**Transcript**

1

00:00:00,630 --> 00:00:03,049

Hello my name is Syed Abbas Ali Safdar. At the end

2

00:00:03,049 --> 00:00:04,930

the topic that I would be explaining

3

00:00:04,930 --> 00:00:06,811

today is the K nearest neighbours or

4

00:00:06,811 --> 00:00:09,590

commonly known as the KNN algorithm. Can

5

00:00:09,590 --> 00:00:12,091

an algorithm is a supervised learning

6

00:00:12,091 --> 00:00:14,591

classifier which uses proximity to

7

00:00:14,591 --> 00:00:16,675

make classifications or predictions about

8

00:00:16,675 --> 00:00:19,592

the grouping of an individual data point

9

00:00:19,592 --> 00:00:22,509

K nearest algorithm is a very simple,

10

00:00:22,509 --> 00:00:24,593

yet a very powerful classification

11

00:00:24,593 --> 00:00:27,383

algorithm. In a real life example,

12

00:00:27,383 --> 00:00:29,979

it is very similar to taking a majority

13

00:00:29,979 --> 00:00:32,841

opinion from friends. So basically

14

00:00:32,841 --> 00:00:35,497

how does Kane and work? So Kanan

15

00:00:35,497 --> 00:00:37,772

basically takes the distance between the

16

00:00:37,772 --> 00:00:40,428

new point an existing points an compares

17

00:00:40,428 --> 00:00:43,083

it and whatever the closet point is

18

00:00:43,083 --> 00:00:45,359

it would identify in that particular

19

00:00:45,359 --> 00:00:48,157

class. I would

20

00:00:48,157 --> 00:00:50,628

explain this with an example. This is a

21

00:00:50,628 --> 00:00:53,408

very small data set. A snippet of a small

22

00:00:53,408 --> 00:00:55,879

data set from IMDb. There are 4 movies.

23

00:00:55,879 --> 00:00:58,351

We have ratings for the movies, we have

24

00:00:58,351 --> 00:01:00,822

durations for the movies and we have the

25

00:01:00,822 --> 00:01:03,602

genre for the movies. If you can see we

26

00:01:03,602 --> 00:01:05,456

have one missing 1 missing column.

27

00:01:06,206 --> 00:01:08,777

Which for genre and.

28

00:01:09,367 --> 00:01:12,192

They we do we do not know the genre for

29

00:01:12,192 --> 00:01:14,833

for movie 4. Week if we

30

00:01:14,833 --> 00:01:17,455

closely see that we have complete data

31

00:01:17,455 --> 00:01:20,452

for the first 3 movies an 1 incomplete.

32

00:01:21,012 --> 00:01:23,811

Attribute for this one. And now

33

00:01:23,811 --> 00:01:26,563

we have to predict what could the genre

34

00:01:26,563 --> 00:01:28,971

before movie 4, movie 123 are

35

00:01:28,971 --> 00:01:31,723

learning data sets. This is the data set

36

00:01:31,723 --> 00:01:34,476

that already exists. We can use this to

37

00:01:34,476 --> 00:01:35,508

make further predictions.

38

00:01:37,388 --> 00:01:39,716

There are different methods how Kanan

39

00:01:39,716 --> 00:01:41,657

could calculate distances. However, what

40

00:01:41,657 --> 00:01:44,373

we would be using is the Euclidean

41

00:01:44,373 --> 00:01:47,089

distance. This is the formula for you

42

00:01:47,089 --> 00:01:49,977

cleared in distance. If you

43

00:01:49,977 --> 00:01:51,964

closely sleep, see the.

44

00:01:52,744 --> 00:01:55,598

X1 attribute would be this one. The

45

00:01:55,598 --> 00:01:57,334

rating would be X1.

46

00:01:58,835 --> 00:02:01,679

So the value for X1 and Y1 would

47

00:02:01,679 --> 00:02:04,523

be static, X1 would be 8.1, Y one

48

00:02:04,523 --> 00:02:07,012

would be fun 55 now we would.

49

00:02:07,682 --> 00:02:10,007

Compare with. Calculate the distances

50

00:02:10,007 --> 00:02:12,795

from these existing points for. For

51

00:02:12,795 --> 00:02:15,584

the distance for movie 1 would

52

00:02:15,584 --> 00:02:18,373

be. Using this formula X1 is

53

00:02:18,373 --> 00:02:21,162

8.1 Y one is 155 depicting

54

00:02:21,162 --> 00:02:23,951

this and rating for 8, Rating

55

00:02:23,951 --> 00:02:26,740

for movie 1 and duration would

56

00:02:26,740 --> 00:02:29,673

be X2. And why 2 steps?X2

57

00:02:29,673 --> 00:02:32,484

and Y2 moving on. Using

58

00:02:32,484 --> 00:02:35,428

this formula, the distance that we have

59

00:02:35,428 --> 00:02:37,111

accumulates to approximately 5.

60

00:02:37,111 --> 00:02:40,055

Similarly, we calculate the data at the

61

00:02:40,055 --> 00:02:42,999

distance for movie 2 from movie 2.

62

00:02:42,999 --> 00:02:45,943

An movie 3 this is 15.12 and

63

00:02:45,943 --> 00:02:48,888

this would be 13.03 now the smallest

64

00:02:48,888 --> 00:02:51,720

distance. For our current data point.

65

00:02:52,651 --> 00:02:55,394

Is 2 movie 1?Which is in

66

00:02:55,394 --> 00:02:58,039

fact an action movie and based on this we

67

00:02:58,039 --> 00:03:00,651

could say. That movie 4

68

00:03:00,651 --> 00:03:03,239

is most likely an action movie.

69

00:03:03,239 --> 00:03:05,827

Is is also an action movie.

70

00:03:05,827 --> 00:03:08,415

We concluded this by finding it,

71

00:03:08,415 --> 00:03:11,003

finding the distance, and using the

72

00:03:11,003 --> 00:03:13,996

attributes of rating and. Rating

73

00:03:13,996 --> 00:03:16,927

and duration we can see that the only

74

00:03:16,927 --> 00:03:19,858

that this was a very small data set

75

00:03:19,858 --> 00:03:21,689

because this post explanation. However

76

00:03:21,689 --> 00:03:24,254

the attributes that we had were relating

77

00:03:24,254 --> 00:03:27,018

and duration we compared. The debt

78

00:03:27,018 --> 00:03:29,074

that we compared the existing data with

79

00:03:29,074 --> 00:03:32,011

the data we are about to find, and we say

80

00:03:32,011 --> 00:03:34,067

that since the rating and duration of

81

00:03:34,067 --> 00:03:36,123

these are very similar or these data

82

00:03:36,123 --> 00:03:38,179

points are very similar, it is very

83

00:03:38,179 --> 00:03:40,529

likely that the genre would be similar as

84

00:03:40,529 --> 00:03:40,823

well.

85

00:03:43,765 --> 00:03:46,759

Moving on, as stated

86

00:03:46,759 --> 00:03:49,752

earlier, Canon can use multiple

87

00:03:49,752 --> 00:03:52,568

methods. To use to calculate the

88

00:03:52,568 --> 00:03:55,391

data. Do creating distance is the is

89

00:03:55,391 --> 00:03:58,163

what we use and these are the other two

90

00:03:58,163 --> 00:04:00,627

methods which can also be used to find

91

00:04:00,627 --> 00:04:02,783

the. The purpose is same. We're basically

92

00:04:02,783 --> 00:04:04,939

trying to calculate the distance and then

93

00:04:04,939 --> 00:04:06,787

compare the distance and the minimum

94

00:04:06,787 --> 00:04:08,943

distance would be where our data point

95

00:04:08,943 --> 00:04:09,867

lies. Moving on.

96

00:04:11,801 --> 00:04:13,887

Feature scaling is a pre processing

97

00:04:13,887 --> 00:04:15,625

technique that transformed the features

98

00:04:15,625 --> 00:04:17,711

feature values to a similar scale

99

00:04:17,711 --> 00:04:19,797

ensuring that all the feature contributed

100

00:04:19,797 --> 00:04:22,230

equally to the model. In our previous

101

00:04:22,230 --> 00:04:25,011

example we could see that we had two

102

00:04:25,011 --> 00:04:27,097

features, namely the rating and the

103

00:04:27,097 --> 00:04:29,530

duration. When applying the Canon on

104

00:04:29,530 --> 00:04:32,311

a data set we have to make sure.

105

00:04:32,611 --> 00:04:35,537

That the features are scaled so that what

106

00:04:35,537 --> 00:04:38,097

happens is if one of the featured

107

00:04:38,097 --> 00:04:40,291

dominates the KNNwould focus

108

00:04:40,291 --> 00:04:43,217

more on it. And if that happens, the

109

00:04:43,217 --> 00:04:45,411

readings order output would be incorrect.

110

00:04:45,411 --> 00:04:48,337

So what we have to do is make

111

00:04:48,337 --> 00:04:51,263

sure that the that all the features are

112

00:04:51,263 --> 00:04:54,163

scaled and they're very similar. They are

113

00:04:54,163 --> 00:04:56,848

very similar in range so that when Kanan

114

00:04:56,848 --> 00:04:59,534

focuses on all of them, it produces

115

00:04:59,534 --> 00:05:00,541

very good result.

116

00:05:03,739 --> 00:05:05,870

Moving on to explanation of the code.

117

00:05:06,690 --> 00:05:09,451

R code firstly creates a synthetic data

118

00:05:09,451 --> 00:05:11,818

set containing 300 samples and two

119

00:05:11,818 --> 00:05:14,604

features. We will

120

00:05:14,604 --> 00:05:16,737

then visually implement the data set

121

00:05:16,737 --> 00:05:19,226

before scaling and after scaling and then

122

00:05:19,226 --> 00:05:22,070

see how it was improved or how it

123

00:05:22,070 --> 00:05:24,914

was changed how how it affected the KC.

124

00:05:24,914 --> 00:05:27,048

Firstly, this is the implementation of

125

00:05:27,048 --> 00:05:29,181

data set without before scaling. The

126

00:05:29,181 --> 00:05:30,958

different colour shows different classes.

127

00:05:30,958 --> 00:05:33,803

It can be seen that feature 2 has

128

00:05:33,803 --> 00:05:36,395

a larger range which means. Which means

129

00:05:36,395 --> 00:05:39,035

in case of KNN algorithm feature 2

130

00:05:39,035 --> 00:05:40,921

will dominate the calculations since

131

00:05:40,921 --> 00:05:43,562

Canon depends upon the calculation of the

132

00:05:43,562 --> 00:05:44,693

nearest K values.

133

00:05:46,224 --> 00:05:48,326

We basically apply feature scaling in

134

00:05:48,326 --> 00:05:50,779

such cases. In cases where one feature

135

00:05:50,779 --> 00:05:53,231

dominates so that the features the range

136

00:05:53,231 --> 00:05:55,684

of the features are very closely aligned

137

00:05:55,684 --> 00:05:58,487

and when can and works, none of the

138

00:05:58,487 --> 00:06:00,589

feature dominates and the results are

139

00:06:00,589 --> 00:06:03,210

not. And the results are impartial.

140

00:06:04,720 --> 00:06:07,246

However, in this case the accuracy was

141

00:06:07,246 --> 00:06:09,772

not affected after scaling. In some cases

142

00:06:09,772 --> 00:06:12,659

the accuracy may not change, but in cases

143

00:06:12,659 --> 00:06:14,824

where more numerical stuff is involved,

144

00:06:14,824 --> 00:06:16,628

where currencies are involved, the

145

00:06:16,628 --> 00:06:18,432

currency comparison involved are involved

146

00:06:18,432 --> 00:06:20,958

or where how properties like number of

147

00:06:20,958 --> 00:06:23,483

rooms, how size involved in such cases

148

00:06:23,483 --> 00:06:25,287

features killing is very important.

149

00:06:26,448 --> 00:06:29,232

Moving on. This is

150

00:06:29,232 --> 00:06:31,976

the. This is the. This is the

151

00:06:31,976 --> 00:06:34,719

comparison of K values and the accuracy

152

00:06:34,719 --> 00:06:37,463

and the accuracy of accuracy against K

153

00:06:37,463 --> 00:06:40,206

values. It can be seen when we

154

00:06:40,206 --> 00:06:42,950

K value is one. This shows over

155

00:06:42,950 --> 00:06:45,476

fitting an. This gives

156

00:06:45,476 --> 00:06:48,232

a very relatively lower accuracy

157

00:06:48,232 --> 00:06:51,065

of 9091% with. Is

158

00:06:51,065 --> 00:06:53,337

not very good in this case.

159

00:06:55,098 --> 00:06:58,059

The optimal K value would be three or

160

00:06:58,059 --> 00:07:00,650

five where the accuracy is highest, 96%

161

00:07:00,650 --> 00:07:03,241

and we can also see some overfitting.

162

00:07:03,821 --> 00:07:06,508

In the later values of K as well, optimal

163

00:07:06,508 --> 00:07:08,897

value of K would entirely depend upon the

164

00:07:08,897 --> 00:07:10,988

data set. However, in our data set,

165

00:07:10,988 --> 00:07:13,377

optimal value of K would be between 3:00

166

00:07:13,377 --> 00:07:13,974

and 5:00.

167

00:07:16,305 --> 00:07:18,629

Moving on to the confusion metrics, the

168

00:07:18,629 --> 00:07:21,150

confusion metrics basically shows. One

169

00:07:21,150 --> 00:07:23,831

case of false, +3 cases of false

170

00:07:23,831 --> 00:07:26,513

-41, cases of two positive and 45

171

00:07:26,513 --> 00:07:29,194

cases of two negative. This is a

172

00:07:29,194 --> 00:07:31,876

very good. This is relatively a good

173

00:07:31,876 --> 00:07:34,174

accuracy as the accuracy when we

174

00:07:34,174 --> 00:07:36,856

calculated that you see it would be

175

00:07:36,856 --> 00:07:39,537

around 96% and there have been only

176

00:07:39,537 --> 00:07:42,234

a few. I don't classifications but

177

00:07:42,234 --> 00:07:42,874

the algorithm.

178

00:07:46,611 --> 00:07:49,340

So basically that is how we implement the

179

00:07:49,340 --> 00:07:52,187

Canaan code. However, in our case,

180

00:07:52,187 --> 00:07:55,099

we generated a data set. We can use an

181

00:07:55,099 --> 00:07:57,363

existing data set or whatever is the

182

00:07:57,363 --> 00:08:00,062

requirement of our job. So basically

183

00:08:00,062 --> 00:08:02,680

first we generated the data set, then we.

184

00:08:03,390 --> 00:08:05,636

Visualise the data's data set without

185

00:08:05,636 --> 00:08:08,256

scaling. We observe the accuracy, then we

186

00:08:08,256 --> 00:08:10,876

scale the features, and then we compare

187

00:08:10,876 --> 00:08:13,496

how they were before and after scaling.

188

00:08:14,957 --> 00:08:17,405

Stating that this again in in our

189

00:08:17,405 --> 00:08:19,504

example. Curious. It didn't really change

190

00:08:19,504 --> 00:08:21,952

after and before scaling, but in mostly

191

00:08:21,952 --> 00:08:24,942

in many cases. Scaling would affect

192

00:08:24,942 --> 00:08:27,622

the results. Moving

193

00:08:27,622 --> 00:08:30,252

on, what were the key findings?

194

00:08:30,252 --> 00:08:32,883

We it when understanding about the

195

00:08:32,883 --> 00:08:35,513

Canaan algorithm? We can conclude that

196

00:08:35,513 --> 00:08:38,143

Kanan performs way better when scaling

197

00:08:38,143 --> 00:08:39,020

is performed.

198

00:08:43,283 --> 00:08:46,059

The best way K value depends on the data

199

00:08:46,059 --> 00:08:48,834

set. There is no fixed value. There is no

200

00:08:48,834 --> 00:08:51,636

best fixed value of K. It really

201

00:08:51,636 --> 00:08:54,470

depends on the data set. Moving on to

202

00:08:54,470 --> 00:08:56,950

the conclusion, can Canon is an intuitive?

203

00:08:57,861 --> 00:09:00,071

Is an intuitive algorithm, but it

204

00:09:00,071 --> 00:09:01,912

requires proper preprocessing and pre

205

00:09:01,912 --> 00:09:03,385

process. What is preprocessing?

206

00:09:03,385 --> 00:09:05,227

Preprocessing is feature scaling and

207

00:09:05,227 --> 00:09:08,204

stuff. We need to understand the distance

208

00:09:08,204 --> 00:09:09,785

metrics and scaling analysis performance.

209

00:09:09,785 --> 00:09:12,398

We just. We need to

210

00:09:12,398 --> 00:09:15,196

we also before before going onto the

211

00:09:15,196 --> 00:09:17,595

code, I explain how Kanan basically

212

00:09:17,595 --> 00:09:20,394

works, how you clear in distance is

213

00:09:20,394 --> 00:09:22,393

calculated and everything. So basically

214

00:09:22,393 --> 00:09:25,192

when we understand how cannon works and

215

00:09:25,192 --> 00:09:27,990

we also discussed how can canin readings

216

00:09:27,990 --> 00:09:30,389

improved and that is true preprocessing.

217

00:09:32,590 --> 00:09:34,924

And we can further experiment with K

218

00:09:34,924 --> 00:09:37,590

values and we can further scale values in

219

00:09:37,590 --> 00:09:40,257

stuff to make sure we have better an

220

00:09:40,257 --> 00:09:42,257

accurate K values. These are the

221

00:09:42,257 --> 00:09:44,257

references and these are the links.

222

00:09:46,108 --> 00:09:48,882

For this presentation. And the

223

00:09:48,882 --> 00:09:51,321

code in this tutorial. Thank you.