UNIVERSITY OF WOLLONGONG
FACULTY OF ENGINEERING AND INFORMATION SCIENCES

MACHINE LEARNING ASSIGNMENT 1:
PERFORMANCE OF SUPPORT VECTOR MACHINE AND LOGISTIC REGRESSION ON DIABETES DATASET

Student name: Ngoc Hai Nguyen

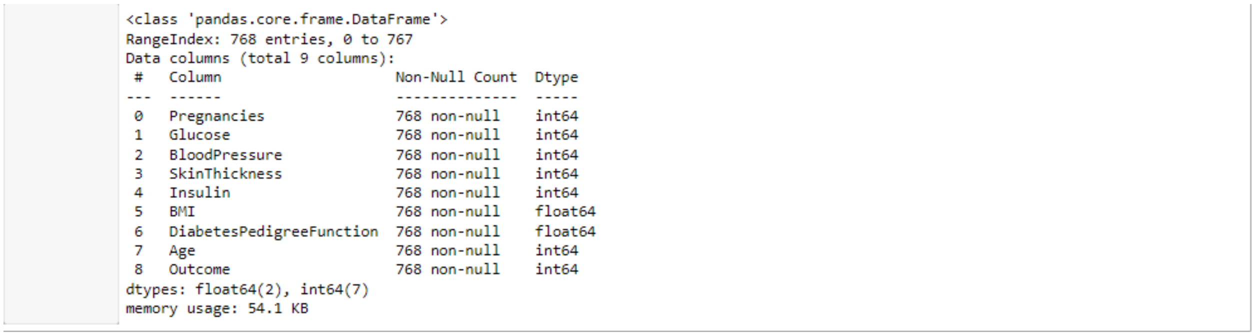
Student ID: 7548953

Wollongong, April 2022

I. INTRODUCTION

The Diabetes dataset contains the key features that are thought to influence a person’s diabetes. Eight columns feature are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, and all types of them are continuous data. The outcome column contains binary values and categorizes the data (768 observations) into 2 parts: 268 positives for diabetes are marked as 1 and 500 negatives for diabetes are labeled as 0.

- Checking missing value:

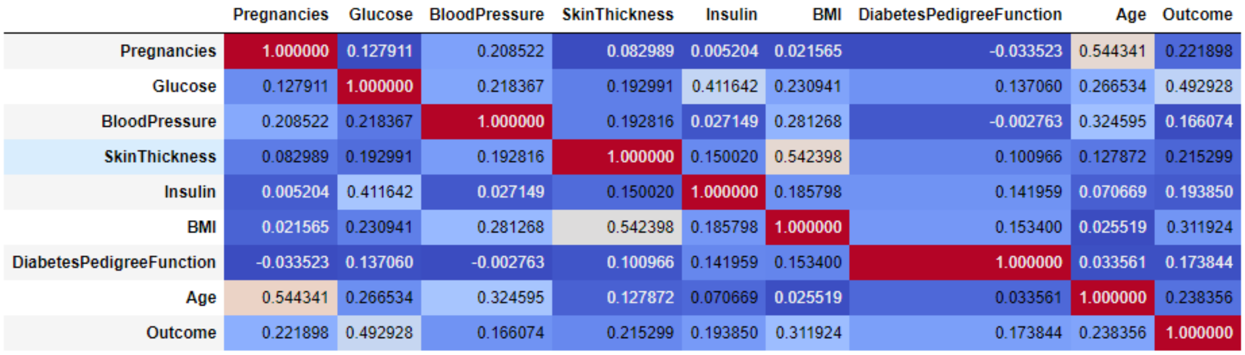


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 1.1: Missing value

From figure 1.1, the diabetes data does not have any missing value. However, features like Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI contain 0 values, but, these human indicators can not be zero, according to some medical science journals. Therefore, I assumed that these missing data had been replaced with the value 0, so I decided to convert these data from 0 to null, to statistic the central tendency of the data fields in order to replace that value in the missing data.

- Statistic summary:
  - Correlation matrix:



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.127911	0.208522	0.082989	0.005204	0.021565	-0.033523	0.544341	0.221898
Glucose	0.127911	1.000000	0.218367	0.192991	0.411642	0.230941	0.137060	0.266534	0.492928
BloodPressure	0.208522	0.218367	1.000000	0.192816	0.027149	0.281268	-0.002763	0.324595	0.166074
SkinThickness	0.082989	0.192991	0.192816	1.000000	0.150020	0.542398	0.100966	0.127872	0.215299
Insulin	0.005204	0.411642	0.027149	0.150020	1.000000	0.185798	0.141959	0.070669	0.193850
BMI	0.021565	0.230941	0.281268	0.542398	0.185798	1.000000	0.153400	0.025519	0.311924
DiabetesPedigreeFunction	-0.033523	0.137060	-0.002763	0.100966	0.141959	0.153400	1.000000	0.033561	0.173844
Age	0.544341	0.266534	0.324595	0.127872	0.070669	0.025519	0.033561	1.000000	0.238356
Outcome	0.221898	0.492928	0.166074	0.215299	0.193850	0.311924	0.173844	0.238356	1.000000

➔ Pregnancies and Age, BMI and SkinThickness are positively correlated with $\rho > 0.5$. Gluscose has the highest correlation to the outcome with approximately 0,49

Feature	Mean	Standard Deviation
Pregnancies	3.85	3.37
Glucose	121.69	30.44
BloodPressure	72.41	12.10
SkinThickness	29.15	8.79

Insulin	130.93	88.70
BMI	32.46	6.88
DiabetesPedigreeFunction	0.47	0.33
Age	33.24	11.76

Table 1.1: Central Tendency of features

From table 1.1 and figure 2.1, we can see that the data of Insulin distribute far the most from the mean, meanwhile, Skin Thickness and Blood Pressure data are the best distribution in all columns when it is closed to the shape of normal distribution.

II. DATA PREPARATION

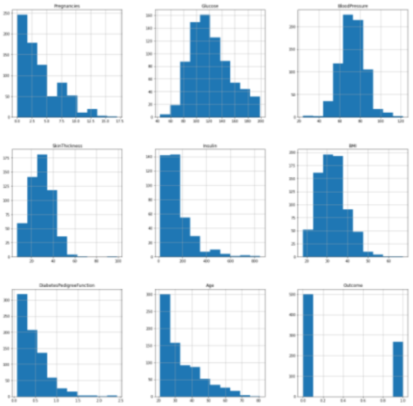


Figure 2.1: Histogram of each feature in the diabetes dataset

- Replace null value

Considering 5 fields that have the missing value (Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI), Glucose, Blood Pressure, SkinThickness, BMI are the feature that is in the shape of normal distribution, so I decided to replace the null value with mean, and insulin will be replaced by median because of its skew distribution.

- Rescale data

All of our features vary from a different wide range, and machine learning models need the attribute to have the same scale. Often the feature will be scaled into the range from 0 to 1, and I will arrange all of the diabetes feature values between 0 and 1.

- Train test split

Train test split is a part of a machine learning project, that allows us to assess our model on the test data. In this case, I set aside 20% data for the test set and use 80% data for the training phase.

III. CLASSIFIERS

1. Logistic Regression

Logistic Regression is a classification model in supervised learning, especially it is used the most for binary output – which is suited to the diabetes dataset.

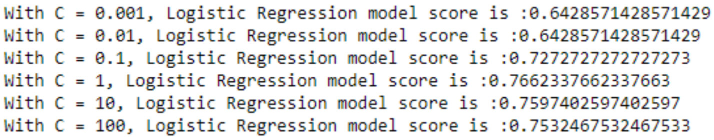
$$\log \frac{p(x)}{1 - p(x)} = \alpha_0 + \alpha . x$$

The idea behind this logistic regression is just like Linear Regression, however, it consists of a logarithm of the odds ratio and passed through a sigmoid function:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

It is hard to tune a logistic regression model. In this experimentation, I will change the hyperparameter C to record the correlation between C and model accuracy. C is regularization parameters, which can provide the penalty strength, and the result is shown below:

$$C = \frac{1}{\gamma}$$



```
With C = 0.001, Logistic Regression model score is :0.6428571428571429
With C = 0.01, Logistic Regression model score is :0.6428571428571429
With C = 0.1, Logistic Regression model score is :0.7272727272727273
With C = 1, Logistic Regression model score is :0.7662337662337663
With C = 10, Logistic Regression model score is :0.7597402597402597
With C = 100, Logistic Regression model score is :0.7532467532467533
```

As we can see from the results above, a small change in C will impact significantly the accuracy of the model. This is because small C values will raise the regularization strength which infers the creation of basic models that tend to underfit the data. By using greater C values, the model can increase its complexity and alter superior to the data.

2. Support Vector Machine

Support Vector Machine is a supervised machine learning model that will find a hyperplane in multi-dimensional vector space that can separate 2 labeled datasets. GridSearchCV is a library that helps us change the hyperparameter of the model in order to assess the accuracy of the model with different hyperparameters.

- The Radial Basis function kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$$

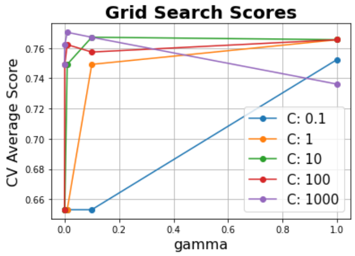
C and Gamma are parameters for a nonlinear support vector machine (SVM) with a Gaussian radial Basis Function kernel. C is the parameter for the soft margin cost function, which controls the effect of each support-vector. This process involves penalizing transactions for stability.

- C=∞ - No deviation, synonymous with hard margin
- Large C - Small deviation allowed, small margin obtained.
- Small C - Allow large deviations, and obtain large margins.

Gamma is a hyperparameter used with non-linear SVMs. The gamma parameter of the RBF controls the influence distance of a single training point.

A low value of Gamma indicates a large similarity radius which results in many points being grouped. For high values of gamma, the points need to be very close to each other to be considered in the same group (or class).

Experimenting with these 2 parameters: C changes from 0.1 to 1000, Gamma changes from 0.0001 to 1, and we have the result as below:



After experimenting with different C and Gamma, we get the best parameters is C = 1000 and gamma =0.01, with the accuracy obtained at approximately 76.5%.

IV. EVALUATION

In this section, I will use the following metrics to assess the model's performance

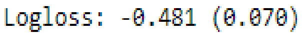
1. Classification accuracy

The number of correct predictions made as a percentage of all predictions made is known as classification accuracy. This is the most commonly used evaluation metric for classification tasks.

Model	Accuracy
Logistic Regression	77.1%
Support Vector Machine	76.4%

2. Logarithmic loss.

Log loss (logistic loss or cross-entropy loss) is the loss function used in logistic regression, and also is the major classification metric based on probability. Log loss is equal to -1* the log of the likelihood function.



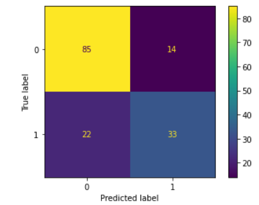
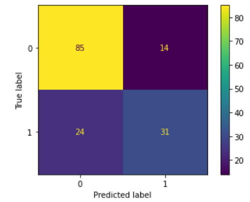
```
Logloss: -0.481 (0.070)
```

3. AUC (Area under the ROC curve)

Classifier	The area under the ROC curve
Logistic Regression	0.831
SVM (kernel: rbf)	0.824

The ROC curve is an evaluation metric commonly used for binary-classifier that has 2 parameters: True Positive Rate and False Negative Rate. The Area under the ROC curve can measure the ability of models in classifying the classes. The larger the area is, the better performance the model provides. Applying this method to this particular diabetes dataset, Logistic Regression performs better than SVM with 0.831 unit square.

4. Confusion Matrix

Logistic Regression	
Support Vector Machine	

Although there is no columns name in the confusion matrix, we can easily recognize that majority of the predictions fall on the main diagonal (which is true prediction).

V. CONCLUSIONS

In conclusion, both Logistic Regression and SVM perform quite well in classifying diabetes on the given dataset (obtaining 77,1% and 76% accuracy respectively). SVM is on the basis of geometrical properties while logistic regression is based on statistical approaches. SVM can perform well on non-linear classification by using kernel trick, in this particular case, is radial basis function, to map the data into separable feature spaces. Meanwhile, Logistic Regression uses maximum likelihood, logit function, and sigmoid function – the statistical method to classify data. Both of the models can work well on both numerical data and categorical data.

- Significant trend:

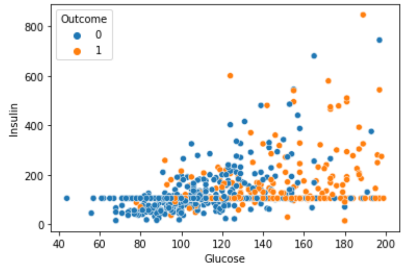


Figure 5.1: Impact of Insulin and Glucose on the diabetes disease

As we can see from figure 5.1, the patients who have 0-130 glucose are less likely to have diabetes, and those who have glucose more than 130 are more likely to be predicted as diabetes.

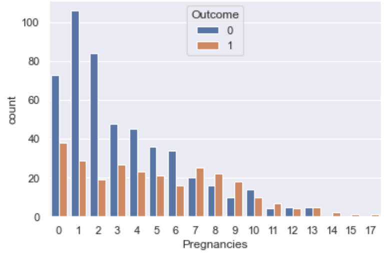


Figure 5.2: Number of Pregnancies and Diabetes

Figure 5.2 shows that patients who have a number of pregnancies less than 7 are less likely to be diabetes and vice versa, who has this index over 7 are more likely to be diagnosed with diabetes.