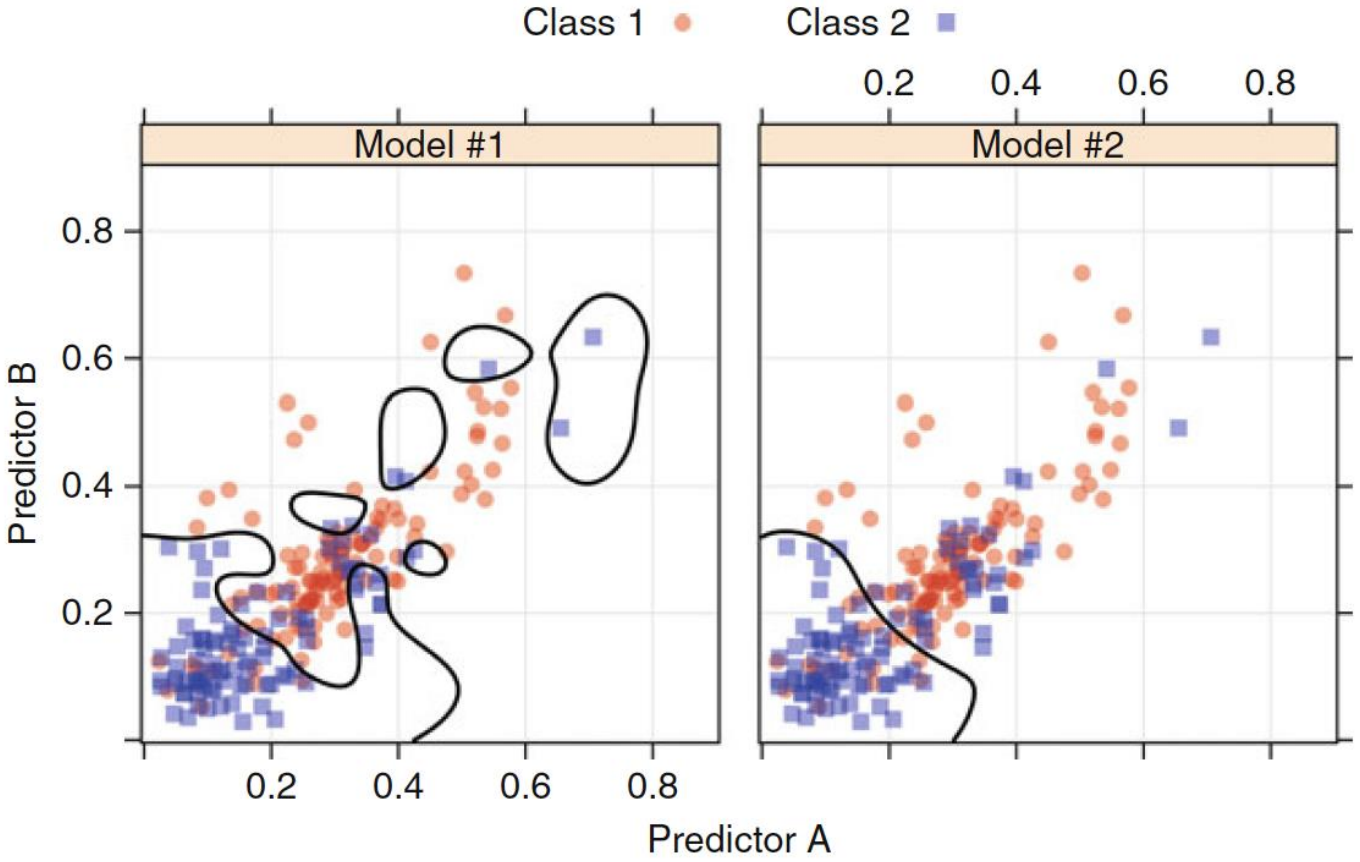
- 208 samples that are designated either as “Class 1” or “Class 2”;
- The classes are fairly balanced; there are 111 samples in the first class and 97 in the second;
- Significant overlap between the classes;



Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Overfitting

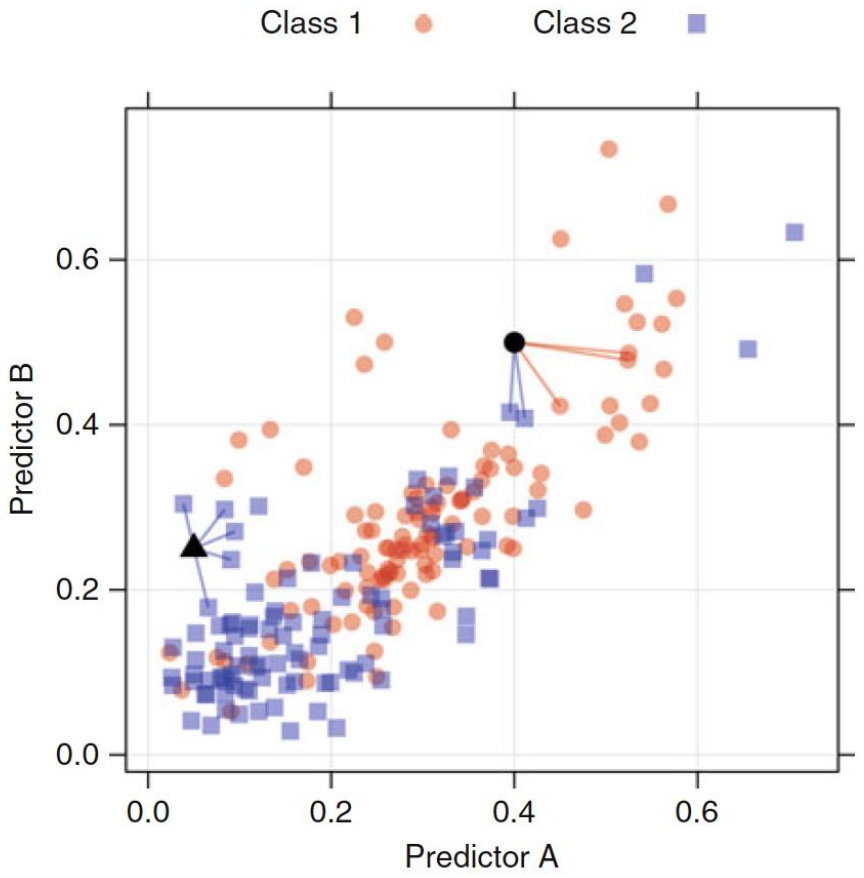
- Class boundaries from two distinct classification models.
- The lines envelop the area where each model predicts the data to be the second class (blue squares).



Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Tuning

- By using K-nearest neighbor classification model, a new sample is predicted based on the K-closest data points in the training set;
- The big question: how many neighbors should be used?
- This type of model parameter is referred to as a tuning parameter because there is no analytical formula available to calculate an appropriate value.



Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Tuning

Define a set of candidate values for tuning parameter(s)

UNIVERSITY OF CAMBRIDGE

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Data splitting

Recap of the common steps in model building:

- Pre-processing the predictor data
- Estimating model parameters
- Selecting predictors for the model
- Evaluating model performance
- Fine tuning class prediction rules (via ROC curves, etc.)

Data splitting

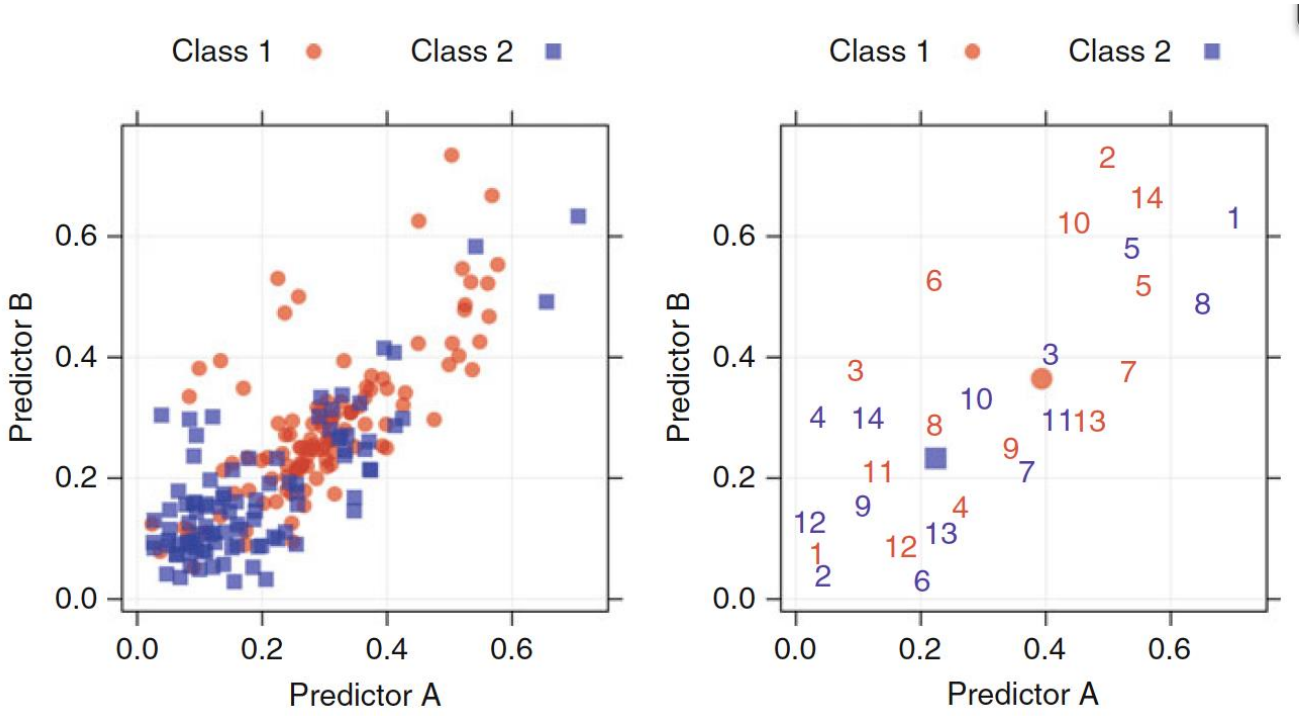
Decisions to be taken:

- Which samples will be used to evaluate **performance**;

- The **“training”** data set is the general term for the samples used to create the model, while the **“test”** or **“validation”** data set is used to qualify performance.

Data splitting

- To split the data into a training and test set is to take a simple random sample.
- Watch for distributions match between classes and values.



Resampling Techniques

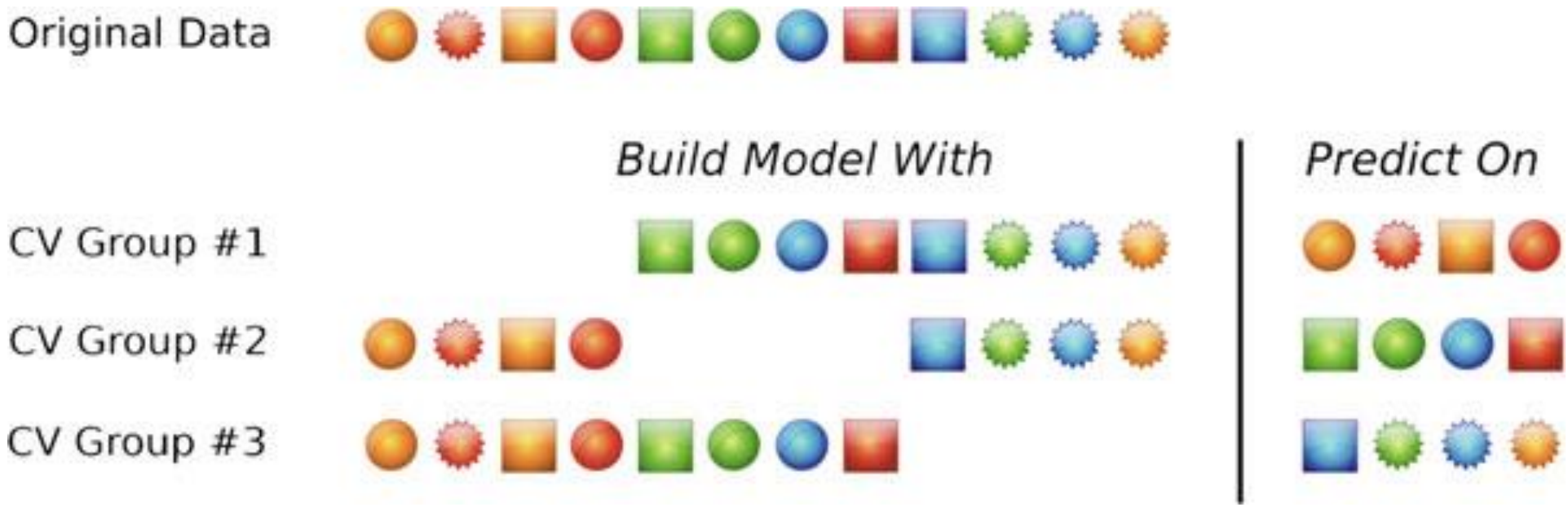
UNIVERSITY OF CAMBRIDGE

- Used for estimating model performance;
- Process: a subset of samples are used to fit a model and the remaining samples are used to estimate the efficacy of the model.
- Some techniques to consider:
  - K-Fold Cross-Validation;
  - Repeated Training/test Splits;
  - Bootstrap;

K-fold Cross-Validation.

UNIVERSITY OF CAMBRIDGE

- The samples are randomly partitioned into k sets of roughly equal size.
- A model is fit using the all samples except the first subset (called the first fold).

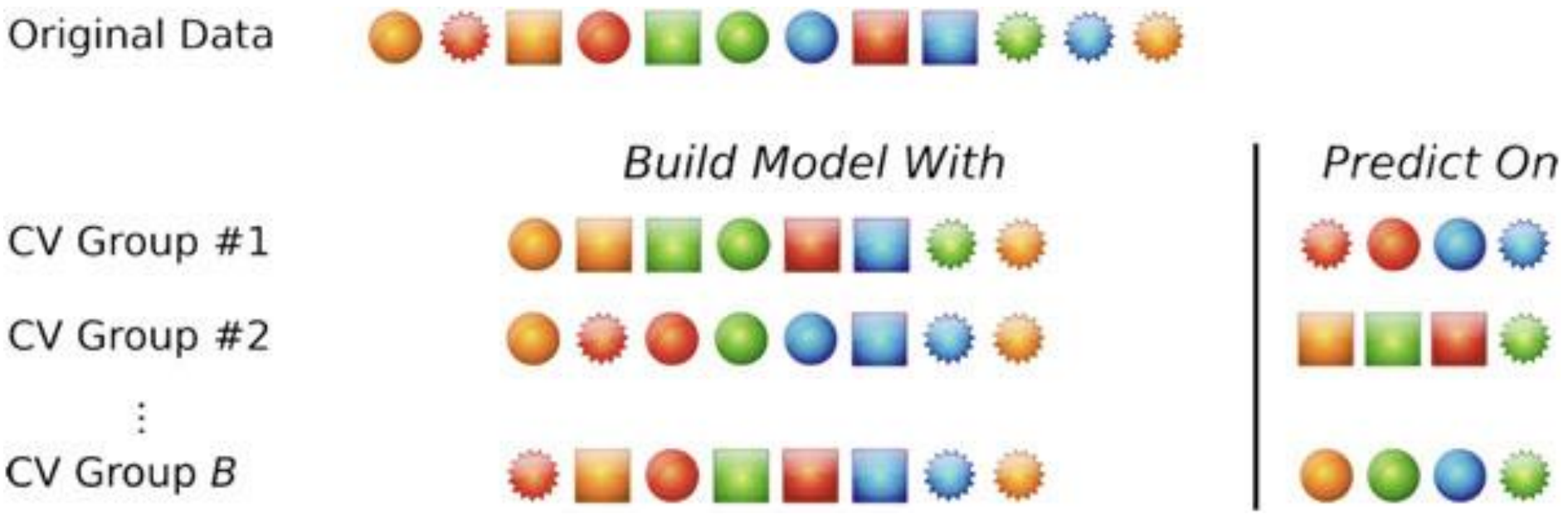


Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

Repeated Training/Test Splits

UNIVERSITY OF CAMBRIDGE

- Also known as “leave-group-out cross validation” or “Monte Carlo cross validation.”
- This technique simply creates multiple splits of the data into modeling and prediction sets.

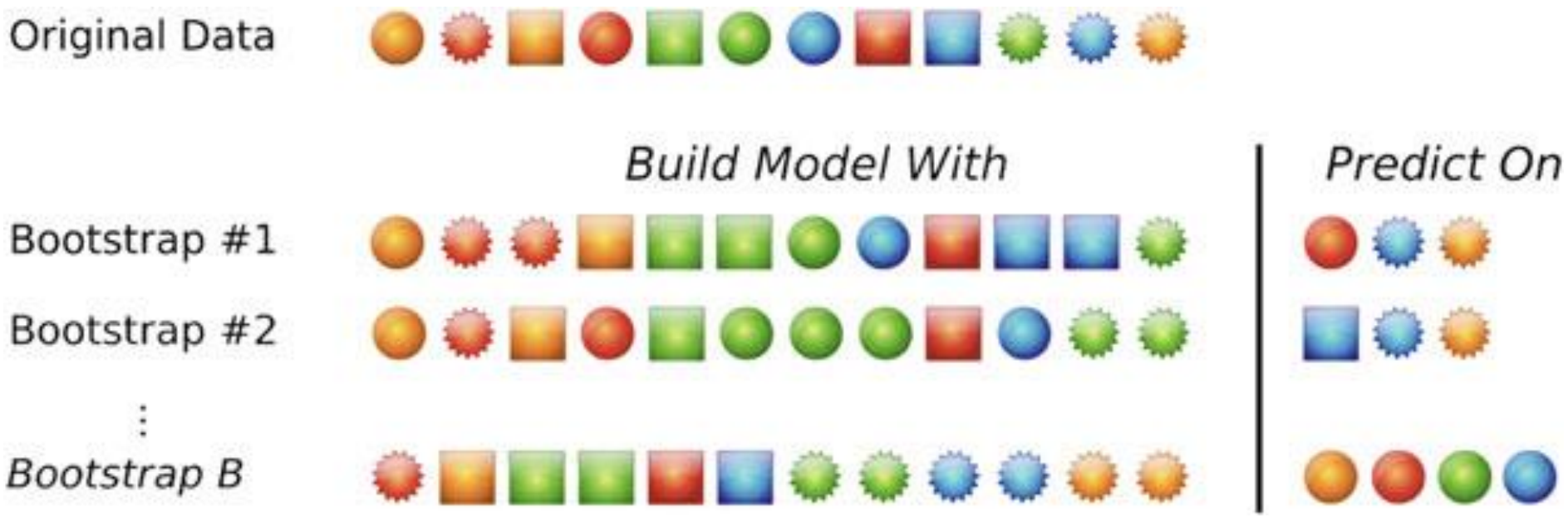


Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

The Bootstrap

UNIVERSITY OF CAMBRIDGE

- A bootstrap sample is a random sample of the data taken with replacement (Efron and Tibshirani 1986).
- After a data point is selected for the subset, it is still available for further selection.
- The bootstrap sample is the same size as the original data set.



Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer.

UNIVERSITY OF CAMBRIDGE

General considerations about data splitting

- A test set is a single evaluation of the model and has limited ability to characterize the uncertainty in the results.
- Proportionally large test sets divide the data in a way that increases bias in the performance estimates.
- Resampling methods can produce reasonable predictions of how well the model will perform on future samples.

UNIVERSITY OF CAMBRIDGE

General considerations about data splitting

- Consider using **10-fold cross validation** if the samples size is small;
- If the goal is to choose between models use **bootstrap**;
- Use **10-fold cross validation** for large sample sizes;

Program

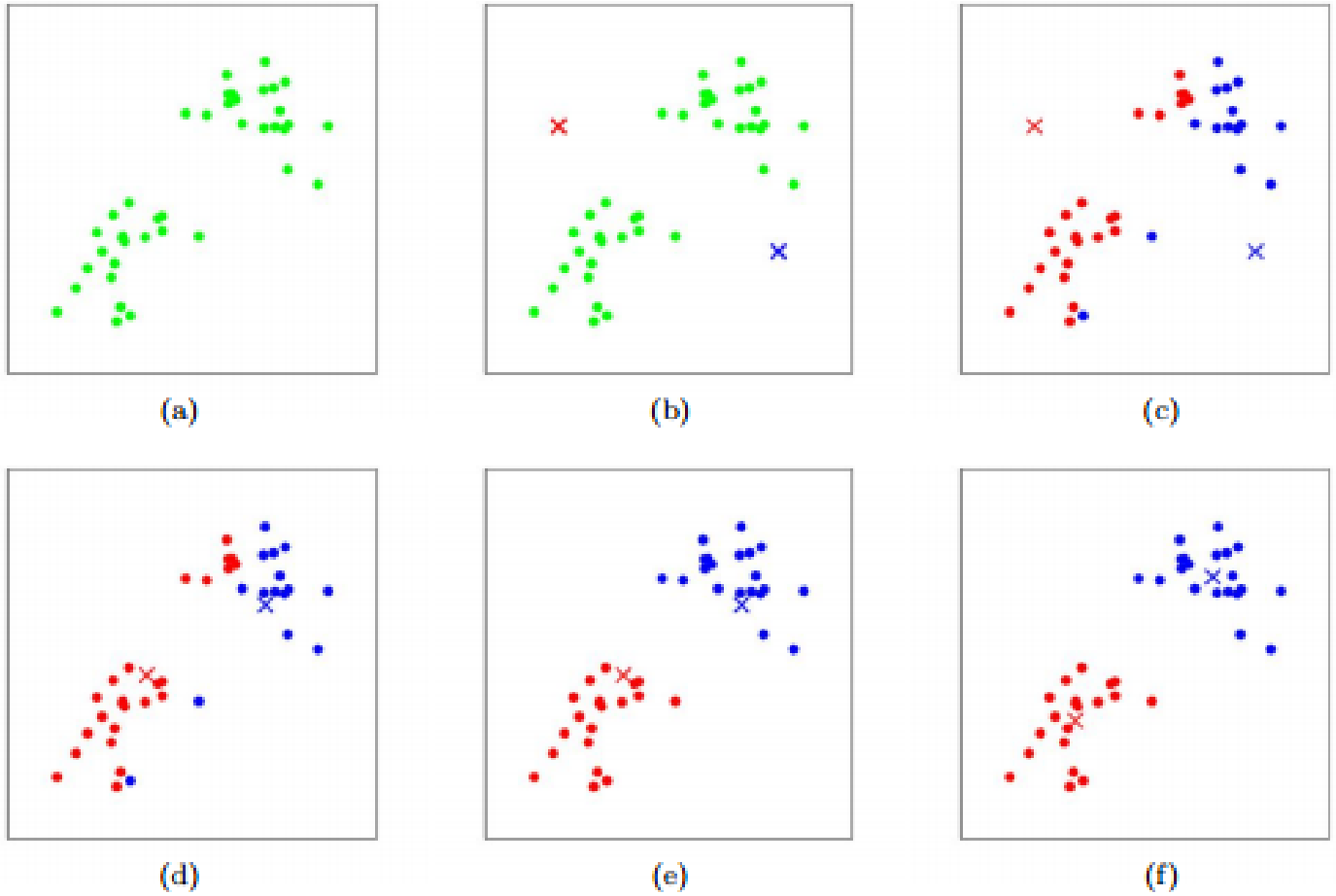
- · **Day 1**
- · 10:30 - 11:00: General Introduction to Machine Learning;
- · 11:00 - 11:30: K-Means Clustering;
- · 11:30 - 12:00: Nearest Neighbor Algorithm;
- · **Lunch**
- · 13:00 - 16:00: Study case: Diabetes Prediction:
- o 13:00 - 13:30: Data exploration;
- o 13:30 - 14:00: Logistic Regression;
- o 14:00 - 14:30: Decision Trees;
- o 14:30 - 15:00: Random Forest Classifiers;
- o 15:00 - 15:30: Support Vector Machines;
- o 15:30 - 16:00: Methods comparison;
- · **Day 2**
- · 09:30 - 10:15: Study case: Heart Disease Prediction;
- · 10:15 - 11:00: Study case: Breast Cancer Prediction;;

K-means

- K-Means falls under the category of centroid-based clustering.
- K-means stores ***k*** centroids that it uses to define clusters.
- A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.
- K-Means finds the best centroids by alternating between
  - (1) assigning data points to clusters based on the current centroids
  - (2) choosing centroids (points which are the center of a cluster) based on the current assignment of data points to clusters.

https://www.datacamp.com/community/tutorials/k-means-clustering-python
https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

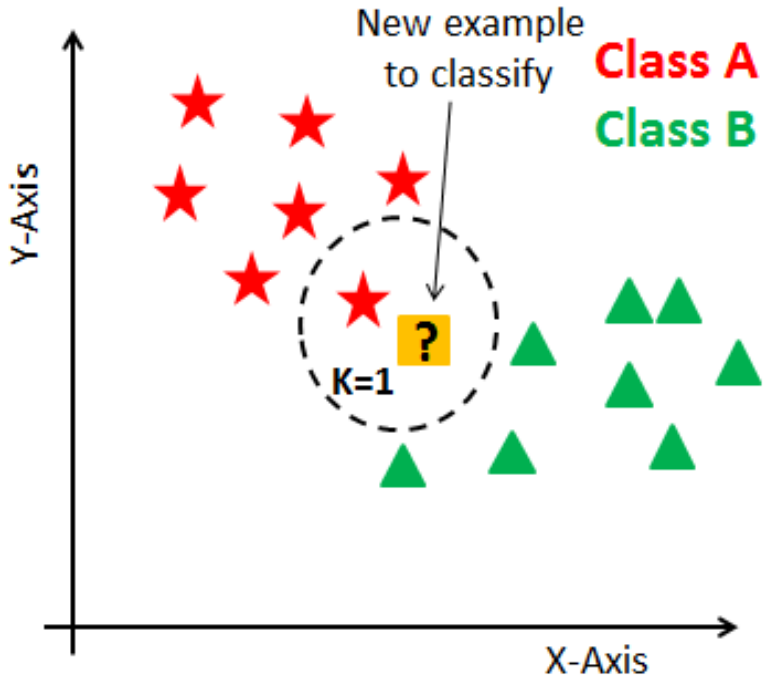
K-means



https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

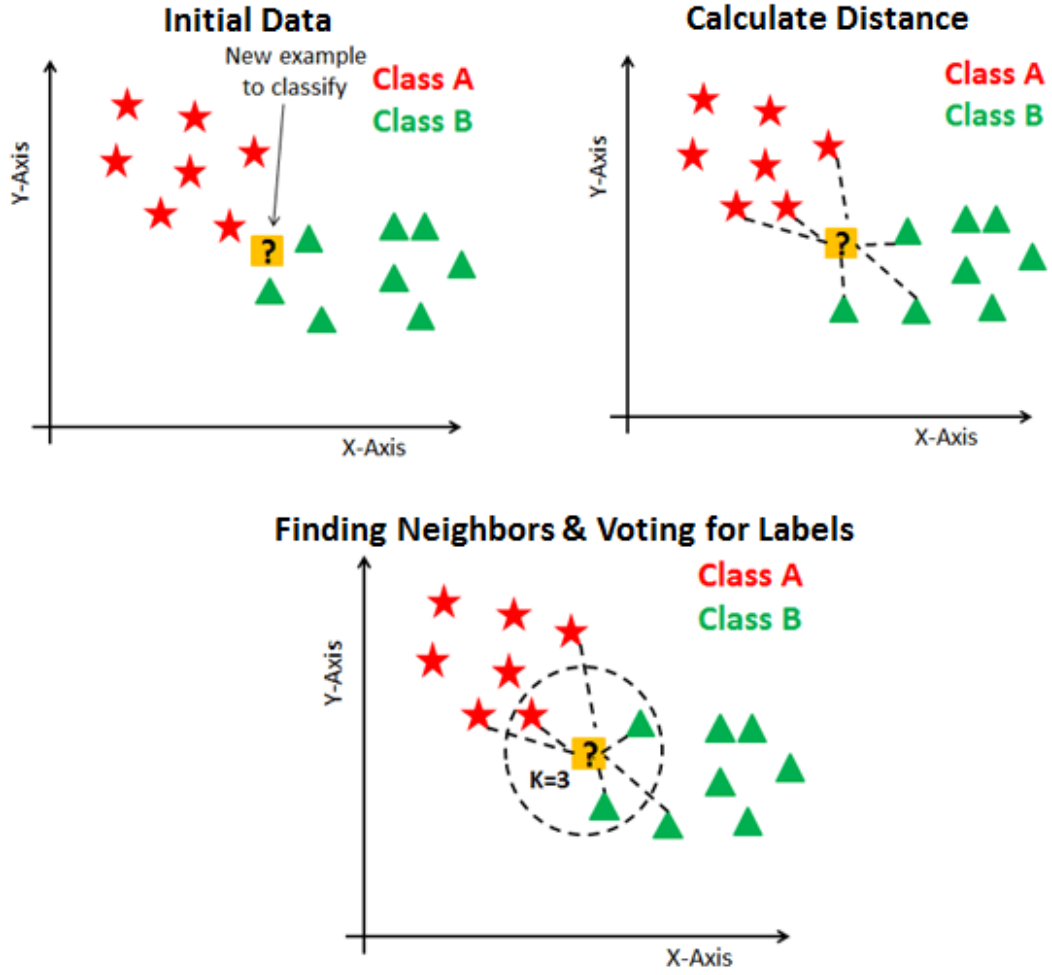
K-Nearest Neighbor

- KNN is a non-parametric and lazy learning algorithm.
- In KNN, K is the number of nearest neighbors.
- How to choose the optimal number of neighbors?



https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

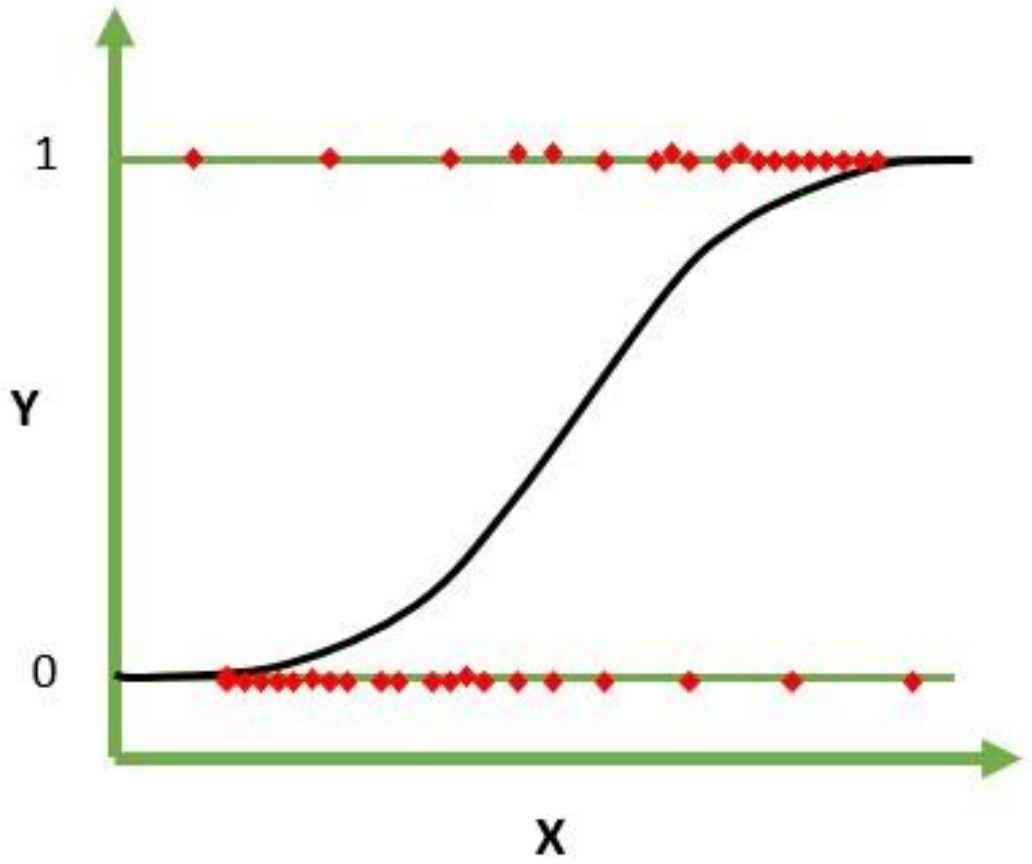
K-Nearest Neighbor



https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

Logistic Regression

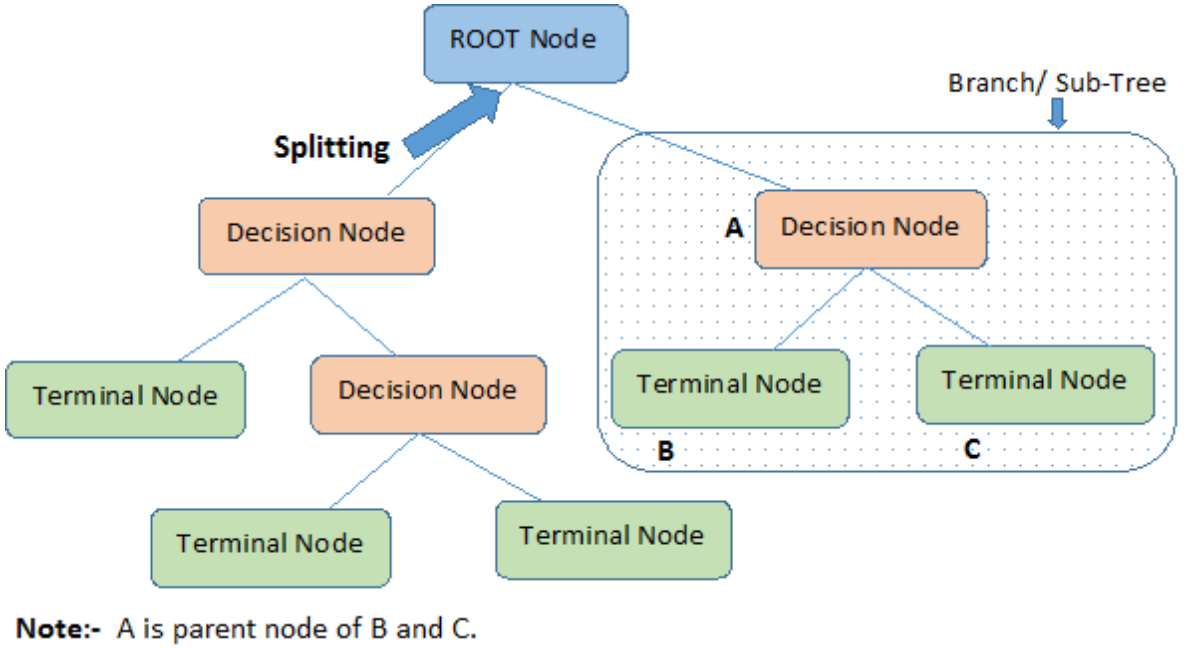
- Logistic regression is useful when you are predicting a binary outcome from a set of continuous predictor variables.



https://techdifferences.com/difference-between-linear-and-logistic-regression.html

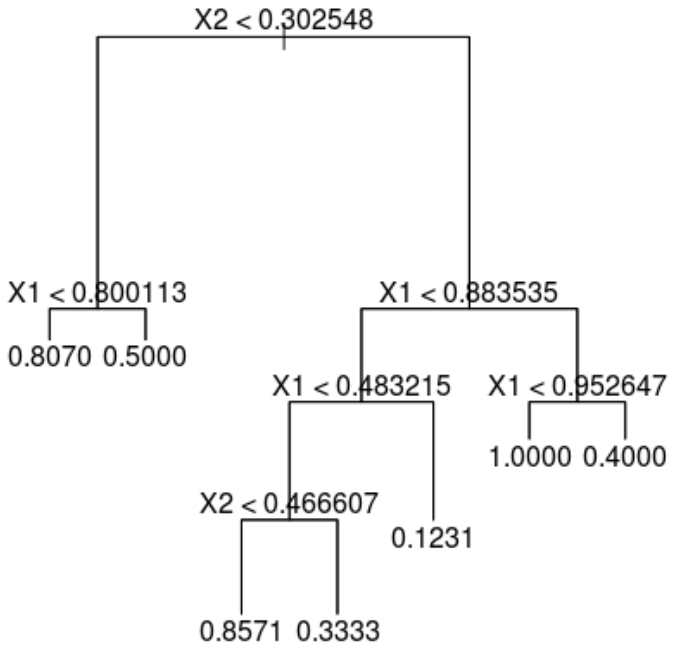
Decision Trees

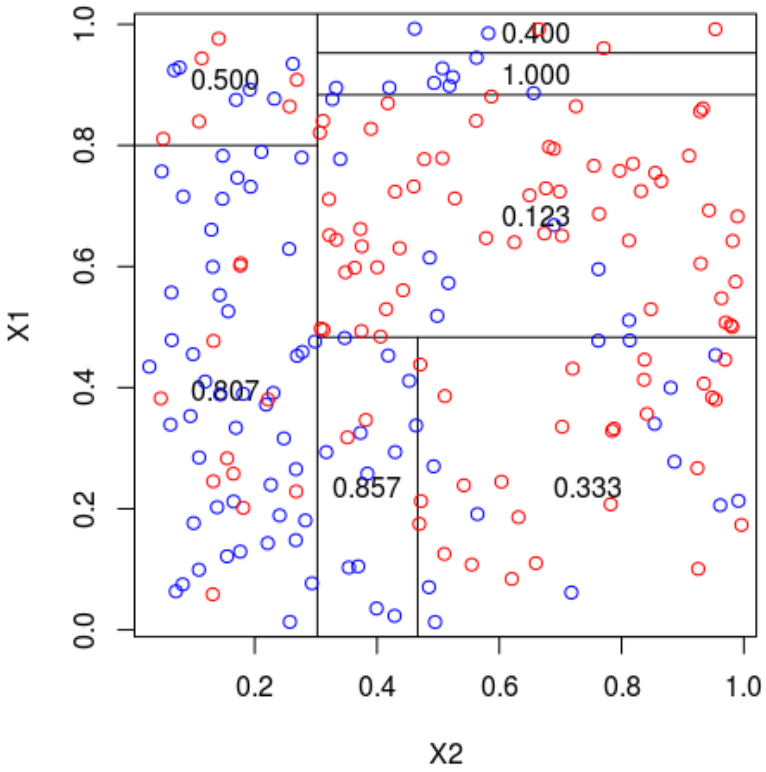
- Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems.
- It works for both categorical and continuous input and output variables.
- Every split of the tree is aligned with one of the feature axes.



https://www.datacamp.com/community/tutorials/decision-trees-R

Decision Trees





https://www.datacamp.com/community/tutorials/decision-trees-R

Random Forest

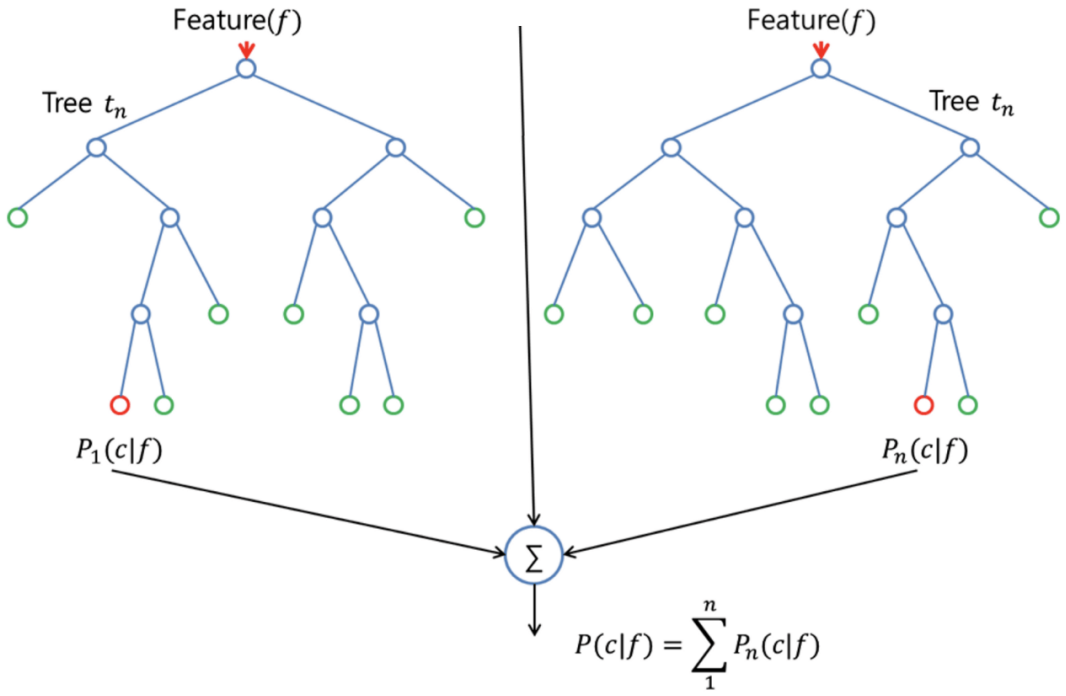
- **Random Forests** is a versatile machine learning method capable of performing both regression and classification tasks.

- Works by building a number of decision trees on bootstrapped training samples.

- But when building these decision trees, each time a split in a tree is considered, a *random sample of **m** predictors* is chosen as split candidates from the full set of ***p*** predictors.

https://www.datacamp.com/community/tutorials/decision-trees-R

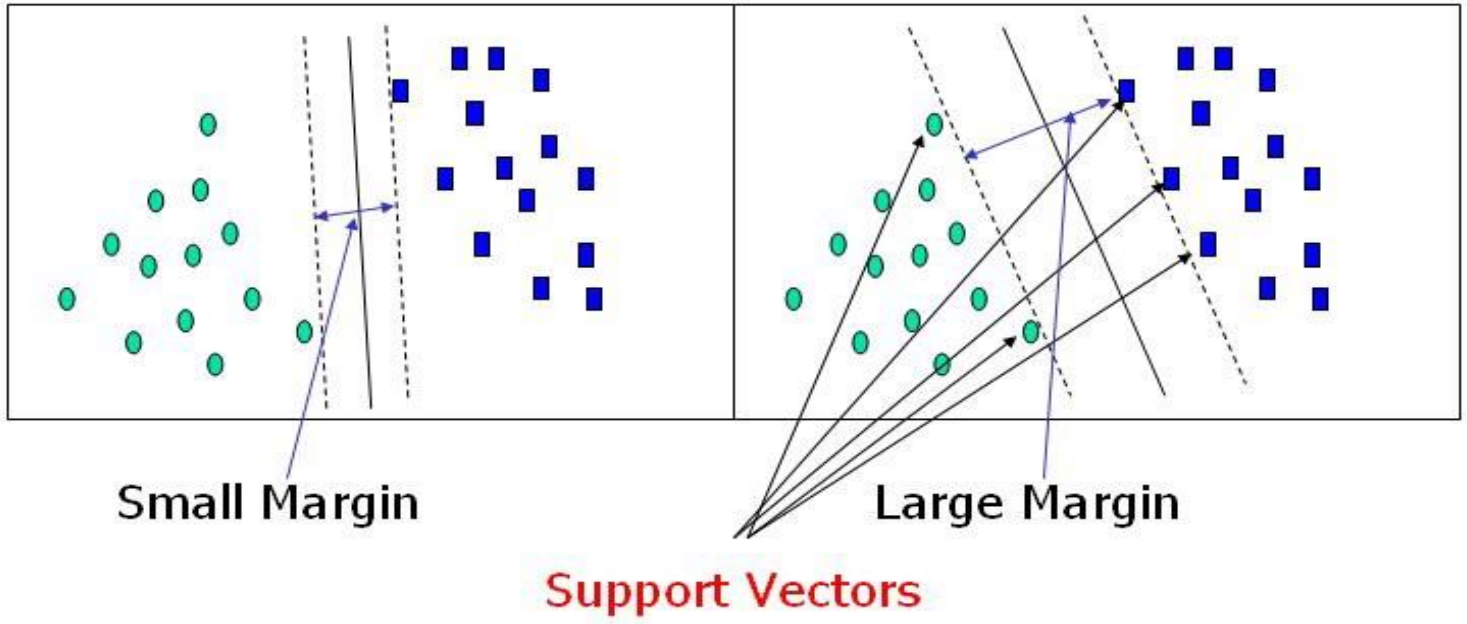
Random Forest



https://www.datacamp.com/community/tutorials/decision-trees-R

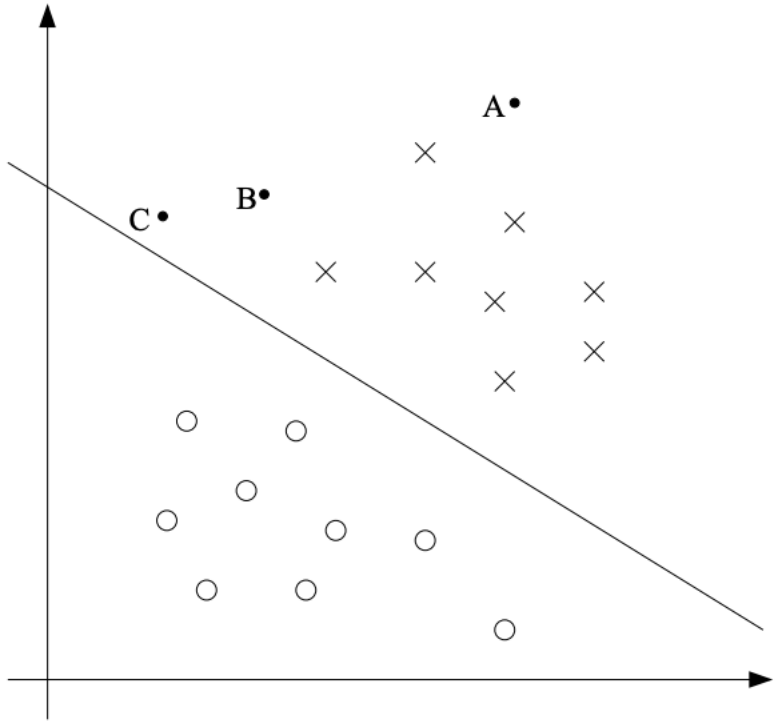
Support Vectors Machine

- The objective of the support vector machine algorithm is to find a hyperplane that distinctly classifies the data points.



https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Support Vectors Machine



http://cs229.stanford.edu/notes/cs229-notes3.pdf

Practical session
UC Irvine Machine Learning Repository
https://archive.ics.uci.edu

UNIVERSITY OF CAMBRIDGE

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Newest Data Sets:		Most Popular Data Sets (hits since 2007):	
07-30-2019:	PPG-DaLiA	2779311:	Iris
07-24-2019:	Divorce Predictors data set	1558049:	Adult
07-22-2019:	Alcohol QCM Sensor Dataset	1208234:	Wine
07-14-2019:	Incident management process enriched event log	1022354:	Car Evaluation
06-30-2019:	Wave Energy Converters	999361:	Wine Quality
06-22-2019:	Query Analytics Workloads Dataset	987612:	Heart Disease
06-17-2019:	Opinion Corpus for Lebanese Arabic Reviews (OCLAR)	978270:	Breast Cancer Wisconsin (Diagnostic)
05-07-2019:	Metro Interstate Traffic Volume	956747:	Bank Marketing
04-22-2019:	Facebook Live Sellers in Thailand	870031:	Human Activity Recognition Using Smartphones
04-15-2019:	Gas sensor array temperature modulation	812456:	Abalone
04-14-2019:	Rice Leaf Diseases	776777:	Forest Fires
04-10-2019:	Parkinson Dataset with replicated acoustic features	545425:	Poker Hand

UNIVERSITY OF CAMBRIDGE

Access the course's repository

https://github.com/adrianobioinfo/autumn_school

UNIVERSITY OF CAMBRIDGE

Practical

K-Means

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	2779352

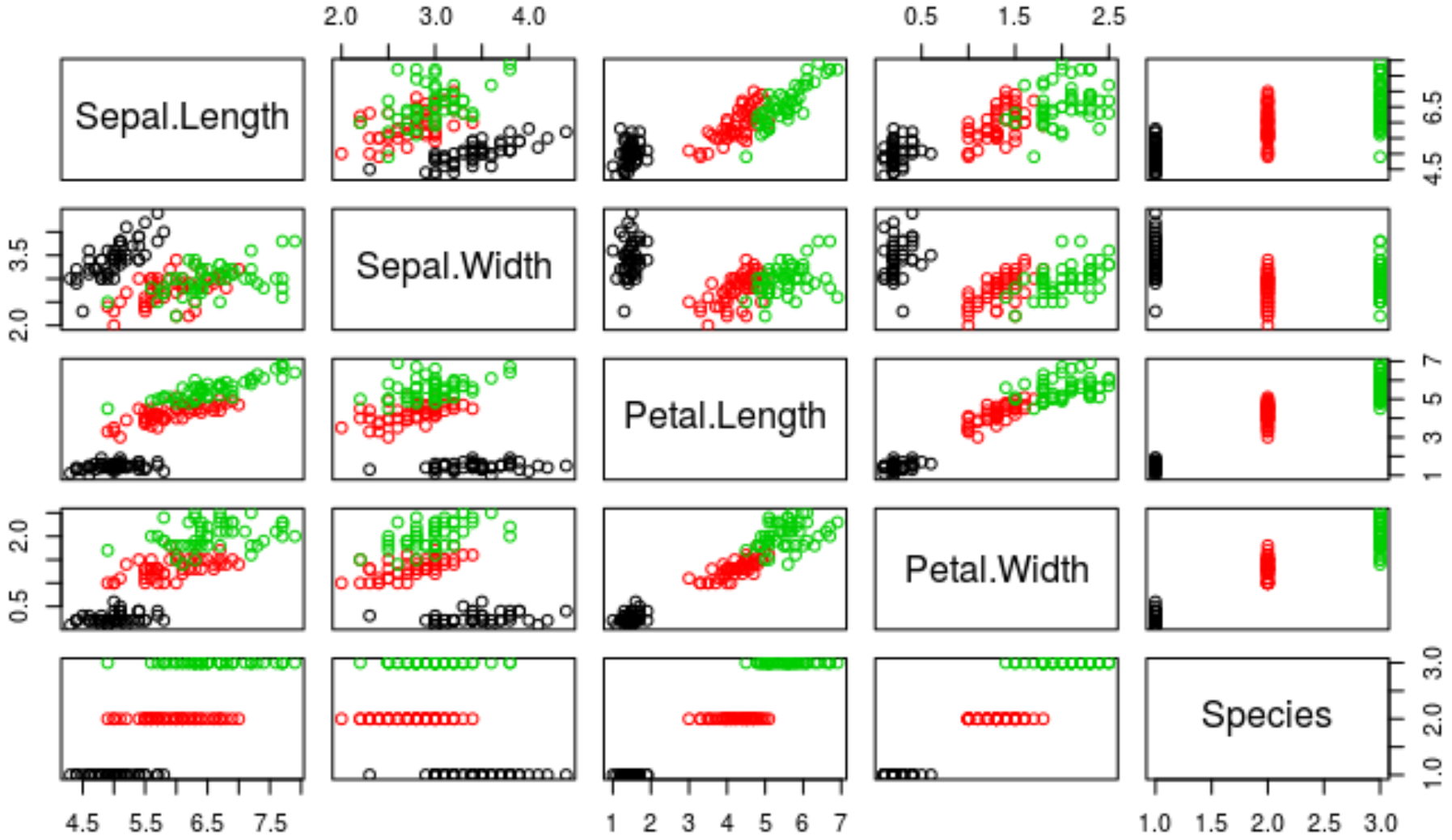
Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, **28**, 100–108. doi: 10.2307/2346830.

kmeans-iris.Rmd

K-Means Clustering

Chunk 3: Plots an overview of the dataset.

UNIVERSITY OF CAMBRIDGE



K-Means Clustering

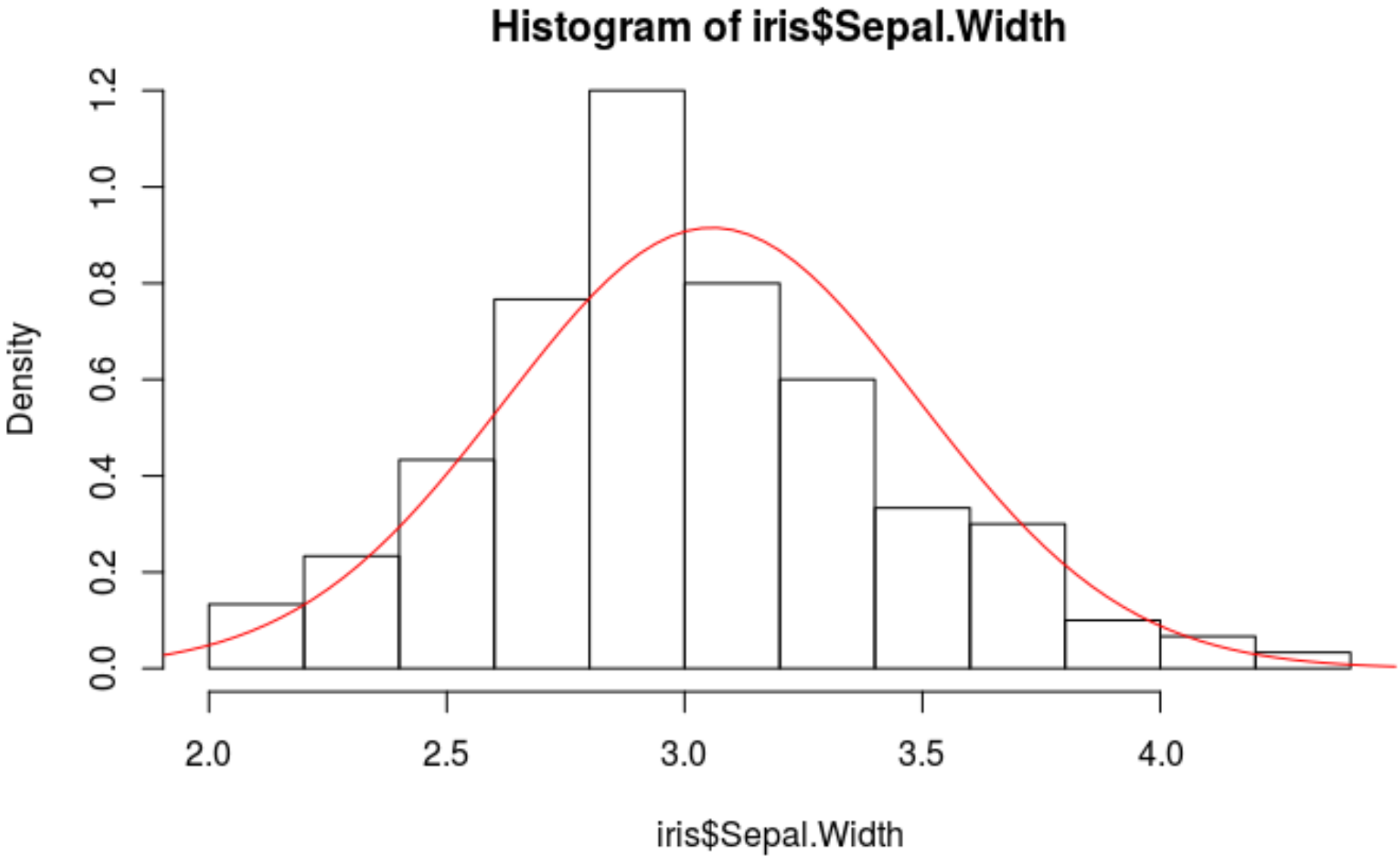
Chunk 3: Plots an histogram of the sepals width

UNIVERSITY OF CAMBRIDGE

K-Means Clustering

Chunk 4:Adds a normal curve to the sepals width histogram

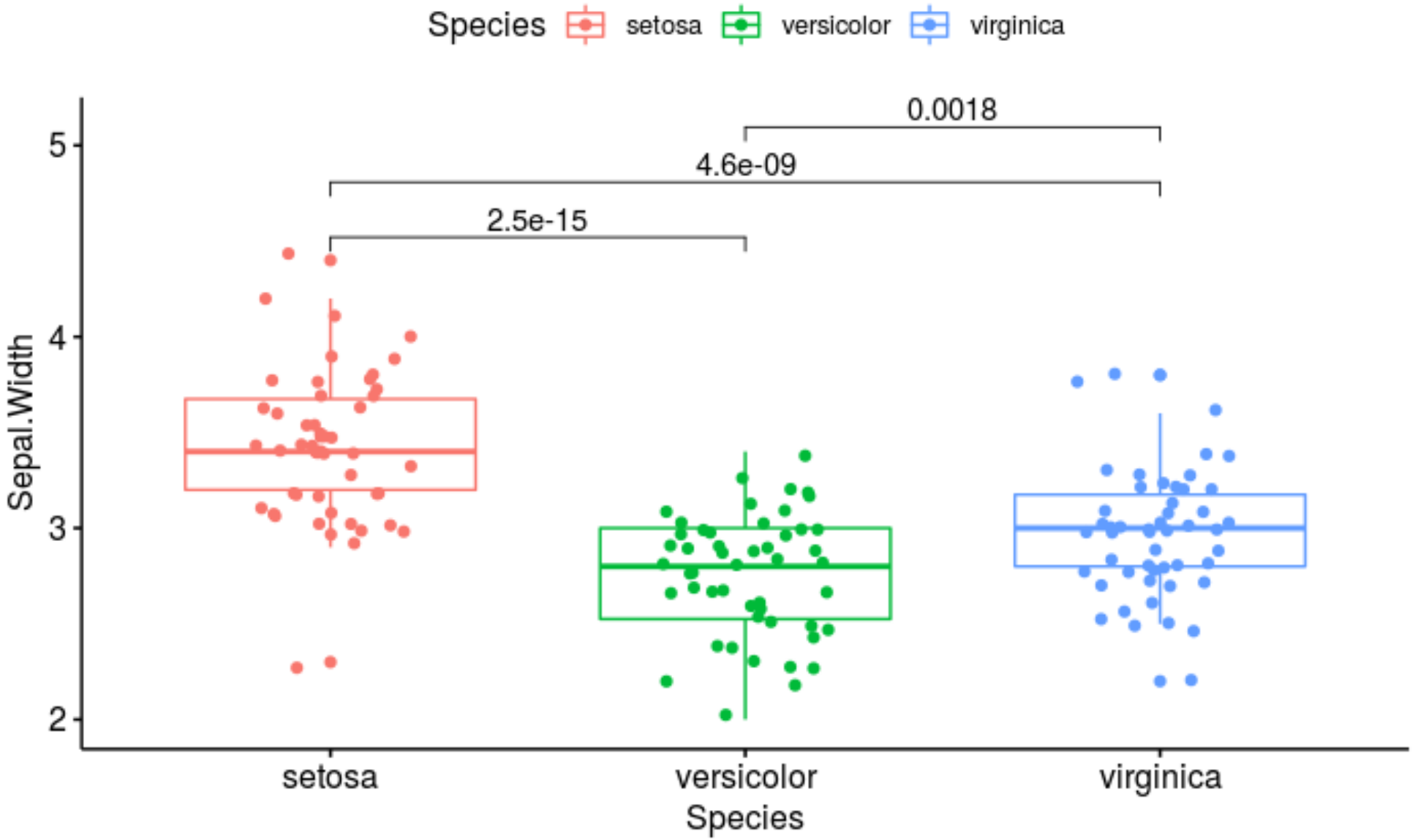
UNIVERSITY OF CAMBRIDGE



K-Means Clustering

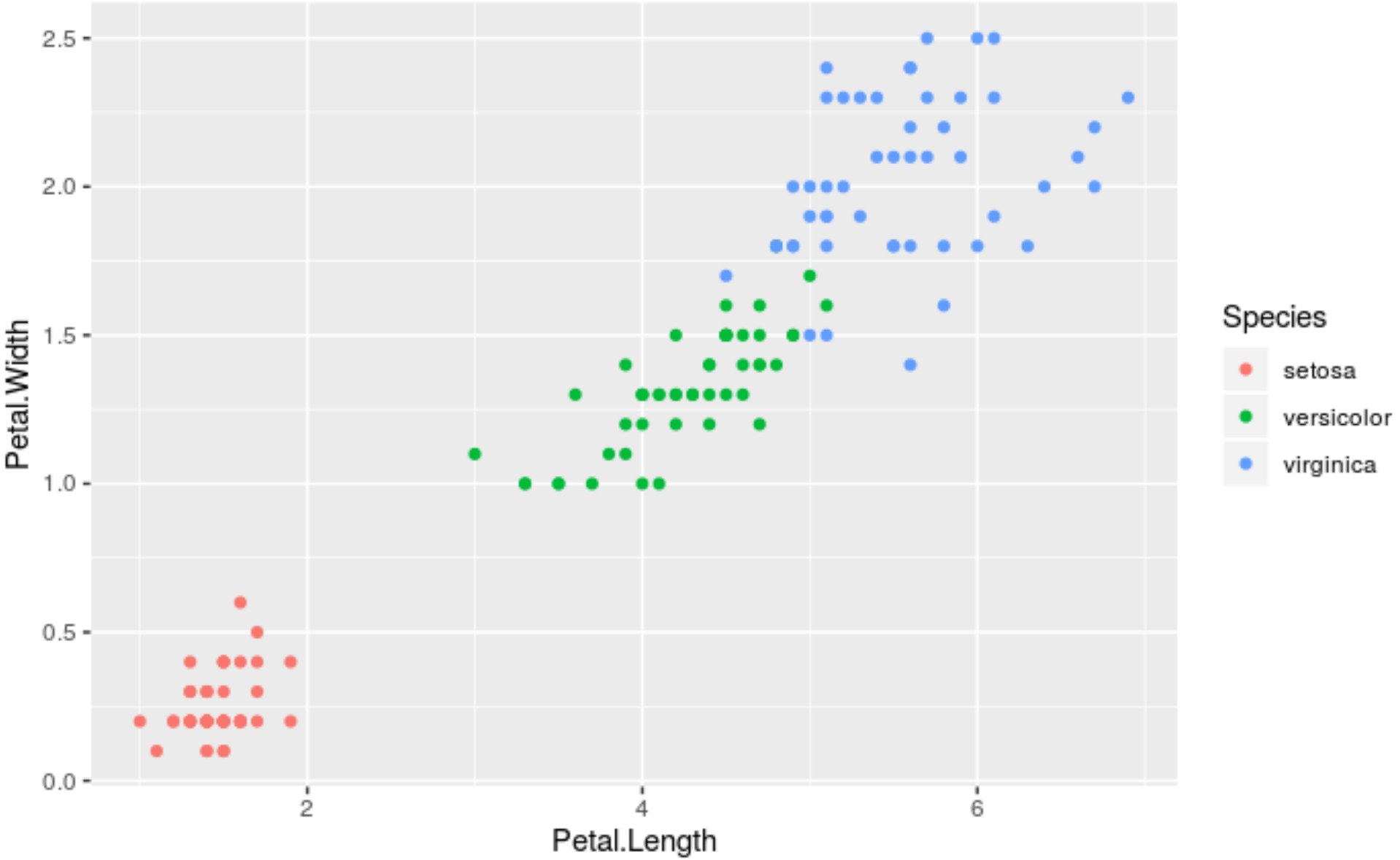
Chunk 5: Compare Species

UNIVERSITY OF CAMBRIDGE



K-Means Clustering

Chunk 6: This plots a scatter plot containing for the species



K-Means Clustering

Chunk 7: Run the K-means function

UNIVERSITY OF CAMBRIDGE

```
K-means clustering with 3 clusters of sizes 52, 48, 50

Cluster means:
  Petal.Length Petal.Width
1     4.269231    1.342308
2     5.595833    2.037500
3     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1
 [52] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2
[103] 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 13.05769 16.29167  2.02200
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

K-Means Clustering

Chunk 8: Compare the clusters with the species.

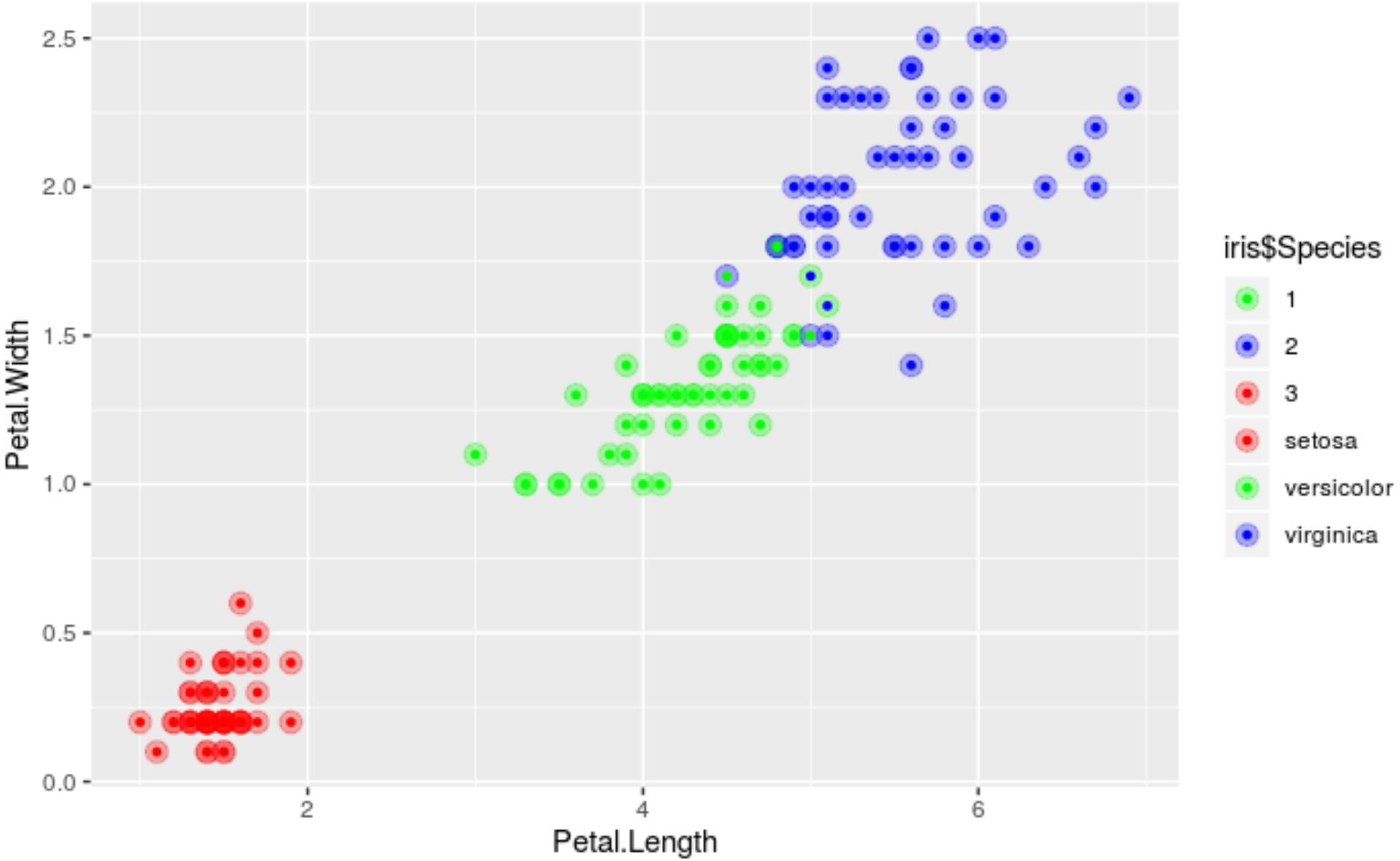
UNIVERSITY OF CAMBRIDGE

	setosa	versicolor	virginica
1	0	48	4
2	0	2	46
3	50	0	0

K-Means Clustering

Chunk 9: Plot the clustering data and compare to the Species distribution

UNIVERSITY OF CAMBRIDGE



UNIVERSITY OF CAMBRIDGE

Practical

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	2779352

k-Nearest Neighbour Classification

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth edition. Springer.

knn-iris.Rmd

k-Nearest Neighbour Classification

Chunk 1: Summarise the dataset

UNIVERSITY OF CAMBRIDGE

```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

k-Nearest Neighbour Classification

Chunk 3: View top rows of the dataset

UNIVERSITY OF CAMBRIDGE

	Sepal.Length <dbl>	Sepal.Width <dbl>	Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

6 rows

UNIVERSITY OF CAMBRIDGE

Practical

Diabetes Dataset – Logistic regression

Dobson, A. J. (1990) An Introduction to Generalized Linear Models. London: Chapman and Hall.

Diabetes Data Set

Download: Data Folder, Data Set Description

Abstract: This diabetes dataset is from AIM '94

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	N/A	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20	**Date Donated**	N/A
Associated Tasks:	N/A	**Missing Values?**	N/A	**Number of Web Hits:**	399057

diabetes-comparison.Rmd

Diabetes prediction: *Logistic regression*

Chunk 3: Looking at the variables

UNIVERSITY OF CAMBRIDGE

```
[1] "pregnant" "glucose"  "pressure" "triceps"  "insulin"  "mass"     "pedigree" "age"
"diabetes"
```

	Pregnant <dbl>	Plasma_Glucose <dbl>	Dias_BP <dbl>	Triceps_Skin <dbl>	Serum_Insulin <dbl>	BMI <dbl>	DPF <dbl>	Age <dbl>	Diabetes <fctr>
1	6	148	72	35	0	33.6	0.627	50	pos
2	1	85	66	29	0	26.6	0.351	31	neg
3	8	183	64	0	0	23.3	0.672	32	pos
4	1	89	66	23	94	28.1	0.167	21	neg
5	0	137	40	35	168	43.1	2.288	33	pos
6	5	116	74	0	0	25.6	0.201	30	neg

6 rows

Diabetes prediction: *Logistic regression*

Chunk 4: Looking at the structure of the data

UNIVERSITY OF CAMBRIDGE

	Pregnant <dbl>	Plasma_Glucose <dbl>	Dias_BP <dbl>	Triceps_Skin <dbl>	Serum_Insulin <dbl>	BMI <dbl>	DPF <dbl>	Age <dbl>	Diabetes <fctr>
1	6	148	72	35	0	33.6	0.627	50	pos
2	1	85	66	29	0	26.6	0.351	31	neg
3	8	183	64	0	0	23.3	0.672	32	pos
4	1	89	66	23	94	28.1	0.167	21	neg
5	0	137	40	35	168	43.1	2.288	33	pos
6	5	116	74	0	0	25.6	0.201	30	neg

6 rows

Diabetes prediction: *Logistic regression*

Chunk 5: Check the number of missing values in each column.

UNIVERSITY OF CAMBRIDGE

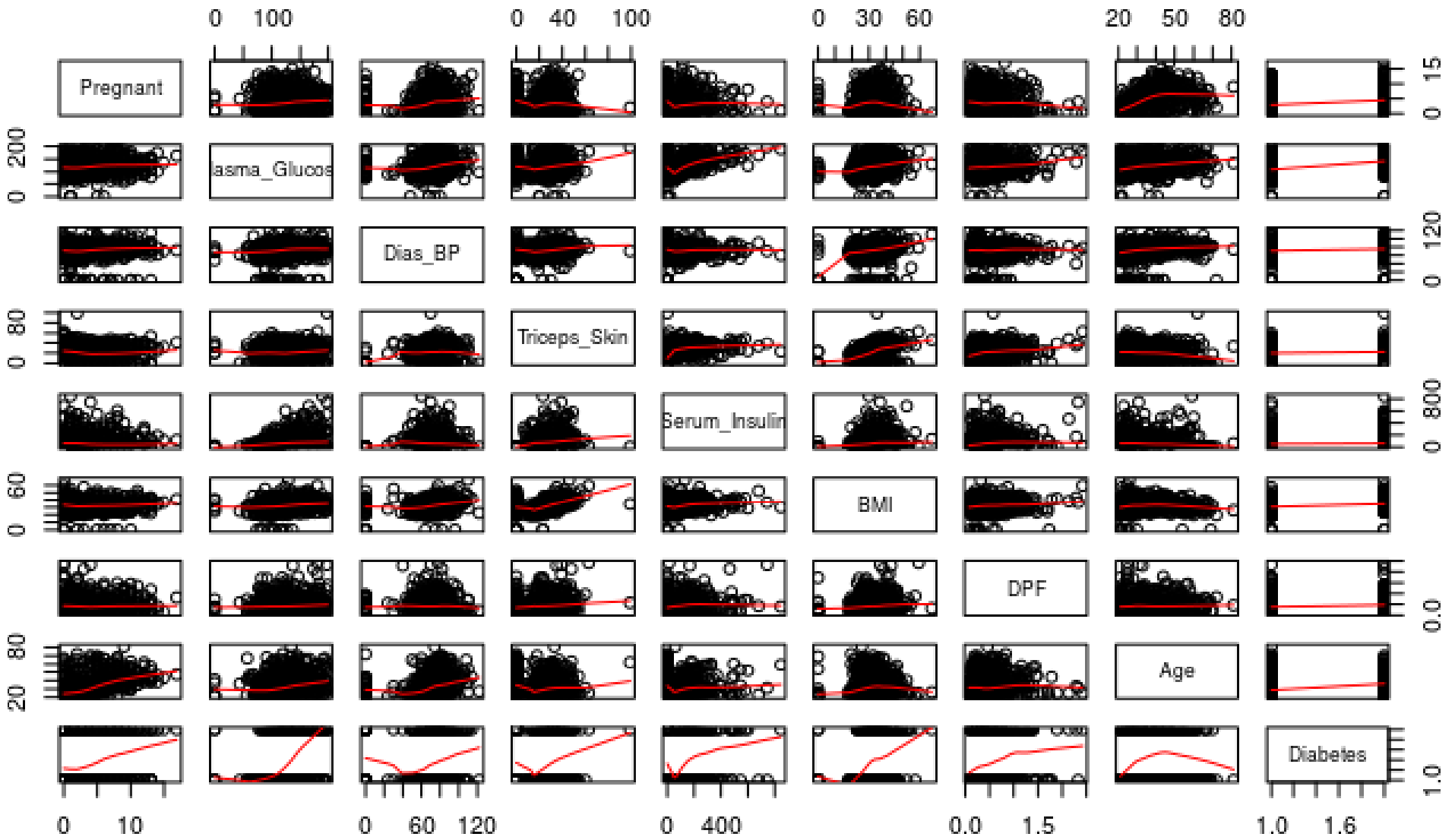
Pregnant	Plasma_Glucose	Dias_BP	Triceps_Skin	Serum_Insulin	BMI	DPF
0	0	0	0	0	0	0

Age	Diabetes
0	0

Diabetes prediction: *Logistic regression*

Chunk 6: Produce a matrix of scatterplots

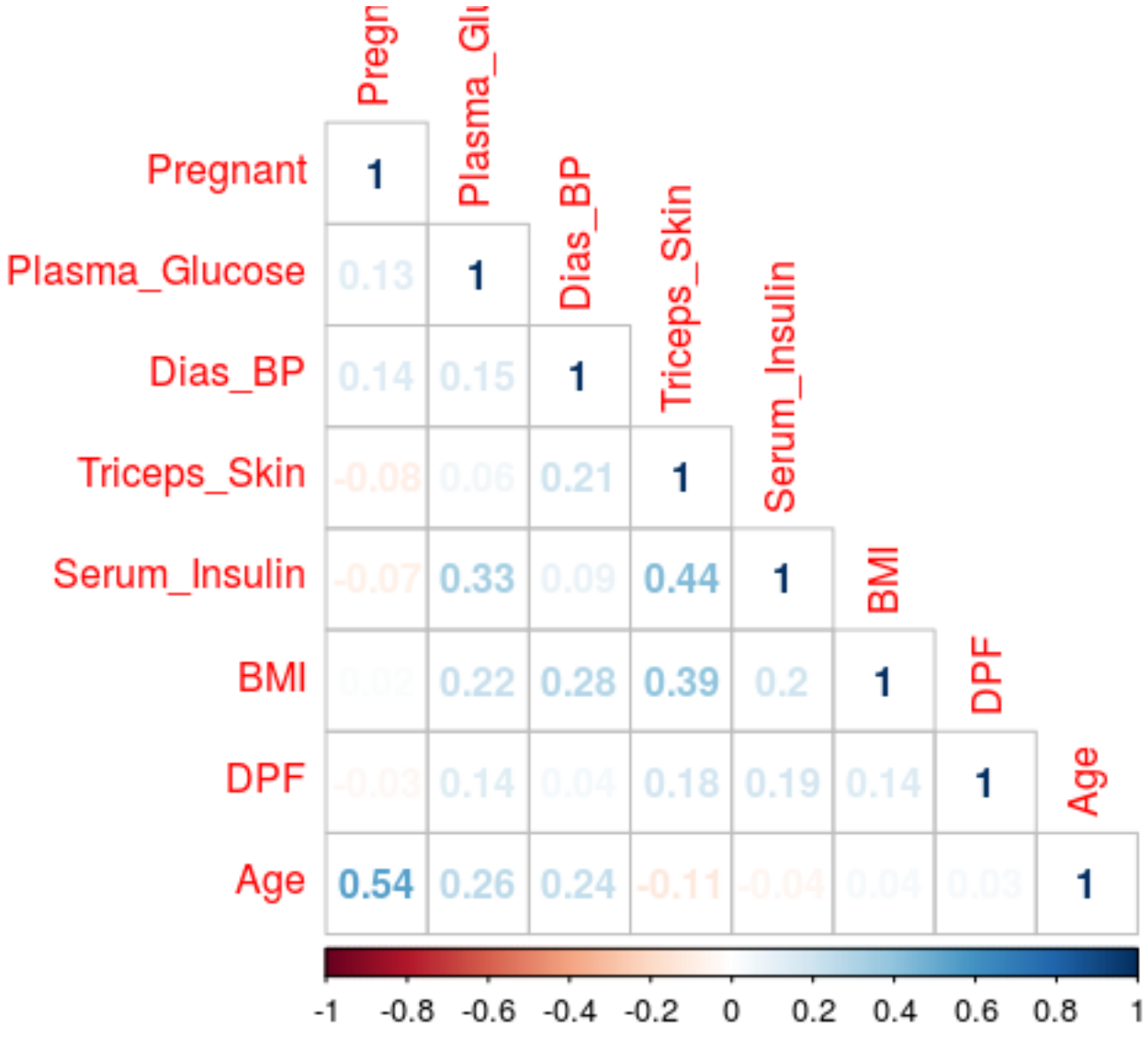
UNIVERSITY OF CAMBRIDGE



Diabetes prediction: *Logistic regression*

Chunk 7: Compute the matrix of correlations between the variables

UNIVERSITY OF CAMBRIDGE



Diabetes prediction: *Logistic regression*

Chunks 8: Apply Logistic Regression model

UNIVERSITY OF CAMBRIDGE

```
Call:
glm(formula = Diabetes ~ ., family = binomial, data = pima_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2424  -0.7256  -0.4283   0.7341   2.9311

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.405409   0.841872  -9.984  < 2e-16 ***
Pregnant        0.103471   0.037973   2.725  0.00643 **
Plasma_Glucose  0.035730   0.004563   7.830 4.89e-15 ***
Dias_BP        -0.012707   0.006057  -2.098  0.03590 *
Triceps_Skin    0.003563   0.008088   0.440  0.65959
Serum_Insulin  -0.001710   0.001060  -1.613  0.10671
BMI             0.088735   0.017954   4.942 7.72e-07 ***
DPF             0.696250   0.334761   2.080  0.03754 *
Age             0.017015   0.011066   1.538  0.12415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 694.17  on 536  degrees of freedom
Residual deviance: 509.76  on 528  degrees of freedom
AIC: 527.76

Number of Fisher Scoring iterations: 5
```

Diabetes prediction: *Logistic regression*

Chunks 9: Update to use only the significant variables

UNIVERSITY OF CAMBRIDGE

```
Call:
glm(formula = Diabetes ~ Pregnant + Plasma_Glucose + Dias_BP +
    BMI + DPF, family = binomial, data = pima_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6175  -0.7389  -0.4472   0.7157   2.9445

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.799784   0.775631 -10.056  < 2e-16 ***
Pregnant        0.138138   0.032715   4.223 2.42e-05 ***
Plasma_Glucose  0.034314   0.004101   8.367  < 2e-16 ***
Dias_BP        -0.011448   0.005844  -1.959   0.0501 .
BMI             0.084610   0.016826   5.028 4.94e-07 ***
DPF             0.676771   0.330840   2.046   0.0408 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 694.17  on 536  degrees of freedom
Residual deviance: 515.01  on 531  degrees of freedom
AIC: 527.01

Number of Fisher Scoring iterations: 5
```

Diabetes prediction: *Logistic regression*

Chunks 10: Testing the Model

UNIVERSITY OF CAMBRIDGE

```
[1] "Confusion Matrix for logistic regression"
          Actual
Predicted neg pos
        0 133  35
        1  17  46
```

Diabetes prediction: *Logistic regression*

Chunks 11-12: Certify that negative is predicted as 0 and positive as 1;

UNIVERSITY OF CAMBRIDGE

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 133  35
         1  17  46

               Accuracy : 0.7749
                 95% CI : (0.7155, 0.8271)
    No Information Rate : 0.6494
    P-Value [Acc > NIR] : 2.434e-05

                  Kappa : 0.4791

 Mcnemar's Test P-Value : 0.0184

            Sensitivity : 0.8867
            Specificity : 0.5679
         Pos Pred Value : 0.7917
         Neg Pred Value : 0.7302
             Prevalence : 0.6494
         Detection Rate : 0.5758
   Detection Prevalence : 0.7273
      Balanced Accuracy : 0.7273

       'Positive' Class : 0
```

Save only the accuracy from the confusion matrix

```
Accuracy
0.7748918
```

UNIVERSITY OF CAMBRIDGE

Practical

Diabetes Dataset – Decision Trees

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth.

diabetes-comparison.Rmd

Diabetes prediction: *Decision Tree*

Chunks 15:

UNIVERSITY OF CAMBRIDGE

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 133  35
         1  17  46

               Accuracy : 0.7749
                 95% CI : (0.7155, 0.8271)
    No Information Rate : 0.6494
    P-Value [Acc > NIR] : 2.434e-05

                  Kappa : 0.4791

 Mcnemar's Test P-Value : 0.0184

            Sensitivity : 0.8867
            Specificity : 0.5679
         Pos Pred Value : 0.7917
         Neg Pred Value : 0.7302
             Prevalence : 0.6494
         Detection Rate : 0.5758
   Detection Prevalence : 0.7273
      Balanced Accuracy : 0.7273

       'Positive' Class : 0
```

Save only the accuracy from the confusion matrix

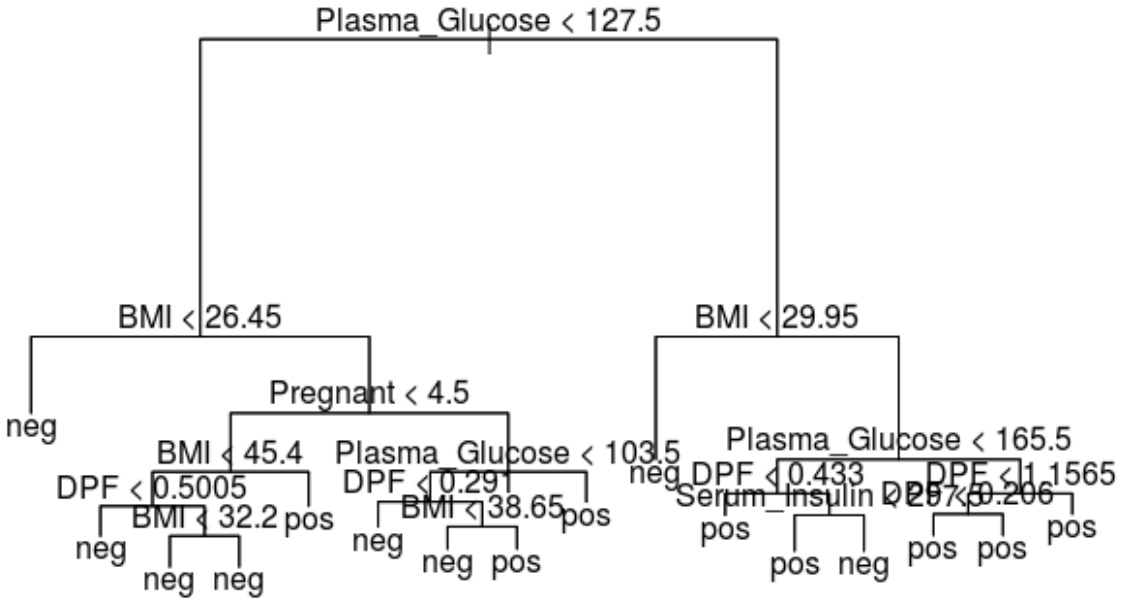
```
  Accuracy
0.7748918
```

Diabetes prediction: *Decision Tree*

Chunks 15-17: Training The Tree Model

UNIVERSITY OF CAMBRIDGE

```
Classification tree:
tree(formula = Diabetes ~ ., data = train)
Variables actually used in tree construction:
[1] "Plasma_Glucose" "BMI"            "Pregnant"       "DPF"
Number of terminal nodes:  16
Residual mean deviance:  0.7515 = 392.3 / 522
Misclassification error rate: 0.1747 = 94 / 538
```




```
1) root 538 696.300 neg ( 0.65056 0.34944 )
  2) Plasma_Glucose < 127.5 337 324.700 neg ( 0.81306 0.18694 )
    4) BMI < 26.45 95  26.640 neg ( 0.96842 0.03158 ) *
    5) BMI > 26.45 242 271.100 neg ( 0.75207 0.24793 )
     10) Pregnant < 4.5 153 129.500 neg ( 0.84967 0.15033 )
       20) BMI < 45.4 146 109.000 neg ( 0.87671 0.12329 )
         40) DPF < 0.5005 92  38.850 neg ( 0.94565 0.05435 ) *
         41) DPF > 0.5005 54  59.610 neg ( 0.75926 0.24074 )
           82) BMI < 32.2 24   8.314 neg ( 0.95833 0.04167 ) *
           83) BMI > 32.2 30  40.380 neg ( 0.60000 0.40000 ) *
       21) BMI > 45.4 7   8.376 pos ( 0.28571 0.71429 ) *
     11) Pregnant > 4.5 89 120.800 neg ( 0.58427 0.41573 )
       22) Plasma_Glucose < 103.5 42  46.110 neg ( 0.76190 0.23810 )
         44) DPF < 0.291 13   0.000 neg ( 1.00000 0.00000 ) *
         45) DPF > 0.291 29  37.360 neg ( 0.65517 0.34483 )
           90) BMI < 38.65 23  24.080 neg ( 0.78261 0.21739 ) *
           91) BMI > 38.65 6   5.407 pos ( 0.16667 0.83333 ) *
       23) Plasma_Glucose > 103.5 47  64.110 pos ( 0.42553 0.57447 ) *
  3) Plasma_Glucose > 127.5 201 266.600 pos ( 0.37811 0.62189 )
    6) BMI < 29.95 51  55.650 neg ( 0.76471 0.23529 ) *
    7) BMI > 29.95 150 167.600 pos ( 0.24667 0.75333 )
     14) Plasma_Glucose < 165.5 96 122.200 pos ( 0.33333 0.66667 )
       28) DPF < 0.433 44  60.910 pos ( 0.47727 0.52273 ) *
       29) DPF > 0.433 52  53.660 pos ( 0.21154 0.78846 )
         58) Serum_Insulin < 297.5 46  35.620 pos ( 0.13043 0.86957 ) *
         59) Serum_Insulin > 297.5 6   5.407 neg ( 0.83333 0.16667 ) *
     15) Plasma_Glucose > 165.5 54  33.320 pos ( 0.09259 0.90741 )
       30) DPF < 1.1565 47  16.540 pos ( 0.04255 0.95745 )
         60) DPF < 0.206 8   8.997 pos ( 0.25000 0.75000 ) *
         61) DPF > 0.206 39   0.000 pos ( 0.00000 1.00000 ) *
       31) DPF > 1.1565 7   9.561 pos ( 0.42857 0.57143 ) *
```

Diabetes prediction: *Decision Tree*

Chunks 18: Testing the model

UNIVERSITY OF CAMBRIDGE

```
Confusion Matrix and Statistics

          Reference
Prediction neg pos
       neg 119  36
       pos  31  44

               Accuracy : 0.7087
                 95% CI : (0.6454, 0.7666)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 0.04036

                  Kappa : 0.3484

 Mcnemar's Test P-Value : 0.62507

            Sensitivity : 0.7933
            Specificity : 0.5500
         Pos Pred Value : 0.7677
         Neg Pred Value : 0.5867
             Prevalence : 0.6522
         Detection Rate : 0.5174
   Detection Prevalence : 0.6739
      Balanced Accuracy : 0.6717

       'Positive' Class : neg
```

UNIVERSITY OF CAMBRIDGE

Practical

Diabetes Dataset – Random Forests

Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32.

diabetes-comparison.Rmd

Diabetes prediction: *Random Forests*

Chunks 20: Training The Model

UNIVERSITY OF CAMBRIDGE

```
Call:
 randomForest(formula = Diabetes ~ ., data = pima_training, mtry = 8,      ntree = 50, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 8

        OOB estimate of  error rate: 24.77%
Confusion matrix:
    neg pos class.error
neg 298  52   0.1485714
pos  81 106   0.4331551
```

Diabetes prediction: *Random Forests*

Chunks 21: Testing the Model

UNIVERSITY OF CAMBRIDGE

```
Confusion Matrix and Statistics

          Reference
Prediction neg pos
       neg 128  34
       pos  22  47

               Accuracy : 0.7576
                 95% CI : (0.697, 0.8114)
    No Information Rate : 0.6494
    P-Value [Acc > NIR] : 0.0002606

                  Kappa : 0.4489

 Mcnemar's Test P-Value : 0.1415789

            Sensitivity : 0.8533
            Specificity : 0.5802
         Pos Pred Value : 0.7901
         Neg Pred Value : 0.6812
             Prevalence : 0.6494
         Detection Rate : 0.5541
   Detection Prevalence : 0.7013
      Balanced Accuracy : 0.7168

       'Positive' Class : neg
```

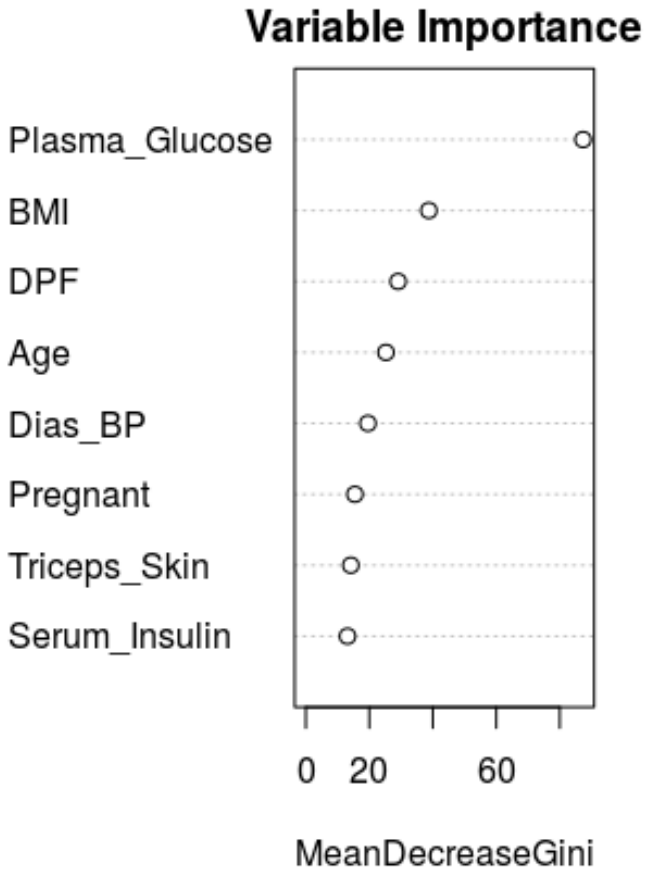
Diabetes prediction: *Random Forests*

Chunks 23-24: The important variables

UNIVERSITY OF CAMBRIDGE

	neg	pos	MeanDecreaseAccuracy	MeanDecreaseGini
Pregnant	2.8535530	1.0928967	3.7002647	15.47504
Plasma_Glucose	10.1287568	11.7275249	14.1368019	87.33836
Dias_BP	3.0475960	0.2337425	2.3148308	19.57903
Triceps_Skin	2.6278950	-0.9940829	1.7848803	14.19321
Serum_Insulin	4.1240729	-3.3179345	0.7842020	13.13958
BMI	4.8846243	6.1544342	7.2488287	38.77288
DPF	-0.6212689	1.7898351	0.9366114	29.07549
Age	2.6260458	1.8448068	3.3753785	25.25592

The Plasma_Glucose is by far the most important variable.



UNIVERSITY OF CAMBRIDGE

Practical

Diabetes Dataset – Support Vector Machines

Support Vector Machines - The Interface to libsvm in package e1071 by David Meyer FH.

diabetes-comparison.Rmd

Diabetes prediction: *Support Vector Machines*

Chunk 26: Tune the model

UNIVERSITY OF CAMBRIDGE

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:

- best performance: 0.222956

- Detailed performance results:
```

gamma <dbl>	cost <dbl>	error <dbl>	dispersion <dbl>
1e-06	0.1	0.3493012	0.06911570
1e-05	0.1	0.3493012	0.06911570
1e-04	0.1	0.3493012	0.06911570
1e-03	0.1	0.3493012	0.06911570
1e-02	0.1	0.3493012	0.06911570
1e-01	0.1	0.2638365	0.05676374
1e-06	1.0	0.3493012	0.06911570
1e-05	1.0	0.3493012	0.06911570
1e-04	1.0	0.3493012	0.06911570
1e-03	1.0	0.3511530	0.06833268

gamma <dbl>	cost <dbl>
0.01	10

Diabetes prediction: *Support Vector Machines*

Chunk 27: Train the model

UNIVERSITY OF CAMBRIDGE

```
Call:
svm(formula = Diabetes ~ ., data = train, kernel = "radial", gamma = 0.01, cost = 10)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10

Number of Support Vectors:  284

 ( 142 142 )


Number of Classes:  2

Levels:
 neg pos
```

Diabetes prediction: *Support Vector Machines*

Chunk 28: Testing the model

UNIVERSITY OF CAMBRIDGE

```
Confusion Matrix and Statistics

          Reference
Prediction neg pos
       neg 136  39
       pos  14  41

               Accuracy : 0.7696
                 95% CI : (0.7097, 0.8224)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 7.748e-05

                  Kappa : 0.4521

 Mcnemar's Test P-Value : 0.0009784

            Sensitivity : 0.9067
            Specificity : 0.5125
         Pos Pred Value : 0.7771
         Neg Pred Value : 0.7455
             Prevalence : 0.6522
         Detection Rate : 0.5913
   Detection Prevalence : 0.7609
      Balanced Accuracy : 0.7096

       'Positive' Class : neg
```

UNIVERSITY OF CAMBRIDGE

Practical

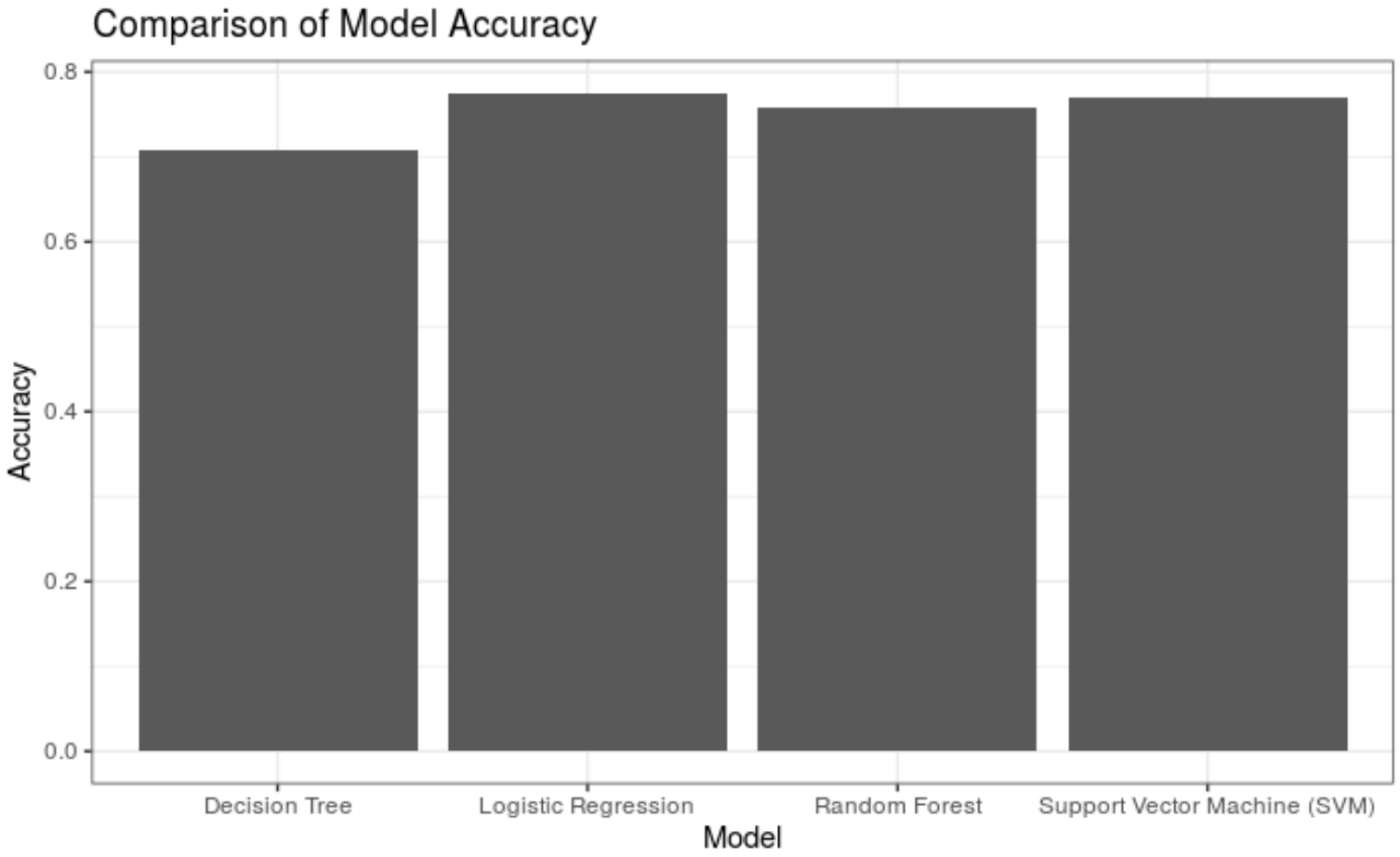
Diabetes Dataset – Comparison of Model Accuracy

diabetes-comparison.Rmd

Diabetes prediction – Pima dataset

Chunks 30-31: Comparison of models accuracy

UNIVERSITY OF CAMBRIDGE



Model <fctr>	Accuracy <dbl>
Logistic Regression	0.7748918
Decision Tree	0.7086957
Random Forest	0.7575758
Support Vector Machine (SVM)	0.7695652

4 rows