* Hello. Hello, everybody. It is 630 and a few seconds.

4:36

* Just to give other people some more time is just a few of us.

4:47

* So the Class B is not particularly dense, meaning that hopefully we will have more time for the in-class assignment and I think we can start that.

5:00

* So the recording is already ongoing. Welcome, everybody.

5:18

* It's a 631 up.

5:24

* And let me start as usual with what is going to happen.

5:27

* So this is the live session that we have.

5:38

* We will discuss the homework.

5:42

* We will review a little bit all the elements of machine learning.

5:45

* It's some of the basic algorithms for machine learning between machine learning and data science and knowledge discovery.

5:54

* And then we will do an exercise and classics of science.

6:08

* I will introduce the midterm, and that will be pretty much the end of the day.

6:15

* So the previous assignment was not the midterm.

6:25

* The previous assignment was about to see if they have.

6:32

* It's your. As far as well.

6:36

* Okay. So the previous assignment that was on you, but as far back as two parts.

6:44

* So the first part, that was a question, sir, we will not discuss that.

6:54

* The second part was on coding and was about working on COVID comorbidities.

6:59

* And the idea was basically to account.

7:09

* I mean, the file had the cases in each case with the comorbidity and the class of the comorbidity, the age group, and the number of deaths.

7:13

* And a few other information. So we focus on the comorbidity, the age group, and I mean the number of casualties.

7:30

* So we counted the people per class age.

7:44

* Then we have a bar chart with class age where the axis will be the class age and the Y, the number of deaths.

7:49

* Then we have the we have a pie chart with the same data and this will show the distribution of deaths per our age group.

8:00

* And then we will bring the comorbidity with the highest number of COVID 19 deaths for the population of less than over 35 years of age.

8:14

* So let me go into the code. The code is relatively straightforward.

8:24

* So you have the library, the CSP file to open, to open the file, and then you have my block LIBOR for plotting and the new PI for some calculation.

8:33

* So the assignment was asking to create a function that will be named get index and

8:49

* the input that will be the dataset or the rule under the all data set and the rule.

9:03

* So this function will return the the dataset, the index meaning zero.

9:12

* If it's a zero 24 one.

9:28

* If is a 23, 34.

9:32

* And so on. And if none of them meaning blank, I said the default value to 1000 could be zero, could not be zero, could be nine.

9:35

* It could be in any other number, but just a number that is not between zero and seven.

9:49

* So that's the function. Then the counter and I created a counter that will contain.

10:01

* I mean a little counter to each one is a counter for the specific age group.

10:08

* So where the first one would be zero 24 and the last one would be 85 and plus.

10:15

* Then. I created a blank dictionary that will contain the condition and the number and then account for the maximum number of deaths.

10:24

* The nephew's professor? Yeah. Did you please share your screen, if you don't mind, like you did for the last homework?

10:40

* Absolutely. I'm so sorry. I was supposed to do it.

10:47

* Yeah, no problem. Yep. Yeah.

10:51

* All right. So my apologies. So let's go back very quickly.

10:57

* So sorry about that. Okay.

11:02

* So the.

11:07

* The goal of the assignment was to create, again, two visualizations.

11:13

* One that is an is a bar chart the mighty be because I can believe that they didn't remember who shared the screen by my apologies.

11:20

* So bachata with the classes of age and the number of casualties and then the distribution.

11:31

* So going back again here.

11:39

* So we have the library that we want to import as of this year for reading the file

11:44

* to plot libor to visualizer and non pi that I will use for some calculation.

11:53

* Reading the file and placing the FILA into data, that is a list of elements.

12:01

* Then I created this function get index that basically based on the value in the the age category,

12:13

* the age group is adding it, returning zero up to seven where zero will be the category is a zero 24 seven.

12:28

* The category is more than 85.

12:42

* Then I initialize the list of counters where each counter will contain the counter for the specific age group.

12:47

* I initialize a dictionary that will contain a condition and number of that.

13:07

* Initialize the counter for the maximum number of deaths and the name will host the name of the co-morbidity with the highest number of deaths.

13:16

* So then I'm looping into the list and getting, let's say, the first one.

13:30

* I call the gate the index passing the road that I'm reading to.

13:41

* And then. I will get the index if the index is 1000.

13:51

* That means that is a blinker or is not one of the categories.

13:59

* And I will pass. Also I will extract the number of deaths.

14:10

* And then I accumulate the number of deaths in the proper index.

14:20

* So basically adding over year whatever is the.

14:26

* Appropriate to her age category. So let's say that from this line 61, I get zero a.

14:33

* The counter in position zero will be increased by the number over that's.

14:48

* So the first one that is the age category in zero 24 is what would be added.

14:57

* And what's. And as for this one, then I calculate the condition with more deaths.

15:05

* So I extracted the condition then that if the condition that is is in the is already in the counter the dictionary.

15:15

* Then I adding the number of that.

15:30

* That they could have used the probably that's that's fine it's the same value then that if the condition is

15:35

* greater equal than the max that's an initially we would not because we'll see it then I will replace it.

15:45

* Uh, otherwise, I'm not going to do anything.

15:55

* And then if it's not in the dictionary, then I will add that key.

15:59

* And body is pretty much the same that we did in a in class assignment and was also in one of the slides.

16:03

* So then I will print the accounts of conditions.

16:15

* Sales by age groups. Print in the condition with more COVID deaths.

16:22

* Creating the charts. And then, uh.

16:31

* I mean, the legend that took me a little bit more time because my clock LIBOR is making it a little bit more complicated.

16:42

* In particular, if you want to avoid overlaps.

16:54

* So but anyway, I citing where I took the this piece of code,

16:57

* then they strongly encourage you to do something similar when you use sources that are external.

17:06

* And then they plotted. So if I run it.

17:13

* So that's the account that by age the condition with more common deaths, injuries is respiratory disease with that much and those are the charts.

17:20

* So that's basically it for the.

17:33

* Exercise. Stop shedding for a second.

17:43

* Check if you have any questions. Before we move on.

17:48

* All right, Professor? Yes, sure. I just had a quick question.

17:58

* I must have misinterpreted some of the instructions on the line.

18:02

* So I ended up actually counting the number of deaths per age group, per co-morbidity.

18:07

* So I ended up making it harder on myself. Yeah, I then I then did come back and get, like, total deaths and then less than 35.

18:17

* There's not going to be a penalty for doing more work, is there? No.

18:26

* There is no reward, but no pain out there. Okay. That's right.

18:30

* I just want to make sure I must have been tired reading the instructions somewhere along the line and thought it said I needed it per co-morbidity.

18:33

* Yeah. I will make sure that she knew. We will just do that.

18:42

* That meaning not interpreting exactly what my what the what was asked the and

18:48

* matching the results and saying that is not a match and meaning it is not good,

18:55

* but in reality, you need more work. So don't worry about.

19:00

* All right. Thank you. Yep. All right.

19:04

* Okay. So let's move on. Let me share the screen again.

19:07

* Again, pretty much each assignment is a little bit more complex than the previous one.

19:16

* So that's basically the spirit, the know no is not going to be a major jump in the next one.

19:23

* That is the meter. But the meter is pretty straightforward.

19:32

* The only thing that you need to be aware of is making sure that you will allocate the proper time because there is nothing complicated.

19:37

* But there are several parts and you will need some time.

19:52

* I mean, I will go back to the midterms in a bit.

19:58

* Okay so machine learning we mentioned last time that be sure that we are.

20:03

* The right thing. Yeah.

20:14

* So. We mentioned that at the very end the machines do not really learn in the proper way.

20:21

* So I'm writing several things on the machine learning next week.

20:30

* That is a spring break, but there is unfortunately no spring break for you.

20:39

* Not for me, but the.

20:44

* And we will have a conference here in Hoboken organized by students, by the School of Systems and Enterprises and these on software engineering.

20:48

* And I will present a paper on natural language processing.

21:05

* It will be actually a workshop that they will give on natural language processing.

21:09

* And I will talk about machine learning in general following the big coverage that we have for Chad.

21:16

* That is a piece of machine learning in the real world by all natural language processing.

21:30

* Bottom line again, machines do not learn.

21:39

* There is no artificial intelligence because there is no intelligence in the systems that we create, or so it appears.

21:45

* And so including the fact that we don't know what intelligence is.

21:54

* So this is something that we already said that. One thing that is really essential in the way we do things in.

21:57

* Data science in a broad sense, including machine learning, is the type of learning that you can provide to assist them.

22:10

* So there are basically two types. One that is supervised and one that is unsupervised.

22:24

* So the supervised the that means you have previous cases are the same event that you want to predict or classify and you use the experience,

22:32

* meaning the patterns from the past, the occurrences of the event to either classify or forecast the future.

22:50

* So that's supervised learning. An example could be the weather.

23:04

* I know the weather in the past X number, number of years and I use that to predict the weather for tomorrow.

23:10

* That's not going to work well, but that's the overall idea.

23:21

* That's supervised learning.

23:25

* Unsupervised learning is when I do not have in the history of the occurrence of the event that I am having to predict or classify.

23:28

* That's the case when I want to just cluster her, my clients.

23:46

* And so I want to launch a new product and I want to be sure that or I want to

23:53

* launch a certain number of new products or a different flavor of the same product.

23:59

* And I want to sell the product or propose the product to the group that is

24:05

* kind of more similar to the potential target for the products that they have.

24:13

* In that case, all you do is to partition the data set in that subset.

24:24

* So that will be as much homogeneous as possible inside the each of the subsets and as much different as possible one from the other.

24:32

* So supervised learning, unsupervised learning is something that is sort of in between is a reinforcement learning.

24:45

* Reinforcement learning either when you do not have supervision, meaning information about the past,

24:54

* but you have at its core that you want to maximize and still think about the game.

25:05

* So you want to reach the highest number of points possible, so you start playing and you want to maximize that value.

25:13

* So in this case. You keep playing till you will get the high value of the score.

25:23

* The cases of failure, meaning the cases where you reach the very low score will be the supervision.

25:36

* The information about the past that you didn't have to begin with.

25:46

* So reinforcement learning at the very end is a self generated sort of supervised learning.

25:51

* So the two main categories are unsupervised or unsupervised,

26:02

* where reinforcement learning is a sort of self supervised learning, but is at the very end a form of unsupervised.

26:07

* When you have a supervised learning, you need to have samples that you want to learn from.

26:20

* So you have your initial dataset with all the cases.

26:30

* The weather for the past.

26:38

* And. Then you split the dataset into two portions.

26:42

* One that you will use for training and one that you will use for testing.

26:48

* So the largest of the two is the one for training.

26:54

* This model for testing typically is 70, 30, 80, 20, something like that.

26:58

* So you use the 70% to create the model.

27:05

* Once you have the model using the proper algorithms, you test the model using the remaining 30% or 20% of the dataset, and you measure the accuracy.

27:09

* Models are a combination of two elements that the data and the algorithm are good.

27:23

* Algorithms with no data means not great models and vice versa.

27:35

* Meaning if you have good data, either no algorithms or inappropriate algorithms, the model will not do much.

27:45

* As an example, Chad GP is using an unbelievable amount of data, so the algorithms are pretty much standard.

27:57

* They are open source.

28:12

* But the data is what is making a difference.

28:17

* So the quality of the result of the jeopardy in this case is high because the data is so big.

28:21

* Then you have the problem of the training model.

28:32

* That will take quite a long time. So we will talk later on about that.

28:36

* I mean, obviously, the largest, the larger the training dataset is, the more complex of the algorithm.

28:43

* So the more resources you will need for training.

28:53

* And then the resources that you need for training are pretty much just once in a while.

28:57

* You do not you don't retrain your model that frequently.

29:04

* So those models are pretty much in tend to be operating in a better way.

29:11

* So they are not operating in real time. So you have the dataset.

29:18

* My GP is working on data out to 2021.

29:23

* So if you ask Chad GP for something that is happening from 2022 to today, you will not get any answer because it has no data on that.

29:30

* But is a massive amount of data relatively complex how good it is?

29:44

* So the combination of the two making the training phase, we need resource intensive.

29:51

* When you use this training, set the you and then you have the mobile load, let's say with a good accuracy.

30:02

* So the accuracy is the number of correct classifications divided by the total number of cases.

30:12

* When you have your model that is performing well and then you apply the model to whatever is your target.

30:20

* So you want to predict the weather today. So you need to be sure that the training.

30:31

* I mean, the dataset that you used to create,

30:40

* the model meaning to train and tested the model is the same or is similar to the data set to the addiction that you want to do.

30:43

* So if you have the data for winter and you want to predict the weather in a day in summer,

30:57

* most likely is not going to work well because you don't have that in that conditions.

31:08

* So the way you select the dataset to create the model that it's really important,

31:14

* then the way you split between the training and the testing is also important.

31:22

* So if you have a dataset with the sales of apparel during the year and you have the training and you're split, I don't know, summer and spring.

31:28

* And then the testing is on winter.

31:46

* So obviously it's not going to work well because the Saints or some of them are different based on the season.

31:50

* So. We mention that the dataset has to be representative by even either even the way we do the split between training and testing.

32:03

* It's really important.

32:17

* So we want to make sure that when you do the split, there is a good representation of all the possible options so that you can have a for the dataset.

32:18

* So you want to be sure that you have the same distribution.

32:31

* And going back to the umbrella, all the seats for the four seasons in both of them.

32:35

* Because only in this case the measurement of the accuracy would be reasonable and as much accurate as possible.

32:41

* So again, when you do this type of learning, you need to to be sure that the distribution of values in the creation of the model,

32:55

* in a broad sense, meaning the combination of training and testing and this beta between training and testing is similar to that.

33:12

* This is part I mean, that the animal that you want to predict is part of that distribution, because otherwise it is not going to work.

33:23

* We mentioned reinforcement learning. So again, reinforcement learning is not, uh, anything completely different from.

33:35

* Supervised learning, but is basically self supervised, meaning the system is generating cases and is getting the results as a training dataset.

33:47

* And then we use it pretty much the same way as it would have done if it was a full supervised learning.

34:01

* So that's basically the bishop is representing that.

34:13

* Obviously, when you can apply that, you can apply that when you have a score.

34:21

* So if you are playing a game, if you are doing any task where you can measure the outcome.

34:25

* If you don't have that, then there is no way that you can do a reinforcement learning.

34:32

* Some of the examples.

34:39

* So you want to move into the board where each move will give you a reward and you want to reach the destination point with the maximum reward.

34:42

* So based on the moves that you do, you will get a certain value.

34:58

* So you can have a system that will generate randomly options.

35:06

* And then we use randomly generating options to create the amount to maximize.

35:11

* Really the same thing. If you are in a network, you want to reach the node aside from the energy or vice versa.

35:20

* Let's say have a supplemental side to the node.

35:32

* That's an example. And there are many options that you want to minimize the number of steps you do.

35:37

* So I will keep those. So you need to define your value and you need to maximize the value.

35:46

* Oh, pretty much see me later or an application.

35:54

* All of the reinforcement learning are a genetic algorithms.

35:59

* Genetic algorithms means that the system is actually generating scenarios randomly and then based on the result is applying.

36:04

* I mean, using that as a supervision, it's they, they named it the genetic because it's kind of mimicking the survival of the fitness.

36:18

* And that is a function of the fitness function that is measuring how far you are from the goal that you want to reach.

36:38

* Deep learning is something that is becoming relatively popular.

36:50

* So. You have machine learning.

36:57

* You have artificial neural networks.

37:02

* Artificial neural networks are algorithms that are based on linear algebra.

37:07

* You have like in this chart here, you have nodes that are equivalent of neurons.

37:15

* So they will get an input along with the wait for the input.

37:25

* Meaning that at the very end,

37:32

* each one of those layers will be a matter that extends and you go from one matrix to the other by multiplying the different Madrick seasons,

37:34

* using the weights and using the input stats.

37:44

* So you have an input layer and output layer and you have one or more hidden limits.

37:49

* So in what is called a shallow neural network, so you have only one key delay in the deep learning.

37:57

* You have a very large number of dominions.

38:07

* Now, if you consider that, you move from one layer to the next one with linear algebra, multiplying matrixes,

38:13

* any, I mean, they can be with a thousand neurons, meaning a thousand elements multiplied by a thousand elements.

38:22

* You'll have a quite a large combination. You cannot really know what's happening inside the when you have a million of you building.

38:31

* So it's called deep learning,

38:41

* but it's pretty much a neural network with a lot of hidden layers and is one of the problems in modern artificial intelligence.

38:42

* Because what is happening inside a deep learning artificial neural network.

38:55

* We really cannot be exact because there is no memory of the different states.

39:04

* It is just a continuous multiplication. So you only have the input and the output.

39:13

* You could have another scene in a network check in the first.

39:23

* But then you are not really doing much in terms of explaining what's going on.

39:30

* There are quite a lot of investments on explainable artificial intelligence realities that when you have deep learning,

39:36

* when you have a lot of those hidden layer, there is no way that you can really explain what's going on.

39:45

* That is a theoretical problem.

39:54

* You can learn what is happening if you have memory, if you don't have memory, and there is no memory at each stage of this.

39:58

* Large neural networks. You cannot really get it.

40:11

* So but anyway, they are effective in particular.

40:17

* I mean, obviously they are all nomadic, although they are they were originally created for recognizing images.

40:21

* Emojis are points are adults, meaning you have large mattresses with the peak cells that are the elements in a border.

40:36

* Somehow that is the picture that you want to imagine that you want to recognize.

40:50

* So initially they been created for that and they were a nomadic outlet.

40:58

* And still there is no way that you can do a neural network for nonmedical elements.

41:05

* So eventually you need to do a transformation from a. non-American to an American.

41:12

* Typical example is the use of machine learning to understand language.

41:20

* So languages, text so is not number.

41:28

* There is no number in a text. I mean, it could be, but it's words, it's alphabetical characters.

41:32

* That's what I'm out of it.

41:41

* So when you have that and if you want to create a representation that is in America, you need to use an algorithm to transform text into numbers.

41:42

* Those algorithms are called the vector is H. There are several of them that are some of them.

41:58

* And basically, it's transforming based on some assumptions.

42:05

* We will talk about that each word in a vector.

42:10

* And that point you can work on numbers and that's the point where you can use numerical only algorithms.

42:15

* So we spoke about deep learning. And again, that's what it is.

42:28

* Now, regardless, what is the algorithm that you are using?

42:36

* You will work with data that can be from all the possible venues.

42:44

* Some of the data can be cleaner because they are generated by machines that are collecting.

42:52

* Events can be messy because they are collected from human beings in different human beings.

42:58

* Eventually, meaning you can have some of the day that with the same meaning but different classification.

43:09

* You can have street as a street as to the other characters or as the daughter or as capital or s as Mola.

43:16

* So with the same meaning, that's an example.

43:30

* So there are quite a lot of not clean data around us, and if you want to play with it,

43:35

* creating your models, you need to transform, you need to clean it.

43:47

* So the data preparation, it's a not so fine step in data mining and machine learning.

43:53

* Keep in mind that. So I don't know how familiar you are if you ever knew the child GPT two train model.

44:05

* They use the algorithms but they also use humans.

44:20

* So humans did quite a lot in different states of this, of the process of creating all the training, the system.

44:24

* One of the steps is to remove inappropriate content so that something that a

44:35

* new man can do and is a terrible job but a human can do and a machine cannot.

44:41

* And then tagging it to make sure that you had the right classification for the different code, for the different elements that you are considering.

44:48

* So large systems like Chop GPT.

45:00

* What they do is basically they create, they detect the pattern in the data and then they match the request you have with the available patterns,

45:06

* compiling them with a layer of compensation.

45:24

* So it's pattern recognition, pattern matching and compensation element.

45:31

* So those three are the three components of something like a child that is part of the GPT class of machine learning systems.

45:38

* That is right now some of the most advanced I mean, for large distribution.

45:53

* Obviously there is always the issue of quality data.

46:01

* Quality. It's a relative term. So if you have a dataset that is several terabytes.

46:06

* If you have a few thousand elements that are not clean enough, don't really matter.

46:14

* But if you have a dataset that is several thousand and you have a few thousand there that are good enough, but then you have a problem.

46:21

* So quality is always in relative terms and is pretty much a function of the size of the dataset you have.

46:30

* The pre-processing is pretty much three phases of cleaning, integration and transformation.

46:39

* So the cleaning air is what we mentioned. You want to remove outlets, you want to remove missing data, things that are damp integration.

46:48

* You may want to have different sources for the target that that you have.

46:59

* You are in marketing,

47:07

* you are targeting potential customers and you want to be sure that you have as many points of view of your customers as possible.

47:09

* So you want to have the information about the the financials, the credit cards, what is doing in terms of easy passage GPUs.

47:20

* So all of those are all repeats visited on the Web.

47:33

* So all of those are with the layer of machine learning can really or activation intelligence

47:38

* can really or a data science can really recreate the profile of the individuals.

47:46

* But the problem with having multiple sources is that the integration is rarely an easy task.

47:54

* So the integration means you have elements that are that can be different but related to the same topic or individual or whatever it is.

48:02

* The difference it can be just formula is a different format or can be structural.

48:13

* So you have one data set with the entities that are covered in a certain way

48:21

* and the other that way are the same entities are called in a different way,

48:28

* or sometimes you don't even have the names of those entities, meaning you need to extract the,

48:35

* the name, the label from the context, the where the data is.

48:43

* And then you have that transformation. Transformation means you may want to combine variables if this may be appropriate.

48:49

* You want to eliminate variables that are not relevant for the analysis that you are doing.

48:57

* You want to normalize eventually. They need to have a more readable and usable dataset.

49:03

* So all of those. Are elements of the pre-processing.

49:12

* Now. If we go into the other algorithms, we mention the unsupervised learning.

49:21

* So the main algorithm in the unsupervised learning category is clustering.

49:31

* Clustering is a technique for finding similarity in data and putting together the elements that are similar.

49:39

* So those buckets are the class sets.

49:49

* So you want the goal of an algorithm.

49:53

* Doing the clustering is basically you wanna one up,

49:57

* you want to create groups that are it's one with two subtasks groups that are as much

50:01

* homogeneous as possible inside and it's much differentiated one from the other.

50:10

* So that's the goal of a clustering.

50:19

* I'll go to a typical level. Clustering algorithm is cleaning to where you define the number of class that you want.

50:23

* So that is key is and this would be decided with the combination of the inputs,

50:33

* the data scientists, but primarily the marketer or whoever is the client that you are serving.

50:43

* Because if you want to clustered, you are your clients,

50:52

* you need to define how many groups you want and you define how many groups you want based on, for example, the number of products you have.

51:01

* I have ten net new products, meaning I need to have ten groups.

51:12

* If I have three products. And in three groups I cannot have ten groups with three products because then it wouldn't make any sense.

51:19

* So but then there are ways to maximize or to suggest the ideal number of clusters.

51:29

* And so now there are measures of that, but the most commonly used is called the elbow or metal, but it is not part of what we are doing.

51:41

* So you define key and then you place randomly the elements into those key markets,

51:53

* and then you measure the distance between the elements and you rearrange the placement of

52:01

* the elements in the buckets to minimize the distance between the centroid of those packets.

52:10

* So we are centroid. So that, I mean that you can have two dimension or n dimension, but the average of the packet.

52:20

* So if is two elements that will be the element in the middle.

52:35

* If the number of dimensions it's 100, then is going to be three, go to see it in a hundred dimensional space, but that's pretty much it.

52:39

* So you move elements to minimize the distance between the centroid and then there is a moment where you cannot move

52:52

* any farther the elements because you are already at the maximum or the minimum or the distance in each group.

53:07

* So and at the end of the process. So this K means and that's what the algorithm is doing.

53:18

* Another very popular algorithm is a decision tree.

53:24

* A decision tree is kind of mimicking what humans are doing.

53:29

* So that's an example. Hey, you want to go to a restaurant, you step into the first one to check how many people inside.

53:34

* Then if the restaurant has no people, you may decide not to go because probably is not a great restaurant.

53:48

* If there are some available tables then you say if is full, then you ask what is the waiting time and based on your tolerance.

53:58

* So for the way that you decide to see or to go based on other conditions, like there are other options in the nearby how hungry I am.

54:09

* So those are. Things that you do consciously or unconsciously when you take a position.

54:20

* So you mentally create a sort of a decision tree, and that's pretty much what the algorithm is doing.

54:29

* You have all the possible options. So let's say that we are in a binary condition.

54:37

* Go, no, go. So you have a combination of options, meaning you have some conditions for going,

54:44

* some conditions for not going, and then you apply the different conditions and you reduce the uncertainty.

54:51

* So like in the previous case, you ask, is the restaurant cooler?

55:03

* What is the waiting time? So those are reducing your options.

55:11

* Somehow the representation that you have on your screen.

55:16

* So you have two options that are represented by the green circles.

55:22

* And the Red Cross is a reddish cross.

55:28

* This pink probably, I don't know.

55:32

* So that initial stage, you have pretty much a random combination of the two.

55:36

* Then you applied a force condition and you had more one type and less of the other type.

55:42

* And then you keep applying conditions.

55:50

* Still, you have no impurity, meaning all of the elements are on one type.

55:54

* So measuring the impurity is also measuring the entropy.

56:02

* So the entropy is higher when you have more chaos, like in the case on the left side,

56:10

* the entropy would be higher and the entropy on the right side would be zero.

56:17

* So if you measure the entropy, you will have a measurement of the entirety of the dataset and the goal of the.

56:24

* A decision tree algorithm is to minimize the impurity, meaning to minimize the entropy.

56:36

* Neural network. I mentioned before that artificial neural networks are a combination of single

56:47

* processing units that are kind of mimicking the neurons we have in our brain.

56:58

* So you have a bunch of input to the same number of weights.

57:05

* They will go into this processing unit that will do summation, and based on their value,

57:10

* there will be a function that will be on or off meaning of fighting or not fighting.

57:17

* So that's pretty much how a single neuron is working.

57:24

* So each neuron as a summation function and then transfer function function.

57:29

* And then you can have an input layer. And he then layered in an output leg.

57:37

* When we have a model, one of the key points is to measure the accuracy.

57:47

* So we generally use two main methods for evaluating the accuracy.

57:55

* One is called the error matrix or confusion matrix, where we measure the number of cases that the model is saying are positive.

58:04

* Ah, yes. And let's say in the cases that based on its testing, the subset are really positive.

58:17

* So measuring pretty much the true positive and the false positive as well as the false negative and the true negative.

58:30

* So in this example, the question was buying a computer, yes or no?

58:40

* So you have a total of 10,000 cases and the model is predicting yes for almost 7000 cases.

58:46

* And those are actual yes and is predicting yes.

58:55

* While the reality is no for 400, the meaning in this case, you have a very good accuracy in predicting yes for no computer.

59:01

* No. Buying a computer? No. Your model is say no.

59:10

* And it is yes. Meaning it's false.

59:15

* Negative. What is six know and is actual no for about 25 to 2600 cases.

59:19

* Now, most of the cases, you don't have the same balanced performance in predicting yes or no,

59:30

* because at the very end it really depends on the dataset you have again.

59:41

* But if you want to predict the cases of cancer in the population of the United States,

59:46

* you definitely have more cases of no cancer for centuries than cases of cancer,

59:54

* meaning you have data set up in terms of representative mass of the two categories.

1:00:00

* It is it's really skewed toward the cancer.

1:00:06

* It would be the same if you want to predict cases of terrorism,

1:00:12

* if you want to predict the number of days of rain in the Sahara Desert, if you were and so on.

1:00:16

* So in that case, it's like human beings.

1:00:27

* If you don't have experience in something, you cannot talk much about that.

1:00:32

* Then we know that there are people that are talking a lot anyway.

1:00:38

* But generally speaking, you cannot probably talk if you don't know about the same metaphor, if you don't have enough cases.

1:00:43

* So in predicting cases of cancer that you can predict better cases of no cancer than the cases of cancer,

1:00:54

* the days of no rain in a day or rain in tomorrow.

1:01:02

* So don't be discouraged if you see that there is a disparity between predicting one or the category compared to the other.

1:01:10

* Go back to your dataset and probably you will see the reason why there.

1:01:22

* Another way to to evaluate the accuracy is the receiver operating characteristics.

1:01:28

* So it's comparing the true positive and the false positive.

1:01:36

* So now you have the main diagonal that is measuring the irrelevance, the flipping the coin.

1:01:42

* So you want to have a curve that is going a bulb where the flipping the coin line.

1:01:49

* So you measure the line under the curve, meaning a desire to acquire the error.

1:01:57

* And the larger the area, the more accurate that is going to be the TV channel the moment if you have cases that are all up to the top,

1:02:03

* meaning an accuracy that is hundred percent most of the time there is something wrong.

1:02:19

* So no model can predict the well hundred percent unless you are using a variable that already has a solution to the problem.

1:02:24

* So if I want to measure, I don't know Ukraine or no rain.

1:02:38

* And one of the variables is the millimeters of rain, obviously.

1:02:45

* When the mm. The value of the variables.

1:02:53

* Millimeters of rain is above zero.

1:02:59

* That is great though. Meaning that variable is giving it up.

1:03:03

* So then yes, you have an accuracy of hundred percent.

1:03:09

* But you already knew that if there is a rain in the street, if there is water in the street, that is rain.

1:03:13

* So that's an example. When you see 100% accuracy, this is a red flag.

1:03:24

* Something is wrong in the dataset. When you have text, things are more complicated because while the data is numbered, its text is not.

1:03:30

* So you need to find a way to create a metaphor and somehow to mine the text.

1:03:47

* So there are techniques. Then that what is main text?

1:03:56

* You can mine text to extract information, but it really depends which type of information you want.

1:04:02

* You need. If you want just to measure the most frequent words like we did, then we would do in next assignments and in particular in the midterm.

1:04:10

* That's an insight that you can get from the attacks, but statistical.

1:04:22

* The main problem is when you want the semantics inside them.

1:04:29

* How similar are two documents?

1:04:35

* How can I extract a summary from a text?

1:04:40

* So those are or what is the opinion of people on something?

1:04:45

* What is the sentiment of people? What are the emotions generated by a text?

1:04:53

* So those are semantic meaning is the meaning of the text.

1:05:00

* So those are obviously more difficult to evaluate and you need to have a mode that you need to have a metaphor to.

1:05:07

* I mean, evaluate that to calculate those volumes.

1:05:20

* But what work is being created in 2013 then?

1:05:25

* Now there are other approaches, but pretty much is a a good starting point to explain how those things are done.

1:05:29

* So the process is called victory zation.

1:05:44

* You basically transform each word into a sequence of numbers in vectors, and then you work with vectors.

1:05:50

* So vectors are points in a m dimensional space.

1:06:02

* So if you have a vector with two components, then this vector would be a point in a regular two dimensional space.

1:06:08

* If you please mute your microphone malfunction.

1:06:20

* So if you have like most of the time we use a vector is Asian, we did that 200 or 300 dimension.

1:06:26

* Then those words will be points in a three dimensional.

1:06:34

* So in 200 or 300 dimensional space,

1:06:39

* we cannot really see that the but it's I mean logically similar to like the we're in a in a 2D space because they are two points in the space.

1:06:45

* You can measure how fast they are and based on how far they are, you can determine how close semantically they are, meaning how similar they are.

1:07:02

* Those two words,

1:07:15

* the way we create the vectors is based on the calculation or the conditional probability of one word appearing in the text because of the other.

1:07:18

* So the way those vectors are calculated that are using shallow neural networks,

1:07:35

* meaning the neural networks with one layer and is already measuring what is after.

1:07:44

* So there are different methods.

1:07:54

* A few years after that, Google created another method that is called the transformer to the transformer.

1:08:00

* Our victories in the bays not only on the conditional probability of one appealing because of the other, but also what was before and what was after.

1:08:11

* So it's I mean, at that point, you need a neural network that is not more shallow but more complex,

1:08:26

* meaning resources for training the or getting the model for a system that is based on the transformer is way much more than one for work to bank.

1:08:38

* So let me skip that and let me go to this one.

1:09:02

* So tools for data science, machine learning.

1:09:07

* It really depends on who you are or what you do.

1:09:13

* How long, how big is your experience?

1:09:21

* So it's a matter of fact that the people who are in data analysis since decades,

1:09:25

* they use less by doing R and similar a lot primarily Python today and they use more tools like SAS or similar that are.

1:09:37

* SPSS did are more with the user with a graphical user interface like Excel on Steroids.

1:09:54

* We do have faculty that are in the same condition.

1:10:07

* So I graduated as a Ph.D. like five or six years ago and to me is kind of normal to write code for a faculty who graduated 15, 20, 30 years ago.

1:10:12

* For them, coding is not that natural.

1:10:31

* So they probably coded at that time in Fortran or C C++, but now they are not much into that.

1:10:35

* So the professional age group, there is a correlation on the professional age and the I mean,

1:10:46

* it is not discrimination and eventually I would discriminate myself and 65 are going 66 or so, but that's a matter of fact.

1:11:01

* And then obviously it depends on what you are doing.

1:11:12

* If you are a computer scientist, obviously you have no problem using coding.

1:11:16

* If you are an economist, then you may be more reluctant in using those.

1:11:21

* When you work with a python, you can do everything with basic python,

1:11:32

* but you may want to use libraries or a framework that will make your life easier.

1:11:40

* When I started working. With machine learning.

1:11:47

* I mean, at the time there was not much of machine learning or artificial intelligence in general.

1:11:56

* So it was a 1986. Pretty much we didn't have by done that.

1:12:01

* So we had languages that were very basic.

1:12:08

* Meaning if you wanted to do an algorithm for, I don't know, decision tree, you need to build it from scratch.

1:12:14

* So now you run a python, you call a skill learner,

1:12:23

* and you use one of the functions that is one function for generating a decision tree and you're good to go.

1:12:29

* So life is easier now. So you have a. Languages that are easier to begin with.

1:12:38

* The bite is relatively easy. You have a million libraries of medium, but you have a several thousand libraries that will make your life easier.

1:12:47

* And then eventually you have a collection of libraries that are that can help you building the pipeline that you will need for your task.

1:12:58

* So when we go into Python and the libraries in data science, believe it or not, the most frequent,

1:13:12

* the most commonly used library is pandas, because at the very end, you have data and you need to represent the data.

1:13:20

* So that's why you use PANDAS to create the somehow the data structure.

1:13:28

* Some people are using them.

1:13:35

* There are other options, but they are very niche.

1:13:40

* So pandas are definitely the go to solution for data science.

1:13:44

* It's basic, so it's not adding a lot of functionality.

1:13:51

* There are no, I don't know, regression clustering, a decision tree algorithm in it.

1:13:56

* But in order to do them, you need to have the data in a data structure.

1:14:07

* And that's what's pandas doing. Elena is basically the next step.

1:14:12

* So there are libraries for all the algorithms that we mentioned before and many other end is built on known by site,

1:14:18

* by locally, meaning all of those are embedded in the library.

1:14:31

* Just a warning. When you work with library says that it's so big.

1:14:37

* Try to import the new code. Only the portions that you need then not the entire library.

1:14:43

* Because the more you import in your code, the bigger your code will become.

1:14:50

* If you then have also a large dataset that you may run out of memory.

1:14:57

* So be mindful and you can import only the portions that you need.

1:15:04

* But again, a skill learner is the way to go.

1:15:10

* Being based in Mumbai, that means that is based on arrays as a data structure.

1:15:14

* And so the combination of arrays and pandas is definitely the way to go.

1:15:26

* But that's a comparison between some of the libraries and.

1:15:32

* Now there are other options. One of the things that I'm not regretting, but right now we have a new Mac that is based on Apple silicon.

1:15:44

* I mentioned that last class.

1:16:00

* The New Macs, they have a graphical processing unit that is working along with the central processing unit,

1:16:05

* meaning that they cannot be addressed individually.

1:16:19

* Most of the libraries in machine learning are based on TensorFlow.

1:16:24

* That is a great library, but is addressing a separately GPU in a CPU, meaning all the libraries are based on TensorFlow.

1:16:33

* I'm not really working on my neck, so there are some way around.

1:16:45

* You can build a virtual machine, but there are not many virtual machines that are running on those Macs or that are.

1:16:50

* Sort of an adapter. But those are affecting the performances quite by the law.

1:17:02

* So I mean by torture.

1:17:10

* It's growing in popularity. Another big fan of Facebook.

1:17:13

* Facebook created by torture. But the very end, I'm using it because I can not use TensorFlow anyway.

1:17:20

* So that's pretty much it in terms of slides.

1:17:29

* Let me go for a second here. Let me show you just a brief example of.

1:17:35

* If I don't create that using some of those algorithms that I mentioned in class.

1:17:46

* So that K means. Uh.

1:17:56

* And then it's decision trees. So those are the two algorithms that I'm using.

1:18:00

* Knew script. I use the quite a lot of comments.

1:18:08

* Where I basically I copied from the Python library and the parameters that you can provide to the different modes.

1:18:16

* So parameters are really important when you do this job because changing parameters would change quite a lot of results.

1:18:28

* So I'm importing the different libraries, so I importing again and using Chrome because it's a killer and it's pretty big.

1:18:37

* I want to import only K means I want to import the data sets and then I don't pay pandas closely.

1:18:48

* But I was getting warnings and I imported the library to ignore the warnings is not elegant, so I didn't have time to go back.

1:18:56

* And the reason primarily the warning was generated by one of the libraries and then was using that.

1:19:09

* So I cannot go into the I could buy but it wouldn't be anything intensive so I'm loading the dataset is

1:19:19

* I this is one of the most commonly used dataset for a class setting is basically how to create class.

1:19:26

* That's all different types of hierarchies based on the length and the width of the sample and the length of the petal.

1:19:34

* So the model itself is basically two lines.

1:19:48

* So it's not that much, again.

1:19:54

* And if you do it manually, it would be probably a good between 102 hundred lines that are in the library.

1:19:57

* In this case means library from a skill learner, and it would be super easy in this case.

1:20:07

* So then the representation decision tree similar think.

1:20:17

* Importing the library is a. And I ordered imported.

1:20:24

* The dataset. So same thing. Defining all the parameters.

1:20:34

* Splitting the dataset into training and testing.

1:20:41

* And then same thing, the representation. So you will have the file and you will play with it.

1:20:48

* So I'm running it. So is generating a PDF with a decision tree.

1:20:55

* So on the decision tree you have. This is the root node.

1:21:03

* Her. And then you apply the different options and you will get that at the very end.

1:21:09

* Am not saying that we are running out of time. So you will have the difference anyway.

1:21:17

* Then. This is the cluster.

1:21:26

* Oops. Three clusters for.

1:21:30

* The dataset that I was using.

1:21:42

* All right. Okay. So let me go back here and let me go to the in-class exercise.

1:21:48

* So the reason why why I created this exercise is because back in time, I let students pick their own final intense problem.

1:21:57

* And that does it for a strange reason.

1:22:08

* A huge number of students in a relatively new age decided to go with analyzing us baby names at the very end.

1:22:12

* I couldn't be sure that what they did was original.

1:22:27

* So I decided to take this out and making it an in-class exercise.

1:22:33

* So us baby names. You want to read the file into a pandas data structure?

1:22:41

* Read the structure, delete the call on unnamed zero and IED.

1:22:50

* Determine if there are more female or male being the top five in terms of number of occurrences.

1:22:56

* But name meaning or frequency of the name, the number of names in the dataset,

1:23:04

* the standard deviation of the name occurrence, and some basic descriptive statistics in the dataset.

1:23:09

* That is a one line in pandas.

1:23:18

* All right. So let me make sure that this assignment.

1:23:24

* It's been posted. So I. Published the.

1:23:32

* Q that size. And if I.

1:23:39

* Okay. You have it. Let me stop shedding.

1:23:53

* Let me start. So there are four breakout rooms with two or three participants per room.

1:24:01

* So in creating it, let's say 10 minutes.

1:24:09

* All right. See you in a bit. All right.

1:24:14

* So it's 801, and we don't want to keep you too long.

1:34:57

* So I will post the solution, but we will not discuss it.

1:35:04

* So my apologies for talking too much during the lecture about I really hope that was useful or at least interesting for you.

1:35:09

* Um. Let me, uh, just.

1:35:21

* For 5 minutes about the meter mount.

1:35:27

* So let me go. Right.

1:35:36

* Everything. Yep.

1:35:43

* Okay. So let me share the screen. And.

1:35:49

* All right. So can we close? Let me go over your.

1:36:04

* So the midterm the midterm a. One think that is essential.

1:36:11

* For the regular on campus classes four and 624.

1:36:23

* Uh, they do the meter, uh, in 2.5 hours.

1:36:30

* So, I mean, it's both a regular on campus and regular online.

1:36:37

* So you will have more time. So you will have until Sunday.

1:36:48

* But that's less than the normal time. So if you think about asking for an extension, it is should be I mean,

1:36:54

* a serious reason for that because you are already getting more than I should I normally give for the midterms.

1:37:07

* All the other modalities we run AM 624.

1:37:21

* So the. The assignment has two parts.

1:37:27

* I mean, apart from liquids. One part that is a choking code.

1:37:36

* So you have some scripts.

1:37:43

* And I also copied those scripts in a note by FILA, so it would be easier for you instead of copying and pasting those,

1:37:49

* and then there would be issues with the indentation.

1:38:04

* So you have three different scripts that may work or may not work.

1:38:08

* So you have the description and the actual script.

1:38:13

* You will run it, you would check it, and eventually you would fix it.

1:38:17

* So those are the three, uh, smaller pieces of code to check.

1:38:23

* Then there are two more, uh, that are new to.

1:38:30

* Right. They are smaller. So you read the, the file transfer fee into a list of words eliminate from the list the words that are in the stop word file.

1:38:36

* So the so called a stop order is a thought is a word with no semantic meaning articles, pronouns, things like that.

1:38:52

* So you want to read the modified and eliminate what is in the so-called the from the original text.

1:39:03

* Then you will calculate the least frequent words.

1:39:11

* The average occurrence of words where occurrence is the frequency is the number of times a word is appearing in the text.

1:39:16

* The longest word. The average word length.

1:39:25

* Again is based on unique words, so each word counts as one.

1:39:30

* The last of the writing scripts is using the file cards.

1:39:38

* They'll see as we you read the file into the structure and then he's about cards.

1:39:44

* So they have several characteristics, including horsepower and average mileage.

1:39:52

* So you want to print the three cards with the lowest average mileage,

1:39:59

* the highest average mileage and the highest ratio horsepower divided by average mileage.

1:40:04

* So. You have the list of files you have also this one with the files that you want to check and eventually fix.

1:40:14

* And you should be good to go. So for.

1:40:30

* Section one and two, meaning the part with the checking code and writing code, you will write one single dot by file.

1:40:36

* So not too far. Not enough files for each of the state of the parts, but one single file with everything.

1:40:48

* And that's basically it. So you should have her think.

1:40:58

* So I'm sorry if I always staying longer than 8:00.

1:41:05

* Questions? I have one.

1:41:12

* I have to write one on the bequest.

1:41:17

* That was for this module. I.

1:41:22

* It was on. I was the first one. It was. Which one of, you know, A, B, C or D is not a software development model.

1:41:26

* And I happen to like I chose the V model because that's traditionally a systems engineering model and

1:41:35

* also was the one that wasn't in your software development PDF also it was the one it was the CNN.

1:41:41

* But arguably like E model is a systems engineering principle, not a software engineering principle.

1:41:49

* Okay. I will review that. And then I keeping in mind that that system engineering and software engineering, they share several principles.

1:41:55

* So I generally call software engineering system engineering applied to computer science, but I would definitely check it.

1:42:04

* Okay. And then my other one was. Let's do Sunday is that, you know, it's not measurable.

1:42:14

* I just feel nervous about the midterm date being due Sunday being that I know for most of us.

1:42:22

* Working full time, having less like Saturday and Sunday.

1:42:30

* Only two full days we would have to. Yeah, you are going to have both the days.

1:42:34

* So the due date is the end of Sunday.

1:42:40

* Right. That's that's what I meant.

1:42:45

* So having Monday evening for the homeworks is a nice cushion just in case working over the weekend isn't enough.

1:42:48

* I know for me these homeworks take me at least like 8 hours to get fully correct, but I just figured it was worth the shot.

1:42:55

* Well, I mean, I understand.

1:43:05

* I, I really need to be, uh, consistent across the different modality.

1:43:08

* So in this case, it was. It was a frustrating one time.

1:43:19

* I understand that. But, I mean, I kind of did a sort of a compromise,

1:43:22

* meaning it's kind of giving us wood and our fireworks and giving several days, including the weekend.

1:43:27

* So you can work during the weekend because I mean,

1:43:36

* you are professionals and you work during the week, but the weekend, so you have the entire weekend.

1:43:40

* So I thought that was. Not saying genitals because that seems to be not appropriate.

1:43:47

* You've got the sentence, I guess. Yeah. All right. Thank you for answering those.

1:43:56

* Absolutely. Other questions.

1:44:01

* Yeah. Gavin, I have a question regarding the grading on this.

1:44:05

* Oh, what do you call that? You know, at the end of a typical homework, it asks us to interpret the results.

1:44:10

* How does the grading work for that? Is there like a template? If we missed certain words, like, let's say, for the city situation,

1:44:16

* if I didn't mention that people would often ride bikes during the spring because it's a warmer climate versus the versus winter because it's colder.

1:44:24

* Is having points taken off? What was your thoughts on that point there?

1:44:37

* Well, when you have questions that are more on the qualitative side, I mean, things that are always questionable.

1:44:42

* So think about sport. If you have a gymnastics.

1:44:57

* So how are you apart from things that that obvious the other kids fail but how you evaluate one

1:45:02

* execution compared to another one so you have a human being the very end judging the results.

1:45:11

* Human beings are fallible, as we know, and pretty much every one has a different opinion about things.

1:45:21

* And in this case, when you when we evaluate the interpretation,

1:45:35

* we want to be sure that you are using all the metrics that you extracted in the proper way,

1:45:43

* meaning you use the metrics or the metrics to make the story.

1:45:51

* So one of the most common way to do the interpretation and is the wrong one is to describe the process.

1:45:58

* So I prepared the data, I loaded the data, prepare the data, I did this operation.

1:46:08

* And then the other operation does the wrong way because we are asking for an interpretation, not an explanation of the results.

1:46:15

* So. Uh, most of the points that we take out are for two reasons.

1:46:26

* One reason is because there is no inside sense or elements of inside, so that can be derived from the matrix.

1:46:34

* Most of the time, again, that is students describing the process instead of the results.

1:46:50

* And the second is when you have, let's say, ten matrix, but you are actually using two.

1:46:58

* So yes, it is insightful, but it is not as insightful as it could be or it should be.

1:47:04

* So I don't know, in this case in particular, if you send me an email just to remind me, I will review it.

1:47:13

* But typically when we take points off because of one of those two reasons.

1:47:23

* Thank you, Professor Atmosphere. Thanks for the. Sure.

1:47:33

* All right. Other questions. All right.

1:47:38

* So that's the end of the class.

1:47:45