SEATTLE PACIFIC
UNIVERSITY

DATA MINING PROJECT

ISM 6359

Data mining

Professor Ryan LaBrie

Ngoc Oanh Nguyen

Winter, 2023

1. State a business reason for selecting your tools (problem you would like to solve).

Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, which could result in severe complications including heart disease, vision loss, kidney disease, etc. As a result, it would lead to reduced quality of life and life expectancy. Moreover, diabetes is among the most prevalent chronic diseases in the United States. According to the 2022 National Diabetes Statistics Report, it is stated that more than 130 million adults are living with diabetes or prediabetes in the United States. Moreover, it is the most expensive chronic condition in the US and exerts a significant financial burden on the economy when $1 out of every $4 in US healthcare costs is spent on caring for people with diabetes, which totals up to $237 billion on direct medical costs and another $90 billion on reduced productivity. There is no cure for diabetes, hence early diagnosis toward lifestyle changes including strategies like losing weight, eating healthily, being active, and receiving medical treatments can effectively mitigate the harms of this disease leading to more efficient treatment in many patients.

Therefore, making predictive models for diabetes risk important tool for the public and public health officials. In addition, there are three types of diabetes, but the type II one is the most common form and its prevalence varies by age, education, income, location, race, and other social determinants of health.

I aim to try to build a model that predicts whether a patient does or does not have type II diabetes disease based on certain characteristics. This is a classification task so I will use the classification algorithms.

2. Document how/where you got your data (if it is publicly available, or internal for a work project).

The Diabetes Health Indicators Dataset is publicly available on Kaggle.com. Dataset Link

It is the dataset for the year 2015 collected by The Behavioral Risk Factor Surveillance System (BRFSS) which is a health-related telephone survey that is conducted annually by The Centers for Disease Control and Prevention (CDC).

There are three versions of the processed dataset on Kaggle and I choose to use the third one which is a clean dataset of 253,680 survey responses. Among them, the target variable Diabetes_binary has 2 classes: 0 is for no diabetes, and 1 is for prediabetes or diabetes. The other 21 feature variables are the important risk factors for diabetes as below.

Table 1: Feature Attributes Description

No.	Feature variables	Description	Values
1	HighBP	Who have been told they have high blood pressure a health professional	0=No 1=Yes
2	HighChol	Who have been told by a health professional that their blood cholesterol is high	0=No 1=Yes
3	CholCheck	Who have checked cholesterol within the past five years	0=No 1=Yes
4	BMI	Body Mass Index	Numeric
5	Smoker	Who has smoked at least 100 cigarettes in your entire life	0=No 1=Yes
6	Stroke	Who has ever had a stroke	0=No 1=Yes
7	HeartDiseaseorAttack	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	0=No 1=Yes
8	PhysActivity	Who does physical activity or exercise during the past 30 days	0=No

		other than their regular job	1=Yes
9	Fruits	Who consumes fruit 1 or more times per day	0=No 1=Yes
10	Veggies	Who consumes vegetables 1 or more times per day	0=No 1=Yes
11	HvyAlcoholConsump	Who are the heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)	0=No 1=Yes
12	AnyHealthcare	Who has any kind of health care coverage (health insurance, prepaid plans, or government plans)	0=No 1=Yes
13	NoDocbcCost	Who had a time in the past 12 months when you needed to see a doctor but could not because of cost	0=No 1=Yes
14	GenHlth	The rate you would say your health is in general	1-5
15	MentHlth	The number of days during the past 30 days your mental health is not good (stress, depression, and problems with emotions)	0-30
16	PhysHlth	The number of days during the past 30 days your physical health is not good (physical illness and injury)	0-30
17	DiffWalk	Who have serious difficulty walking or climbing stairs	0=No 1=Yes
18	Sex	The sex of respondent	0=Female 1=Male
19	Age	Fourteen-level age categories	1-14
20	Education	The highest grade or year of school that you completed	1-6
21	Income	Their annual household income from all sources	1-8

3. **Document how you used the tool. Many tools are super rich in features and you probably will not be exploring them all, explain the parts you did use**

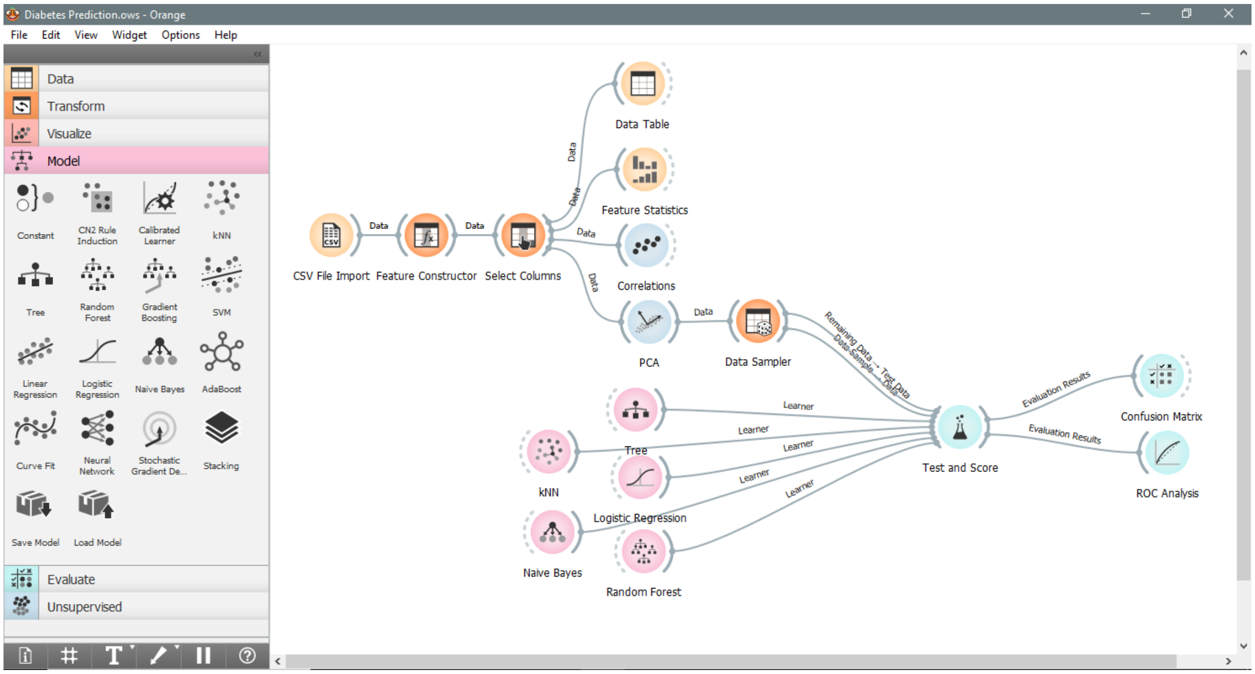
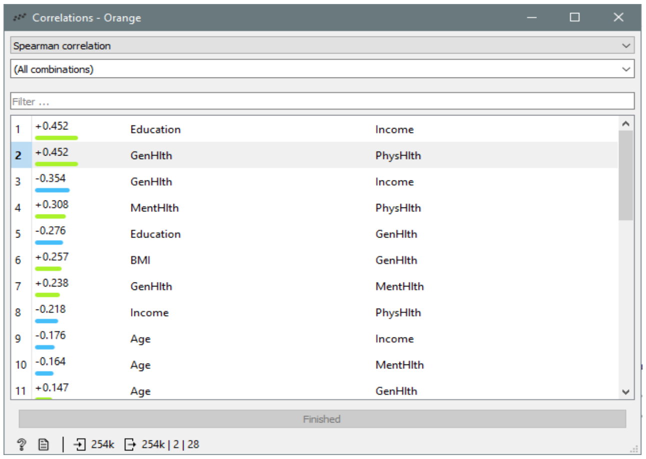


Figure 1: The Workflow

Firstly, I have to load the csv dataset file named "diabetes_binary_health_indicators_BRFSS2015.csv" into the Orange tool and create the ID attribute for this dataset. Then I use the Select Columns widget to set "Diabetes_binary" as the target attribute and "ID" as the meta one. After that, by using the Feature Statistics widget, I check that there is no missing value in each attribute and the unbalanced distribution of the dataset

I also check the correlations between attributes and spot that (Education; Income), (GenHlth; PhysHlth), and (MenHlth; PhysHlth) are highly positively correlated. There are also tight negative correlations in (GenHlth; Income) and (Education; GenHlth).


Correlations - Orange

Spearman correlation

(All combinations)

Filter ...

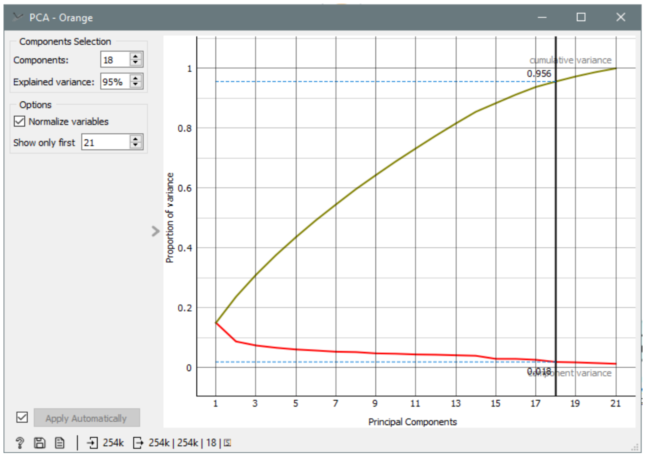
1	+0.452	Education	Income
2	+0.452	GenHlth	PhysHlth
3	-0.354	GenHlth	Income
4	+0.308	MentHlth	PhysHlth
5	-0.276	Education	GenHlth
6	+0.257	BMI	GenHlth
7	+0.238	GenHlth	MentHlth
8	-0.218	Income	PhysHlth
9	-0.176	Age	Income
10	-0.164	Age	MentHlth
11	+0.147	Age	GenHlth

Finished

254k 254k | 2 | 28

Figure 2: Correlations between attributes

The Principal Component Analysis (PCA) shows that the first 18 attributes account for more than 95% of the variance so I just use them for the model.


PCA - Orange

Components Selection

Components: 18

Explained variance: 95%

Options

☑ Normalize variables

Show only first 21

☑ Apply Automatically

254k 254k | 254k | 18 |

Figure 3: The Principal Component Analysis (PCA)

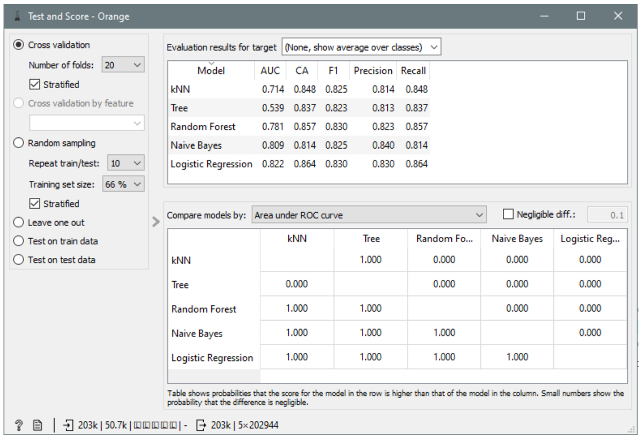
Next, I use the Data Sampler to deal with the unbalanced data and put them into "Test and Score".

4. **When you choose a data mining algorithm(s) for you mining model, tell us why you chose that one (or that category of algorithms)**

I need to classify whether one patient is diabetic or not, so I choose the typical classification algorithms including Decision tree, Random Forest, Naïve Bayes, Logistic Regression, and kNN. I also use the Cross-validation function with 20 folds inside the "Test and Score" widget for this model.

5. Given an explanation/analysis of the output (What did you learn or uncover)

After running, the "Test and Score" provides a table that helps to evaluate the best model which uses the Logistic Regression. It has the highest evaluation results with the Classification Accuracy of 86.4% and the AUC of 0.822.



Test and Score - Orange

● Cross validation
Number of folds: 20
☑ Stratified
○ Cross validation by feature
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.714	0.848	0.825	0.814	0.848
Tree	0.539	0.837	0.823	0.813	0.837
Random Forest	0.781	0.857	0.830	0.823	0.857
Naive Bayes	0.809	0.814	0.825	0.840	0.814
Logistic Regression	0.822	0.864	0.830	0.830	0.864

Compare models by: Area under ROC curve ☐ Negligible diff.: 0.1

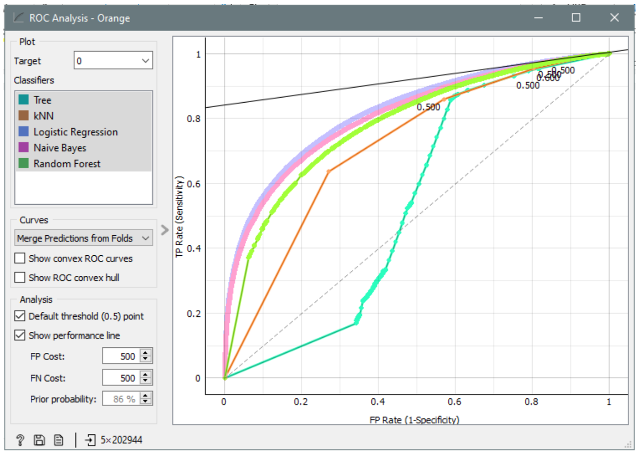
	kNN	Tree	Random Fo...	Naive Bayes	Logistic Reg...
kNN		1.000	0.000	0.000	0.000
Tree	0.000		0.000	0.000	0.000
Random Forest	1.000	1.000		0.000	0.000
Naive Bayes	1.000	1.000	1.000		0.000
Logistic Regression	1.000	1.000	1.000	1.000	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

203k | 50.7k | - 203k | 5×202944

Figure 4: Evaluation results

Moreover, its ROC (Receiver Operating Characteristic) Curves is the top one so it can distinguish more accurately if a patient has type II diabetes disease or not.



ROC Analysis - Orange

Plot
Target: 0
Classifiers
🟦 Tree
🟫 kNN
🟦 Logistic Regression
🟪 Naive Bayes
🟩 Random Forest

Curves
Merge Predictions from Folds
☐ Show convex ROC curves
☐ Show ROC convex hull

Analysis
☑ Default threshold (0.5) point
☑ Show performance line
FP Cost: 500
FN Cost: 500
Prior probability: 86 %

5×202944

Figure 5: ROC Analysis

6. Conclude with the 3 W's (What Went Well, What Did NOT go Well, What Would you do Differently Next Time).

- What Went Well

The data preparation did not take so much time for me since the dataset is a clean one and all the attributes are in numeric or binary forms, so I did not have to transform them. Moreover, there is no missing data and confusingly classified attributes. Everything is in good condition which facilitates me to focus on manipulating the dataset.

The Orange tool is quite similar to the RapidMiner so it did not take me a decent time to learn about the Orange tool in terms of different names and operators. In addition, it is also rich in features and visualizations, thus it is easy for me to adapt and explore the data.

- What Did NOT go Well

Finding a good dataset is extremely time-consuming. In the first place, I chose the Coffee Reviews dataset to make recommendations. However, that dataset is not cleaned with a lot of missing values in important attributes. Therefore, I had to switch to other datasets before deciding to deal with this Diabetes Prediction.

My computer crashed a lot of times due to the weak CPU even though the dataset is not really large (at 253,680 rows).

- What Would you do Differently Next Time

I want to spend more time exploring the data to understand each attribute and their distribution to the target one more comprehensively.

I will try to detect the outliers of the data to enhance the classification accuracy. Moreover, I would try to discretize the Income and Age attributes into specific ranges.

I also want to use a different tool such as KNIME or Python with a larger size of dataset broaden my knowledge in this field.