Diabetes and Possible Interventions

Project 8 Group Members:

1. S2174087/1 JUNHAO XU
2. S2175919/1 HAOJIE CHU
3. S2174898/1 ZHENG WANG
4. S2157792/1 PENGWEI LI
5. S2180377/1 YING MING TANG
6. S2164604/1 ELAINE LI

github：https://github.com/forAlgorithm/WQD7003

CRISP-DM

Business Understanding

Problem Statement

Diabetes is a chronic disease characterized by elevated blood sugar levels in the body. Its chronic complications are also extremely serious diseases. The disease has the characteristics of a long treatment cycle, a high recurrence rate and many complications. Physical and mental health have brought great harm (Wang Y, et al., 2020). In 2021, the International Diabetes Federation (IDF) (Saeedi P, Petersohn I, Salpea P, et al., 2019) published a new version of the map of the number of people diagnosed with diabetes worldwide.According to the latest data on this map, there are approximately 463 million adults (aged 20-79 years) with diabetes worldwide, and the number is expected to continue to grow over the next 10 to 20 years. The number of adults with diabetes (aged 20-79 years) is estimated to be 463 million worldwide and is expected to continue to grow over the next 10 to 20 years. 24% of Americans now meet the criteria for the metabolic syndrome, a risk factor for both type 2 diabetes and cardiovascular disease (CVD). And diabetic complications are not easy to be found in the early stage, so many people miss the best treatment time. Diabetes and its complications have become a serious threat to people's health and life killers. How to screen out the "hidden" diabetes group and early-stage diabetes group, find out the most likely complications and pathogenesis of diabetes, and let such groups receive drug treatment in advance and scientifically control and intervene in their lifestyles, which will not only make diabetes more likely Returning to normal life in the early stage can also effectively reduce the probability of complications in the later stage.

The diabetes epidemic is associated with changes in lifestyle-most notably increased energy intake, changes in diet composition and decreased levels of physical activity-and the development of overweight and obesity. (Edward S Hoston., 2017). In recent years, with the advent of the era of smart healthcare, there has been a great improvement in the application of diabetes and complications to the field of data mining. The most commonly used classifiers for diabetes prediction were the Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree, and Naive Bayes.(Larabi-Marie-Sainte S, et al. , 2019)

Business objective

1. To predict the risk of having diabetes based on lifestyle.
2. To predict the level of diabetes (Type 1 diabetes, Type 2 diabetes, Prediabetes, Gestational Diabetes) based on lifestyle.

3. To predict the chance of having other diseases due to diabetes.

Data Understanding

Briefing

The data we use for the project is huge in columns. There are 700 patients' data but attributes reach 106. Here are the basic information about these attributes.

Attributes	Explanation
CodeCentre	Places acquired this da 1 = Seri Kembangan 2 = Dengkil 3 = Salak
Dengkil1	Places acquired this da 1 = Seri Kembangan 2 = Dengkil 3 = Salak
CodeNumber	Similar to patient ID
Age	Age classified in 3 interv 1 = age < 50 2 = age between 50-6 3 = age > 60
AgeGroups	Age classified in 3 interv 1 = age < 50 2 = age between 50-6 3 = age > 60
DiabetesDuration	Diabetes duration in ye All patients were diagnosed
DiabDuration3Cat	Diabetes duration classified in 1 = year< 5 2 = year between 5-9. 3 = year > 10
Gender	The gender of the patie 0 = Female 1 = Male
Ethnics3Cat	The main 3 ethnicity grc
Religion6Ca	The main religious grc

Attributes	Explanation
Religiosity3Cat	The main religiosity into 3
Martial4Cat	Martial status in 4 gro
Education3Cat	Educational status in 3 g 1 = never 2 = primary & seconda 3 = teritary
Employment3Cat	Job status in 3 group 1 = retired/home mana 2 = employed 3 = unemployeed
Exercise	Whether the patient exercise 0 = no 1 = at most 3 times in a 2 = more than 3 times in a
Smoke3Cat	Smoking conditions in 3 g 0 = never 1 = stop > 5 years 2 = stop < 5 years and active
Severe DDS	Moderate diabetes distress as 0 = score < 3 1 = score >= 3
SeverEB	Moderate diabetes distress as 0 = score < 3 1 = score >= 3
SeverePD	Moderate physician dist 0 = score < 3 1 = score >= 3
SevereRD	Moderate regimen distr 0 = score < 3 1 = score >= 3
SevereIPD	Moderate interpersonal di 0 = score < 3 1 = score >= 3
DDS2	The patient feels that if diabetes is taking up too much r 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob

Attributes	Explanation
DDS4	Whether the patient feels angry, scared and/or depressed v 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS7	Whether the patient thinks he/she will end up with serious long-ter 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS10	Whether the patient thinks the diabetes 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS14	Whether the patient feels overwhelmed by the de 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
TotalEmotionalBurden	Sum of the 5 DDS abo
MeanEmotionalBurden	Men cumulative score on the 5
DDS1	Whether the patient thinks his/her doctor doesn't know en 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS5	Whether the patient thinks his/her doctor doesn't give him/her eno 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob

Attributes	Explanation
DDS11	Whether the patient thinks his/her doctor doesn't take 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS15	Whether the patient thinks he/she doesn't have a doctor who he/she 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
TotalPhysicianDistress	Sum of the 4 DDS abc
MeanPhysicianDistress	Men cumulative score on the 4
DDS3	Whether the patient feels unconfident in his/her day- 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS6	Whether the patient feels he/she is not testing the l 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS8	Whether the patient thinks he/she is often failir 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob
DDS12	Whether the patient thinks he/she is not sticking clo 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pr 5 = A serious probler 6 = A very serious prob

Attributes	Explanation
DDS16	Whether the patient thinks he/she is not feeling motivated to 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pro 5 = A serious probler 6 = A very serious prob
TotalRegimenDistress	Sum of the 5 DDS abo
MeanRegimenDistress	Men cumulative score on the 5
DDS9	Whether the patient thinks his/her family or friends are not su 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pro 5 = A serious probler 6 = A very serious prob
DDS13	Whether the patient thinks his/her family or friends don't apprec 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pro 5 = A serious probler 6 = A very serious prob
DDS17	Whether the patient thinks his/her family or friends don't p 1 = Not a problem 2 = A slight problem 3 = A moderate proble 4 = Somewhat serious pro 5 = A serious probler 6 = A very serious prob
TotalInterpersonalDistress	Sum of the 3 DDS abo
MeanInterpersonalDistress	Men cumulative score on the 3
TotalDDS	Sum of all DDS questio
MeanTotalDDS	Mean cumulative of all DDS q
DistressDepress	Distress and depress 1 = have 0 = not have
PHQ1	The patients has little interest on 0 = Not at all 1 = Several days 2 = More than half the d 3 = Nearly everyday

Attributes	Explanation
PHQ2	The patients feels down, depresse 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ3	The patient has trouble falling down or staying a 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ4	The patient feels tired or has li 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ5	The patient is poor appetite or 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ6	The patient feels bad about him 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ7	The patient has trouble concerning on things, such as 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ8	The patient moves or speaks slowly and other people r 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ9	The patients thinks he/she would be better off d 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
PHQ10	Information missing 0 = Not at all 1 = Several days 2 = More than half the c 3 = Nearly everyday
TotalPHQ	Sum of all PHQ abov

Attributes	Explanation
DepressSeverity3Cat	Depression Severity in 3 g
YearDiagonsed	The year diagnosed diabetes, fori
Weight	Weight in kg
Height	Height in cm
HbA1c6.5	Glycated hemoglobin is a form of hemoglobin linked to a suga The lacking of HbA1c might due to In this dataset, the missing value is no 0 = HbA1c <= 6.5 1= HbA1c > 6.5
HbA1c7.0	The same meaning with the attribute above bu 0 = HbA1c <= 7.0 1= HbA1c > 7.0
HbA1c7.0	The same meaning with the attribute above bu 0 = HbA1c <= 7.0 1= HbA1c > 7.0
BPTarget	Blood pressure classified in 0 = bp > 130/80 1 = bp <= 130/80
LDLC2.6	Low-Density Lipoprotein Cholesterol. It's the main lipoprotein in fasting plasma, accounting for ⅔ of plasma lipoprotein, and the main ca 0 = LDL-C > 2.6 1 = LDL-C <= 2.6
HDL1.1	High-Density Lipoprotein Cholesterol. It is mainly synthesized in the liver, it is an anti-atherosclerotic lipoprotein that transports cholesterol from extrahepatic tiss Cannot be lacking in huma 0 = HDL-C < 1.1 1 = HDL-C >= 1.1
TG1.7	Triglyceride, regarded as fa 0 = TG > 1.7 1 = TG <= 1.7
TotalC4.5	Total Cholesterol 0 = TotalC > 4.5 1 = TotalC <= 4.5
SBP	Systolic blood pressure in
DBP	Diastolic blood pressure in
SBP130	SBP classified in 2 grou 0 = SBP > 130 1 = SBP <= 130
DBP80	DBP classified in 2 grou 0 = SBP > 80 1 = SBP <= 80
HbA1c	HbA1c in %

Attributes	Explanation
CBG	CBG in mmol/L
LDL	LDL-C in mmol/L
HDL	DHDL-C in mmol/L
TG	Triglyceride in mmol/
TotalC	Total Cholesterol in mm
HPT	Hypertension 0 = No 1 = Yes
Dyslipid	Dyslipidaemia 0 = No 1 = Yes
DiabetesCx1	Any diabetes complicat 0 = No 1 = Yes
MicroCx1	Any microvascular compli 0 = No 1 = Yes
MacroCx1	Any macrovascular compl 0 = No 1 = Yes
Stroke	A kind of cerebrovascular c 0 = No 1 = Yes
IHD	Ischaemic Heart Disea 0 = No 1 = Yes
Retino	Retinopathy 0 = No 1 = Yes
Nephro	Nephropathy 0 = No 1 = Yes
DFP	Diabetic Foot Problems: Peripheral Neuropathy, Peripheral Vascular Di 0 = No 1 = Yes
Diet	Diet therapy only 0 = No 1 = Yes

Attributes	Explanation
OHA	Oral Hypoglycaemic Ag 0 = No 1 = Yes
Biguanide	Information missing 0 = No 1 = Yes
Sufonylureas	Information missing 0 = No 1 = Yes
AGI	Alpha glucosidase inhib 0 = No 1 = Yes
OHAOthers	Information missing 0 = No 1 = Yes
Insulin	Information missing 0 = No 1 = Yes
Insulin2	Information missing
AHAnumber	Number of AHA agen 0 = No 1 = Yes
AHAZeroto3	AHA 0 to 3 and abov
LLAnumber	Number of LLA agent 0 = No 1 = Yes
LLA2Cat	LLA 2 categories
APAnumber	Number of APA agent 0 = No 1 = Yes
AHA0to3	AHA with highest 3
APA2Cat	APA 2 categories

Discussion

As shown in the table above, there are 106 attributes about the features of diabetes patients. Obviously the dimensions are too many to launch an analysis, so we need to remove some redundant attributes out of the dataset.

- DDS is a survey consisting of 17 questions. There are the detail answers to the questions which provides 17 dimensions, and also some basic statistic result for different kinds of questions(5 categories), consisting of 10 dimensions. Thus there are three methods for these dimensions. We can remain these questions and remove all the Total and Mean total of different type of questions. Or we can remove all the questions and only pay attention to the statistic results. Or we can even create a new feature to describe these existing features, which may perform better. Since we don't know if any of these methods outperforms the others, I would consider to try them all.
- PHQ is another questionnaire contributing more than 10 dimensions. The solutions for these dimensions are the same as DDS.
- Codecentre and Dengkil1 has the same meaning so we can delete one of them.
- Code number looks like the ID of different patients, which is not useful in the analysis. This feature shall be dropped.
- HbA1c, LDL-C, HDL-C, TG, Total cholesterol, BP have both the classified categories and the detail number. Considering removing categories or numbers is all feasible and the performances remain to be tested.

The described missing value format is not always right. There are missing values using a mere blank instead of 999. Thus the process of filling missing values needs to be cautious.
By observing the dataset, we can easily recognize that the statistical result of a questionnaire from a patient will be 'missing' if there is at least one missing value in the questionnaire data. Thus, considering the relationship between them, it will be better to drop all statistical data rather than questionnaire data so that more features can be better maintained. For the missing value, we can use the answers from the related type of questions to make a regression to estimate. Here is a new method about regrading the questionnaire. For DDS, they are categorized into 5 types of questions, and each question contains 6 grading options. If we consider–3, -2, -1, 1, 2, 3 referring to the 6 options, we can finally calculate the unhealthy level(0 means unhealthy, minus grade means not so unhealthy, and the bigger number means unhealthier). This can also be applied on PHQ which contains 4 options and can be graded as -2, -1, 1, 2 and calculate.

Basic Data Report Creation

Before all exploration is intiated, a simple report will be generated to have a browse of the basic situatoin. The code is as followed:

```
# Install packages
! pip install('pandas_profiling')
! pip install('pandas')
```

```
1
```

```
1 #  Load  packages
2 import  pandas  as  pd
3 from  pandas_profiling  import  ProfileReport
4 import  numpy  as  np
```

```
1 #  Load  the  Data
2 dt  =  pd.read_csv('Diabetes  Dataset.csv')
3
4 display(dt.head())
```

	CodeCentre	Dengkil1	CodeNumber	Age	AgeGroups	DiabetesDuration	DiabDuration3Cat	Gender	Ethnic3Cat	Religion6Cat	R
0	1	2	275	68	3	5	2	1	3	4	
1	1	2	112	65	3	33	3	1	2	3	
2	1	2	141	56	2	9	2	0	1	2	
3	1	2	295	61	3	5	2	0	2	1	
4	1	2	5	58	2	20	3	0	1	2	

5 rows × 106 columns

```
1 dt.info(verbose=True,  null_counts=True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Data columns (total 106 columns):
```

#	Column	Non-Null Count	Dtype
0	CodeCentre	700 non-null	int64
1	Dengkil1	700 non-null	int64
2	CodeNumber	700 non-null	int64
3	Age	700 non-null	int64
4	AgeGroups	700 non-null	int64
5	DiabetesDuration	700 non-null	object
6	DiabDuration3Cat	700 non-null	object
7	Gender	700 non-null	int64
8	Ethnic3Cat	700 non-null	object
9	Religion6Cat	700 non-null	object
10	Religiosity3Cat	700 non-null	object
11	Marital4Cat	700 non-null	object
12	Education3cat	700 non-null	object
13	Employment3Cat	700 non-null	object
14	Exercise	700 non-null	object
15	Smoke3Cat	700 non-null	object
16	SevereDDS	700 non-null	object
17	SevereEB	700 non-null	object
18	SeverePD	700 non-null	object
19	SevereRD	700 non-null	object
20	SevereIPD	700 non-null	object
21	DDS2	700 non-null	int64
22	DDS4	700 non-null	int64
23	DDS7	700 non-null	int64
24	DDS10	700 non-null	int64
25	DDS14	700 non-null	int64
26	TotalEmotionalBurden	700 non-null	object
27	MeanEmotionalBurden	700 non-null	object
28	DDS1	700 non-null	object
29	DDS5	700 non-null	int64
30	DDS11	700 non-null	object
31	DDS15	700 non-null	int64
32	TotalPhysicianDistress	700 non-null	object
33	MeanPhysicianDistress	700 non-null	object
34	DDS3	700 non-null	int64
35	DDS6	700 non-null	int64
36	DDS8	700 non-null	int64
37	DDS12	700 non-null	int64
38	DDS16	700 non-null	int64

```
39  TotalRegimenDistress        700 non-null    object
40  MeanRegimenDistress         700 non-null    object
41  DDS9                        700 non-null    object
42  DDS13                       700 non-null    int64
43  DDS17                       700 non-null    int64
44  TotalInterpersonalDistress  700 non-null    object
45  MeanInterpersonalDistress   700 non-null    object
46  TotalDDS                    700 non-null    object
47  MeanTotalDDS                700 non-null    object
48  DistressDepress             700 non-null    object
49  PHQ1                        700 non-null    int64
50  PHQ2                        700 non-null    int64
51  PHQ3                        700 non-null    int64
```

```
# Create a report file
profile = ProfileReport(dt, title='statistics report')
# The report file lies in the same directory with your dataset.
profile.to_file("report.html")
```

Data Preparation

```
# find the NA data
print(dt.info())
null=dt.isnull().sum()
print('Missing values: ',sum(null))

dt.head(5)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Columns: 106 entries, CodeCentre to APA2Cat
dtypes: float64(7), int64(47), object(52)
memory usage: 579.8+ KB
None
Missing values:  0
```

	CodeCentre	Dengkill	CodeNumber	Age	AgeGroups	DiabetesDuration	DiabDuration3Cat	Gender	Ethnic3Cat	Religion6Cat	R
0	1	2	275	68	3	5	2	1	3	4	
1	1	2	112	65	3	33	3	1	2	3	
2	1	2	141	56	2	9	2	0	1	2	
3	1	2	295	61	3	5	2	0	2	1	
4	1	2	5	58	2	20	3	0	1	2	

We can't find the NA value and from the observation of data set we can find 999 and ' ' represent the null data. So we replace them with 'NaN' in order to ensure consistency of null data which benefits our process.

```
# replace the 999 ' ' with Nan and count
data = dt.replace(' ',np.nan)
data = dt.replace(999,np.nan)

# find the NA data
print(data.info())
null=data.isnull().sum()
print(null)
print('Missing values: ',sum(null))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Columns: 106 entries, CodeCentre to APA2Cat
dtypes: float64(51), int64(3), object(52)
memory usage: 579.8+ KB
None
```

```
CodeCentre     0
Dengkill       0
CodeNumber     0
Age            2
AgeGroups      2
               ..
LLAnumber      6
LLA2Cat        0
APAnumber      8
AHA0to3        0
APA2Cat        0
Length: 106, dtype: int64
Missing values:  851
```

```
#Getting  categorical  data
df_cat  =  data.loc[:,dt.dtypes==np.object]
#Getting  Numeric  data
df_num  =  data.loc[:,dt.dtypes!=np.object]
```

```
<ipython-input-19-ea2a6622c761>:2: DeprecationWarning: `np.object` is a deprecated alias for the builtin `object`. To silence this warning,
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  df_cat = dt.loc[:,dt.dtypes==np.object]
<ipython-input-19-ea2a6622c761>:4: DeprecationWarning: `np.object` is a deprecated alias for the builtin `object`. To silence this warning,
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  df_num = dt.loc[:,dt.dtypes!=np.object]
```

Data set description

Our dataset has many features

1. Up to 106 columns
2. Only have 700 rows
3. Many missing values filled with 999 and " "
4. Some rows have many null columns

Due to the above characteristics , We will focus on data cleaning, data transformation and data reduction.

The data have 106 columns and we divide it into categorical and numerical Data.

- Categorical Data: 54 rows exist the null data value is ' '
- Numerical Data: 52 rows exist the null data value is 999

Data cleaning

We replace the value 999 and ' ' with NaN in order to ensure consistency of null data which benefits our process.

Our dataset has fewer rows so we must delete the data carefully. However we can see some rows have many null columns which could affect our analytics . So What we will do is check whether we should drop the row to ensure we have enough data to provide an accurate analysis and also avoid serious biases. Then what we should do is impute different columns with different imputation methods depending on different attributes' characteristics.

The approach we will use is as follows

- cold-deck
- hot-deck
- predictive imputation

For some features that have a stronger correlation, we can use a classifier model to predict the missing value. As the dataset contains patients, most of the patients have the same lifestyle. Also, hot-deck imputation is a best way for our case. We can identify the most similar case to the case with a missing value and substitute the most similar case's value for the missing case's value .

Data Transformation

We will focus on data smoothing to remove noise from data, and data normalization to scale the attribute fall within the small and specified range.

From our dataset, some features have lots of outliers and some are complete so we will attach the following approach to process the data normalization.

min-max normalization:

Guarantees all features will have the exact same scale but does not handle outliers well.

z-score normalization:

Handles outliers, but does not produce normalized data with the exact same scale.

Data Reduction

We can see the original dataset has 106 columns so we must process the data reduction to obtain a reduced representation of the data set which has the same analysis results.

For our dataset, reducing the column is the main task. As the data introduction mentioned. DDS,PHQ both are survey data that cover more than 10 attributes. Data cube aggregate is suited for our project. If we can aggregation them and use a new feature to replace them. We can get fewer columns. As our project objective is to predict the level of diabetes based on lifestyle, the feature selection is the best way So for the data reduction, we mainly use the following approach to process it.

Data cube aggregation:

to find the smallest representation which can solve our objective

Attribute subset selection(Feature selection)

Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features.

Modelling

Select ModelingTechnique

The goal of our project is to do prediction tasks for diabetes, mainly classification tasks, so we can choose a class of classifier models within machine learning. For example: logistic regression, decision tree, XGBoost, etc.

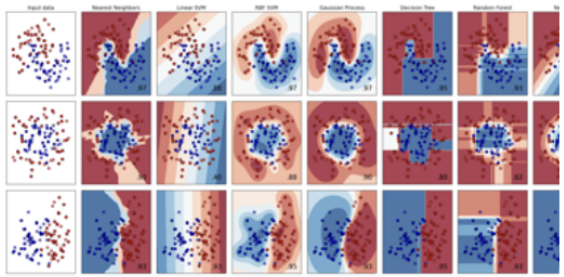
We will use scikit-learning for model building. scikit-learn is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license. it is a complete computational library for machine learning algorithms, with built-in support for common traditional machine learning algorithms and rich documentation and cases.

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

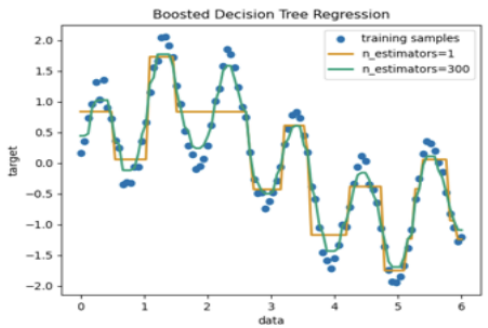


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

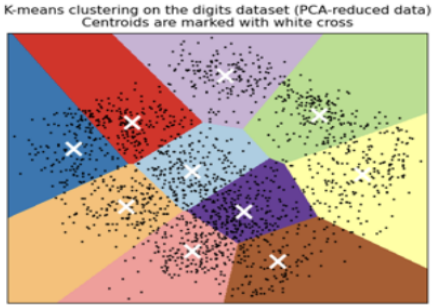


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Model selection

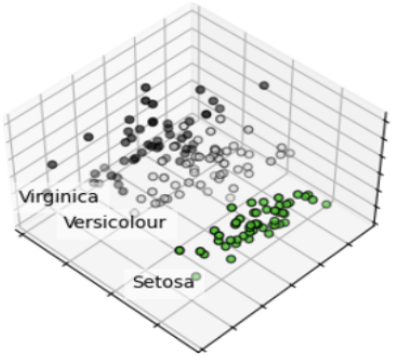
Comparing, validating and choosing parameters and models.

Preprocessing

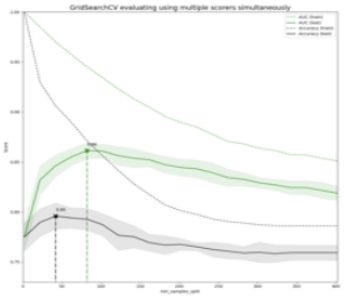
Feature extraction and normalization.

Applications: Transforming input data such as text

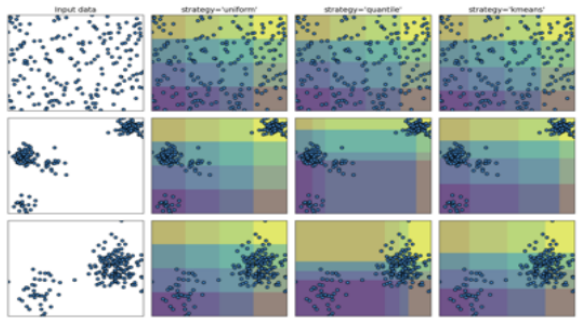
Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization, and more...



Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...



for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...



Build Model

Cross-validation

To evaluate how well a given algorithm performs in generalization after training on a given dataset, we introduce cross-validation. Cross-validation is used to evaluate the predictive performance of a model, especially the performance of a trained model on new data, and can reduce overfitting to some extent. It also allows to obtain as much valid information as possible from a limited amount of data.

Grid search

In order to find the optimal parameters, we introduce grid search. Grid search is a common means of transferring parameters, and is an exhaustive method. Given a series of hyperparameters, the optimal set of hyperparameters is selected from all combinations by exhaustive traversal, which is actually a violent method to find the optimal solution among all solutions.

Logistic regression

Logistic regression is a generalized linear regression analysis model, which belongs to supervised learning in machine learning. Its derivation process and calculation is similar to that of regression, but it is actually mainly used to solve dichotomous classification problems (and also multi-classification problems). The model is trained with a given n sets of data (training set), and the given set or sets of data (test set) are classified at the end of the training. Each of these sets of data is composed of p indicators.Introduce the softmax function. The function form is as follows

$$\mathrm{Softmax}(k, x_1, x_2 \ldots, x_n) = \frac{e^{x_k}}{\sum_{i=1}^{n} e^{x_i}}$$

Here again, we put the k classes with the numbers 1, 2 .... .k to represent the function in which the function value represents the probability. x is processed by the function to obtain the value of the corresponding position (classification) inside the vector is the probability of taking the corresponding position (classification).

Decision tree

Decision tree is a basic classification and regression method. The decision tree model has a tree-like structure and represents the process of classifying instances based on features in a classification problem. It can be thought of as a collection of if-then rules or as a conditional probability distribution defined over the feature space and class space. Its main advantages are the readability of the model and the speed of classification. For learning, a decision tree model is built based on the principle of minimizing the loss function using the training data. For prediction, the decision tree model is used to classify new data, using the decision tree model. Decision tree learning usually consists of 3 steps: feature selection, decision tree generation and decision tree pruning.

XGBoost

XGBoost (eXtreme Gradient Boosting), also known as extreme gradient boosting tree, is an implementation of the boosting algorithm. It works very well for classification or regression problems. It shines in various data competitions and is also widely used in industry, mainly because of its excellent results, simplicity of use, speed, and other advantages. xgb is an implementation of the boosting algorithm that focuses on reducing bias, that is, reducing the error of the model. Therefore it uses multiple base learners, each of which is relatively simple to avoid overfitting, and the next learner learns the difference between the result of the previous base learner and the actual value. The basic idea is to keep generating new trees, each tree is learned based on the difference between the previous tree and the target value, thus reducing the bias of the model

Assess Model

The results of different trainers are compared and evaluated, a comparison table of the model is drawn, and the underfitting and overfitting of the model is evaluated based on the learning curve and validation curve, and whether to introduce regularization is considered. Finally, model fusion can be done to obtain better prediction results.

Evaluation

Evaluate Results

We will proceed to assess the outcomes from each model and review business understandings to ensure the chosen model meets the business objectives. If there is deficiency happened, we may need to review the process from business understanding to data modelling in order to provide an accurate model. We may use confusion matrix method to evaluate the model and further with checking on classification error, recall, accuracy, precision, false positive rate and F1-score.

Review Process

With enough confidence level, we shall work on reviewing the process of developing the model to avoid overlooking on some important variables or factors that may cause the inaccuracy of the model and challenge the quality assurance of the model.

Determine Next Steps

Once we have the chosen model, we will proceed to make decision on whether to move to the next stage which is deployment or back to business understanding to further enhance on the model. Before deployment, we shall look into resources such as time and budget as well to determine if it is worth to be deployed.

Deployment

Plan Deployment

When our model is ready to use, we will undergo the deployment plan which we need to. At this stage, we will need to take the evaluation results to conclude a strategy for deployment. This is to create a summary of our strategy for deployment, the required steps, and the instructions for carrying out those steps.

Plan Monitoring and Maintenance

Data-mining work is a cycle, so we will need to stay actively to monitor and ready for any coincident happen with our model. Thus, we will need to ensure that the model is being used properly on an ongoing basis, and any decline in performance of the model will be detected. This is to avoid any unnecessarily long periods of incorrect or inappropriate usage of data mining results.

Produce Final Report

Deliverables for this task include two items which are final report and final presentation.

1. Final report: The final report summarizes the entire project by assembling all the reports created up to this point, and adding an overview summarizing the entire project and its results.
2. Final presentation: A summary of the final report of data mining outcome is presented in a meeting with the team and the stakeholder. This is also an opportunity to address any open questions.

Review Project

At this stage, the team has to meet to discuss what worked and what didn’t, what would be good to do again, and what should be avoided. For example, is this model the best fit model for these tasks? If the answer is no, then what is the best model to replace the current model. After review and discussion, the team can generate an experience documentation report. This is where we should outline any work methods that worked particularly well, so that they are documented to use again in the future, and any improvements that might be made to the process. It is also the place to document problems and bad experiences, with our team’s recommendations for avoiding similar problems in the future.