CLASSIFICATION OF DIABETES DISEASE USING
NAIVE BAYES
Case Study :
SITI KHADIJAH HOSPITAL

1Ida Lailatul Qurnia, 2Eko Prasetyo, 3Rifki Fahrial Zainal
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Bhayangkara
Jl. A. Yani 114 Wonocolo Surabaya 60231
Email:1IdaLailatulQurnia@ubhara.ac.id, 2eko@ubhara.ac.id, 3rifki@ubhara.ac.id

ABSTRACT

Less knowledge about symptoms and how to treat the disease of diabetes mellitus as well as a number
of specialist diabetes mellitus which is still limited is one of the causes of the growing number of people
affected by the disease. Diabetes disease classification system development aims to predict the type of
diabetes patient or user who already suffer from diabetes mellitus. Therefore this system is made to
diagnose the type of diabetes through laboratory test results, namely in the form of gender, age,
disease history, family history, systolic, diastolic tensi tensi, temperature, pulse, blood sugar, fasting
blood sugar JPP and Random blood sugar. That is by using the method of naive bayes as a method to
process data on the patient’s diagnosis. Test results of this system indicates that the system is able to
predict the type of diabetes in patients, from the amount of data as much as 200 patient data, with an
output that is the form of Diabetes Without Complications, Diabetes Type II and Normal but obtained
the lowest accuracy rating of 39% and the value of the highest accuracy of 80%.

Keyword: Classification, Naive Bayes, Diabetes Mellitus, Random Blood Sugar, A History Of The
Disease In The Past

1. INTRODUCTION

Diabetes is a disease in which the body cannot produce insulin (blood sugar balance hormones) or the
insulin produced is insufficient insulin or don't work well. Therefore will cause increased blood sugars
while review. Therefore, this study aims to help patients in order to know the problems of early
diagnosis of the disease diabetes mellitus, so patients can find out his condition is being affected by the
disease of diabetes, that is immediately checked himself into the hospital to get the handling medically.
The methods used for classification of diabetes is naive bayes method that is one of the algorithms
found in the technique of classification. Naive Bayes classification is by the method of probability and
statistics. This method aims to conduct classification data on a particular class. Based on fact, the naive
bayes algorithm will be applied in this study to determine the type of diabetes in patients who are
already affected by diabetes namely diabetic without complications or type II diabetes (insulin), using a
predefined data include gender, i.e. the value of blood sugar (limited to the value of blood sugar blood
sugar values, JPP fasting and random blood sugar value), age, systolic tensi tensi, diastolic, temperature, pulse, a history of the disease in the past and family history as input variables of the
system.

2. RESEARCH METHODS

This research uses Bayes algorithm is naive. Naive Bayes algorithm is one of the algorithms found in
the technique of classification. Naive Bayes classification is by the method of probability and
statistics. Naive Bayes algorithm is a classification algorithm techniques that are easy to apply
and quick process
3. ANALYSIS AND SYSTEM DESIGN

As for the several stages in conducting classification with naive bayes algorithm diabetes i.e. enter data for calculating each training class. In table 4.1 there are eleven feature to be diagnosed to know types of diabetes in patients, namely gender, age, disease history, family history, systolic, diastolic tensi, temperature, pulse, blood sugar, fasting blood sugar JPP and Random blood sugar. And there are three classes of output that would result in mind i.e. Diabetes Without Complications, Diabetes Type II (insulin) and Normal. In this study, the patients against the predictions do not yet know the diabetes class has namely criteria gender male, age 30 years, a history of the disease in the past instead of DM, no family history, tensi tensi 100 systolic, diastolic 70, temperature of 36.5 ° C, 67/min pulse, blood sugar blood sugar 120 jpp, fasting and random blood sugar is 80 to 90, with training data data 15 hospital patients Siti Khadijah.

1) patients at the hospital Siti Khadijah

<table>
<thead>
<tr>
<th>no</th>
<th>name</th>
<th>gender</th>
<th>age</th>
<th>disease history</th>
<th>family history</th>
<th>systolic</th>
<th>diastolic</th>
<th>temperature</th>
<th>pulse</th>
<th>blood sugar</th>
<th>fasting blood sugar</th>
<th>random blood sugar</th>
<th>conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Siti</td>
<td>M</td>
<td>30</td>
<td>Yes</td>
<td>Yes</td>
<td>120</td>
<td>80</td>
<td>37.5</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Without Complications</td>
</tr>
<tr>
<td>2</td>
<td>Mira</td>
<td>F</td>
<td>30</td>
<td>Yes</td>
<td>No</td>
<td>140</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>3</td>
<td>Angga</td>
<td>L</td>
<td>30</td>
<td>Yes</td>
<td>No</td>
<td>130</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>4</td>
<td>Eka</td>
<td>F</td>
<td>30</td>
<td>Yes</td>
<td>No</td>
<td>110</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>5</td>
<td>Reza</td>
<td>L</td>
<td>30</td>
<td>Yes</td>
<td>Yes</td>
<td>110</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>6</td>
<td>Lala</td>
<td>F</td>
<td>30</td>
<td>Yes</td>
<td>Yes</td>
<td>110</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>7</td>
<td>Siti</td>
<td>F</td>
<td>30</td>
<td>Yes</td>
<td>Yes</td>
<td>110</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>8</td>
<td>Mira</td>
<td>F</td>
<td>30</td>
<td>Yes</td>
<td>Yes</td>
<td>110</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>9</td>
<td>Angga</td>
<td>L</td>
<td>30</td>
<td>Yes</td>
<td>Yes</td>
<td>110</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
<tr>
<td>10</td>
<td>Eka</td>
<td>F</td>
<td>30</td>
<td>Yes</td>
<td>Yes</td>
<td>110</td>
<td>70</td>
<td>37.2</td>
<td>65</td>
<td>120</td>
<td>80</td>
<td>120</td>
<td>Diabetes Type II</td>
</tr>
</tbody>
</table>

1) Perform calculations of the number of each class
P (H) = (Number Of Each Class)/(Number Of Overall Grade)
P (Diabetes Without Complications) = 0.3333
P (Type II Diabetes) = 0.3333
P (Normal) = 0.3333

2) Perform calculations and probability for each feature class
Next calculate the number of features and the probability for each class, for only kategorikal data calculated based on how much the same data on the features in one class and then divided by the number of classes as for numerical data needed for the calculation of the mean average value – knowing the median, probability calculation variants and calculation features.

Formula Gaussian distribution:

\[ P(X_i = x_j | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right) \]

Gaussian distribution, which is used to calculate the probability of numerical data
Next calculate the initial probability that is the multiplication of the values of the probabilities of each feature in each class, whereas to calculate the probability of the end that is the calculation of the probability multiplied by the beginning of class.

The latter compares the results from each class of probability that is looking for the greatest value among the class of Diabetes Without Complications, Diabetes Type II class or classes Normal, due to the value found in the largest class of Diabetes Without Complications so the output is "Diabetes Without Complications".

4. RESULTS AND DISCUSSION

System testing is testing in entering data into the form – form that has been provided. At this stage of testing is done by randomly:

a. testing with training data and test data.
b. testing with training data 45 and test data45.
c. testing with training data 90 and test data 90.
d. testing with 150 training data and test data 60

Based on the test results of 60 test data obtained the results that there are 24 data corresponding to the actual class.

Table 2. The Probability Of Each Feature

<table>
<thead>
<tr>
<th>No.</th>
<th>class</th>
<th>gender</th>
<th>age</th>
<th>History of Disease</th>
<th>Family History</th>
<th>Tensi/Systol</th>
<th>Tempe-rature/ C</th>
<th>Sa02</th>
<th>Satc</th>
<th>Results Laboratory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DWC</td>
<td>1</td>
<td>0.043</td>
<td>0.2</td>
<td>1</td>
<td>0.0071</td>
<td>0.0178</td>
<td>0.434</td>
<td>0.0141</td>
<td>0.0008</td>
</tr>
<tr>
<td>2</td>
<td>DT II</td>
<td>0.6</td>
<td>0.0101</td>
<td>0</td>
<td>1</td>
<td>0.0052</td>
<td>0.019</td>
<td>0.405</td>
<td>0.00009</td>
<td>0.003</td>
</tr>
<tr>
<td>3</td>
<td>Normal</td>
<td>0.4</td>
<td>0.000000000000</td>
<td>0.4</td>
<td>1</td>
<td>0.0011</td>
<td>0.0016</td>
<td>0.0049</td>
<td>0.0220</td>
<td>0.00102</td>
</tr>
</tbody>
</table>

Table 3. Results of Probability beginning and end

<table>
<thead>
<tr>
<th>No.</th>
<th>Class</th>
<th>early Probability</th>
<th>Final Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diabetes Without Complications</td>
<td>1,99766E-18</td>
<td>6,65888E-19</td>
</tr>
<tr>
<td>2</td>
<td>Diabetes Type II</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Normal</td>
<td>1,06103-21</td>
<td>3,53677E-22</td>
</tr>
</tbody>
</table>

Table 4. The Results Of The Testing System

<table>
<thead>
<tr>
<th>testing</th>
<th>Number Of correct predictions</th>
<th>Number Of incorrect predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DWC</td>
<td>DT II</td>
</tr>
<tr>
<td>Test I</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Test II</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Test III</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Test IV</td>
<td>16</td>
<td>8</td>
</tr>
</tbody>
</table>
Testing I:
From training data and 15, test data obtained as a result of a correct prediction value of as many as 12 of the overall data i.e. 5 data from the class of Diabetes Without Complications, 5 data from class of Diabetes Type II and 2 data from Normal classes. From the calculations it can be concluded that the system has an accuracy of 80% while the rate of error of 20%.

Testing II:
From training data and 45, test data obtained as a result of a correct prediction value of as much as 22 of the total data, namely data from the 13th class of Diabetes Without Complications, 9 data from class of Diabetes Type II and 0 data from Normal classes. From the calculations it can be concluded that the system has an accuracy of 49% and 51% error rate.

Testing III:
From training data and the 90 test data obtained as a result of a correct prediction value of as much as 35 of the overall data i.e. 15 data from classes of Diabetes Without Complications, 20 data from class of Diabetes Type II and 0 data from Normal classes. From the calculations it can be concluded that the system has an accuracy of 39% and the rate of error of 61%.

Testing IV:
From training data and 150 test data obtained as a result of a correct prediction value of as much as 24 of the overall data i.e. 16 data from classes of Diabetes Without Complications, 8 data from classroom Diabetes Type II and 0 data from Normal classes. From the calculations it can be concluded that the system has an accuracy of 40% and the rate of errors by 60%.

For performance assessment process is carried out by the concept of precision that is the metric for measuring system performance in obtaining relevant data and metrics to measure i.e. recall system performance in obtaining relevant data unreadable (E. Prasetyo, 2014).

Table 5. Performance Appraisal System

<table>
<thead>
<tr>
<th></th>
<th>Performance</th>
<th>DWC</th>
<th>DT II</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing I</td>
<td>Precision</td>
<td>1</td>
<td>0.625</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>1</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>Testing II</td>
<td>Precision</td>
<td>0.52</td>
<td>0.45</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.8667</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>Testing III</td>
<td>Precision</td>
<td>0.375</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.5</td>
<td>0.6667</td>
<td>0</td>
</tr>
<tr>
<td>Testing IV</td>
<td>Precision</td>
<td>0.3333</td>
<td>0.6667</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.8</td>
<td>0.4</td>
<td>0</td>
</tr>
</tbody>
</table>

Process calculation of precision do with TP (True Positive) divided by the TP (True Positive) plus FP (False Positive). Whereas the calculation of the recall is done by means of TP (True Positive) divided by the TP (True Positive) plus FN (False Negative).

4. CONCLUSION
Of such research gives the conclusion that the system can help to classify the type of diabetes based on the results of the laboratory examinations on patients affected by diabetes mellitus. That is by using the method of naive bayes classification, but in a working model built have insufficient performance when dujikan with larger training data, because of...
some performance testing shows the lowest accuracy of 39% and the highest accuracy of 80%, then need to add the features – features more in order to get better results.

REFERENCES