Maksymilian Romańczuk

nr albumu GD41582

**Algorytm Drzewa decyzyjnego klasyfikacyjnego**

**Przy użyciu języka Python i biblioteki sklearn**

Gdańsk 14.12.19

**Wstęp:**

**Cel:**

Sprawdzenie czy dany pacjent ma cukrzyce czy nie.

**Technologie:**

**Drzewo decyzyjne**- graficzna metoda wspomagania procesu decyzyjnego, stosowana w uczeniu maszynowym do pozyskiwania wiedzy na podstawie przykładów.

**Scikit-learn**(znane także, jako sklearn) - jest top darmowa biblioteka, którą można wykorzystać przy uczeniu maszynowym stworzona dla języka Python, która pozwala na stworzeniu algorytmów klasyfikacji, regresji, i analizy skupień. Jest tak zaprojektowana żeby interpretery Pythona mogły używać bibliotek numerycznych takich jak NumPy czy SciPy mogły z nim działać. (W moim wypadku wykorzystanie NumPy).

**Pandas**- biblioteka stworzona dla języka Python, aby analizować i manipulować danymi. Oferuje użytkownikowi, operacje na strukturach danych i manipulowanie tabelami numerycznymi.

**Graphviz**- zestaw narzędzi do tworzenia diagramów za pomocą grafów. Diagramy tworzy się w języku DOT.

**Train\_test\_split**-metoda do podzielenia bazy na grupy trenujące i grupy testujące w tym wypadku jest to 70% procent trenujących i 30% testujących

**PyCharm**- IDE dla języka programowania Python firmy JetBrains.

**Baza Danych**- W tym wypadku wykorzystanej bazy danych o nazwie diabietesV2.csv składającej się z 9 kolumn i 768 wierszy. Baz danych przechowuje dane składające się z kilku medycznych zmiennych takich jak ilość przebytych ciąży, zawartość glukozy wykrytej w testach, ciśnienie krwi, grubość skóry, zawartość insuliny w ciele po 2 godzinach od podania dawki, BMI, wiek, możliwość rozwoju cukrzycy, i wynik u pacjenta(0/1), gdzie zero oznacza brak wykrycia choroby, a jedynka wykrycie.

**Budowa drzewa:**

Drzewo decyzyjne klasyfikacyjnego używa obiektów i trenuje model tworzący strukturę drzewa żeby przewidzieć przyszłe wartości widziane, jako liczby na wyjściu. Wskutek tego przypadku przewiduje czy pacjent jest chory na cukrzyce.

**Kod:**

import pandas as pd #imort pandas  
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier  
from sklearn.model\_selection import train\_test\_split # Import train\_test\_split  
from sklearn import metrics #import metric do obliczeń  
  
#nazwanie kolumn  
col\_names = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'label']  
  
#import bazy z pliku  
dataset = pd.read\_csv("diabetesV2.csv", header=None, names=col\_names)  
  
#wyświetlenie bazy  
print(dataset)  
  
#podzielenie bazy na obiekty i cechy  
feature\_cols = ['pregnant', 'insulin', 'bmi', 'age','glucose','bp','pedigree']  
X = dataset[feature\_cols] #cechy  
y = dataset.label #obiekty  
  
#podział bazy na grupy testujące 30% i trenujące 70%  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=1)  
  
#stworzenie klasyfikatora drzewa  
clf = DecisionTreeClassifier()  
  
#testowanie klasyfikatora drzewa  
clf = clf.fit(X\_train,y\_train)  
  
#przewidywanie wyniku  
y\_pred = clf.predict(X\_test)  
  
#wyświetlenie jak często model klasyfikacji jest poprawny(0.67)  
print("Accuracy:",metrics.accuracy\_score(y\_test, y\_pred))  
  
#import graphviz  
from sklearn.tree import export\_graphviz  
  
#eskport drzewa w postaci txt  
export\_graphviz(clf,out\_file ='tree.txt')  
  
#optymalicazcja drzewa  
clf = DecisionTreeClassifier(criterion="entropy", max\_depth=3)  
clf = clf.fit(X\_train,y\_train)  
y\_pred = clf.predict(X\_test)  
  
#wyświetlenie jak czesto model klasyikacji jest poprawan po optymalizacji(0.77)  
print("Accuracy:",metrics.accuracy\_score(y\_test, y\_pred))  
  
#eksport po optymalizacji  
export\_graphviz(clf,out\_file ='treeOP.txt')

#http://www.webgraphviz.com/ wyświetlenie grafów

**Kod grafu Drzewa Dot:**

digraph Tree {

node [shape=box] ;

0 [label="X[4] <= 129.5\ngini = 0.449\nsamples = 537\nvalue = [354, 183]"] ;

1 [label="X[2] <= 26.3\ngini = 0.329\nsamples = 357\nvalue = [283, 74]"] ;

0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;

2 [label="X[2] <= 9.1\ngini = 0.06\nsamples = 97\nvalue = [94, 3]"] ;

1 -> 2 ;

3 [label="X[4] <= 114.5\ngini = 0.444\nsamples = 6\nvalue = [4, 2]"] ;

2 -> 3 ;

4 [label="gini = 0.0\nsamples = 4\nvalue = [4, 0]"] ;

3 -> 4 ;

5 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

3 -> 5 ;

6 [label="X[6] <= 0.669\ngini = 0.022\nsamples = 91\nvalue = [90, 1]"] ;

2 -> 6 ;

7 [label="gini = 0.0\nsamples = 76\nvalue = [76, 0]"] ;

6 -> 7 ;

8 [label="X[6] <= 0.705\ngini = 0.124\nsamples = 15\nvalue = [14, 1]"] ;

6 -> 8 ;

9 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

8 -> 9 ;

10 [label="gini = 0.0\nsamples = 14\nvalue = [14, 0]"] ;

8 -> 10 ;

11 [label="X[3] <= 27.5\ngini = 0.397\nsamples = 260\nvalue = [189, 71]"] ;

1 -> 11 ;

12 [label="X[2] <= 45.4\ngini = 0.243\nsamples = 120\nvalue = [103, 17]"] ;

11 -> 12 ;

13 [label="X[5] <= 12.0\ngini = 0.212\nsamples = 116\nvalue = [102, 14]"] ;

12 -> 13 ;

14 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

13 -> 14 ;

15 [label="X[0] <= 7.0\ngini = 0.201\nsamples = 115\nvalue = [102, 13]"] ;

13 -> 15 ;

16 [label="X[6] <= 1.272\ngini = 0.188\nsamples = 114\nvalue = [102, 12]"] ;

15 -> 16 ;

17 [label="X[2] <= 30.95\ngini = 0.165\nsamples = 110\nvalue = [100, 10]"] ;

16 -> 17 ;

18 [label="gini = 0.0\nsamples = 43\nvalue = [43, 0]"] ;

17 -> 18 ;

19 [label="X[5] <= 53.0\ngini = 0.254\nsamples = 67\nvalue = [57, 10]"] ;

17 -> 19 ;

20 [label="X[6] <= 0.264\ngini = 0.5\nsamples = 6\nvalue = [3, 3]"] ;

19 -> 20 ;

21 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

20 -> 21 ;

22 [label="X[4] <= 125.5\ngini = 0.375\nsamples = 4\nvalue = [3, 1]"] ;

20 -> 22 ;

23 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

22 -> 23 ;

24 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

22 -> 24 ;

25 [label="X[6] <= 0.652\ngini = 0.203\nsamples = 61\nvalue = [54, 7]"] ;

19 -> 25 ;

26 [label="X[1] <= 36.5\ngini = 0.15\nsamples = 49\nvalue = [45, 4]"] ;

25 -> 26 ;

27 [label="X[1] <= 34.0\ngini = 0.32\nsamples = 20\nvalue = [16, 4]"] ;

26 -> 27 ;

28 [label="X[4] <= 111.5\ngini = 0.266\nsamples = 19\nvalue = [16, 3]"] ;

27 -> 28 ;

29 [label="X[6] <= 0.458\ngini = 0.142\nsamples = 13\nvalue = [12, 1]"] ;

28 -> 29 ;

30 [label="gini = 0.0\nsamples = 11\nvalue = [11, 0]"] ;

29 -> 30 ;

31 [label="X[0] <= 1.0\ngini = 0.5\nsamples = 2\nvalue = [1, 1]"] ;

29 -> 31 ;

32 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

31 -> 32 ;

33 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

31 -> 33 ;

34 [label="X[2] <= 34.5\ngini = 0.444\nsamples = 6\nvalue = [4, 2]"] ;

28 -> 34 ;

35 [label="X[5] <= 66.0\ngini = 0.444\nsamples = 3\nvalue = [1, 2]"] ;

34 -> 35 ;

36 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

35 -> 36 ;

37 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

35 -> 37 ;

38 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

34 -> 38 ;

39 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

27 -> 39 ;

40 [label="gini = 0.0\nsamples = 29\nvalue = [29, 0]"] ;

26 -> 40 ;

41 [label="X[1] <= 65.5\ngini = 0.375\nsamples = 12\nvalue = [9, 3]"] ;

25 -> 41 ;

42 [label="gini = 0.0\nsamples = 7\nvalue = [7, 0]"] ;

41 -> 42 ;

43 [label="X[4] <= 115.0\ngini = 0.48\nsamples = 5\nvalue = [2, 3]"] ;

41 -> 43 ;

44 [label="gini = 0.0\nsamples = 3\nvalue = [0, 3]"] ;

43 -> 44 ;

45 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

43 -> 45 ;

46 [label="X[6] <= 1.496\ngini = 0.5\nsamples = 4\nvalue = [2, 2]"] ;

16 -> 46 ;

47 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

46 -> 47 ;

48 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

46 -> 48 ;

49 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

15 -> 49 ;

50 [label="X[3] <= 22.5\ngini = 0.375\nsamples = 4\nvalue = [1, 3]"] ;

12 -> 50 ;

51 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

50 -> 51 ;

52 [label="gini = 0.0\nsamples = 3\nvalue = [0, 3]"] ;

50 -> 52 ;

53 [label="X[6] <= 0.563\ngini = 0.474\nsamples = 140\nvalue = [86, 54]"] ;

11 -> 53 ;

54 [label="X[4] <= 101.5\ngini = 0.408\nsamples = 98\nvalue = [70, 28]"] ;

53 -> 54 ;

55 [label="X[5] <= 27.0\ngini = 0.208\nsamples = 34\nvalue = [30, 4]"] ;

54 -> 55 ;

56 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

55 -> 56 ;

57 [label="X[3] <= 42.5\ngini = 0.165\nsamples = 33\nvalue = [30, 3]"] ;

55 -> 57 ;

58 [label="gini = 0.0\nsamples = 23\nvalue = [23, 0]"] ;

57 -> 58 ;

59 [label="X[4] <= 97.5\ngini = 0.42\nsamples = 10\nvalue = [7, 3]"] ;

57 -> 59 ;

60 [label="X[2] <= 35.25\ngini = 0.5\nsamples = 6\nvalue = [3, 3]"] ;

59 -> 60 ;

61 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

60 -> 61 ;

62 [label="X[4] <= 96.0\ngini = 0.375\nsamples = 4\nvalue = [3, 1]"] ;

60 -> 62 ;

63 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

62 -> 63 ;

64 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

62 -> 64 ;

65 [label="gini = 0.0\nsamples = 4\nvalue = [4, 0]"] ;

59 -> 65 ;

66 [label="X[5] <= 67.0\ngini = 0.469\nsamples = 64\nvalue = [40, 24]"] ;

54 -> 66 ;

67 [label="X[5] <= 58.0\ngini = 0.465\nsamples = 19\nvalue = [7, 12]"] ;

66 -> 67 ;

68 [label="X[2] <= 30.05\ngini = 0.245\nsamples = 7\nvalue = [6, 1]"] ;

67 -> 68 ;

69 [label="X[6] <= 0.182\ngini = 0.5\nsamples = 2\nvalue = [1, 1]"] ;

68 -> 69 ;

70 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

69 -> 70 ;

71 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

69 -> 71 ;

72 [label="gini = 0.0\nsamples = 5\nvalue = [5, 0]"] ;

68 -> 72 ;

73 [label="X[6] <= 0.425\ngini = 0.153\nsamples = 12\nvalue = [1, 11]"] ;

67 -> 73 ;

74 [label="gini = 0.0\nsamples = 11\nvalue = [0, 11]"] ;

73 -> 74 ;

75 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

73 -> 75 ;

76 [label="X[2] <= 43.1\ngini = 0.391\nsamples = 45\nvalue = [33, 12]"] ;

66 -> 76 ;

77 [label="X[4] <= 102.5\ngini = 0.337\nsamples = 42\nvalue = [33, 9]"] ;

76 -> 77 ;

78 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

77 -> 78 ;

79 [label="X[2] <= 34.6\ngini = 0.289\nsamples = 40\nvalue = [33, 7]"] ;

77 -> 79 ;

80 [label="X[3] <= 35.0\ngini = 0.434\nsamples = 22\nvalue = [15, 7]"] ;

79 -> 80 ;

81 [label="gini = 0.0\nsamples = 6\nvalue = [6, 0]"] ;

80 -> 81 ;

82 [label="X[4] <= 123.5\ngini = 0.492\nsamples = 16\nvalue = [9, 7]"] ;

80 -> 82 ;

83 [label="X[6] <= 0.375\ngini = 0.426\nsamples = 13\nvalue = [9, 4]"] ;

82 -> 83 ;

84 [label="X[6] <= 0.258\ngini = 0.298\nsamples = 11\nvalue = [9, 2]"] ;

83 -> 84 ;

85 [label="gini = 0.0\nsamples = 5\nvalue = [5, 0]"] ;

84 -> 85 ;

86 [label="X[2] <= 32.0\ngini = 0.444\nsamples = 6\nvalue = [4, 2]"] ;

84 -> 86 ;

87 [label="X[4] <= 118.0\ngini = 0.444\nsamples = 3\nvalue = [1, 2]"] ;

86 -> 87 ;

88 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

87 -> 88 ;

89 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

87 -> 89 ;

90 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

86 -> 90 ;

91 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

83 -> 91 ;

92 [label="gini = 0.0\nsamples = 3\nvalue = [0, 3]"] ;

82 -> 92 ;

93 [label="gini = 0.0\nsamples = 18\nvalue = [18, 0]"] ;

79 -> 93 ;

94 [label="gini = 0.0\nsamples = 3\nvalue = [0, 3]"] ;

76 -> 94 ;

95 [label="X[0] <= 8.5\ngini = 0.472\nsamples = 42\nvalue = [16, 26]"] ;

53 -> 95 ;

96 [label="X[4] <= 97.0\ngini = 0.5\nsamples = 33\nvalue = [16, 17]"] ;

95 -> 96 ;

97 [label="gini = 0.0\nsamples = 5\nvalue = [5, 0]"] ;

96 -> 97 ;

98 [label="X[5] <= 87.0\ngini = 0.477\nsamples = 28\nvalue = [11, 17]"] ;

96 -> 98 ;

99 [label="X[4] <= 116.5\ngini = 0.413\nsamples = 24\nvalue = [7, 17]"] ;

98 -> 99 ;

100 [label="X[6] <= 1.395\ngini = 0.165\nsamples = 11\nvalue = [1, 10]"] ;

99 -> 100 ;

101 [label="gini = 0.0\nsamples = 10\nvalue = [0, 10]"] ;

100 -> 101 ;

102 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

100 -> 102 ;

103 [label="X[1] <= 78.0\ngini = 0.497\nsamples = 13\nvalue = [6, 7]"] ;

99 -> 103 ;

104 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

103 -> 104 ;

105 [label="X[4] <= 123.5\ngini = 0.42\nsamples = 10\nvalue = [3, 7]"] ;

103 -> 105 ;

106 [label="X[4] <= 119.5\ngini = 0.5\nsamples = 6\nvalue = [3, 3]"] ;

105 -> 106 ;

107 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

106 -> 107 ;

108 [label="X[6] <= 0.712\ngini = 0.375\nsamples = 4\nvalue = [3, 1]"] ;

106 -> 108 ;

109 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

108 -> 109 ;

110 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

108 -> 110 ;

111 [label="gini = 0.0\nsamples = 4\nvalue = [0, 4]"] ;

105 -> 111 ;

112 [label="gini = 0.0\nsamples = 4\nvalue = [4, 0]"] ;

98 -> 112 ;

113 [label="gini = 0.0\nsamples = 9\nvalue = [0, 9]"] ;

95 -> 113 ;

114 [label="X[2] <= 27.85\ngini = 0.478\nsamples = 180\nvalue = [71, 109]"] ;

0 -> 114 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;

115 [label="X[4] <= 145.5\ngini = 0.375\nsamples = 36\nvalue = [27, 9]"] ;

114 -> 115 ;

116 [label="X[3] <= 59.5\ngini = 0.1\nsamples = 19\nvalue = [18, 1]"] ;

115 -> 116 ;

117 [label="gini = 0.0\nsamples = 16\nvalue = [16, 0]"] ;

116 -> 117 ;

118 [label="X[3] <= 61.0\ngini = 0.444\nsamples = 3\nvalue = [2, 1]"] ;

116 -> 118 ;

119 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

118 -> 119 ;

120 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

118 -> 120 ;

121 [label="X[0] <= 1.5\ngini = 0.498\nsamples = 17\nvalue = [9, 8]"] ;

115 -> 121 ;

122 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

121 -> 122 ;

123 [label="X[2] <= 23.1\ngini = 0.49\nsamples = 14\nvalue = [6, 8]"] ;

121 -> 123 ;

124 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

123 -> 124 ;

125 [label="X[2] <= 25.8\ngini = 0.444\nsamples = 12\nvalue = [4, 8]"] ;

123 -> 125 ;

126 [label="gini = 0.0\nsamples = 6\nvalue = [0, 6]"] ;

125 -> 126 ;

127 [label="X[4] <= 157.0\ngini = 0.444\nsamples = 6\nvalue = [4, 2]"] ;

125 -> 127 ;

128 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

127 -> 128 ;

129 [label="gini = 0.0\nsamples = 4\nvalue = [4, 0]"] ;

127 -> 129 ;

130 [label="X[4] <= 158.5\ngini = 0.424\nsamples = 144\nvalue = [44, 100]"] ;

114 -> 130 ;

131 [label="X[3] <= 30.5\ngini = 0.487\nsamples = 88\nvalue = [37, 51]"] ;

130 -> 131 ;

132 [label="X[5] <= 23.0\ngini = 0.49\nsamples = 42\nvalue = [24, 18]"] ;

131 -> 132 ;

133 [label="gini = 0.0\nsamples = 4\nvalue = [0, 4]"] ;

132 -> 133 ;

134 [label="X[5] <= 88.0\ngini = 0.465\nsamples = 38\nvalue = [24, 14]"] ;

132 -> 134 ;

135 [label="X[5] <= 72.0\ngini = 0.444\nsamples = 36\nvalue = [24, 12]"] ;

134 -> 135 ;

136 [label="X[2] <= 33.75\ngini = 0.5\nsamples = 20\nvalue = [10, 10]"] ;

135 -> 136 ;

137 [label="X[2] <= 31.05\ngini = 0.397\nsamples = 11\nvalue = [8, 3]"] ;

136 -> 137 ;

138 [label="X[6] <= 0.374\ngini = 0.5\nsamples = 6\nvalue = [3, 3]"] ;

137 -> 138 ;

139 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

138 -> 139 ;

140 [label="X[5] <= 67.0\ngini = 0.375\nsamples = 4\nvalue = [3, 1]"] ;

138 -> 140 ;

141 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

140 -> 141 ;

142 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

140 -> 142 ;

143 [label="gini = 0.0\nsamples = 5\nvalue = [5, 0]"] ;

137 -> 143 ;

144 [label="X[6] <= 0.535\ngini = 0.346\nsamples = 9\nvalue = [2, 7]"] ;

136 -> 144 ;

145 [label="gini = 0.0\nsamples = 6\nvalue = [0, 6]"] ;

144 -> 145 ;

146 [label="X[5] <= 66.0\ngini = 0.444\nsamples = 3\nvalue = [2, 1]"] ;

144 -> 146 ;

147 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

146 -> 147 ;

148 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

146 -> 148 ;

149 [label="X[4] <= 157.5\ngini = 0.219\nsamples = 16\nvalue = [14, 2]"] ;

135 -> 149 ;

150 [label="X[5] <= 85.5\ngini = 0.124\nsamples = 15\nvalue = [14, 1]"] ;

149 -> 150 ;

151 [label="gini = 0.0\nsamples = 13\nvalue = [13, 0]"] ;

150 -> 151 ;

152 [label="X[0] <= 2.5\ngini = 0.5\nsamples = 2\nvalue = [1, 1]"] ;

150 -> 152 ;

153 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

152 -> 153 ;

154 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

152 -> 154 ;

155 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

149 -> 155 ;

156 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

134 -> 156 ;

157 [label="X[2] <= 34.05\ngini = 0.405\nsamples = 46\nvalue = [13, 33]"] ;

131 -> 157 ;

158 [label="X[5] <= 75.0\ngini = 0.49\nsamples = 21\nvalue = [9, 12]"] ;

157 -> 158 ;

159 [label="gini = 0.0\nsamples = 4\nvalue = [4, 0]"] ;

158 -> 159 ;

160 [label="X[4] <= 143.0\ngini = 0.415\nsamples = 17\nvalue = [5, 12]"] ;

158 -> 160 ;

161 [label="X[3] <= 47.5\ngini = 0.496\nsamples = 11\nvalue = [5, 6]"] ;

160 -> 161 ;

162 [label="X[2] <= 33.0\ngini = 0.375\nsamples = 8\nvalue = [2, 6]"] ;

161 -> 162 ;

163 [label="X[1] <= 147.5\ngini = 0.245\nsamples = 7\nvalue = [1, 6]"] ;

162 -> 163 ;

164 [label="gini = 0.0\nsamples = 5\nvalue = [0, 5]"] ;

163 -> 164 ;

165 [label="X[2] <= 31.4\ngini = 0.5\nsamples = 2\nvalue = [1, 1]"] ;

163 -> 165 ;

166 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

165 -> 166 ;

167 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

165 -> 167 ;

168 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

162 -> 168 ;

169 [label="gini = 0.0\nsamples = 3\nvalue = [3, 0]"] ;

161 -> 169 ;

170 [label="gini = 0.0\nsamples = 6\nvalue = [0, 6]"] ;

160 -> 170 ;

171 [label="X[6] <= 1.088\ngini = 0.269\nsamples = 25\nvalue = [4, 21]"] ;

157 -> 171 ;

172 [label="X[1] <= 306.5\ngini = 0.172\nsamples = 21\nvalue = [2, 19]"] ;

171 -> 172 ;

173 [label="X[6] <= 0.222\ngini = 0.1\nsamples = 19\nvalue = [1, 18]"] ;

172 -> 173 ;

174 [label="X[5] <= 74.0\ngini = 0.444\nsamples = 3\nvalue = [1, 2]"] ;

173 -> 174 ;

175 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

174 -> 175 ;

176 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

174 -> 176 ;

177 [label="gini = 0.0\nsamples = 16\nvalue = [0, 16]"] ;

173 -> 177 ;

178 [label="X[1] <= 356.0\ngini = 0.5\nsamples = 2\nvalue = [1, 1]"] ;

172 -> 178 ;

179 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

178 -> 179 ;

180 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

178 -> 180 ;

181 [label="X[1] <= 147.5\ngini = 0.5\nsamples = 4\nvalue = [2, 2]"] ;

171 -> 181 ;

182 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

181 -> 182 ;

183 [label="gini = 0.0\nsamples = 2\nvalue = [0, 2]"] ;

181 -> 183 ;

184 [label="X[6] <= 1.157\ngini = 0.219\nsamples = 56\nvalue = [7, 49]"] ;

130 -> 184 ;

185 [label="X[6] <= 0.343\ngini = 0.147\nsamples = 50\nvalue = [4, 46]"] ;

184 -> 185 ;

186 [label="X[3] <= 48.5\ngini = 0.332\nsamples = 19\nvalue = [4, 15]"] ;

185 -> 186 ;

187 [label="X[4] <= 177.0\ngini = 0.219\nsamples = 16\nvalue = [2, 14]"] ;

186 -> 187 ;

188 [label="gini = 0.0\nsamples = 9\nvalue = [0, 9]"] ;

187 -> 188 ;

189 [label="X[0] <= 8.0\ngini = 0.408\nsamples = 7\nvalue = [2, 5]"] ;

187 -> 189 ;

190 [label="X[4] <= 187.5\ngini = 0.278\nsamples = 6\nvalue = [1, 5]"] ;

189 -> 190 ;

191 [label="gini = 0.0\nsamples = 4\nvalue = [0, 4]"] ;

190 -> 191 ;

192 [label="X[5] <= 69.0\ngini = 0.5\nsamples = 2\nvalue = [1, 1]"] ;

190 -> 192 ;

193 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

192 -> 193 ;

194 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

192 -> 194 ;

195 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

189 -> 195 ;

196 [label="X[4] <= 184.5\ngini = 0.444\nsamples = 3\nvalue = [2, 1]"] ;

186 -> 196 ;

197 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

196 -> 197 ;

198 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]"] ;

196 -> 198 ;

199 [label="gini = 0.0\nsamples = 31\nvalue = [0, 31]"] ;

185 -> 199 ;

200 [label="X[0] <= 3.5\ngini = 0.5\nsamples = 6\nvalue = [3, 3]"] ;

184 -> 200 ;

201 [label="X[6] <= 1.191\ngini = 0.375\nsamples = 4\nvalue = [1, 3]"] ;

200 -> 201 ;

202 [label="gini = 0.0\nsamples = 1\nvalue = [1, 0]"] ;

201 -> 202 ;

203 [label="gini = 0.0\nsamples = 3\nvalue = [0, 3]"] ;

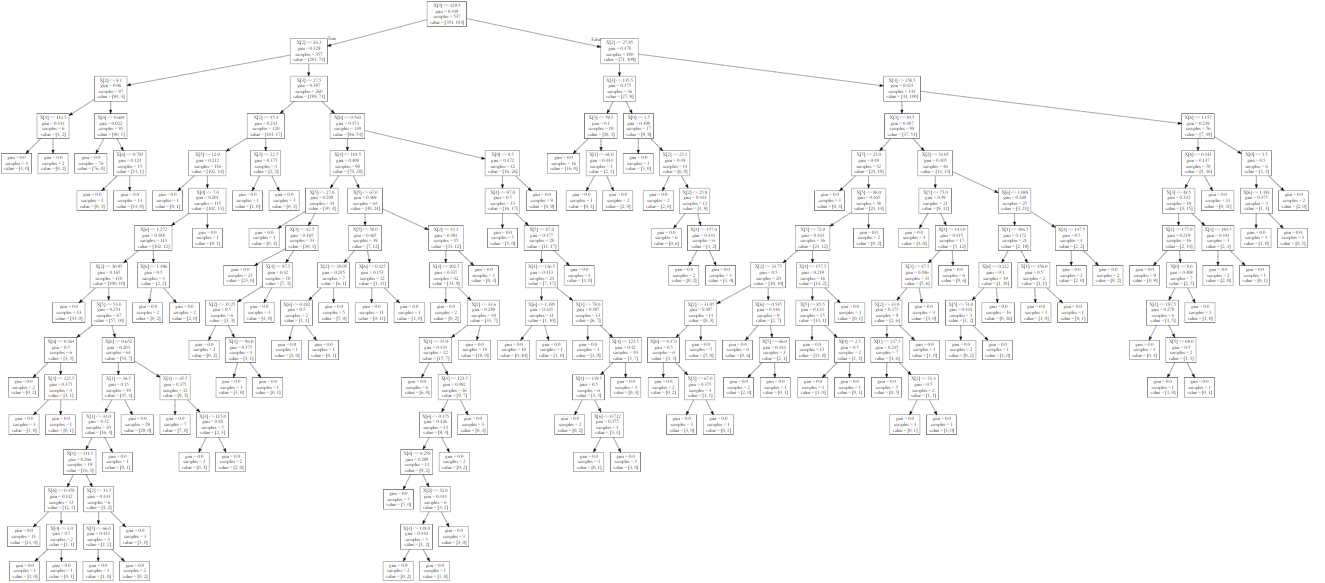
201 -> 203 ;

204 [label="gini = 0.0\nsamples = 2\nvalue = [2, 0]"] ;

200 -> 204 ;

}

**Drzewo Decyzyjne Wykres:**



**Kod grafu Drzewa Dot po optymalizacji:**

digraph Tree {

node [shape=box] ;

0 [label="X[4] <= 127.5\nentropy = 0.926\nsamples = 537\nvalue = [354, 183]"] ;

1 [label="X[2] <= 26.45\nentropy = 0.72\nsamples = 342\nvalue = [274, 68]"] ;

0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;

2 [label="X[2] <= 9.1\nentropy = 0.201\nsamples = 96\nvalue = [93, 3]"] ;

1 -> 2 ;

3 [label="entropy = 0.918\nsamples = 6\nvalue = [4, 2]"] ;

2 -> 3 ;

4 [label="entropy = 0.088\nsamples = 90\nvalue = [89, 1]"] ;

2 -> 4 ;

5 [label="X[3] <= 27.5\nentropy = 0.833\nsamples = 246\nvalue = [181, 65]"] ;

1 -> 5 ;

6 [label="entropy = 0.544\nsamples = 112\nvalue = [98, 14]"] ;

5 -> 6 ;

7 [label="entropy = 0.958\nsamples = 134\nvalue = [83, 51]"] ;

5 -> 7 ;

8 [label="X[2] <= 28.15\nentropy = 0.977\nsamples = 195\nvalue = [80, 115]"] ;

0 -> 8 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;

9 [label="X[4] <= 145.5\nentropy = 0.82\nsamples = 43\nvalue = [32, 11]"] ;

8 -> 9 ;

10 [label="entropy = 0.402\nsamples = 25\nvalue = [23, 2]"] ;

9 -> 10 ;

11 [label="entropy = 1.0\nsamples = 18\nvalue = [9, 9]"] ;

9 -> 11 ;

12 [label="X[4] <= 158.5\nentropy = 0.9\nsamples = 152\nvalue = [48, 104]"] ;

8 -> 12 ;

13 [label="entropy = 0.985\nsamples = 96\nvalue = [41, 55]"] ;

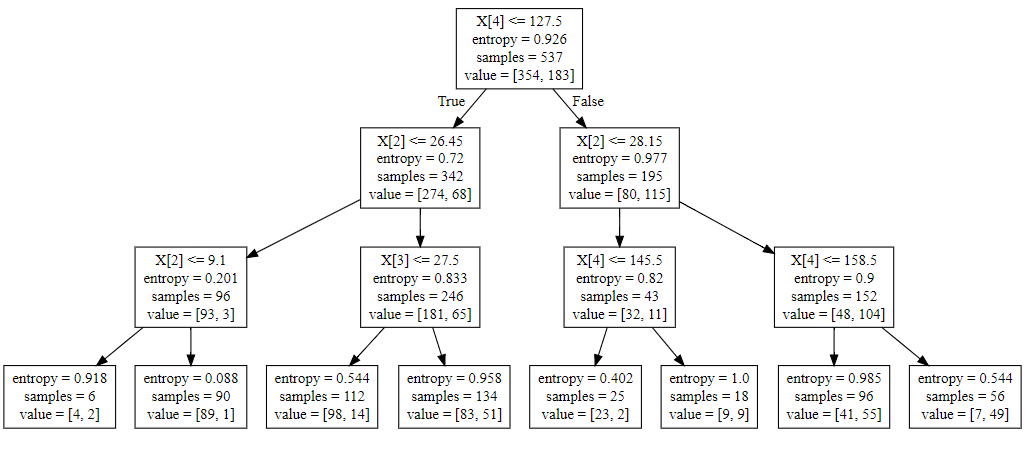
12 -> 13 ;

14 [label="entropy = 0.544\nsamples = 56\nvalue = [7, 49]"] ;

12 -> 14 ;

}

**Drzewo Decyzyjne Wykres po optymalizacji:**



**Wnioski:**

Podsumowując projekt, drzewo przed optymalizacją przedstawia 67% dobrych diagnoz u pacjentów. Jednakowoż po optymalizacji polegającej na zmniejszeniu wielkości drzewa(teraz maksymalną wielkością poszukiwań jest 3 poziom głębi) wykazuje 10% wzrost niż przed optymalizacją, co daje nam większą precyzje w wykryci choroby.

Algorytm można zastosować do innych dziedzin takich jak, porównywanie danych statystycznych, dalsza pomoc w celach naukowych, które mogą pomóc ludziom w przyszłości

Bibliografia:

<https://scikit-learn.org/stable/index.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html?highlight=decision%20tree#sklearn.tree.DecisionTreeClassifier>

<https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>

<https://moodle2.e-wsb.pl/pluginfile.php/2619181/mod_resource/content/0/drzewa_decyzyjne.pdf>

<https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article>

<https://en.wikipedia.org/wiki/Scikit-learn>

<https://en.wikipedia.org/wiki/Pandas_(software)>

<https://pl.wikipedia.org/wiki/PyCharm>

<https://en.wikipedia.org/wiki/Decision_tree>

<https://www.kaggle.com/uciml/pima-indians-diabetes-database><https://en.wikipedia.org/wiki/Graphviz>