Predicting the type of ESD Disease

DA5030 Intro to Machine Learning & Data Mining
Jia Yi (Terri) Shen

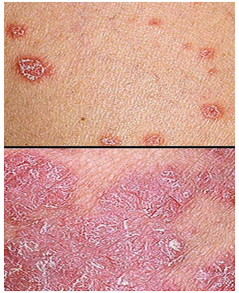
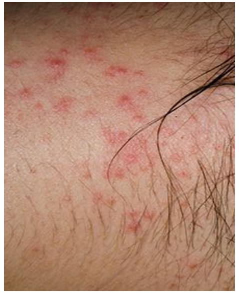
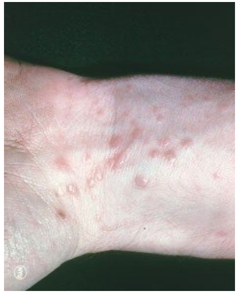
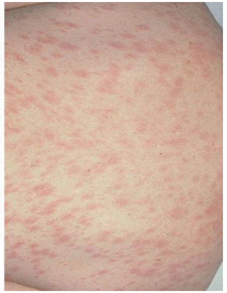
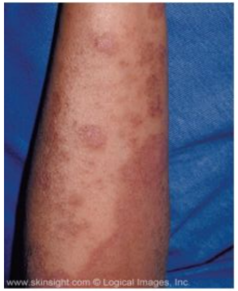
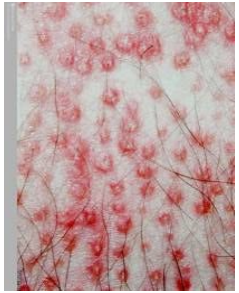
Agenda

CRISP-DM Model

1. Business Understanding
2. Data Understanding - Data Acquisition & Exploration
3. Data Preparation
4. Data Modelling & Evaluation
5. Summarize Performance
6. Deployment

CRISP-DM: Business Understanding

There are a total of six possible phenotypes (physical characteristics) of Erythematous-squamous disease (ESD) disease:

psoriasis	seborrhoeic dermatitis	lichen planus	pityriasis rosea	chronic dermatitis	pityriasis rubra pilaris
			 © MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED	 www.skinsight.com © Logical Images, Inc.	

ESD disease are common in the population. This frequent skin disease share some of the clinical features of scaling and symptom with very small difference which make the differential diagnosis very difficult.

CRISP-DM: Business Understanding

The objective of this project is to construct and compare models that will automatic detecting the type of ESD diseases in order to:

- Reduce the unnecessary biopsy cost
- Help physician for decision making
- Shorten the diagnosis time length
- Assign effective treatment for the patients
- Further enhance drug development efforts

CRISP-DM: Data Understanding

Data Acquisition

- The dermatology data set used in this study is downloaded from UCI Machine Learning Repository.
- The data set was provided by Gazi University School of Medicine, and Bilkent University Department of Computer Engineering and Information Science from January 1998.
- Patients were first evaluated clinically with 12 features.
- The skin samples were then taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

CRISP-DM: Data Understanding

Data Exploration and Data Preparation (Cleaning and Shaping)

- INSPECT
  - Structure and data type
  - Exploratory data plots - Histogram to detect outliers
  - Explore missing data
- TRANSFORM
  - Missing data imputation
  - Transform data type & binning for best performance
- EXPLORE CORRELATION AND PCA
  - Correlation
  - Principal component analysis
- FEATURE ENGINEERING
  - Make a dummy coded data frame for neural network modeling.
  - Convert the diagnosis index to unique string for frequency table using to construct Naive Bayes modeling.

CRISP-DM: Data Understanding

Structure and data type

- Total of 34 attributes, 366 instances (observations), and 6 Classes.
- Clinical and histopathological attributes are mostly ordinal categorical variables ranging from (0-3) with two exceptions:
  - Family history is a categorical 0/1 variable
  - Age is a continuous variable

6 Classes of ESD Diseases

Class	Features
1	psoriasis
2	seboreic dermatitis
3	lichen planus
4	pityriasis rosea
5	cronic dermatitis
6	pityriasis rubra pilaris

12 Clinical Attributes

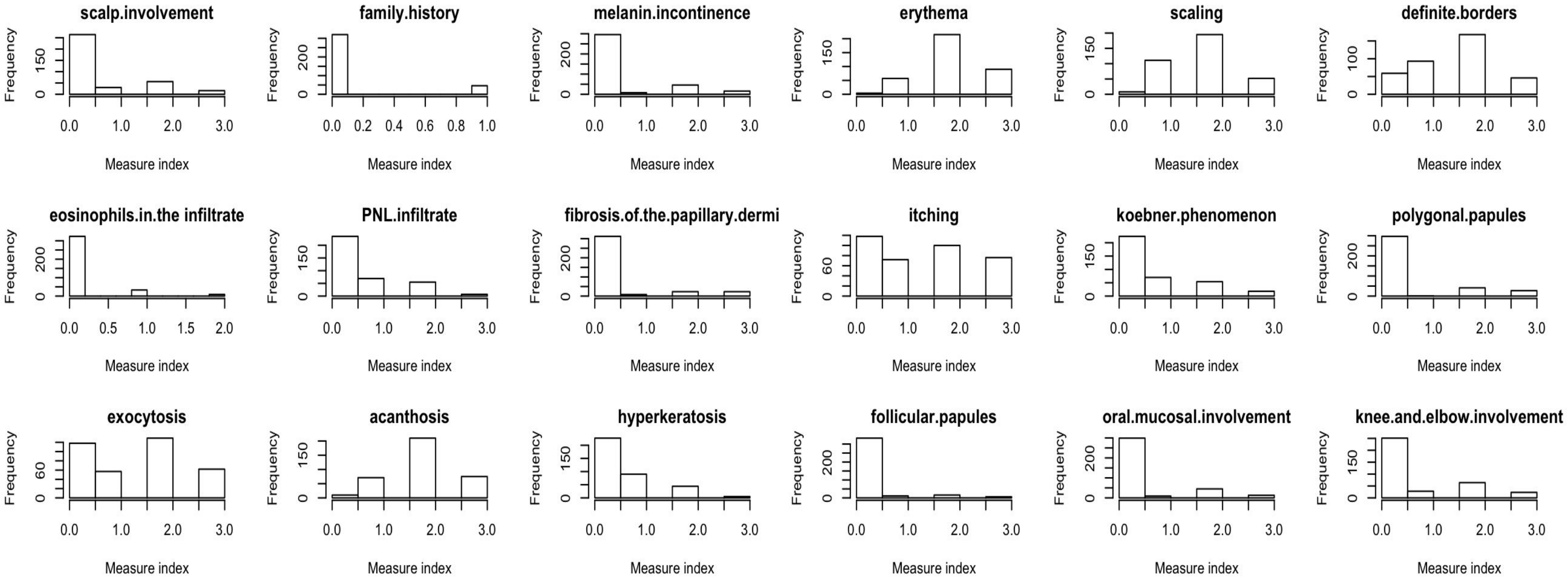
Column	Features
1	erythema
2	scaling
3	definite borders
4	itching
5	koebner phenomenon
6	polygonal papules
7	follicular papules
8	oral mucosal involvement
9	knee and elbow involvement
10	scalp involvement
11	family history (0/1)
34	Age (Linear)

22 Histopathological Attributes

Column	Features
12	melanin incontinence
13	eosinophils in the infiltrate
14	PNL infiltrate
15	fibrosis of the papillary dermis
16	exocytosis
17	acanthosis
18	hyperkeratosis
19	parakeratosis
20	clubbing of the rete ridges
21	elongation of the rete ridges
22	thinning of the suprapapillary epidermis
23	spongiform pustule
24	munro microabcess
25	focal hypergranulosis
26	disappearance of the granular layer
27	vacuolisation and damage of basal layer
28	spongiosis
29	saw-tooth appearance of retes
30	follicular horn plug
31	perifollicular parakeratosis
32	inflammatory monoluclear inflitrate
33	band-like infiltrate

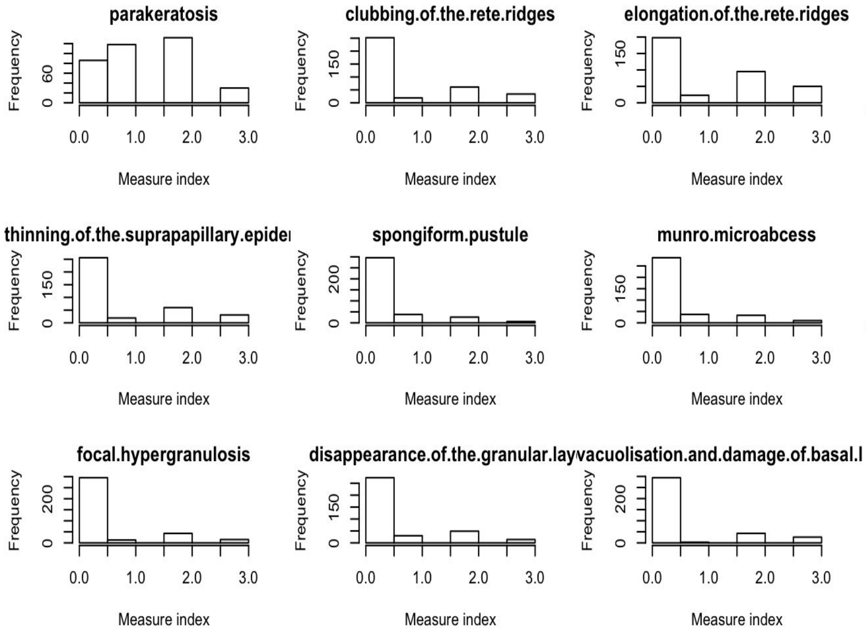
CRISP-DM: Data Understanding

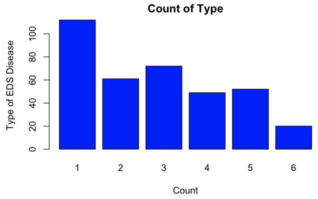
Exploratory data plots - Histogram to detect outliers



CRISP-DM: Data Understanding

Exploratory data plots - Histogram to detect outliers





Notice that we are unable to plot the age column because it will require some data transformation of the data type. I will deal with that in the transformation section later.

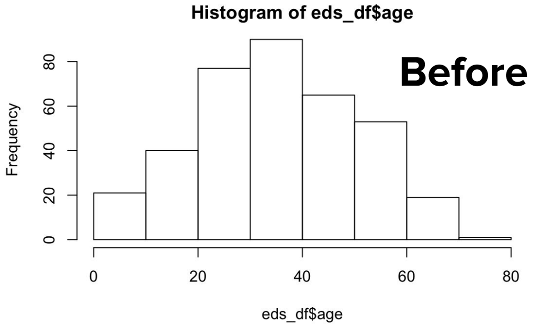
CRISP-DM: Data Understanding

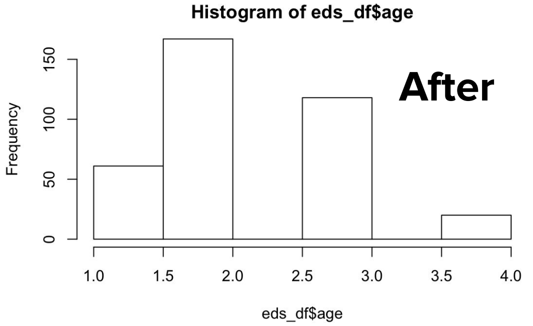
Explore and impute missing data

- There are total of 8 missing data recorded as ? in the age column. It is imputed by mode.

Transform data type & binning

- Column 1-33 as ordinal categorical variable.
- As we state previously about the age column, it is recorded as "character" which is not a categorical variable but discrete numeric. From the histogram, I can see that the age group can be categorized into 4 groups. Therefore, I separate this dataset into 4 bins (1:0-20, 2:21-40, 3:41-60, and 4:61-80).
- After binning, the data type for column 34 and 35 is transformed to factors.

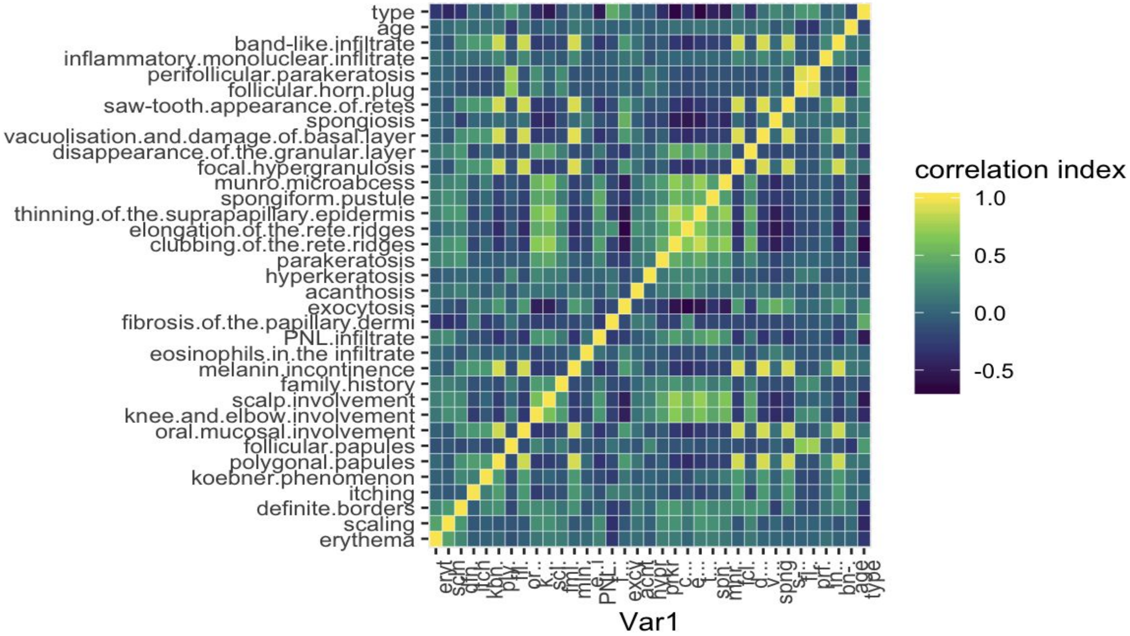

Before


After

CRISP-DM: Data Understanding

Explore correlation and PCA

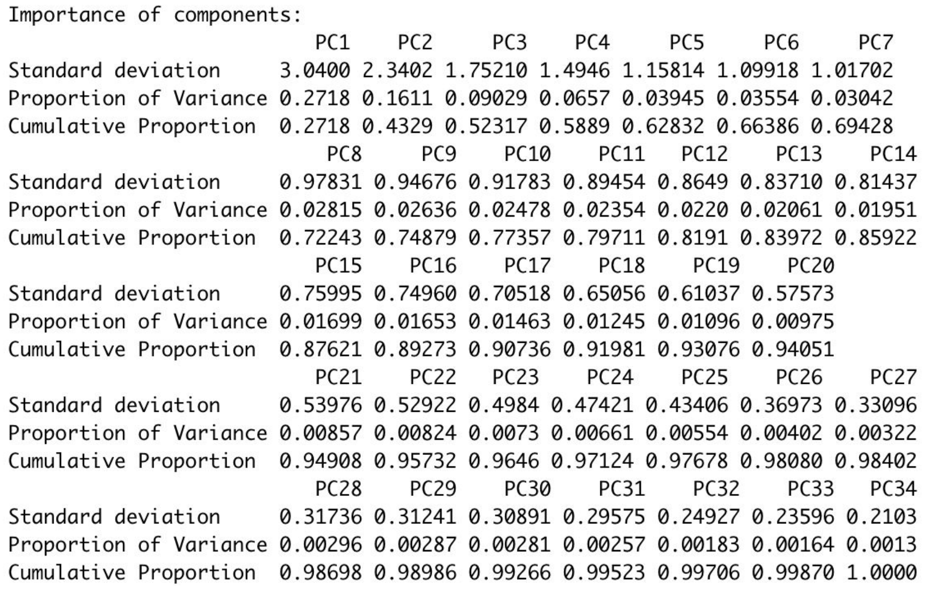
- Correlation
  - Using the Kendall statistic because it estimates rank-based measure of association.
  - Most of them do have some level of dependency on others and it is quite hard to draw a very definitive pattern.



CRISP-DM: Data Understanding

Explore correlation and PCA

- Principal component analysis
  - Apply PCA by scaling the features using z standard score of the sample in all the columns apart from the y value (type) which we try to predict.
  - We then show the explained variance which is the measure of the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set.

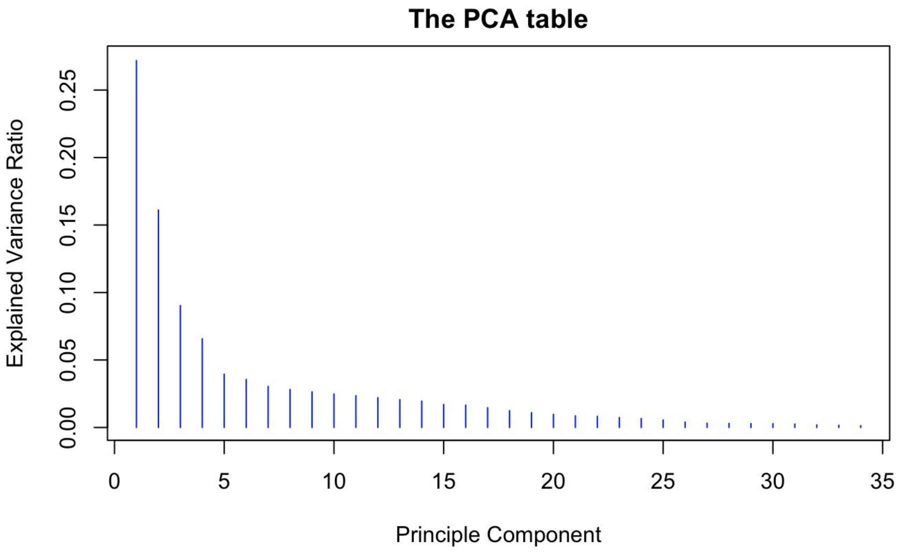


```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7
Standard deviation     3.0400 2.3402 1.75210 1.4946 1.15814 1.09918 1.01702
Proportion of Variance 0.2718 0.1611 0.09029 0.0657 0.03945 0.03554 0.03042
Cumulative Proportion  0.2718 0.4329 0.52317 0.5889 0.62832 0.66386 0.69428
                           PC8     PC9    PC10    PC11   PC12    PC13    PC14
Standard deviation     0.97831 0.94676 0.91783 0.89454 0.8649 0.83710 0.81437
Proportion of Variance 0.02815 0.02636 0.02478 0.02354 0.0220 0.02061 0.01951
Cumulative Proportion  0.72243 0.74879 0.77357 0.79711 0.8191 0.83972 0.85922
                          PC15    PC16    PC17    PC18    PC19    PC20
Standard deviation     0.75995 0.74960 0.70518 0.65056 0.61037 0.57573
Proportion of Variance 0.01699 0.01653 0.01463 0.01245 0.01096 0.00975
Cumulative Proportion  0.87621 0.89273 0.90736 0.91981 0.93076 0.94051
                          PC21    PC22   PC23    PC24    PC25    PC26    PC27
Standard deviation     0.53976 0.52922 0.4984 0.47421 0.43406 0.36973 0.33096
Proportion of Variance 0.00857 0.00824 0.0073 0.00661 0.00554 0.00402 0.00322
Cumulative Proportion  0.94908 0.95732 0.9646 0.97124 0.97678 0.98080 0.98402
                          PC28    PC29    PC30    PC31    PC32    PC33   PC34
Standard deviation     0.31736 0.31241 0.30891 0.29575 0.24927 0.23596 0.2103
Proportion of Variance 0.00296 0.00287 0.00281 0.00257 0.00183 0.00164 0.0013
Cumulative Proportion  0.98698 0.98986 0.99266 0.99523 0.99706 0.99870 1.0000
```

CRISP-DM: Data Understanding

Explore correlation and PCA

- Principal component analysis
  - Plot the graph to learn which features carry maximum information.



CRISP-DM: Model Construction & Evaluation

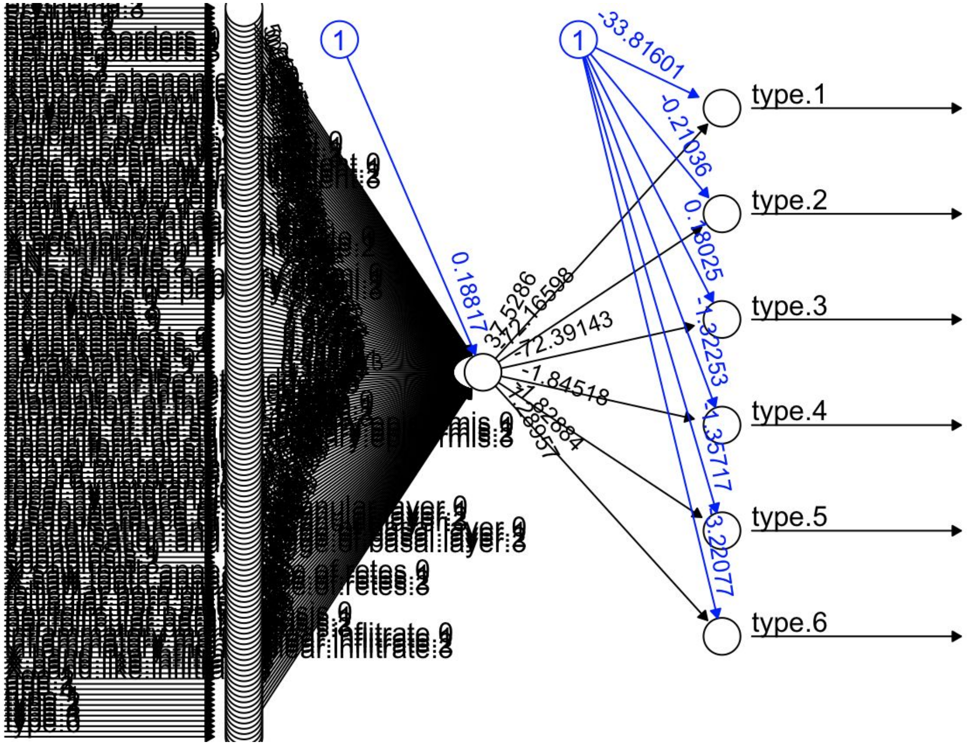
The training and testing set are parted with 75/25 ratio. Four model will be constructed to predict the class of ESD disease:

1. Neural Network via neuralnet
2. Decision Tree via C5.0
3. Naive Bayes via e1071
4. Alternative Naive Bayes

CRISP-DM: Model Construction & Evaluation

Neural Network via neuralnet

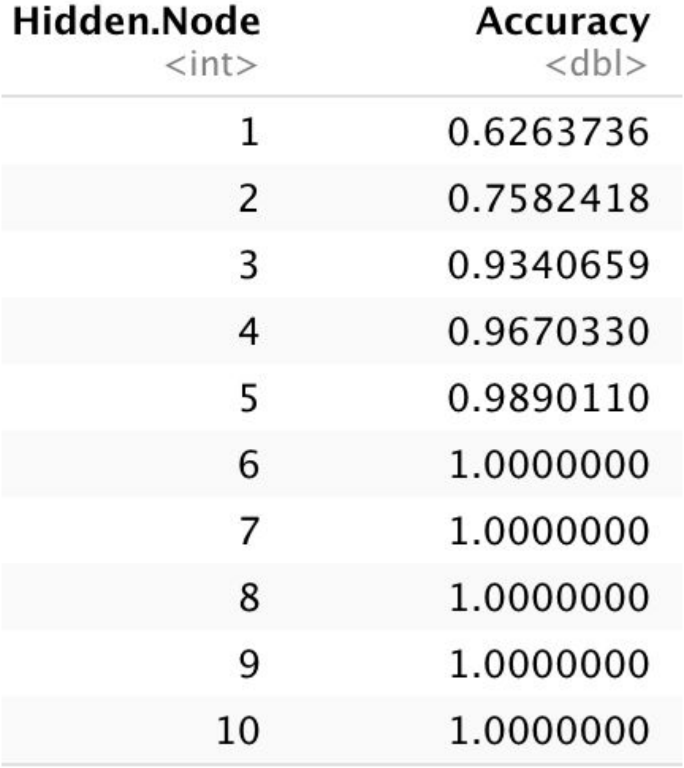
- The dummy data frame has been trained with logistic neural network via neuralnet package to predict the class of the ESD disease.



CRISP-DM: Model Construction & Evaluation

Neural Network via neuralnet

- With 1 hidden node, the accuracy is quiet low about 0.6263736. However, the accuracy increases dramatically as the hidden node increases.

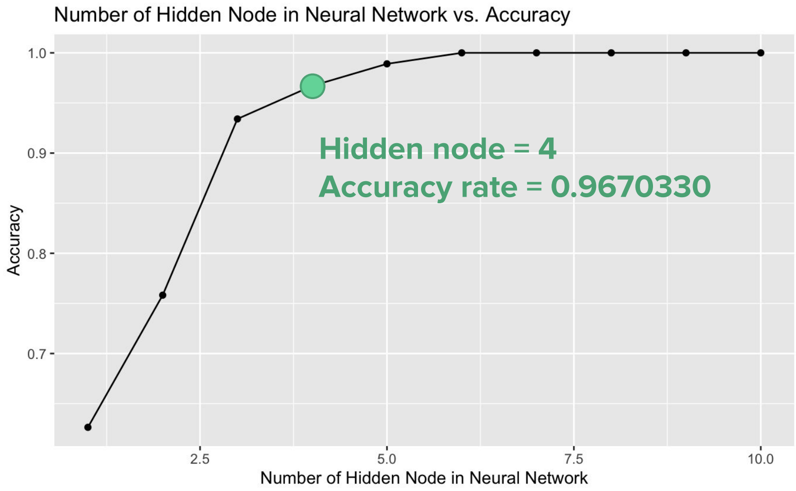


Hidden.Node	Accuracy
<int>	<dbl>
1	0.6263736
2	0.7582418
3	0.9340659
4	0.9670330
5	0.9890110
6	1.0000000
7	1.0000000
8	1.0000000
9	1.0000000
10	1.0000000

CRISP-DM: Model Construction & Evaluation

Neural Network via neuralnet

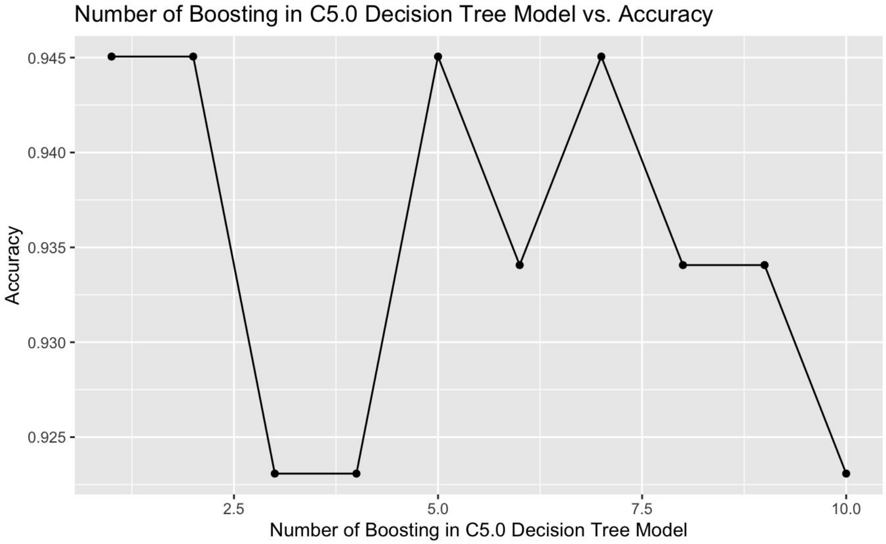
- Networks with more complex topologies are capable of learning more difficult concepts, I increases the number of hidden node to improve the model. Using the elbow law, I concluded that the hidden node of 4 which results in accuracy of 0.9670330 is the best performance improved model.



CRISP-DM: Model Construction & Evaluation

Decision Tree via C5.0

- The original data frame has been trained with C5.0 decision tree model via C50 package to predict the class of the ESD disease. Without boosting, the accuracy is about 0.9450549. However, the accuracy punctuate as the boosting increases.



CRISP-DM: Model Construction & Evaluation

Naive Bayes via e1071

- The original data frame has been trained with Naive Bayes model via e1071 package to predict the class of the ESD disease. Without laplace, the accuracy is about 1. However, this might shows a overfitting problem.
- In order to solve the overfitting problem, laplace smoothing parameter is added. Laplace smoothing solves the overfitting problem by adding 1 to every count to the combination of factors that never occur so it's never zero probability. The final accuracy after adding the Laplace smoothing is calculated to be 0.989011.

CRISP-DM: Model Construction & Evaluation

Alternative Naive Bayes

- The diagnosis index to unique string data frame has been trained with Naive Bayes model construct by myself from studying the Naive Bayes rule to predict the class of the ESD disease.
- My naive bayes model shows an accuracy of 0.8571429, which suggest that the naiveBayes function from e1071 did a lot of fine tuning and model optimization.
- To further improve my naive bayes model, I will actually apply some sort of classifier combination such as ensembling, boosting, and bagging. It also makes sense to explore further at the data quality.

CRISP-DM: Model Construction & Evaluation

Summary

	Initial.model.accuracy	Final.model.accuracy	Note
Neural Network	0.6263736 with 1 hidden node	0.9670330 with 4 hidden node 2	Stable trend when hidden node increase
C5.0 Decision Tree	0.9450549 without boosting	0.9450549 without boosting 3	Unstable trend when boosting increase suggest the noise in the data that will cause model judgement
e1071 Naive Bayes	1 without laplace smoothing	0.989011 with laplace smoothing 1	Laplace smoothing solves the overfitting problem
Alternative Naive Bayes	0.8571429 4	N/A	Need further improvement such as classifier combination via ensembling, boosting, and bagging. It also makes sense to explore further at the data quality.

CRISP-DM: Deployment

The derived model can have a few different implementations:

- Medical information providers such as WebMD can inform accurate information to their users so their users can determine if visiting a clinic is necessary.
- The clinical decision-maker can determine the correct treatment in the earlier stage to prevent delay in treatment.
- The pharmaceutical companies can analyze the data of most frequent occur types in order to define their project scope and pull in research to develop more effective drugs.

References

- https://www.webmd.com/skin-problems-and-treatments/psoriasis/ss/slideshow-psor-overview
- http://www.desimd.com/?q=health-education/skin-and-subcutaneous-disorders/seborrhic-dermatitis
- https://www.aad.org/public/diseases/rashes/lichen-planus
- https://www.mayoclinic.org/diseases-conditions/pityriasis-rosea/symptoms-causes/syc-20376405
- https://nationaleczema.org/eczema/types-of-eczema/atopic-dermatitis/
- http://www.pcds.org.uk/clinical-guidance/pityriasis-rubra-pilaris
- https://archive.ics.uci.edu/ml/datasets/Dermatology
- https://machinelearningmastery.com/better-naive-bayes/