0:0:0.0 --> 0:0:1.590  
PG-Kadri, Khalid  
Okay, um.

0:0:4.650 --> 0:0:6.430  
PG-Kadri, Khalid  
So retention.

0:0:8.140 --> 0:0:11.30  
PG-Kadri, Khalid  
Good evening, um, this is my.

0:0:12.140 --> 0:0:25.0  
PG-Kadri, Khalid  
A poster for my MATLAB project and and this project I was doing a comparison between Decision Tree and naive base applied to breast cancer database.

0:0:25.320 --> 0:0:55.50  
PG-Kadri, Khalid  
Um and the description of the problem. The description on the motivation of the problem basically was to compare the performance of Decision tree and naive based algorithm when applied to a press cancer database. The purpose was to call to do a comparison to evaluate the effects of the algorithm in predicting the presence or absence or of breast cancer in patients. Now I did some initial preprocessing.

0:0:55.230 --> 0:0:57.500  
PG-Kadri, Khalid  
Of the data before.

0:0:59.290 --> 0:1:18.960  
PG-Kadri, Khalid  
Before processing the um, clean data on MATLAB, this was done on, UM, Python. Um, using Jupyter notebooks and I was able to derive a heat map to solve. Choose the top 10 features that correlate with the.

0:1:19.640 --> 0:1:24.790  
PG-Kadri, Khalid  
The class database, which was diagnosis because it's a binary.

0:1:25.890 --> 0:1:36.940  
PG-Kadri, Khalid  
It's a binary classification. We have malignant and biennial cancer, so we're comparing between both and seeing how close our model and can accurately depict that.

0:1:37.360 --> 0:1:37.830  
PG-Kadri, Khalid  
Um?

0:1:39.260 --> 0:1:53.980  
PG-Kadri, Khalid  
You can also see over here we've got the features listed. The means the standard deviation, the mean and Max values, and a brief definition and pros and cons of both models that I was using in this.

0:1:55.240 --> 0:2:1.330  
PG-Kadri, Khalid  
In this research or this, um, research for machine learning?

0:2:3.230 --> 0:2:4.60  
PG-Kadri, Khalid  
We also

0:2:6.200 --> 0:2:14.110  
PG-Kadri, Khalid  
talk about the methodology that I used because the data set was largely small compared to other data sets available.

0:2:15.10 --> 0:2:18.240  
PG-Kadri, Khalid  
It consisted of just 570 rows.

0:2:18.700 --> 0:2:40.30  
PG-Kadri, Khalid  
Um and um 32 features of which I chose 10 features for this particular experiment, and I split my data into 290% for training and 10% for data to test the accuracy of that. So I'll switch over to MATLAB now so I can demonstrate.

0:2:40.370 --> 0:2:40.920  
PG-Kadri, Khalid  
Um?

0:2:42.560 --> 0:3:6.650  
PG-Kadri, Khalid  
The model, um I a quick demo of my model and see what results we get. Like I said, the data was already loaded and reprocessed on Python And that's what this code over here is essentially showing us, so I'll just run the programme now so we can sort of see what we've generated.

0:3:7.30 --> 0:3:20.960  
PG-Kadri, Khalid  
Um, in terms of how well the model is doing in terms of its accuracy, I also use the auto optimization tool in MATLAB to solve find the.

0:3:22.540 --> 0:3:25.910  
PG-Kadri, Khalid  
To find the accurate the the most accurate.

0:3:25.990 --> 0:3:36.450  
PG-Kadri, Khalid  
From Hyper Farm Hyper parameter to use with the tool and over here. That's what the programme is doing. Now is trying to find a the.

0:3:37.310 --> 0:3:39.740  
PG-Kadri, Khalid  
The hyperparameters that is most.

0:3:41.780 --> 0:3:45.530  
PG-Kadri, Khalid  
Accurate to to to solve.

0:3:46.770 --> 0:4:2.600  
PG-Kadri, Khalid  
Determine how we how the data is processed. Um, we also generated the correlation matrix. Um, for the data sets which you see shortly once the system finishes its evaluation, and we can also see a quick.

0:4:4.270 --> 0:4:5.350  
PG-Kadri, Khalid  
We can see a quick.

0:4:6.330 --> 0:4:18.270  
PG-Kadri, Khalid  
Demo ohh we can cheque the results once this process in terms of the accuracy. So when compared to our test score the model accuracy was actually quite surprisingly 1.

0:4:18.370 --> 0:4:29.240  
PG-Kadri, Khalid  
Um, which can be attributed to the fact that the the data set is not as large, so it might not be providing a an accurate.

0:4:29.570 --> 0:4:36.750  
PG-Kadri, Khalid  
Um, an accurate description of the data. But when you compare that to the training data because the training data.

0:4:37.660 --> 0:4:39.210  
PG-Kadri, Khalid  
Was, uh.

0:4:40.510 --> 0:4:57.200  
PG-Kadri, Khalid  
The modular currency and the training data was 93, about 93% so when compared to the test data, it's actually done relatively well, and it's relatively accurate in predicting whether a patient would have money, glance, or benign cancer.

0:4:57.840 --> 0:5:13.10  
PG-Kadri, Khalid  
Um and similar thing also was done. This is the correlation matrix after the test done, the model was applied to the test data. We can see as well the.

0:5:14.860 --> 0:5:17.460  
PG-Kadri, Khalid  
You can see also the.

0:5:18.820 --> 0:5:26.730  
PG-Kadri, Khalid  
The correlation matrix of the training data um before the auto um optimization has taken place.

0:5:28.40 --> 0:5:42.960  
PG-Kadri, Khalid  
And we also have the hyper parameter. This basically is telling us how well we're closed, how well the data is doing and we can see on here that because boots are test and training data are actually sorry.

0:5:43.240 --> 0:5:46.150  
PG-Kadri, Khalid  
I'm in minimum observed value and are estimated.

0:5:47.470 --> 0:5:57.700  
PG-Kadri, Khalid  
Minimal objectives are relatively close. It shows that the data is doing relatively well, so that's just the quick demo of my model.

0:5:58.300 --> 0:6:10.300  
PG-Kadri, Khalid  
Im on naive base. We've also got a similar thing for Decision Tree. Same process, loaded data, split it into my training set and my test set and.

0:6:13.350 --> 0:6:13.790  
PG-Kadri, Khalid  
Um?

0:6:14.930 --> 0:6:23.700  
PG-Kadri, Khalid  
Try to predict get DF-1 scored the model accuracy to overall recall. Um and some of the others. Thats go um.

0:6:24.860 --> 0:6:32.20  
PG-Kadri, Khalid  
Methods that could be used to tell how well the module is doing, um, so if we go back to my.

0:6:34.260 --> 0:7:4.670  
PG-Kadri, Khalid  
My poster which we have have generated over here. We can see some of the results I explained earlier, listed on the and some of the lessons learned over the course of the project. You can learn that the naive base model has had an I accuracy on the test set than the decision tree model. Despite having a lower accuracy on training sets on the training sets and that basically just suggests that the naive model naive based model may be more.

0:7:4.760 --> 0:7:9.120  
PG-Kadri, Khalid  
Of a robust and generally to be on scene data.

0:7:9.500 --> 0:7:29.990  
PG-Kadri, Khalid  
Um, both datas, although I didn't run the second one in this in the short demo I showed earlier, both models achieved relatively high precision recall and F1 scores, indicating that they could effectively identify the samples and the positive class while minimising false positives and false negatives and the naive base features are independent.

0:7:31.130 --> 0:7:46.150  
PG-Kadri, Khalid  
Like a probabilistic model on the. On the other hand, decision trees and non probabilistic model and just only basic model for classification which tends to create complex trees leading to poor performance in real world.

0:7:46.570 --> 0:7:56.860  
PG-Kadri, Khalid  
I'm just in data, so there are some of the things that we can we we can learn from from running both models and test. And I also noted some feature works that could be done.

0:7:58.460 --> 0:8:13.590  
PG-Kadri, Khalid  
To solve, improve our our our our test and our our to improve our model and one of such is to investigate why the Decision tree performs so well on the test train set but not on.

0:8:14.50 --> 0:8:36.650  
PG-Kadri, Khalid  
I'm not in the sorry on the train set, but not on the test set. Um, and we can use other method like the K fold crossvale to cheque to cheque down. That could act as a another sort of method to verify our results and thank you very much for listening and my name is Khalid Kadri and Macquarie Master Student.

0:8:37.290 --> 0:8:46.360  
PG-Kadri, Khalid  
Uh, of data Science City University of London. I hope that you've enjoyed this short presentation and thank you very much.