Rutuja Kore  
helps payers manage their **provider network.**

So basically a **payers have different kinds of networks with their providers and they want to optimize their network.**

which **provider is causing them higher cost**

bad reviews by the members so that when **next year they renewed their contracts**,

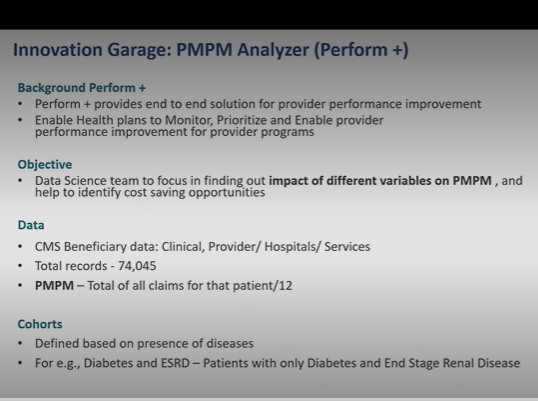
they're aware or other than that, if they know, OK, this performer, this particular **provider does not have any particular kind of facility**, but many of the Members are reaching out there for this facility.

Citiustech already had a solution called as **perform Plus** which is like an **interactive dashboard kind of solution where graphs have been shown** and like **payer can view how their providers are behaving at a glance.**

However, the important factor for which the data science team was approached by **perform plus team was to identify the PMPM.**

**They wanted to know PMPM analyzer utilization**

Rutuja Kore  
Let's try to understand **what PMPM analyzer**.



PM PMis nothing but per member per month cost.

**the cost that is incurred by payer for a particular provider particular member in entire one year** like. total of all claims for that particular patient divided by 12 that is per member per month cost so using.

OK, so uh, so why did we want to do this?

To **enable health plans to monitor, prioritize and enable provide a performance improvement for provider programs.**

So data science team was to focus in **finding out impact of different variables on PMPM** and help to **identify cost saving opportunities**.

**list of Variables which are affective affecting my PMPM may positively negatively**

ex:**Geographical area, so particular states are the ones which are causing you high PMPM**. Then what are the characteristics of that particular geographical area. So this much was there ask give us only features which are affecting PMPM.

important factor is suppose **we cannot compare PMPM forTwo different people who have two different diseases or age group.**

For example, if somebody is between the **age group of 20 to 25 and does not have any comorbid condition,** then obviously we **that person will not have high PMPM because here she will be only occasionally utilizing the services**.

**Whereas if there is another person who is 60 plus maybe and has said diabetes, hypertension or cancer along with that, then we would expect that persons.**

those **Members who belong to similar characteristics in terms of clinical progress--** we had defined cohorts and then we started comparing their PMPM

Analyzing and so on to **identify what are the variables that are affecting PMPM**, how are they affecting and then we will understand how we approach the problem.

But to begin with it was very clear that it will be different for different groups of people.

**So we defined cohorts as per the diseases**.

so **we had clubbed few diseases together according** to after discussing with clinicians \

cohorts were basically defined based on the presence of disease,

example diabetes and end stage renal disease and so on.

So these are the patients who had only these two decisions there.

There was another cohort where people had only diabetic,

there was another one where people had only cancer and

then in that cohort we tried to find what are the factors that are affecting PMP.

OK. So on the high level, let's try to see what was your approach to deal with this problem so?

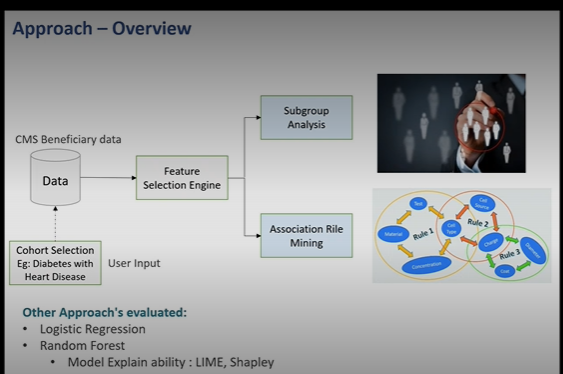
0:6:37.0 --> 0:6:55.670  
**CMS beneficiary data in hand and cohort selection was done as per the disease present.**

They just wanted to know **which are the features that are affecting most considering the interrelationship between the variables**.

**Classic statistical model is logistic regression**.

So **whether this person will have high, high PMPM yes or no**.

tried random forest and tried doing like model explain ability.



**It was not really linear and then doing any kind of predictive analytics** would be like if the data changes, it will entirely change and we are **not sure what kind of data we will encounter**. That is why this we **decided to go ahead with unsupervised techniques** which will be **more flexible** and there can be a scope to the domain expert in order to understand or interpret the rules that we get out of any algorithm.

So since **we had multiple features, we did feature selection**.

\ finalized top features were consumed by these two algorithms.

**Subgroup analysis and association rule mining**.

we have done for 2/3 disease groups. So for example you are the **user from business team who doesn't understand statistics at all or any machine learning anything**. But you know I'm interested in diabetes and hypertension people and what are the factors affecting their PMPM. So there will be a **drop down menu in the dashboard where they can select diabetes and hypertension**. They can select disease and then that only that much data will be consumed by our algorithm for engine we can see.

feature selection engine.: taking out only top features which have higher contribution- Python library called as feature whiz.

What this does is **it removes highly correlated features to remove noise from the data** and it **utilizes xgboost algorithm to talk find top features so.**

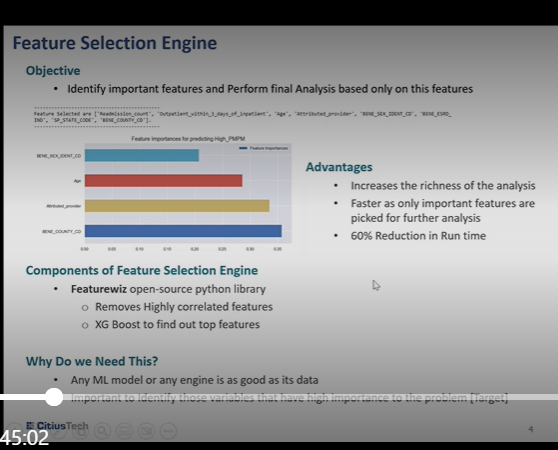
Objective of adding this feature selection engine was to **identify important features and perform final analysis based on these features** we will do **subgroup analysis and association rule mining** and these are **kinds of algorithm which will give you as many rules as many features** you will provide to them. A

nd then like if suppose **I give 5 features to these algorithms but in that if I have in each of them five categories.**

\ association rule mining treats :**each category of each variable as a separate variable altogether**, so it **will be like 25 features**.

So if imagine if each **feature has only 5 categories and we are giving only 5 features**, then also we are in. Actually it is being **consumed as 25 different features** so it is very difficult to interpret at a later stage and it is computationally expensive as well.

Similar is the case for subgroup analysis. That is **why we are actually doing feature selection**.



\ only **important features are picked up and 60% reduction in runtime adds**. Both of these algorithms are computationally quite expensive.

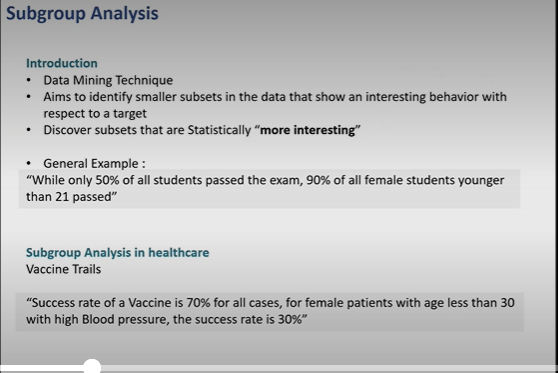
\ important to **identify those features that have high importance in the target identification**.

So in our data we found **that beneficiary sex age attributed provider** who is like **specific provide like a family docto**r we have in India. Similarly they have **attributed providers** there who is assigned to each member and then **county code**. These were the top features too like give you high idea.

we have utilized both the algorithms. We have **kept it open as in whichever gives more feasible rules, user can utilize them.**

**Both of them are like independent algorithms** and we are you like running both of them. It is up to the user which one to utilize because it was like difficult to choose one of them given that we were working on a synthetic data, it was not a real data. So we have kept both of them.

\ subgroup analysis: is basically a **data mining techniqu**e which **aims to identify smaller subsets** in the data that show **an interesting behavior with respect to a target**. Now subgroup analysis is a **usupervised technique, but.**



\  
Well, it is not **really unsupervised in a way, because we have to provide target variable** in here. So like it is we can call it as a **supervised clustering kind of algorithm**. We will go in the details of the algorithm later on. But.

to start the algorithm, **you have to give a target column**. Here it was **PMPM which was categorized into high, medium and low based on the quantiles** and then it **discovers subsets that are statistically more interesting** finds out groups in the given data.

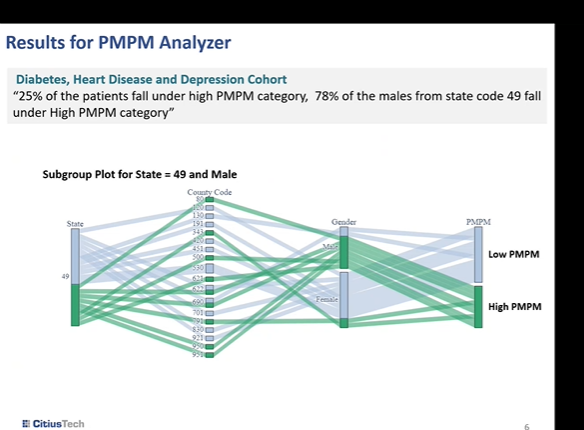
While only **50% of all student passed the exam, 90% of all female students younger than 21 past**. So this is like more information in like 50% of students who have passed among them like 90% of all students were of age greater than like lesser than 21. So similarly another example in Healthcare is vaccine trial so.

\  
**Success rate of the vaccine is 70% for all cases, but for female patient with age less than 30 with high blood pressure, the success rate is 30%.**

\f subgroup analysis and so you get multiple rules with different support or say confidence. \

But then they also **have freedom of choosing those rules**, which might come up as with lesser confidence or with lesser support, because lesser number of.

\PM PM analyzer, the results of were like one of the example you can see is we had utilized it for diabetes, hypertension and depression cohort and one of the observation was

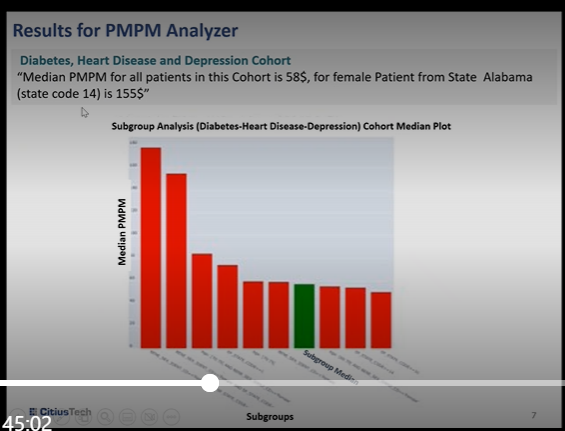


**25% of patient which fall under high PMPM category. 75% of the male from state code 49 fall under high PMPM category. So of all of these 25% of all high PMPM 75.**

78% of the mails from state 49 are them. S

o maybe then the user can go ahead and see **what is the state 49 and why** are the **male from this particular region are from this particular cohor**t they are \

Another view of, like uh looking at overall results is a median PMPM for all patient in the cohort, the same cohort as before is **$58 for female patients from State Alabama. That is state code 14, it is 155**. So maybe they can go and look into Alabama State and why are **female population is having such a high utilization** and then they can **look into the features that are caught contributing towards high PMPM utilization**



association rule mining:

common example like you see whenever you go to **a supermarket**, there is usually a pattern in which people buy certain things. So it was like it started with a it another name of this algorithm is **Market Basket analysis, Market Basket.**

\ **supermarket survey where they tried to find what are the items that are byte together.**

\ **milk and butter is already in a card**

milk, butter, bread and other items and then frequency.

**Members frequency of buying those items and using this data this algorithm**

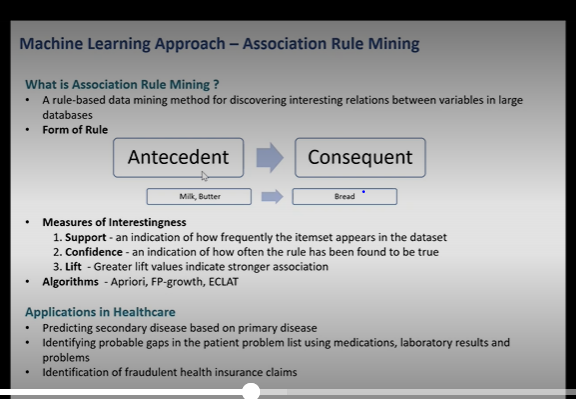
Essentially association rule interesting relationship between variables in larger databases.

**support of milk would be if it is bought by eight people out of 10**,

Take it will see how many times **milk and butter occur when milk and butter was bought**

**greater lift values indicate stronger association and lift will not be in between zero** and one.

**association is found to be occurring multiple times and it is quite stronger association.**



So there are multiple algorithms -**Co occurrences, most oldest one and famous and common one** is **apriori algorithm**.

However, it is very **computationally expensive** because it will treat each of these features like **milk, butter, bread** and whatever items that are present as **a separate entity** and then for each like it will calculate. First of all it **will calculate support for each item**.

milk and butter, milk and bread, milk and sauce, milk and jam

Then it will consider 3 pairs at a milk, butter bread, milk, butter.

Ohh toast and like this so it is **computationally very expensive**, FP growth algorithm.

**FP Growth algorithm** works is it is kind of a **tree algorithm**.

unimportant or less frequent items are deleted.

computationally it is more feasible and easy.

Ality is small developed version of FP growth.

**Ohh PMPM analyzer, we have utilized FP growth algorithm.**

how does **association rule mining work**

predicting **secondary diseases based on primary diseases** like many times people go to the doctor and patient might not be able to tell s**ymptoms properly which can be guessed using the primary symptoms in or symptoms of primary disease**, so that treatment or care could be started for those diseases in like advance.

Another application is **identifying probable gaps in patient problem** list using medication laboratory results and problems because patients are not really aware of what all is going on with them.

something to the doctor, which is **noted in the EMI data**. We have laboratory results and medications, but the things which are lost which are not really covered, that **can be guessed using this association rule mining technique**, 30s **identification of fraudulent health insurance claims**

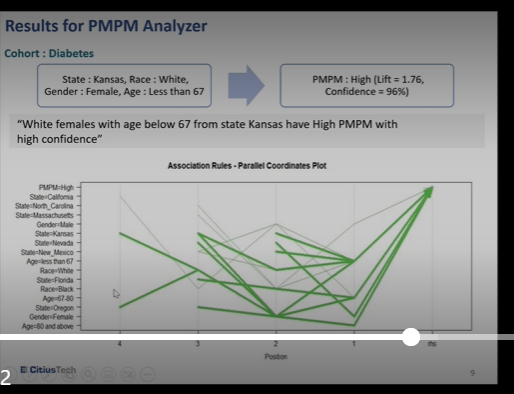
We have used **this subgroup analysis which is identifying the which provider is taking like expensive** kind of thing.

**association rule mining**. particular Disease what all possible disease can be

**we are using it for same thing like subgroup analysis**.

diabetes cohort then how our model has like a association rule mining has given us is **state consists occurred in the like along with that race white gender female is less than 67** leads **to high PM PM** and the **confidence is 96%** and **lift is 1.76** like how how we can interpret it?

woman with white rays with age less than 67 belonging to Kansas State have high PMPM with very high confidence.



You can see like this one had the this one was with the very high confidence. Another rule that you can look here is **State Oregon here in the top like. Then that person was with white race again female then New Mexico no age less than 67 had high PMPM.** This is how you can interpret you will get multiple rules but it is up to user which one.

Like which **one of them they find appropriate for their analysis**, and then it is up to them which one to utilize, which one are actually true because this **this algorithm will give you all possible occurrences**, some of them could be just by chance or they have occurred multiple times because of some reason.

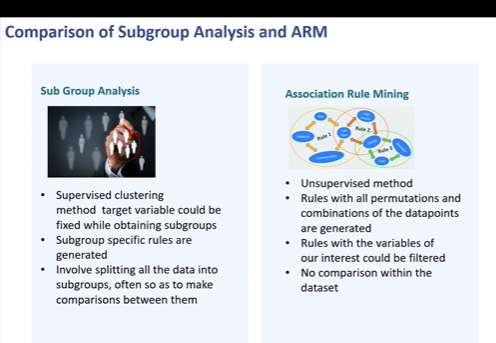
I'll tell you another example in this supermarket chain itself, one thing that they had noticed was if on **Saturdays male patient stands to buy Diapers and beers together**, and supermarket chain was not able to identify why this is occurring.

What is the correlation between them? So later on they found out because on for some reason on Fridays.

Uh, for others used to go for **supermarket uh purchase and those who had babies for them**, they they had to buy diapers. But since more male population was going to do this purchase, more beers were also sold.

here there is no **correlation between diaper and beer**. But then they found out this OK, this was occurring. And then they arranged beer near.

could be some other reason lying behind it and that was another reason why we have kept both of these algorithms because we might not be sure that which one of them will give you better results at any given time. So that's why we have kept it open.



But again, to compare these two algorithms, first we'll see like **subgroup analysis is supervised clustering** method, **where target variable could be fixed while obtaining subgroups**. So you can fix OK, I want **only high PMPM related rules** or only low PMPM related rules. So that's how you can fix it. **Association rule mining is totally unsupervised method subgroup specific rules are generated**.

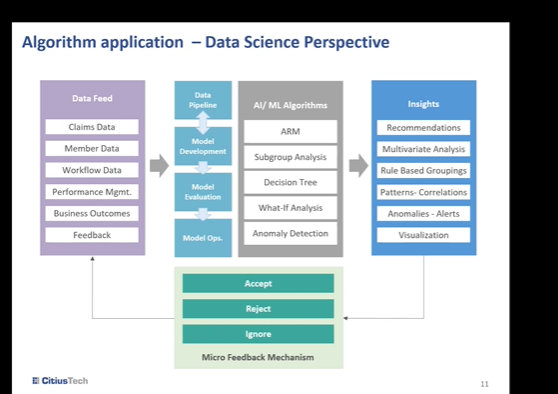
Even though, like **we are giving cohort of say diabetes and hypertension and depression** and so on under that also if you want OK I want **only to particular state** so that you can **specificity could be maintained**

whereas in **association rule mining rules with all permutation combinations of the data points are generated.** It is **not very specific to any specific subgroup**.

Then **subgroup analysis involves splitting all the data into subgroups often so as to make comparison between them.**

Whereas in **association rule mining rules with the variables of our interest could be filtered** **and no comparison within dataset occurs in this association** like in this algorithm.

For example, if I'm interested in **geographical location according to my domain knowledge**, I think that in **particular state I find more cancer patient** then I can look at geographical variable and filter accordingly **which facility might not be available with subgroup analysis**.



Applications

you have different kinds of datasets like member data, claims data, workflow data and so on, you can utilize it in the model development or model evaluation.

along with decision tree. What if analysis and alarmingly detection ARM and subgroup analysis can be utilized

insights we get is **recommendation, rule based grouping, pattern correlation analysis**.

**fraudulent claims detection ARM** can be used so **anomaly detection** and **sending alerts or just high level visualizations.**

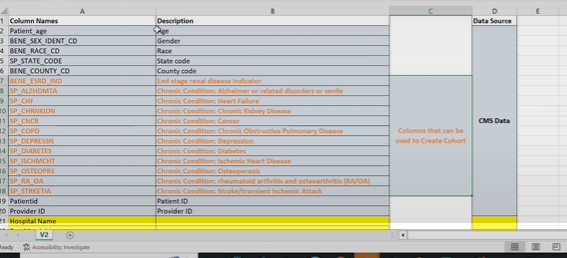
Any kind of **multivariate analysis**, all these places where we can utilize these **two algorithms as well**,

interactive dashboard with different rules for different cohorts and identifying

data:

data dictionary for our data

. This is CMS data which you might be aware of **patient age, gender, race, state, county code.** Then these red colored ones were.



Red coloured: used for cohot creations

Ones which were used for cohort creation. yes or no kind of.

These are kind of **binary variables**, then patient ID provider ID. Some of the **features we had included from data dot Word hospital datasets** because that information was not available in CMS data like **provider information for example hospital name, provider address, provider, City, state, ZIP code, hospital type emergency.**

0:33:52.340 --> 0:34:7.440  
Rutuja Kore  
Services if available, yes or no,

then rating for that particular hospital overall readmission, national comparison.

This this is something which is calculated by U.S. government and it is stored in the data like it was.

I think it was done on provider ID itself, but if not then.

0:34:44.550 --> 0:34:45.40  
Varsha Khamitkar  
Uh.

0:34:57.300 --> 0:34:57.560  
Varsha Khamitkar  
Umm.

0:35:3.140 --> 0:35:3.700  
Varsha Khamitkar  
OK.

0:34:42.910 --> 0:35:12.780  
Rutuja Kore  
Ohh code was you guessed then this information was used from CMS provider data. OK now I understand we did not join this data dot world information directly with CMS data. We also had CMS provider data. So we first joined these two datasets and then using provider ID we have joined it with member data. So in provider dataset we had a total beds in the given hospital psychiatric unit tests different numbers of like.

Uh, like blood bank? Yes or no cat label? Yes or no? How many cardiac catheterization laboratory services are provided by a hospital? So basically, information giving on what what all kinds of services that are available in given hospital like pharmacy.

**PMPM was calculated by us using the claims data** as I explained earlier. So this is how the overall data look like. Now I'll show you the actual data as well.

OK, so this is provider ID attributed provider

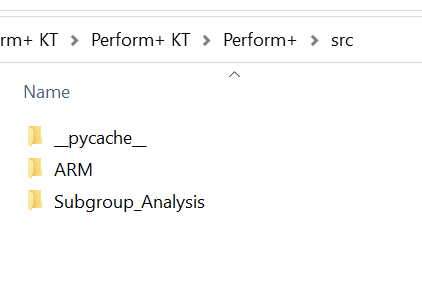
**Family doctor, they have attributed provider** so each provider will have attributed provider and as of now we have mapped each patient with attributed providers details because it could be possible that same patient had different claims with different providers. But to simplify it at this moment we have used attributed providers details then PMPM calculation then in patient we did it's outpatient visits age.

reason why provider is causing high PMPM. It is possible that.

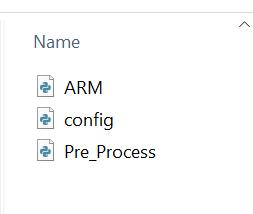
have built totally **automated pipeline** **where you just have to give the information about paths of datasets, original data,** then preprocess data and then where you want to store the rules that you get out of it. We have two different.

Src-

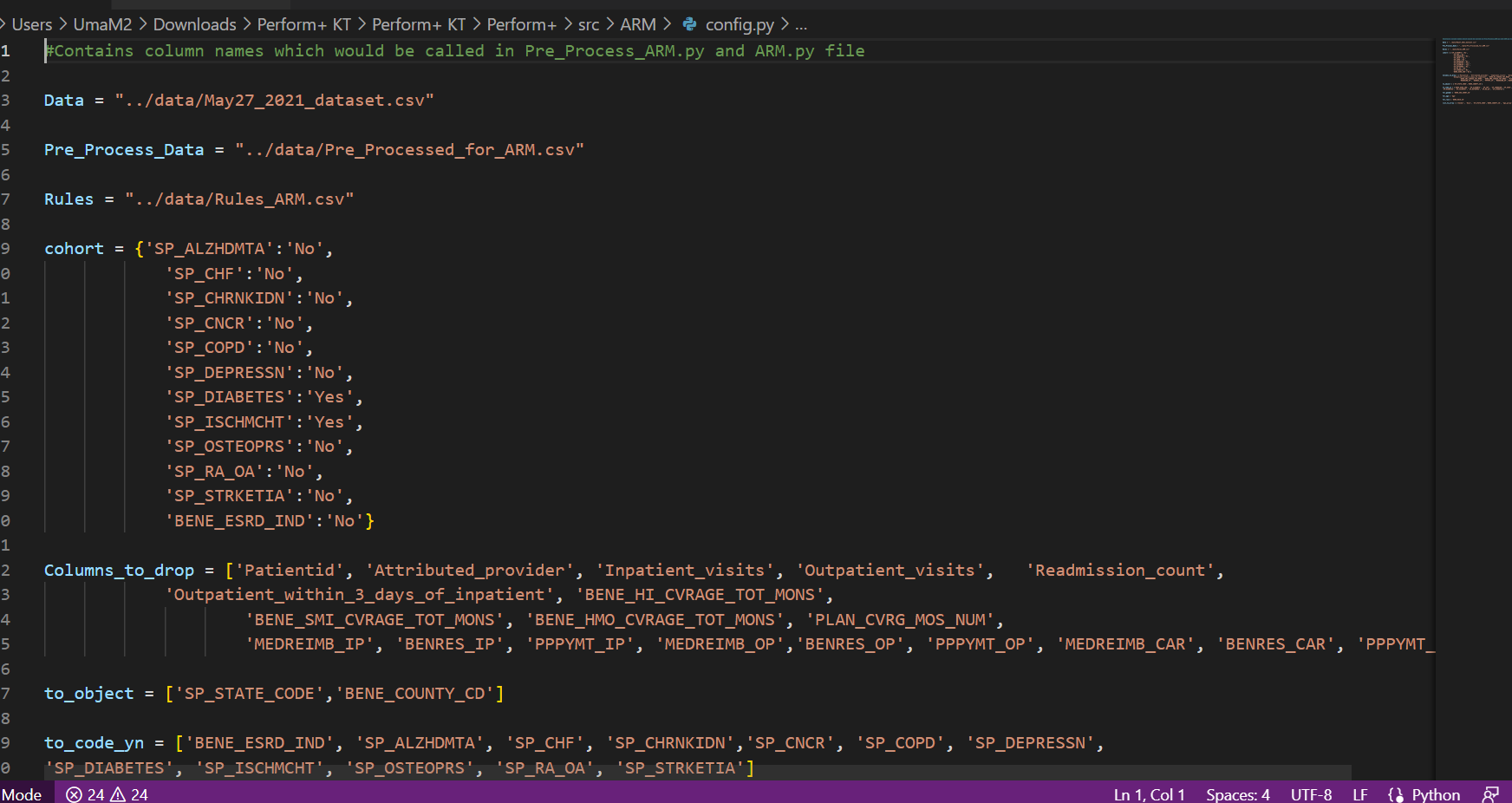
Arm



In Arm folder



Config file



We found that some of the columns will **not so important** according to the like feature selection, we have done so we **have dropped patient a**nd also they're **redundant as well since say in patient visit outpatient visits these are already utilized well calculating PMPM**.

state county codes are in numbers, so they were considered as a numerical quantities. **So we have changed them to object** and so on. S

o **yes, no type of we have changed for these particular things** because it was 1 0, we kept changed it to **yes, no** to make it simple to understand.

So this is the **config file where you can change things**

**then in preprocess file we** **have written functions to change as per our need**. For example for mail.

You mean 10? How it will like one and 012 was there in the original data. **We will change it to male. Female. Then racists were also recoded. Age groups were created then**. This is core of the disease like how to this is **create a disease cohort, prepare a cohort** then all the **preprocessing like like** I said in the config file it will take it will be done in the preprocessing steps. Then for creating the data each cohort will be joined. How?

Then we have we are joining **PMPM column with it and then PMPM category** is also created based on the quantiles. So then **feature selection using that feature wiz**.