Cancer prediction for type-2 diabetes using machine learning

V Malathy1, N Shilpa2, M Anand3, N Swathi4 and R Elavarasi5

1Department of Electronics and Communication Engineering, SR University, Warangal
2Department of Electronics and Communication Engineering, S R Engineering College, Warangal
3Department of Electronics and Communication Engineering, DR M.G.R Educational and Research Institute University, Chennai
4Department of Electronics and Communication Engineering, Sumathi Reddy Institute of Technology for Women, Warangal
5Department of Electronics and Communication Engineering, Priyadarshini Engineering College, Vaniyamadi, Vellore
1malathy_vanniappan@srecwarangal.ac.in

Abstract. Type-2 diabetic patients may easily get suffered from cancer. The presented work in this paper predicts whether the diabetic patient has got cancer problem or not. If cancer is identified in the patient, the type of cancer such as breast cancer, colon cancer and liver cancer is classified. The parameters associated with cancer and diabetes are recognized and collected. Then the collected dataset is preprocessed. During preprocessing the null values in the dataset are replaced with mean values. The dataset is now ready for training. Cancer disease prediction is done by breast cancer, colon cancer and liver cancer. To classify the cancer type, random forest algorithm is used.

1. Introduction

1.1. Mellitus
The first type of diabetes disease Type-1 disease is a long term disorder and it is insulin-independent diabetes. When the insulin supply is less, the level of blood sugar that is glucose level in the body rises and it may cause diabetes mellitus type 2. Pancreas secretes insulin hormone in the human body. Pancreas permits the human body to consume glucose for energy. If there is more sugar level in the body than required, the sugar is stored in the liver with the help of insulin. Later excess sugar is released when the blood sugar level comes to low level or when body requires sugar when hungry or during exercise. Thus insulin helps to balance the blood sugar level. It keeps the sugar in a normal range. Type 2 diabetic patient body may not exhaust the insulin properly. This is called insulin resistance. So, pancreas secretes extra insulin to have a balance. If it continues the pancreas will not be able to secrete insulin and keep the normal sugar level. Type-2 diabetic patients may have high blood glucose level.

1.2. Primary Symptoms
The disease is due to obesity and lack of exercise.
Figure 1 shows the healthy body and Type 2 diabetic patient body.
1.3. Breast Cancer
Any change in hormone level may induce the variation in blood sugar level. The response of the body to insulin will be affected by the hormone estrogen and progesterone. General risk factors related to the breast cancer are age, hereditary, obesity and alcohol use.

1.4. Colon Cancer
People with diabetes will suffer from gastrointestinal problems. Diabetes is related to high insulin level. The risk factors associated with colon cancer are old age, hereditary, obesity, smoking and diet.

1.5. Liver Cancer
Obesity, high blood pressure, increase insulin and NAFLD (nonalcoholic fatty liver disease) cause liver cancer. This may build up excess fat in the liver.

The advances in biomedical technology result in low cost data generation. Thus the technology of biology is introduced in big data area [1,2]. The machine learning is to develop the software system to be flexible so that it can learn from the history [3]. The machine learning is described by Mitchel [4]. The symptom of diabetes is weight loss. Diagnosis of disease is related to the blood glucose level (the level of glucose during fasting is 7 mmol/L) [5].

In paper [6], the authors narrated about anticancer. In paper [7] to paper [11] the authors explained the detection of diseases due to various reasons and also the diseased regions separation.

2. System architecture
The details of Type 2 diabetic patients were collected and stored in .csv file format. The format contains 200 entries. 15 features were taken as symptoms. The dataset (80% training and 20% testing) was preprocessed. Using Navie Bayes algorithm the presence of cancer is identified. If cancer exists, using the random forest algorithm the type of cancer is classified. Figure 2 shows the methodology.

Figure 1. Healthy body and Type 2 diabetic patient body
2.1. UML diagram
The persons involved in this work are scientists and patients. The primary humans are the patients and the secondary humans are data scientists. Figure 3 shows the humans and their duties and the figure is called as USECASE diagram.

3. Prediction
3.1. Activity diagram
After the condition of data, the raw data has been transformed into processed data. Then processed input data has been divided to have train data 80% and test data 20%. During training phase the diabetic patient is checked for cancer condition. If cancer condition is false, it shows that the patient is not affected by the cancer disease. If true, the type of cancer is to be confirmed.
4. Results and discussion

4.1. Preprocessing
The desired libraries in python are to be imported. Type2 diabetic dataset should be in .csv file format. It must also be imported. The null values in the dataset are identified to replace with the mean and median values. Thus an approximate value is obtained. The command LabelEncoder class in the algorithm is used to convert categorical data YES/NO in the training set to numerical values 0/1. From the processed dataset, 80% is considered as training data and 20% is considered as testing data.

4.1.1. Coding
# import the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as pyplot
import seaborn as sns
import warnings
from sklearn.preprocessing import Imputer
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
#from sklearn.cross_validation import train_test_split
from sklearn.model_selection import train_test_split
warnings.filterwarnings('ignore')

# import the dataset
dataset = pd.read_csv('cancer_d.csv')
dataset.shape
print("dataset shape : ")
print(dataset.shape)
dataset.index
print("dataset index : ")
print(dataset.index)

# check for the missing values
print("checking the missing values")
dataset.isnull().sum()
print(dataset.isnull().sum())
print(dataset.info())

# filling of null values
dataset["BP"] = np.where(dataset["BP"] == "120-80", 0, np.where(dataset["BP"] == "140-90", 1, 2))
dataset = dataset[['BMI', 'TRIGLYCERIDES', 'HDL', 'LDL', 'INSULIN', 'BP', 'AGE', 'SMOKING', 'POLYPS HISTORY', 'POST MENSURAL STATUS', 'FAMILY HISTORY OF BREAST CANCER', 'NO.OF PREGNANCY', 'AGE OF 1ST DELIVERY', 'MENSURATION AGE', 'ALCOHOL']].fillna(dataset.median())

# splitting dependent and independent variables
X = dataset.iloc[:, :-1].values
print("array of independent variable")
print(X)
y = dataset.iloc[:, -1].values
print("array of dependent variable")
print(y)

# filling missing values in csv file
print("fillinf missing values")
dataset.to_csv('cancer.csv')
print("data written to csv file")

# filling categorial values
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)
print("filling categorial values")
print(y)

# training data set and test dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
print(X_train, y_train)
print(X_test, y_test)

4.1.2. Output

Figure 5 shows the dataset before preprocessing after collecting the dataset.

Figure 6 shows the dataset after preprocessing. The collected dataset is preprocessed before any operation.
Figure 7 shows the checking of missing values. The dataset is checked for the missing values.

Figure 7. Checking missing values

Figure 8 shows the independent and dependent variable.

Figure 8. Independent and dependent variable
Figure 9 shows the numerical data.

```
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"
"LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME" "LIVER_CANCER_OUTCOME"

No No No No No No No No No No No No No No No No No No No No
No No No No No No No No No No No No No No No No No No No No
No No No No No No No No No No No No No No No No No No No No
No No No No No No No No No No No No No No No No No No No No
Colorectal Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
Colon Colon Colon Colon Colon Colon Colon Colon Colon
No No No No No No No No No No No No No No No No No No No No
No No No No No No No No No No No No No No No No No No No No
No No No No No No No No No No No No No No No No No No No No
No No No No No No No No No No No No No No No No No No No No
"Breast" "Breast" "Breast" "Breast" "Breast" "Breast" "Breast" "Breast"
"Breast" "Breast" "Breast" "Breast" "Breast" "Breast" "Breast" "Breast"
"Breast" "Breast" "Breast" "Breast" "Breast" "Breast" "Breast" "Breast"

filling missing values  
data written to csv file

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
```

**Figure 9.** Categorical to numerical data

Figure 10 shows the dataset for testing the disease and training.

```
[[2.257000e+01, 5.400000e+01, 7.572000e+01, ... , 0.000000e+00, 9.100000e-03, 0.000000e+00],
 ...
[3.726000e+01, 4.935000e+02, 3.767000e+01, ... , 1.000000e+01, 6.062000e-01, 1.000000e+00],
[3.891000e+01, 6.631700e+02, 3.215000e+01, ... , 0.000000e+00, 4.775500e-01, 1.000000e+00],
[3.792000e+01, 3.983000e+02, 3.820000e+01, ... , 0.000000e+00, 7.331100e-01, 1.000000e+00]],
array([[2.257000e+01, 5.400000e+01, 7.572000e+01, ... , 0.000000e+00, 9.100000e-03, 0.000000e+00],
 ...
[2.257000e+01, 5.400000e+01, 7.572000e+01, ... , 0.000000e+00, 9.100000e-03, 0.000000e+00],
[3.592000e+01, 3.938000e+02, 3.813000e+01, 1.948700e+02, 1.918500e+02, 0.000000e+00, 5.100000e-01, 0.000000e+00, 1.000000e+00, 1.000000e+00, 1.000000e+00, 1.000000e+00],
[4.432000e+01, 5.394000e+02, 3.164000e+01, 1.793800e+02, 1.909100e+02, 0.000000e+00, 5.000000e-01, 0.000000e+00, 1.000000e+00, 1.000000e+00, 1.000000e+00],
[4.521000e+01, 3.970000e+02, 3.838000e+01, 1.942200e+02, 1.915000e+02, 0.000000e+00, 5.000000e-01, 0.000000e+00, 1.000000e+00, 1.000000e+00, 1.000000e+00],
[5.180000e+01, 3.160900e+02, 4.620000e+01, 1.860000e+02, 1.738000e+02, 0.000000e+00, 8.900000e+00, 0.000000e+00, 0.000000e+00, 9.000000e+00, 3.000000e+00, 8.400000e+00, 4.700000e-03, 0.000000e+00],
```

**Figure 10.** Test and train data
Testing and training data are given in Figure 11.

![Figure 11. Test and train data](image)

4.2. Prediction

This step tells in advance whether the diabetic patient under test is prone to cancer. Naive Bayes Theorem is made use of for predicting the cancer. There are several Naive Bayes Algorithms. One among them is GaussianNB algorithm. It is used because the values are continuous. After detecting the cancer, the type of cancer is recognized by using Random forest algorithm.

The steps for prediction:
1. GaussianNB is first initialized.
2. The fit function (X, Y) suits to the Naive Bayes based on X,Y values. It is used to train the dataset where the algorithm learns the values to prediction.
3. The prediction (X_test) predicts the outcome for the given input test data.
4. And finally the accuracy of the algorithm is tested.
5. As the dataset size increases, the accuracy also increases.

4.2.1 Coding

```python
#prediction of cancer
import pandas as pd
import numpy as np
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB

data = pd.read_csv("cancer_final.csv")
X=data.iloc[:,:-2].values
```
y=data.iloc[:,-2].values
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=0)
gnb = GaussianNB()
gnb.fit(X_train, y_train)
y_pred = gnb.predict(X_test)
print(y_pred)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

4.2.2. Output
Figure 12 shows the output for the prediction of cancer

Figure 12. Prediction of cancer
Figure 13 shows the output screen of WEKA tool.

![WEKA tool output](image)

Figure 13. WEKA tool

5. Conclusion

From this paper it is understood that type 2 diabetic patients or patients with disease history are likely to be diagnosed with any type of cancer when compared to others without such disease like diabetes. The work in this paper predicts whether a type-2 diabetic patient may get affected by cancer or not. If the patient is detected with cancer, the next step is to find out the type of cancer namely, breast cancer, colon cancer and liver cancer.

References

[1] Marx V 2013 Biology: The big challenges of big data Nature Jun 13 vol 498(7453) 255-60
[2] Mattmann C A 2013 Computing: a vision for data science Nature. 493(7433) 473–5
[3] Wilson R A and Keil F C 1999 The MIT encyclopaedia of the cognitive sciences MIT Press.
[4] Mitchell T 1997 Machine learning McGraw Hill 0-07-042807-7
[5] Cox E M and Elelman D 2009 Test for screening and diagnosis of type 2 diabetes Clin Diabetes vol 4(27) 132–8
[6] Ramaiah K Srishailam K Laxma Reddy K Reddy B V and Ramana Rao G 2019 Synthesis, crystal and molecular structure, and characterization of 2-((2-aminopyridin-3-yl)methylene)-N-ethyl hydrazine carbothio amide using spectroscopic (1 H and 13 C NMR, FT-IR, FT-Raman, UV–Vis) and DFT methods
and evaluation of its anticancer activity *Journal of Molecular Structure*. 1184 405-417

[7] Harshavardhan A Babu S and Venugopal T 2017 Brain tumor segmentation methods-a survey *Journal of Advanced Research in Dynamical and Control Systems* 9(11) 240-245

[8] Malathy V and Kamali S M 2019 Integrating Fuzzy C-Means Algorithm with Level Set Methods for Segmentation of Injured Human spinal Cord in Computed Tomography Images *Journal of Advanced Research in Dynamical and Control Systems* 11(2) 794-801

[9] Malathy V and Kamali S M 2019 Brain Tumor Segmentation from Brain Magnetic Resonance Images using Clustering Algorithm in International Journal of Innovative Technology and Exploring Engineering 8(8) 625-629

[10] Malathy V Shilpa N and Anand M 2019 Detection of brain tumor using modified centroid k-means clustering algorithm *Journal of Advanced Research in Dynamic Control Systems*. 11(7) 749-754

[11] Umadevi G Malathy V and Anand M 2019 Diagnosis of Diabetes from Tongue Image Using Versatile Tooth-Marked Region Classification *Test Engineering & Management* 81 5953–5965