

## **1 - 1 - Course Background and Practical Information**

Welcome to this course, process mining, data science in action. The goal of this short movie is to provide you with some background information on the instructors, the organizations, organizing this course, and some practical information. TU/e is the university organizing this MOOC. TU/e is one of the leading technical universities in Europe and it has a rich history in engineering. For example, in computer science Dyskra a famous touring award winner, worked here. He was, for example, the person that invented the shortest path algorithm. But there are not only scientists in this region. In the broader Eindhoven region, which is called Brainport, there are many high tech companies like Phillips, ASML and many others. So it's a vibrant region with many organizations producing and using data. That is the reason why this university started the Data Science Center in 2013. The Data Science Center is a virtual organization consisting of 25 research groups covering all aspect of data science. So in the yellow box you see the disciplines that are involved to produce data and to handle large volumes of events. In the red box you see all kinds of disciplines related to analysing event data. And of course, process mining, the topic of this course, is a very important ingredient in this. In the green box, you can see the context part of data signs. So in the end, we would like data signs to be used in practice. So we need to think about ethical aspects and the business value of data. Let's tell a bit about myself. So my name is Will van der Aalst. You can see here a picture of my favorite hobby, walking and climbing in the Alps. I'm 48 years old. I, I'm married. I have four children. And if I'm not walking in the mountains and spending time with my family, I'm having these roles. So, I'm a professor at the Technical University of Eindhoven. I also have part-time appointments off at the other side of the world. For example I am a joint professor at QUT, that's Queenswood University of Technology in Brisbane, and I am the Scientific Director of a lab in Moscow at the High School of Economics. I am also the scientific director of the data science

center at Intel. So in the last 25 years, I've been working on these topics. So I started working more in the area of modeling and using models toward analysis. And in recent years, I've moved more to data driven types of analysis, like process mining, an essential ingredient for data sites. Now I would like to hand over to Joos Buijs, who will tell you much more about the practical things, related to this course. Thank you very much Will, so my name is Indie Joos Buijs, I did my master's here in Angova, and I recently defended my PHD thesis on process mining also here in Angova. And currently I'm a postdoctoral researcher in a group of So a little bit about my hobbies outside of work and so I like to watch Pixar movies and everything involved. I like to visit theme parks and one of my reasons, most recent hobbies is playing with my eight month old daughter and together with my wife and my daughter explore the world a little bit. But this is not about me. Let's talk about the course. So in the next couple of slides I'll explain a little bit more about the details of the course. You can also find this in the study guide and in the grading policy. This video is just a brief introduction, more details are in these documents. And if there are any, if there's any inconsistency the documents prevail. But this is just to get you into the into the course and what you can expect. So there are six weeks of video lectures. Per weeks we have roughly eight videos of eight to 15 minutes each. the, several topics that we discuss. So there's a brief introduction and general data mining things. Then we go into the models and pros of discovery. Then in week three you get little bit of better understanding of the different types of process models there are. This is followed in week four by more discovery techniques and also conformance checking techniques. In week five we look at how you can enrich the process models, and in week six we discuss operational support and we also conclude this lecture, this course. So learning objectives. After this course I hope you see how process mining fits in the big scheme. Of all the techniques like simulation, machine learning machine learning and data mining. but, and also we hope that you see what different

aspects process mining entails. So you can do process discovery, conformance checking, and in the end, hopefully you can also actually do a process mining project from beginning to end, including data gathering and such. In this course you can get two certificates. So the normal certificate you get when you have 35 out of 50 points you can get with the week and the final quiz. So every week there will be a week quiz where you can get five points four, and there's a final quiz worth 20 points, and if you get enough points on these quizzes, you get the normal certificate. If you would like to obtain the certificate with distinction then do, do these week of final quizzes. But also do a Tool Quiz where you experiment with the tools that we use. And also you have to do a Peer Assignment which, which I will explain later. And, if you want to get the Certificate with Distinction, you have to earn at least 80 out of 100 points. So, in more detail that's several quizzes. So every week there's a week quiz covering the topics of that week. Here you can earn five points per quiz. And a quiz contains of ten multiple choice questions. There's no timing, and you get two attempts and the highest score counts. At the end of this course you have a final quiz. This allows you to earn 20 points. It, you get 20 multiple choice questions and there's one timed attempt. Then if you want to go to, for the certificate with distinction you have a tool quiz worth 10 points, which is an untimed quiz of 20 questions. Then next to this is a peer assignment. You can earn 40 points. It will start end of week three early week four, and the goal is to write a process mining report using one of the data sets that we will provide. We will also proved a report template in Boleta and Word document. And you can use this to analyze the data and write your findings. Then you have to submit it. And then you have to also evaluate three other reports using criteria that we provide. And in a similar way, your report is evaluated by three peers based on the same criteria. For the peer assignment, you can use these two tools. So on the one hand we have the ProM framework. This is an academic tool, with many plug-ins and model types supported. It started here at

this University, but many researchers and practitioners from around the globe have contributed to it. And you can obtain it at the processmining.org web site. On the other end, you have Disco, which is a commercial tool. It's simple, fast and easy and it mainly supports fuzzy models. And in another video I'll explain more about Disco, but you can install it from fluxicon.com. Next to this we have also a forum, please use this to ask for any help and report issues that you have with tools. So use this to discuss any things or insights that you've gained also with fellow students. Even also use this to arrange meetups so arrange a local meetup where you'll meet with other students to discuss things in person. We will monitor this forum, so we hope that you among yourself start a discussion and answer the questions. But three of our PhDs in this group and myself will monitor this. So Eduardo, Shengnan, and Maikel will answer any questions that are remain, that remain open. And I will also have a look on the forum. So please use the forum to kind of meet each other. The scores is mainly based on the process mining book, so in, at the end of each video, we'll, we'll point out which chapter we were covered. Some chapters will be provided to you for free. But you can buy the whole book either physical or digital, using a discount kindly offered by Springer. Only for students participating in this MOOC. Well that's it from the more organizational side I really encourage you to watch all the videos, learn all the tools that we have so can analyze your own data. But please be aware this is only the tip of the iceberg. For each technique that is discussed, we did a lot of reasearch, and there's a lot of mu, more detail available in scientific papers that you can, find online. So if you find something that you want to learn more about outside of this book, then please, look online, and there's much more knowledge that we can share with you. So, to conclude to this introductory lecture have a lot of fun in the coming six weeks. Watch the videos, make the quizzes. Do the assignment. And I hope to meet you a little bit on the forum. Have fun.

## **1 - 1 – Base do Curso e Informações Práticas**

Bem-vindo a este curso, mineração processo, a ciência de dados em ação. O objetivo deste filme é curto para lhe fornecer algumas informações básicas sobre os instrutores, as organizações, organizar este curso, e algumas informações práticas. TU / e é a universidade organização deste Mooc. TU / e é uma das principais universidades técnicas na Europa e tem uma história rica em engenharia. Por exemplo, em ciência da computação Dyskra um famoso vencedor do prêmio de turismo, trabalhou aqui. Ele foi, por exemplo, a pessoa que inventou o algoritmo de caminho mais curto. Mas não são apenas os cientistas nesta região. Em toda a região Eindhoven, que é chamado Brainport, existem muitas empresas de alta tecnologia, como a Phillips, ASML e muitos outros. Portanto, é uma vibrante região com muitas organizações produção e utilização de dados. Essa é a razão pela qual esta universidade começou o Centro de Ciência de dados em 2013. O Centro de Ciência de Dados é uma organização virtual composto por 25 grupos de pesquisa que cobrem todos os aspectos da ciência de dados. Então na caixa amarela você vê as disciplinas que estão envolvidos para produzir dados e para lidar com grandes volumes de eventos. Na caixa de vermelho que você vê todos os tipos de disciplinas relacionadas com a análise de dados de eventos. E, claro, mineração processo, o tema deste curso, é um ingrediente muito importante neste processo. Na caixa de verde, você pode ver a parte contexto de sinais de dados. Então, no final, gostaríamos sinais de dados a ser utilizado na prática. Então, nós precisamos pensar sobre aspectos éticos e o valor de negócios de dados. Vamos falar um pouco sobre mim mesmo. Então, meu nome é Will van der Aalst. Você pode ver aqui uma foto do meu passatempo favorito, caminhada e escalada nos Alpes. Eu sou 48 anos de idade. Eu, eu sou casado. Tenho quatro filhos. E se eu não estou andando nas montanhas e passar o tempo com a minha família, eu estou tendo esses papéis. Então, eu sou um professor da Universidade Técnica de Eindhoven. Eu também tenho

compromissos a tempo parcial fora do outro lado do mundo. Por exemplo, eu sou um professor comum em QUT, que "é Queenswood University of Technology, em Brisbane, e eu sou o diretor científico de um laboratório em Moscou na High School of Economics. Eu também sou o diretor científico do centro de ciência de dados em Intel. Então, nos últimos 25 anos, tenho vindo a trabalhar sobre estes temas. Então eu comecei a trabalhar mais na área de modelagem e uso de modelos para análise. E nos últimos anos, eu mudei mais para os tipos de dados dirigidos de análise , como a mineração processo, um ingrediente essencial para sites de dados. Agora eu gostaria de entregar a Joos Buijs, que lhe dirá muito mais sobre as coisas práticas, relacionadas com este curso. Muito obrigado Will, então meu nome é Indie Joos Buijs, eu fiz o meu mestre está aqui em Angova, e eu recentemente defendi minha tese de doutoramento sobre a mineração processo também aqui em Angova. E atualmente eu sou um pesquisador de pós-doutorado em um grupo de So um pouco sobre meus hobbies fora do trabalho e assim por Eu gosto de assistir filmes da Pixar e tudo que está envolvido. Eu gosto de visitar os parques temáticos e uma das minhas razões, passatempos mais recentes está a brincar com minha filha oito meses de idade e, juntamente com a minha esposa e minha filha explorar o mundo um pouco. Mas isto não é sobre mim. Vamos falar sobre o curso. Assim, no próximo par de lâminas Vou explicar um pouco mais sobre os detalhes do curso. Você também pode encontrar isso no guia de estudo e na política de classificação. Este vídeo é apenas uma breve introdução, mais detalhes estão nesses documentos. E, se houver algum, se há alguma inconsistência prevalecer os documentos. Mas isto é só para obtê-lo no no curso e que você pode esperar. Então, há seis semanas de aulas em vídeo. Por semanas, temos cerca de oito vídeos de oito a 15 minutos cada. os, vários temas que discutimos. Portanto, há uma breve introdução e gerais coisas de mineração de dados. Então vamos nos modelos e profissionais de descoberta. Então, em três semanas você começa pouco de uma melhor

compreensão dos diferentes tipos de modelos de processos que existem. Isto é seguido na semana quatro por mais técnicas de descoberta e também técnicas de controlo de conformidade. Na quinta semana, veremos como você pode enriquecer os modelos de processo, e na semana seis discutimos apoio operacional e também concluir esta palestra, este curso. Assim, os objetivos de aprendizagem. Após este curso eu espero que você ver como mineração processo se encaixa no grande esquema. De todas as técnicas como a simulação, a aprendizagem de máquina de aprendizado de máquina e mineração de dados. mas, e também esperamos que você vê o que os diferentes aspectos de mineração processo implica. Assim você pode fazer a descoberta de processos, verificação de conformidade, e no final, espero que você também pode realmente fazer um projeto de mineração processo do início ao fim, incluindo a coleta de dados e tal. Neste curso, você pode obter dois certificados. Assim, o certificado normal, você começa quando você tem 35 de 50 pontos que você pode obter com a semana e o questionário final. Assim, a cada semana, haverá um quiz semana onde você pode obter cinco pontos de quatro, e não há um teste final, vale 20 pontos, e se você conseguir pontos suficientes sobre esses quizzes, você obter o certificado normal. Se você gostaria de obter o certificado, com distinção, em seguida, fazer, fazer estes semana de testes finais. Mas também fazer um quiz ferramenta onde você experimentar com as ferramentas que usamos. E também você tem que fazer uma atribuição de pares que, o que eu vou explicar mais tarde. E, se você deseja obter o Certificado de Distinção, você tem que ganhar pelo menos 80 de 100 pontos. Assim, de forma mais detalhada que é vários quizzes. Assim, a cada semana há um quiz semana cobrindo os tópicos dessa semana. Aqui você pode ganhar cinco pontos por questionário. E um questionário contém dez questões de múltipla escolha. Não há tempo, e você terá duas tentativas e as maiores contagens de pontuação. Ao final deste curso, você tem um teste final. Isso permite que você ganhe 20

pontos. É, você tem 20 questões de múltipla escolha e há uma tentativa cronometrado. Então, se você quiser ir, para o certificado com distinção você tem um quiz ferramenta vale 10 pontos, o que é um questionário de 20 perguntas de duração indeterminada. Em seguida, ao lado esta é uma atribuição de pares. Você pode ganhar 40 pontos. Ele vai começar a final de semana três cedo semana quatro, e o objetivo é escrever um relatório mineração processo usando um dos conjuntos de dados que irá proporcionar. Vamos também provou ser um modelo de relatório em Boleta e Word documento. E você pode usar isso para analisar os dados e escrever suas descobertas. Então você tem que apresentá-lo. E então você tem que avaliar também outros três relatórios utilizando critérios que nós fornecemos. E de um modo semelhante, o relatório é avaliada por três pares com base nos mesmos critérios. Para a atribuição de pares, você pode usar essas duas ferramentas. Assim, por um lado, temos o quadro de baile. Esta é uma ferramenta acadêmica, com muitos plug-ins e tipos de modelos suportados. Começou aqui a esta Universidade, mas muitos pesquisadores e profissionais de todo o mundo têm contribuído para isso. E você pode obtê-lo no site processmining.org. Na outra ponta, você tem o disco, que é uma ferramenta comercial. É simples, rápido e fácil e que apoia principalmente modelos fuzzy. E em outro vídeo que eu vou explicar mais sobre o disco, mas você pode instalá-lo a partir de fluxicon.com. Próximo a este temos também um fórum, por favor use este para pedir qualquer ajuda e denunciar questões que você tem com as ferramentas. Então, usar isso para discutir todas as coisas ou ideias que você ganhou também com outros estudantes. Mesmo também usar isso para organizar meetups assim organizar um meetup local, onde você vai encontrar-se com outros estudantes para discutir as coisas pessoalmente. Vamos acompanhar este fórum, por isso esperamos que você entre si inicie uma discussão e responder às perguntas. Mas três dos nossos doutores nesse grupo e me vai acompanhar isso. Então

Eduardo, Shengnan, e Maikel irá responder a quaisquer perguntas que são permanecer, que permanecem em aberto. E eu também terá uma olhada no fórum. Então, por favor use o fórum para tipo de conhecer uns aos outros. As pontuações baseia-se principalmente no livro de mineração processo, de modo a, no final de cada vídeo, vamos, vamos apontar qual capítulo, foram cobertos. Alguns capítulos será fornecido a você de graça. Mas você pode comprar todo o livro, quer físico ou digital, usando um desconto gentilmente oferecido pela Springer. Apenas para os estudantes que participam neste Mooc. Bom é isso do lado mais organizacional eu realmente incentivá-lo a assistir a todos os vídeos, aprender todas as ferramentas que temos para que possam analisar seus próprios dados. Mas lembre-se esta é apenas a ponta do iceberg. Para cada técnica que é discutido, fizemos um monte de pesquisas, e há um monte de mu, mais detalhes disponíveis em trabalhos científicos que você pode, encontrar online. Então, se você encontrar algo que você quer aprender mais sobre fora deste livro, então, por favor, procure online, e há muito mais conhecimento que podemos compartilhar com você. Então, para concluir a esta palestra introdutória tem um monte de diversão nas próximas seis semanas. Assista aos vídeos, fazer os testes. Faça a atribuição. E eu espero encontrá-lo um pouco no fórum. Diverta-se.

## **1 - 1 - Course Background and Practical Information (END)**

---

---

## 1 - 2 - Introducing Fluxicon & Disco

Hi there. I'm Anne Rozinat from Fluxicon, and I would like to welcome you to this process mining MOOC. Process mining is a very interesting topic. Very relevant for today's organizations. Based on just data you can automatically discover processes. And this can be done very fast and very quickly and for many different processes in these organizations. There are so many processes and so much data is around. Well, process mining is already picking up in several companies nowadays, and those companies who are using it, they get lots of value out of it and are expanding their usage. But many more companies will start using process mining in the future and will become an indispensable tool for process analysts, process managers, and auditors, and, and many other process roles. So, it's not just a very interesting topic that you chose, but also a very relevant one. And, there's no better person to learn process mining from than professor Bill Finoust, who is also called the godfather of process mining. Well, as for us at Fluxicon, we caught the process mining back more than ten years ago. When we were still studying in Germany we came across this topic and then we went to Eindhoven to do PhD in the process mining group of And this was a really great time. Well, first of all, lots of other colleagues also working on different process mining topics, but also we had the chance to work with industry partners such as and Phillips. And in those projects what we learned first of all is that the processes were really much more complex. And we had the chance to develop our algorithms to meet those, yeah, complex processes in the real world. Back then most things were really just working based on toy examples. So for example, the fuzzy miner was one of the first algorithms that, yeah, really was able to deal with very complex and heterogeneous processes. So that was one thing. But, at the same time, what we notice is that people were always really amazed and enthusiastic about what we could show them when we took their data and made their processes visible for them. They couldn't believe that that was possible. So there was this huge gap between

what we were able to do, and what they thought were possible, or what they were doing in their everyday life. So, for us that was the trigger that we said, well we have to do something with it. And we started Fluxicon as a company to make business, yeah, process mining software for business users. and, yeah, with the goal to really make this technology accessible to businesses around the world. And well, when we started out we just started doing consulting for Jake's with prom, and the academic tool set. yeah, under, under, under our arms, which we knew very well because we had developed large parts of it. And we learned what kind of questions people had and what kind of problems they wanted to solve with process mining. And then, we started developing our own software [COUGH] to, to address those needs and to kind of build the perfect tool for those business users that we met in those projects. So, this is what Disco is about, Disco stands for discovery. And Disco it's on the market now for, yeah, since 2012. And [COUGH].So, our prom is really a great academic

tool set, which makes it very easy for our researchers develop, to develop new algorithms. For example, Disco's really meant for a business professional, who does not need to know anything about how these algorithms works, work, and how the technology looks like under the hood. We have hidden that away from them. But we make it available for them, so that based on their domain knowledge and process knowledge, they can yeah, use these results and draw the right conclusions and interpretations from it. Now with Disco, process mining becomes really easy. You have an Excel sheet for example, with some data, and you can simply import the CSV or Excel sheet in a visual way, map the different columns to case ID activity name and time stamp, and then once you import you are directly taken into the process map. You can interactively determine the level of detail that you want to look at the process map. And then you can interactively very easily answer different questions about the process through filters. For example, performance filter that focuses on the long running cases, taking

longer than an expected time. And focus in on those bottlenecks that you can then find and see where things are really taking very long in the process. Now, you can even animate that process in a, in a very easy way, and this brings the process to life, and helps a lot in communicating the problems at hand and, so that people can point their finger on the place where the problem occurs, and discuss why this is happening, and what they can do about it. Now well, with as a participant of this MOOC, we make Disco available for you for the time of the MOOC with a training license. You can use this license to import data sets up to 1 million records. And this is enough to do all the exercises, but you can also import some of your own data if you have any. And in fact, we would encourage you to do that. Because there's no better way to learn about process mining than to, yeah, to play around with some of your own data and to get a good understanding of what it can really do for your own business. And yeah. So we really hope this will spread the, the, the knowledge about this technology which still very few people know about, that it even exists. And yeah, I just want to mention also that we have an academic initiative with close to 200 universities all around the world where we provide lecture materials and also free academic licenses. So you can take a look at our website to find more about it as well. Now I would like you to realize that process mining is really a very special yeah, data science technique or data analytics technique because yeah. Most of those other data analytics techniques they make assumptions about the process. Or, if you think of, for example BI, business intelligence technology, you have to set up an implementation project. And it's all very heavyweight. With process mining you just need to get some data and you can start mining. It's very lightweight and very powerful. But, what is especially interesting about process mining, is what it makes possible for the business user. So, let's take a look at this graph. Well, the business users what they have access to typically is, well first of all, specific reports and dashboards that are made specifically for them. They're very accessible for them, but it's

also kind of limited, because those reports they include exactly that kind of information that was pre-pro, programmed for them to contain. Well, Excel, on the other hand is a very popular tool among professionals not without reason. Excel is a very powerful tool, and you can do lots of things with it. And it's something that the, the business user can use themselves. On the other end of the spectrum, you have those data science techniques that everyone talks about today in the big data craze. And well, that's, those techniques are very powerful but at the same time you really need to be an engineer or a specialist to use them. Now process mining, on the other hand, is particularly interesting because it's very powerful. It's much more powerful for analyzing processes than Excel, because processes consist of multiple steps, and that's a thing that's very difficult to do in Excel. But it's still accessible for the business user. So the business user who has the knowledge about their business process and about the problems in the process, they can use process mining yeah, in their own hands to really yeah, get great results based on that analysis. And so, if you think that process mining is just a way to automate something that was being done manually before. Like previously we were modeling processes, manually modeling them by hand, now we can do it automatically with process mining. That falls way too short. Process mining can do much more than that. And in a way, you can compare it with spreadsheets because when, before spreadsheets were around, people had to do everything manually using a calculator. So if you wanted to do a projection or some kind of forecast, you had to type it in and you had to manually calculate it. But once spreadsheets were around, it's not just that the things that you did manually before those could now, now be automated, but much more things could be done. Many more things could be done. And yeah, so it gave you much more power to do all kinds of things that nobody thought about were possible before. Process mining works in the same way for processors. And is as revolutionary as spreadsheets for, for numbers. So, for business professionals, who

have the knowledge to draw the right interpretation and make the right decisions based on those results, process mining really brings kind of special powers to them. So we kind of, that's how we like to see our, our Disco users also, that we can now give them super powers for their own processes. Well I'm really glad you're participating in this MOOC. I would like to welcome you to the process mining community and I'm sure you will love this topic as much as we do. And happy mining.

## **1 - 2 - Apresentando Fluxicon & Disco**

Olá. Sou Anne Rozinat de Fluxicon, e gostaria de recebê-lo neste Mooc mineração processo. Mineração processo é um tema muito interessante. Muito relevante para organizações de hoje. Com base apenas dados que você pode descobrir automaticamente processos. E isto pode ser feito muito rapidamente e por muitos processos diferentes nestes organismos. Há tantos processos e tantos dados está ao redor. Bem, mineração processo já está pegando em diversas empresas nos dias de hoje, e as empresas que estão usando, eles obter lotes de valor fora dele e estão expandindo seu uso. Mas muitas mais empresas vão começar a usar o processo de mineração no futuro e irá tornar-se uma ferramenta indispensável para analistas de processos, gerentes de processos, e auditores, e, e muitas outras funções do processo. Portanto, não é apenas um tema muito interessante que você escolheu, mas também muito relevante. E, não há pessoa melhor para aprender mineração processo a partir do que o professor Bill Finoust, que também é chamado o padrinho de mineração processo. Bem, quanto a nós em Fluxicon, nós pegamos o processo de mineração de volta mais de dez anos atrás. Quando ainda estávamos estudando na Alemanha que veio este tema e, em seguida, fomos para Eindhoven para fazer doutorado no grupo de mineração processo de E este foi um tempo muito grande. Bem, em primeiro lugar, muitos outros colegas também trabalham em diferentes temas de mineração processo, mas também tivemos a oportunidade de trabalhar com parceiros da indústria, tais como e Phillips. E naqueles projetos que aprendemos antes de tudo é que os processos foram realmente muito mais complexa. E nós tivemos a oportunidade de desenvolver os nossos algoritmos para atender a essas, sim, processos complexos no mundo real. Naquela época, a maioria das coisas foram realmente apenas trabalhando com base em exemplos de brinquedo. Assim, por exemplo, o mineiro difusa foi um dos primeiros algoritmos que, sim, realmente era capaz de lidar com processos muito complexos e heterogêneos.

Então isso foi uma coisa. Mas, ao mesmo tempo, o que notamos é que as pessoas sempre foram realmente surpreso e entusiasmado com o que poderíamos mostrar-lhes quando levou seus dados e fez os seus processos visível para eles. Eles não podiam acreditar que isso era possível. Portanto, não havia esta enorme lacuna entre o que fomos capazes de fazer, eo que eles pensavam que eram possíveis, ou o que eles estavam fazendo em sua vida cotidiana. Assim, para nós, que foi o gatilho que disse, bem que temos que fazer algo com ele. E começamos Fluxicon como uma empresa para fazer negócio, sim, software de mineração de processo para usuários corporativos. e, sim, com o objetivo de realmente tornar esta tecnologia acessível a empresas de todo o mundo. E assim, quando começamos nós apenas começamos a fazer consultoria para Jake com baile, eo conjunto ferramenta acadêmica. sim, sob, sob, debaixo de nossos braços, que sabíamos muito bem porque tinha desenvolvido grandes partes dele. E aprendemos que tipo de perguntas que as pessoas tinham e que tipo de problemas que eles queriam resolver com a mineração processo. E, em seguida, começamos a desenvolver nosso próprio software [COUGH] para, para atender a essas necessidades e tipo de construir a ferramenta perfeita para os usuários de negócios que conhecemos nesses projetos. Então, isso é o que é de cerca de Disco, Disco significa descoberta. E Disco é no mercado agora para, sim, desde 2012. E [COUGH]. Assim, o nosso baile é realmente um grande acadêmico conjunto de ferramentas, o que torna muito fácil forour os pesquisadores a desenvolver, para desenvolver novos lgorithms. Por exemplo, Disco de realmente significou para um profissional de negócios, que não precisa saber nada sobre como esses algoritmos funciona, trabalho, e como a tecnologia parece sob o capô. Nós ter escondido que longe deles. Mas nós torná-lo disponível para eles, de modo que, com base em seu conhecimento de domínio e conhecimento do processo, eles podem, sim, usar estes resultados e tirar as conclusões e interpretações corretas a partir dele. Agora,

com o Disco, mineração processo torna-se muito fácil. Você tem uma folha de Excel, por exemplo, com alguns dados, e você pode sim, basta importar a folha de CSV ou Excel de uma forma visual, mapear as colunas diferentes para caso nome da atividade ID e carimbo de tempo, e, em seguida, uma vez que você importar você está diretamente tomados em diagrama do processo. Você pode interativamente determinar o nível de detalhe que você quer olhar para o mapa do processo. E então você pode interativamente muito facilmente responder a perguntas diferentes sobre o processo através de filtros. Por exemplo, filtro de desempenho que incide sobre os casos de execução longa, demorando mais do que um tempo esperado. E se concentrar em esses gargalos que você pode, então, encontrar e ver onde as coisas estão realmente tomando muito tempo no processo. Agora, você pode até animar esse processo em um, de uma forma muito fácil, e isso traz o processo para a vida, e ajuda muito em comunicar os problemas na mão e, de modo que as pessoas podem apontar o dedo sobre o local onde o problema ocorre, e discutir por que isso está acontecendo, eo que eles podem fazer sobre isso. Agora bem, com como participante deste Mooc, fazemos Disco disponível para você para o tempo da Mooc com uma licença de formação. Você pode usar essa licença para importar conjuntos de dados de até 1 milhão de registros. E isso é o suficiente para fazer todos os exercícios, mas você também pode importar alguns dos seus próprios dados, se tiver algum. E, de fato, nós incentivamos você a fazer isso. Porque não há melhor maneira de aprender sobre mineração processo do que, sim, para brincar com alguns dos seus próprios dados e para obter uma boa compreensão do que ele realmente pode fazer para o seu próprio negócio. E sim. Então, nós realmente espero que isso vai se espalhar o, o, o conhecimento sobre essa tecnologia, que ainda muito poucas pessoas conhecem, que ele ainda existe. E sim, eu só quero mencionar também que temos uma iniciativa acadêmica, com cerca de 200 universidades em todo o mundo onde nós fornecemos materiais de aula e também

licenças acadêmicas livres. Assim, você pode dar uma olhada no nosso site para saber mais sobre ele também. Agora eu gostaria que você perceber que a mineração processo é realmente uma técnica muito especial sim, técnica da ciência de dados ou dados de análise, porque sim. A maioria dessas outras técnicas de análise de dados que eles fazem suposições sobre o processo. Ou, se você pensar, por exemplo, BI, tecnologia de inteligência de negócios, você tem que criar um projeto de implementação. E é tudo muito pesado. Com mineração processo, você só precisa ter alguns dados e você pode começar a mineração. É muito leve e muito poderoso. Mas, o que é especialmente interessante sobre a mineração processo, é o que torna possível para o usuário corporativo. Então, vamos dar uma olhada neste gráfico. Bem, os usuários de negócios que eles têm acesso a tipicamente seja, bem antes de tudo, relatórios e dashboards específicos que são feitos especificamente para eles. Eles são muito acessíveis para eles, mas é também uma espécie de limitado, porque esses relatórios que incluem exatamente que tipo de informação que foi pré-pro, programado para que eles contêm. Bem, Excel, por outro lado, é uma ferramenta muito popular entre os profissionais não sem razão. Excel é uma ferramenta muito poderosa, e você pode fazer muitas coisas com ele. E isso é algo que o, o usuário pode usar-se. Na outra extremidade do espectro, você tem essas técnicas científicas de dados de que todos falam hoje nessa loucura de dados. E bem, isso é, essas técnicas são muito poderosas, mas, ao mesmo tempo que você realmente precisa para ser um engenheiro ou um especialista para usá-los. Agora mineração processo, por outro lado, é particularmente interessante porque é muito potente. É muito mais poderosa para a análise de processos do que Excel, porque os processos consistem em várias etapas, e isso é uma coisa que é muito difícil de fazer no Excel. Mas ainda é acessível para o usuário corporativo. Assim, o usuário de negócios que tem o conhecimento sobre o seu processo de negócio e sobre os problemas no processo, eles podem usar o processo de

mineração, sim, em suas próprias mãos para realmente sim, obter grandes resultados com base nessa análise. E assim, se você acha que a mineração processo é apenas uma maneira de automatizar algo que estava sendo feito manualmente antes. Como anteriormente estávamos modelar processos, manualmente modelá-las à mão, agora nós podemos fazê-lo automaticamente com a mineração processo. Isso cai muito curta. Mineração processo pode fazer muito mais do que isso. E de certa forma, pode compará-lo com folhas de cálculo porque, quando, antes de planilhas estavam ao redor, as pessoas tinham que fazer tudo manualmente usando uma calculadora. Então, se você queria fazer uma projeção ou algum tipo de previsão, você tinha que escrevê-la no e você tinha que calculá-lo manualmente. Mas uma vez planilhas estavam ao redor, não é apenas que as coisas que você fez manualmente antes aqueles podia agora, agora ser automatizado, mas há muito mais coisas que poderia ser feito. Muitas coisas mais poderiam ser feitas. E sim, por isso deu-lhe muito mais poder para fazer todos os tipos de coisas que ninguém pensou sobre eram possíveis antes. Mineração processo funciona da mesma forma para os processadores. E é tão revolucionário como planilhas, para os números. Assim, para os profissionais de negócios, que têm o conhecimento para desenhar a interpretação correta e tomar as decisões corretas com base nesses resultados, a mineração processo realmente traz tipo de poderes especiais para eles. Então nós meio que, é assim que nós gostamos de ver nossos, nossos usuários Disco também, que agora podemos dar-lhes superpoderes para seus próprios processos. Bem, eu estou realmente feliz que você esteja participando deste Mooc. Eu gostaria de recebê-lo para a comunidade mineira processo e tenho certeza que você vai adorar este tema, tanto quanto nós. Feliz mineração.

## **1 - 2 - Introducing Fluxicon & Disco (END)**

---

---

## 2 - 1 - Lecture 1.1- Data Science and Big Data

Welcome to this first lecture of the course on Process Mining: Data Science in Action. In this first week, we focus on the relationship between Process Mining and other types of data analytics. Today, I start by discussing the broader topic of Data Science and Big Data. Today, many people say data is the big oil. And this is illustrating the incredible amount of data that we are collecting and the corresponding value. If you would think about, how much data was generated from pre-historic times, until 2003. And we think, what we can now do it at. We are able to generate such amounts of data in just 10 minutes. So, in 10 minutes, we are generating, as much data now. As we were doing from prehistoric times until 2003. So this is illustrating the incredible growth of data. So, what kind of data is this? What kind of event data are we generating? Well, we are generating, when we buy a cup of coffee with our credit card. When we make a phone call, we are generating data. When we are getting a speeding ticket. And there are many other examples, as you can see on this slide, showing that we are generating all the time. Even while, you are watching this MOOC, you are generating data. Because all kinds of things are being recorded. If you take another example and you look at the simple telephone that now these days everybody owns, then you can see that such a phone has many different sensors. There is a GPS sensor, telling you where you are, but there are many other sensors, for example, looking at your fingerprint, and stuff like that. These are all generating data, and therefore we are talking about the Internet of Events. All the data that is being recorded in all kinds of ways. Let's look at this Internet of Events in more detail. What does it consist of? One can talk about 4 different sources of event data. The first source of event data. Is the Internet of Content. This is the classical Internet that we know, from Google and Wikipedia and then people typically, talk about Big Data they are talking about this Internet. But next to this classical Internet of Content, we now also have an Internet People. So we have Twitter messages. We have

Facebook. All kinds of events, social events that are generating data. Then we have the Internet of Things. It's another source of event data. Today already many devices are connected to the internet but in the future many more devices will be connected to the Internet. For example your shaving device. Your refrigerator, everything in the future will be connected to the Internet and this will generate large amounts of data. Last but not least, there is the Internet of Places, when you are using your mobile phone, as I just illustrated, the phone contains all kinds of sensors that are recording, where you are and what you are doing and this is another source of information. So, this is why today many people talk, about Big Data. Incredible amounts of event data that are being recorded. When people talk about Big Data. Talk about the exponential growth of data. This picture looks very complicated but don't be scared of by it. It is showing the exponential growth of the number of transistors on a chip. And this was predicted by the founder of Intel, Moore, and we can see that this exponential growth has continued over the last 40 years. Every 2 years, the numbers of transistors on a chip is doubling, so that means that on a chip now there are 1 million times the number of Transistors as there was 40 years ago. We not only see this in terms of the number of transistors but also in terms of computing speed. The capacity of a hard disk and the number of bytes, you get for a Dollar or Euro. And so this is showing this incredible growth. If other fields would have had the same growth of data, we would see very surprising things. For example, if I take the train. From Eindhoven to Amsterdam, 40 years ago, this would take approximately 1.5 hours. The question is now, if transportation would have followed more Moore's law, how fast would we be able to go from Eindhoven to Amsterdam by train today? Please think about this question. The answer to the question can be computed in a very easy way. We take one and a half hours times 60 minutes times 60 seconds and divide it by 2 to the power 20. And we would see that today we would only need 5 milliseconds to travel from Eindhoven to Amsterdam. We can look at another

example, do exactly the same calculation. If we would fly, from Amsterdam to New York 40 years ago it would take seven hours. How long would it take today, if transportation would have followed Moore's law? We can do exactly the same calculation. We take the number of hours. Times 60 minutes, times 60 seconds, divided by 2 to the power 20. And we would see that only in 24 milliseconds, we would be able to fly from Amsterdam to New York. As I said, the capacity on the hard disk is also growing exponentially. And also the costs per byte are also decreasing exponentially. So what does this mean? Let's compare it to fuel consumption. If 40 years ago, we needed 4000 liters of petrol to drive a car around the world, how much petrol would we need today if we would have followed Moore's Law? The answer to this question. Can be computed in exactly the same way. We take 4000 liters divided by 2 to the power 20 and we would see that 4 milliliters of petrol would be sufficient to travel all around the world. These examples are showing the incredible, growth of the data and why people talk about Big Data. The challenge today is not to generate more data, but the challenge today is to turn this data into real value. And this is a crucial topic. People often talk about the four V's of Big Data. The first V. I just explained that is the v of Volume. So we are generating incredible amounts of data, but that's not the only challenge. The second challenge is Velocity. We are not only generating large amounts of data, data is continuously being added. And things are changing very rapidly. So, velocity is the second challenge. The third challenge is Variety. So it's not one type of data. We are confronted with many different types of data. Ranging from text, to images, to audit trails. And we need to combine all these different sources of information. Last but not least, a problem of Big Data is Veracity. And that means, that you can not be completely sure that what you have recorded is completely accurate. For example, your shaving device of the future will have an Internet connection. Somebody has bought that shaving device. And we are recording events from that device, how it is being used. But can we be sure that the person

who purchased the shaving device is the actual person using it? That is a kind of uncertainty that you see if you collect data in a very large scale. And you need to be able to deal with that. I just spoke a lot about Big Data, but data doesn't have to be big to be challenging. Data analytics questions are everywhere and that is why there is a very urgent demand for data scientists. So, what do these data scientists do? What is their profession? What is their task? Well, their goal is to collect, analyze, and interpret data from a variety of sources. I've given you already several examples. That is why this will become a very important profession in the future. And the main goal is to turn data into value. And in this course, we will focus on this particularity. If you look at data science there are 4 generic data science questions, that you can ask in any situation. The first data science question is the question. What happened? If we see a bottle neck, if we see deviations, we record event data and we can actually, see that these things have happened. Then the second logical question, is to ask yourself. Why did it happen? Why was there this delay? Why did people deviate from the expected path? These things are just about the past, but of course data science also aims to answer questions about the future. So, if you look at the future. You also ask yourself the question, what will happen? What can we learn from historic information to make predictions, about what is happening at this point in time. And then last, but not least, the fourth question of data science. Is to ask yourself, what is the best that can happen? So, you want to use analytics to recommend certain things that will improve a situation. Remove a bottleneck. Make sure that people don't deviate or provide a better service. So these are the four questions of data science. And if we look for example at the care flows in a hospital we can ask ourselves some of these questions. So for example, a hospital you could ask yourself the question. Why do patients have to wait so long? What is causing these delays? Do doctors follow the guidelines? We have been doing a lot of analysis in hospitals, and we see that different doctors typically, process things in a

completely different way. So why does one person do something different than another person? Can we predict waiting times for patients? Can we predict how much staff we will need tomorrow? And last but not least, very important in health care, how can we reduce costs without sacrificing quality? So these questions are at, a level of what we call a business process. But even if you look at the technical device, like for example an X-ray machine in exactly the same hospital, we can also see that this X-ray machine is generating lots of data. And again, we can ask all kinds of questions which you can answer by analysing this data carefully. So for example. How are the X-ray machines, really used in the field? There's a very important, question for the people making these machines. Why and when do X-ray machines malfunction? When do they break down, and why? Which components should be replaced? The machine doesn't work anymore. Can we generate diagnostic information telling us which component is broken, so that an engineer can go to the device and repair it instantly. Can we predict, whether a machine will break down? And last but not least, can we learn from existing problems, which parts need to be improved? So these, are all Data science questions and you need very different skills to address all of these questions. So this picture tries to summarize the various skills that are important for data science. There are obvious candidates like Statistics and Data Mining. You also need to be able to visualize large amounts of data. You need to be able to deal with incredibly large databases. But that's not the only thing. You also need, to have knowledge of social sciences to talk about things like ethics and privacy. You need to think about the value of data, how you can extract value from it. So Industrial Engineering is also important. And last but not least, and that the core topic of this course, you want to apply Process Mining techniques to these data, to learn and improve processes in the way that I just described. So the importance of Data science, hopefully is obvious. Many people say the Data scientist will be the most sexy job of this century. And that is why in Eindhoven we have

started The Data Science Center Eindhoven, which groups specialists in the area. And educate students in this exciting new topic. So far, I've been mainly talk about Data science in a very general sense. This course is about Data science, but in particular, we will focus on the analysis on process is based on data. Processes are key. You don't want to improve data. You want to improve processes. They are the thing that matter. Not the data. Not the software. It's also different from data mining. We are not interested in just isolated decisions or low level patterns. We are interested in improving end-to-end processes. That is a key thing. So if you take this Process-centric view on data science, and we return to the examples that I gave before, we can see that in a hospital setting, there are many process related questions that we can ask about care flows. If we go to an X-ray machine, we can see that in such a machine there are many processes unfolding. And you would like to analyze and understand them. And they are generating terabytes of data. So we have a rich source of information to do so. So the focus of this course is on the interplay between event data and processes and process models. And that is what the topic of process mining, or some people refer to it as business process intelligence, is all about. And so that is what we will focus on in the next couple of weeks. Let me sketch some use cases for process mining. The first use case is to ask yourself the question, what is the process that people really follow? What do they really do? Not what they tell they do, but what do they really do? What are the bottlenecks? Where are they? What is causing them? Where and why, do people or machines deviate from an expected or an idealized process? These are key questions and these are just a few examples. There are many more examples. What are the highways in my process? Which factors are influencing a bottleneck, etcetera. And so these are the questions that we will focus on in the next couple of weeks. So, Process Mining is Data Science in Action. We are looking at the dynamics of machines and business processes and try to learn from them and

improve them. So, if you want to read more about this topic of Process Mining, I would advise you to read chapter one of the process mining book, and you will will be able to learn much more about this. Thank you for watching and hope to see you soon.

## **2 - 1 - Palestra 1.1- Ciência dos Dados e Big Data**

Bem-vindo à primeira palestra do curso sobre processo de mineração: Ciência de Dados em Ação. Nesta primeira semana, vamos nos concentrar sobre a relação entre processo de mineração e outros tipos de análise de dados. Hoje, gostaria de começar por discutir o tema mais amplo da Ciência de Dados e Big Data. Hoje, muitas pessoas dizem que os dados é o grande óleo. E isso está ilustrando a incrível quantidade de dados que estamos coletando e o valor correspondente. Se você pensar, a quantidade de dados foi gerado desde os tempos pré-históricos, até 2003. E nós pensamos, o que agora podemos fazê-lo em. Somos capazes de gerar tais quantidades de dados em apenas 10 minutos. Então, em 10 minutos, estamos gerando, o máximo de dados agora. Como estávamos fazendo desde os tempos pré-históricos até 2003. Portanto, este está ilustrando o incrível crescimento de dados. Então, que tipo de dados é isso? Que tipo de dados de eventos que estamos gerando? Bem, estamos gerando, quando compramos uma xícara de café com o nosso cartão de crédito. Quando fazemos uma chamada de telefone, estamos gerando dados. Quando estamos recebendo uma multa. E há muitos outros exemplos, como você pode ver neste slide, mostrando que estamos gerando o tempo todo. Mesmo quando, você está assistindo esse Mooc, você está gerando dados. Porque todos os tipos de coisas que estão sendo gravados. Se você tomar um outro exemplo, e você olha para o telefone simples que agora estes dias toda a gente possui, então você pode ver que um celular tem muitos sensores diferentes. Há um sensor GPS, dizendo-lhe onde você está, mas há muitos outros sensores, por exemplo, olhando para a sua impressão digital, e coisas assim. Estes são todos os dados de geração e, portanto, nós estamos falando sobre a Internet de Eventos. Todos os dados que está sendo gravado em todos os tipos de formas. Vamos olhar para este Internet de Eventos com mais detalhes. O que é que consiste? Pode-se falar de cerca de 4 diferentes fontes de dados do evento. A primeira

fonte de dados do evento. A Internet é de conteúdo. Esta é a Internet clássico que sabemos, a partir do Google e Wikipedia e, em seguida, as pessoas normalmente, falar sobre Big Data estão falando sobre isso Internet. Mas junto a este clássico da Internet Content, agora também temos uma Internet Pessoas. Portanto, temos as mensagens do Twitter. Temos Facebook. Todos os tipos de eventos, eventos sociais que estão gerando dados. Então nós temos a Internet das Coisas. É uma outra fonte de dados do evento. Hoje já muitos dispositivos estão conectados à internet, mas no futuro muitos mais dispositivos serão conectados à Internet. Por exemplo o dispositivo de barbear. O frigorífico, tudo, no futuro, ser ligada à Internet e isso irá gerar grandes quantidades de dados. Por último, mas não menos importante, há a Internet de Places, quando você estiver usando o seu telefone móvel, como acabei de ilustrado, o telefone contém todos os tipos de sensores que são de gravação, onde você está eo que você está fazendo e isso é outra fonte de informações. Então, é por isso que hoje muitas pessoas falam, sobre Big Data. Incríveis quantidades de dados de eventos que estão sendo gravados. Quando as pessoas falam sobre Big Data. Fale sobre o crescimento exponencial de dados. A imagem parece muito complicado, mas não tenha medo de por ele. É mostrando o crescimento exponencial do número de transistores em um chip. E este foi previsto pelo fundador da Intel, Moore, e podemos ver que este crescimento exponencial continuou ao longo dos últimos 40 anos. A cada dois anos, o número de transistores em um chip dobra, o que significa que em um chip agora há 1 milhão de vezes o número de transistores como havia há 40 anos. Nós não só vê isso em termos do número de transistores, mas também em termos de velocidade de computação. A capacidade de um disco rígido e do número de bytes, você começa por um dólar ou euro. E assim, este está mostrando esse crescimento incrível. Se outros campos teria tido o mesmo crescimento de dados, queremos ver as coisas muito surpreendentes. Por exemplo, se eu pegar o trem. A partir de

Eindhoven para Amsterdã, há 40 anos, isso levaria cerca de 1,5 horas. A questão agora é, se o transporte teria seguido mais a lei de Moore, o quão rápido poderíamos ser capazes de ir de Eindhoven para Amsterdam de trem hoje? Por favor, pense sobre esta questão. A resposta para a pergunta pode ser calculado de forma muito fácil. Nós levamos uma hora e meia vezes 60 minutos vezes 60 segundos, e dividi-lo por 2 elevado à potência 20. E veremos que hoje nós só precisa de 5 milissegundos para viajar de Eindhoven para Amsterdã. Podemos olhar para outro exemplo, fazer exatamente o mesmo cálculo. Se quisermos voar, a partir de Amsterdam para Nova York há 40 anos, ele levaria sete horas. Quanto tempo levaria, hoje, se o transporte teria seguido a lei de Moore? Nós podemos fazer exatamente o mesmo cálculo. Tomamos o número de horas. Número de 60 minutos, os tempos de 60 segundos, dividido por 2 elevado à potência 20. E veremos que apenas em 24 milissegundos, que seria capaz de voar a partir de Amsterdam para New York. Como eu disse, a capacidade do disco rígido também está a crescer exponencialmente. E também os custos por byte também estão a diminuir exponencialmente. Então, o que isso significa? Vamos compará-lo ao consumo de combustível. Se há 40 anos, precisávamos de 4000 litros de gasolina para dirigir um carro ao redor do mundo, a quantidade de gasolina que precisamos hoje se teria seguido a Lei de Moore? A resposta para esta pergunta. Pode ser calculado da mesma maneira. Tomamos 4000 litros divididos por 2 elevado à potência 20 e veríamos que 4 mililitros de gasolina seria suficiente para viajar por todo o mundo. Estes exemplos mostram o incrível, o crescimento dos dados e por que as pessoas falam sobre Big Data. O desafio hoje não é para gerar mais dados, mas o desafio hoje é transformar esses dados em valor real. E este é um tema crucial. As pessoas muitas vezes falam sobre os quatro V do de Big Data. A primeira V. eu acabei de explicar que é a v de Volume. Então, nós estamos gerando quantidades incríveis de dados, mas esse não é o único desafio. O segundo desafio é Velocity. Nós não

estamos apenas gerando grandes quantidades de dados, os dados são continuamente sendo adicionado. E as coisas estão mudando muito rapidamente. Assim, a velocidade é o segundo desafio. O terceiro desafio é Variety. Portanto, não é um tipo de dados. Somos confrontados com muitos tipos diferentes de dados. Variando de texto, de imagens, de trilhas de auditoria. E nós precisamos combinar todas essas diferentes fontes de informação. Por último, mas não menos importante, um problema de Big Data é Veracidade. E isso significa que você não pode estar completamente certo de que o que você gravou é totalmente preciso. Por exemplo, o dispositivo de corte do futuro vai ter uma ligação à Internet. Alguém comprou esse dispositivo de barbear. E estamos a gravação de eventos a partir desse dispositivo, como ele está sendo usado. Mas podemos ter certeza de que a pessoa que comprou o aparelho de barbear é a pessoa real usá-lo? Esse é um tipo de incerteza que você veja se você coletar dados em uma escala muito grande. E você precisa ser capaz de lidar com isso. Acabei de falar muito sobre Big Data, mas os dados não tem que ser grande para ser um desafio. Análise de dados questões estão em toda parte e é por isso que há uma demanda muito urgente para os cientistas de dados. Então, o que os cientistas estes dados fazer? Qual é a sua profissão? Qual é a sua missão? Bem, seu objetivo é coletar, analisar e interpretar os dados a partir de uma variedade de fontes. Eu dei-lhe já vários exemplos. É por isso que isso vai se tornar uma profissão muito importante no futuro. E o objetivo principal é o de transformar dados em valor. E neste curso, vamos focar essa particularidade, Se você olhar para a ciência de dados há 4 genéricos questões de ciências de dados, que você pode pedir em qualquer situação. A primeira pergunta a ciência de dados é a questão. O que aconteceu? Se vemos um gargalo de garrafa, se vemos desvios, registramos dados do evento e nós podemos realmente, ver que essas coisas aconteceram. Então a segunda questão lógica, é de se perguntar. Por que isso aconteceu? Por que houve esse atraso? Por que as pessoas se

desviam do caminho esperado? Essas coisas são apenas sobre o passado, mas é claro que a ciência dados também visa responder a perguntas sobre o futuro. Então, se você olhar para o futuro. Você também se fazer a pergunta, o que vai acontecer? O que podemos aprender com a informação histórica para fazer previsões, sobre o que está acontecendo neste momento no tempo. E, em seguida, por último, mas não menos importante, a quarta questão da ciência de dados. É de se perguntar, o que é o melhor que pode acontecer? Então, você quer usar analytics para recomendar certas coisas que vão melhorar a situação. Retirar um gargalo. Certifique-se de que as pessoas não se desviem ou fornecer um serviço melhor. Então, essas são as quatro questões de ciência de dados. E se olharmos para o exemplo no cuidado fluí em um hospital que pudermos como a nós mesmos algumas destas questões. Assim, por exemplo, um hospital que você poderia fazer a pergunta. Por que os pacientes têm de esperar tanto tempo? O que está causando estes atrasos? Os médicos seguem as diretrizes? Temos vindo a fazer um monte de análise em hospitais, e vemos que diferentes médicos tipicamente, processar as coisas de uma maneira completamente diferente. Então, por que uma pessoa fazer algo diferente do que a outra pessoa? Podemos prever o tempo de espera para os pacientes? Podemos prever o quanto a equipe que vai precisar amanhã? E por último mas não menos importante, muito importante nos cuidados de saúde, como podemos reduzir os custos sem sacrificar a qualidade? Então, essas questões são em um nível do que chamamos de um processo de negócio. Mas mesmo se você olhar para o dispositivo técnico, como por exemplo, uma máquina de raio-X em exatamente o mesmo hospital, também podemos ver que esta máquina de raio-X está gerando grandes quantidades de dados. E mais uma vez, podemos pedir a todos os tipos de perguntas que você pode responder por analisar esses dados com cuidado. Assim, por exemplo. Como são as máquinas de raios-X, realmente utilizado no campo? Há uma muito importante, pergunta

para as pessoas que fazem estas máquinas. Por que e quando fazer máquinas de raios-X mau funcionamento? Quando eles quebram, e por quê? Quais componentes devem ser substituídos? A máquina não funciona mais. Podemos gerar informações de diagnóstico nos dizer qual componente está quebrado, de modo que um engenheiro pode ir para o dispositivo e repará-lo instantaneamente. Podemos prever, se uma máquina vai quebrar? E por último mas não menos importante, que podemos aprender com os problemas existentes, quais as partes precisam ser melhorados? Então, essas, são questões de ciência de dados e você precisa de muito diferentes habilidades para lidar com todas estas perguntas. Então esta imagem tenta resumir as várias habilidades que são importantes para a ciência de dados. Há candidatos óbvios, como Estatística e Data Mining. Você também precisa ser capaz de visualizar grandes quantidades de dados. Você precisa ser capaz de lidar com grandes bases de dados incrivelmente. Mas essa não é a única coisa. Você também precisa, para ter conhecimento das ciências sociais para falar sobre coisas como ética e privacidade. Você precisa pensar sobre o valor dos dados, como você pode extrair valor a partir dele. Assim, Industrial Engineering também é importante. E por último mas não menos importante, e que o tema central deste curso, que pretende aplicar técnicas de processo de mineração com estes dados, para aprender e melhorar os processos da maneira que eu acabei de descrever. Assim, a importância da ciência de dados, espero é óbvio. Muitas pessoas dizem que o cientista de dados vai ser o trabalho mais sexy deste século. E é por isso, em Eindhoven, começamos O Data Center Ciência Eindhoven, que agrupa especialistas na área. E educar os alunos neste novo tópico emocionante. Até agora, eu fui falar principalmente sobre a ciência de dados em um sentido muito geral. Este curso é sobre a ciência de dados, mas, em particular, vamos nos concentrar na análise sobre o processo é baseado em dados. Processos são fundamentais. Você não quer para melhorar os dados. Você quer

melhorar processos. Eles são a única coisa que importa. Nem os dados. Nem o software. Também é diferente de mineração de dados. Nós não estamos interessados nas decisões apenas isoladas ou padrões de baixo nível. Estamos interessados em melhorar processos end-to-end. Isso é uma coisa fundamental. Então, se você tomar este ponto de vista do processo centrada na ciência de dados, e voltamos para os exemplos que eu dei antes, podemos ver que em um ambiente hospitalar, há muitos processos relacionados com questões que podemos fazer sobre os fluxos de cuidados. Se formos a uma máquina de raio-X, podemos ver que, em tal máquina há muitos processos que se desenrolam. E você gostaria de analisar e compreendê-los. E eles estão gerando terabytes de dados. Portanto, temos uma rica fonte de informações para o fazer. Assim, o foco deste curso é sobre a interação entre os dados de eventos e processos e modelos de processos. E é isso que o tema da mineração processo, ou algumas pessoas se referem a ele como a inteligência de processos de negócios, é tudo. E é isso que vamos focar no próximo par de semanas. Deixe-me esboçar alguns casos de uso para mineração processo. O primeiro caso de uso é para se fazer a pergunta, o que é o processo que as pessoas realmente seguir? O que eles realmente fazer? Não é o que eles dizem que fazem, mas o que eles realmente fazem? Quais são os gargalos? Onde eles estão? O que está causando-los? Onde e por isso, é que as pessoas ou máquinas afastarem do esperado ou um processo idealizado? Estas são questões fundamentais e estes são apenas alguns exemplos. Há muitos outros exemplos. Quais são as rodovias no meu processo? Quais os fatores que estão influenciando um gargalo, etcetera. E assim, estas são as questões que se centrará nos no próximo par de semanas. Então, processo de mineração de dados é a Ciência em Ação. Nós estamos olhando para a dinâmica de máquinas e processos de negócio e tentar aprender com eles e melhorá-los. Então, se você quiser ler mais sobre esse assunto de processo de mineração, eu aconselho

você a ler o capítulo um do livro de mineração processo, e você será capaz de aprender muito mais sobre isso. Obrigado por assistir e espero vê-lo em breve.

## **2 - 1 - Lecture 1.1- Data Science and Big Data (END)**

---

---

## 2 - 2 - Lecture 1.2- Different Types of Process Mining

Welcome to the second lecture of the course on process mining data science in action. Today I provide an overview of process mining by discussing three types of mining. Process discovery, performance checking, and enhancement. In my last lecture, I showed you that we have large amounts of data, but that it is not just about collecting data, it's about analyzing processes. That is one of the key things. If you think about what process mining is, then the next diagram provides a nice illustration of that. So process mining is bridging the gap between classical process model analysis. And data oriented analysis like data mining and machine learning. Process mining is bridging this gap, because it's focusing on processes. But at the same time, using the real data. In classical data mining, people typically do not look at processes. Especially not end to end processes. In areas where people are concerned with process model analysis, they typically ignore the data. Why are you doing process mining? We are doing process mining to answer performance related questions and compliance related questions. So why are there certain bottlenecks, how can they be removed, why do people deviate. So the starting point for process mining is event data. What you see here is a table, and every row in this table corresponds to an event. An event has different properties, so there are generic types of properties. The first property is a case id. In this case every record refers to a student making an exam. So the case id refers to the student. The activity name refers to the exam. And the timestamp in this case is the date of the exam. There can be additional data like, for example, the mark, or the grade that somebody got. Let's take another example. We are now looking at the handling of orders. Again every line, corresponds to an event. If here we look at the first column, we see the order number. That is the case id. The second column refers to the activity that is being executed. The third column, again refers to the timestamp. Then there is a column referring to the resource. The person executing the corresponding activity. And we can have all kinds of other

columns with other data like the products and quantity and things like that. Let's take a look at the third example. Now we are in a hospital we are treating patients. And we look at an event log and every row refers to a step in the treatment of a particular patient. So what you see is that, again, we see a case id, activity name, and timestamp. So the first column refers to the patient, this case represented by a number. The second column to the activity name, the third column to the timestamp. Again we see the resource, in this case the doctor or the nurse executing a particular activity, and we can have all kinds of other data. Like for example the age of the patient. And all kinds of other properties of the patient or the corresponding step. So these are examples of the kind of event data that we would like to use for analysis. And we are focusing on the relationship between process models and event data. And I will talk about three types of relationships between models and event data. Play-out, Play-in, and Replay. Let's first take a look at Play-out. Here the basic idea is that you start from a model, and from that model, you generate behavior. And let me try to illustrate this using a small example. What you see here is a so-called BPMN model. And don't be scared off by it. In later weeks we will discuss these types of things in detail but this model is describing a particular work flow. How things are executed, and in what order they are being executed. So, if we take one execution of this process model, which is about handling travel requests, then it always starts with an activity a, where we register the travel request. It is done at a particular point in time, and it is done by a particular person. Then we start concurrently two paths. In one path we execute activity d, check budget. And in another path we have a choice between two different activities, b and c. So let's assume that we do activity b, get support from local manager. And this is done at the particular time and by a particular person. Then, we have to do activity d if we follow this model. Check budget by finance. Again, done at a particular point in time by a particular person. And then, we can make a decision doing activity e. Here we

see an XOR-split. So we need to make one of three decisions accept, reject or there is insufficient information and we need to redo part of the process again. If we decide to accept, we execute activity g and the process ends. So this is one possible scenario. Let us now look at another possible scenario following the same process. So we first do an a we then, again start two paths in parallel. But now we do first d and then we do c instead of b. We then have to make a decision again. But now there is not enough information. We need to go back. And if we go backwards, we again need to execute d and b or c. So we go back. We execute b again. We execute d. Then we execute e. And finally we make a decision. And in this case the decision is to reject the request. So this shows you two possible paths. But there are many more. Now if we look at this process model then these are possible ways of playing out this model. These are not the only ones because there is a loop that are infinitely many. So this is playing out a model. And if you are using simulation or if you are building an information system that is driven by such models, play-out is the thing that you are doing. Let's now look at the reverse. We are just reversing the arrow. We now go from even data to the corresponding process model. And in later weeks we will look at different algorithms to do this. The basic idea is you look at the number of example behaviors. And you automatically infer a model from it. So here, if you look at the process, if you look at the traces that you see on top of this diagram, you can see that it always starts with an a. So we infer a model that always starts with an a. All the traces end with a g or and h, so in the process model at the end there is a choice between accept and reject, g and h. This is a more complicated set of traces and if we get this more complicated set of traces, we get this more complicated model. So we are automatically learning a model from examples. And what is very important to understand here is that no modeling is needed. We are not making any models by hand. We automatically infer process models from raw event data. So let me ask you a question. And of course you did not learn yet process

discovery techniques. And later you will look at algorithms to do so. But take a look at these example traces. And I ask you, can you create a process model that allows for the traces that you see here. Think a bit about this for sometime. And I'm quite sure that you probably end up with a model like this. So what is shown in this model. That all the traces started with a. And ends with an e or an f. And in between we again have a choice between B and C and we always do D. And this can be done in any order that is why we have an AND-split in this diagram. So that was a toy example, we can easily apply this to real world data. So here you see an example. That we will revisit several times in the future. Here we look at raw event data from a Dutch housing agency. And every row refers to an event within this housing agency. The first column refers to the number of an apartment, the ID of an apartment. The second column to the activity. The third column to the time stamp. Again, very similar to the things that we have seen before. If we take this raw event data, we infer a process model that is describing what happens when somebody that is renting an apartment, cancels his rent. Until the moment that apartment is rented out to somebody else. This is the full process model. If we zoom into a particular part, then you can see a small fragment which is executed by, let's say, approximately thirty, cases. This case houses or apartments that followed this sequential path. Without modeling you are able to see what is the process that is really being executed. Let's take a look at a more complicated example. We have been applying this in probably 20 to 30 different hospitals where we can take event data about patients. So this is a homogeneous group of patients within a Dutch hospital. We can look at the audit trail of a single patient of that patient, we see all kinds of properties, and we see the events that have been executed. If we take 600 of such cases. And we automatically infer a model. We see this scary diagram that you can see on one side. It is showing that a treatment process of this particular group of patients is incredibly complicated. Don't be scared off by this spaghetti like diagram. Later we will see all kinds

of techniques to simplify such diagrams, so that you only have to look at the highways, or the parts that you're interested in. On the right-hand side of the screen, you see also a graph-based model, but it is now not showing activities, but it is showing departments. So rather than seeing the sequence from one activity to the next activity we now see how work is flowing from one department to another department. So these are some practical examples showing you how process mining can be applied in practice. But it doesn't stop here. It is not enough to just learn the model, the key element is that you can replay reality on top of models. Whether you have made these models by hand, or whether you have automatically learned them using process discovery is irrelevant. What is important is that you try to replay reality on top of the model. So, at the top of the slide, you can see a trace that happened in reality. Again, this is an abstract example, later I will show you a real example. In reality, we saw the sequence a c d e g. Let's try to replay this on this model. And if we do this, we see a path through the process model that is indeed possible. So we replay reality on top of the model, and while doing that we don't encounter any problems. So, nothing surprising yet Let's now take a look at another case. Where we see the sequence a c e g. So this could be the path of a patient. This could be the life cycle of an apartment. It could be somebody applying for a mortgage. It can be anything. So we look at the sequence of events. In this case in reality first a happens. That is possible according to the model. Then c happens, this is also still possible in the model but then in the reality e happened, and according to the process model you can only make a decision if also activity d is being executed it is impossible to execute d at this point in time. So while replaying, we continue, but we record the fact that there is a problem. So here activity d, check budget, is missing. And then we continue and in the remainder of the trace, we don't see any surprising things. So we can see where the deviations are. Here I show you another example. If we look at this example first a then we do a c then we

see that in reality h has happened. So there was a rejection while the process was still in the middle. Again by doing replay we can record a problem. Provide diagnostics and continue our replay. This case, we make a decision and finally we do the acceptance. So again, we have found a deviation between the reality and the model. So, to check whether you have understood these types of things. Take a look at this model and the corresponding tree traces. Are these traces possible in this model or not? And if they are not possible, where does reality deviate from this model? Let's take a look at the answer. If we take a look at the first trace. a, b, d, c, e h. Then we can see that this is impossible. If we tried to replay it, we will see that it is impossible to do both b and c. When we take a look at the second trace, we find kind of a reversed problem. Now we don't have an activity too much we have an activity too little. Activity d is missing. So it doesn't fit. If you try to replay the last trace you will see that you don't encounter any problems so indeed it is replayable. So these examples show you how you can apply yourself. Conformance checking techniques by doing replay. Again, let us take a look at a practical example. This is a fragment of a larger process model, describing the handling of complaints in a Dutch municipality. These are people that are complaining about the valuation of their house. And because of that, they think that they need to pay too much tax. We take a look at this process and we replay a reality on top of this model, we will find interesting deviations. Although the overall process has a very good fitness, the relationship between reality and the model. They are very close to one another. We still find all kind of interesting deviations. So for example here the house was reevaluated 23 times. While it was not supposed to happen. So something happened in reality. Which was not possible according to the model. And by replaying reality on top of the model, we can see these types of problems. Replay is not just about conformance. It's also about performance analysis. We can easily replay an event logged that has timestamps on top of such a model. And each time we execute an activity. We record the

corresponding time as you can see here on this slide. So we just execute reality on top of the model. We record all the times. And at the end we know exactly how much time was spent in all the different parts of the process. And the black circles here indicate delays between individual activities. So we can do this for a single execution, for a single trace. But you can also do that for many traces. If you do that for many traces, you will see how often an activity is being executed. You will see how often a particular path is being followed. But you also see how long activities have taken and you will see what the delays are in between these activities. What are waiting times. What are synchronization times. You can record them for hundreds of thousands or perhaps even millions of cases. And derive probability distribution from them. So again let us take a look at a practical example. Here we look again at a process where people complain about the valuation of their house. We have derived automatically a process model. And these terms expressed as a petri net. And now we can by replaying we can see where the delays take place. So if you look at this slide you can see that certain places are indicated in purple. And these are the places where we see a large delay. So this is an indication where possible performance problems are. It can go one step further. You can point at two arbitrary points in the process. You can see how many cases, in this case how many objections are flowing from this first red activity to the second red activity, and we can see how long it took. And you can see here the corresponding statistics. So by replaying you can see exactly where delays take place. You can start investigating what is causing them. So to provide an overview of the things that I've been talking about we have a real world where things are happening. We have a software system, that is, somehow, recording events of the things that take place. People objecting to the valuation of their house. Patients being treated in a hospital. Then we have these event data. And we can do discovery, automatically learning process models. We can do conformance checking. Comparing the event log to the process model and we

can enrich, it's called enhancement here, we can enrich the process model with information about deviations and performance. For example the bottle necks that I just showed you. How does that relate to the things that I've been talking about? Play-out is about the classical use of process models not involving any event data. Play-in corresponds to discovery. You automatically learn a process model without any modeling from raw event data. And you have replay where you compare a process model to an event log to check conformance, to investigate performance problems, etcetera. So, I hope this gave you a good overview of the three different types of, process mining. If you want to find more high-level information about these three types, please read chapter one. Thank you very much for watching and hope to see you soon.

## **2 - 2 - Palestra 1.2- Diferentes tipos de processo de mineração**

Bem-vindo à segunda palestra do curso no processo científico mineração de dados em ação. Hoje eu fornecer uma visão geral do processo de mineração por discutir três tipos de mineração. Descoberta de processos, verificação de desempenho, e de enriquecimento. Na minha última palestra, eu mostrei que temos grandes quantidades de dados, mas que não é apenas sobre a coleta de dados, trata-se de analisar os processos. Essa é uma das coisas mais importantes. Se você pensar sobre o que a mineração processo é, então o próximo diagrama fornece uma ilustração agradável do que isso. Então mineração processo é fazer a ponte entre a análise do modelo de processo clássica. E análise de dados como a mineração de dados e aprendizado de máquina orientada. Mineração processo é colmatar esta lacuna, porque está com foco em processos. Mas, ao mesmo tempo, utilizando os dados reais. Na mineração de dados clássica, as pessoas normalmente não olhar para os processos. Especialmente não acabar com os processos finais. Em áreas onde as pessoas estão preocupadas com a análise do modelo de processo, que normalmente ignoram os dados. Por que você está fazendo mineração processo? Estamos fazendo mineração processo para responder a questões relacionadas com desempenho e questões relacionados à conformidade. Então, por que existem certos pontos de estrangulamento, como eles podem ser removidos, por que as pessoas se desviam. Assim, o ponto de partida para a mineração processo é dados do evento. O que você vê aqui é uma mesa, e cada linha nesta tabela corresponde a um evento. Um evento tem propriedades diferentes, por isso há tipos genéricos de propriedades. A primeira propriedade é um ID de caso. Neste caso, cada registro refere-se a um estudante fazer um exame. Assim, a id caso refere-se ao aluno. O nome refere-se a atividade do exame. E o timestamp, neste caso, é a data do exame. Não pode haver dados adicionais, como, por exemplo, a marca, ou o grau que alguém tem. Vamos dar outro exemplo. Estamos agora olhando

para o tratamento das encomendas. Mais uma vez a cada linha, corresponde a um evento. Se aqui olhamos para a primeira coluna, vemos o número de ordem. Esse é o ID de caso. A segunda coluna refere-se à atividade que está sendo executado. A terceira coluna, refere-se novamente para a data e hora. Em seguida, há uma coluna referente ao recurso. A pessoa que executa a actividade correspondente. E podemos ter todos os tipos de outras colunas com outros dados, como os produtos e quantidade e coisas assim. Vamos dar uma olhada no terceiro exemplo. Agora estamos em um hospital estamos tratando pacientes. E olhamos para um registo de eventos e cada linha refere-se a um passo para o tratamento de um paciente em particular. Então, o que você vê é o que, mais uma vez, vemos um caso id, nome da atividade, e timestamp. Assim, a primeira coluna refere-se ao paciente, neste caso representados por um número. A segunda coluna para o nome de actividade, a terceira coluna para a data e hora. Novamente vemos o recurso, neste caso, o médico ou o enfermeiro a execução de uma atividade particular, e nós podemos ter todos os tipos de outros dados. Como por exemplo, a idade do paciente. E todos os tipos de outras propriedades do paciente ou a etapa correspondente. Então, esses são exemplos do tipo de dados do evento que gostaria de usar para análise. E estamos focando a relação entre modelos de processos e dados do evento. E eu vou falar sobre três tipos de relações entre os modelos e dados do evento. Play-out, jogue-in e replay. Vamos primeiro dar uma olhada no Play-out. Aqui, a idéia básica é que você começa a partir de um modelo, ea partir desse modelo, você gera comportamento. E deixe-me tentar ilustrar isso usando um pequeno exemplo. O que você vê aqui é um chamado modelo BPMN. E não se assuste com isso. Nas semanas posteriores, vamos discutir esses tipos de coisas em detalhes, mas este modelo está descrevendo um fluxo de trabalho particular. Como as coisas são executadas, e em que ordem elas estão sendo executadas. Então, se tomarmos uma execução deste modelo de processo, que é sobre a manipulação de solicitações de viagens,

então ele sempre começa com uma atividade, onde registramos o pedido de viagem. É feito num ponto particular no tempo, e é feito por uma pessoa em particular. Então nós iniciados simultaneamente dois caminhos. Em um caminho que executar atividade d, verificar orçamento. E em outro caminho, temos uma escolha entre duas atividades diferentes, b e c. Então vamos supor que fazemos atividade b, obter o apoio do gestor local. E isto é feito no momento em particular e por uma pessoa em particular. Então, nós temos que fazer atividade d se seguirmos este modelo. Verifique orçamento pelas finanças. Mais uma vez, feito em um determinado ponto no tempo por uma pessoa em particular. E, em seguida, pode-se tomar uma decisão e fazer atividade. Aqui vemos um XOR-split. Então, nós precisamos fazer uma de três decisões aceitar, rejeitar ou não há informação suficiente e precisamos refazer parte do processo novamente. Se decidirmos aceitar, nós executamos atividade g e o processo termina. Portanto, este é um cenário possível. Vamos agora olhar para um outro cenário possível, seguindo o mesmo processo. Então, primeiro vamos fazer um a nós, então, mais uma vez começar a dois caminhos em paralelo. Mas, agora, o que fazemos primeiro d e depois fazemos c em vez de b. Temos, então, que tomar uma decisão novamente. Mas agora não há informações suficientes. Precisamos voltar. E se formos para trás, mais uma vez precisa executar d e b ou c. Então, vamos voltar. Executamos b novamente. Executamos d. Em seguida, executar e. E, finalmente, tomar uma decisão. E, neste caso, a decisão é rejeitar o pedido. Então, isso mostra dois caminhos possíveis. Mas há muitos mais. Agora, se olharmos para este modelo de processo de estas são as possíveis maneiras de jogar fora este modelo. Estes não são os únicos, porque não é um loop que são infinitamente muitos. Então, isso está jogando fora um modelo. E se você estiver usando simulação ou se você está construindo um sistema de informação que é conduzido por esses modelos, play-out é a coisa que você está fazendo. Vamos agora olhar para o inverso. Estamos apenas invertendo a seta. Vamos

agora até mesmo de dados para o modelo de processo correspondente. E nas semanas posteriores, vamos olhar para diferentes algoritmos para fazer isso. A idéia básica é que você olhar para o número de exemplos de comportamentos. E você inferir automaticamente um modelo a partir dele. Então, aqui, se você olhar para o processo, se você olhar para os traços que você vê no topo desta diagrama, você pode ver que ele sempre começa com um a. Assim, podemos inferir um modelo que sempre começa com um a. Todos os vestígios acabar com ag ou e h, por isso, o modelo de processo no final há uma escolha entre aceitar e rejeitar, g e h. Este é um conjunto mais complexo de traços e se conseguirmos esse conjunto mais complicado de vestígios, temos este modelo mais complicado. Então, nós estamos aprendendo automaticamente um modelo a partir de exemplos. E o que é muito importante para entender aqui é que nenhuma modelagem é necessário. Nós não estamos fazendo alguns modelos com a mão. Nós inferir automaticamente modelos de processo de dados de eventos brutos. Então deixe-me fazer uma pergunta. E é claro que você não aprendeu ainda técnicas de descoberta de processo. E mais tarde você vai olhar para algoritmos para fazê-lo. Mas dê uma olhada nestes exemplos de traços. E eu lhe pergunto, você pode criar um modelo de processo que permite que os traços que você vê aqui. Pense um pouco sobre isso por algum tempo. E eu tenho certeza que você provavelmente acabar com um modelo como este. Então, o que é mostrado neste modelo. Que todos os vestígios começou com um. E termina com uma mensagem ou um f. E entre temos novamente uma escolha entre B e C e sempre fazemos D. E isso pode ser feito em qualquer ordem que é por isso que temos um E-split neste diagrama. Então isso foi um exemplo de brinquedo, podemos facilmente aplicar isso a dados do mundo real. Então aqui você ver um exemplo. Que vamos revisitar várias vezes no futuro. Aqui olhamos para os dados de eventos brutos de uma agência de habitação holandês. E cada linha refere-se a um evento dentro desta agência de habitação. A primeira coluna

refere-se ao número de um apartamento, a ID de um apartamento. A segunda coluna para a atividade. A terceira coluna para a data e hora. Mais uma vez, muito semelhante às coisas que temos visto antes. Se tomarmos esses dados de eventos matérias, inferimos um modelo de processo que está descrevendo o que acontece quando alguém que está alugando um apartamento, cancela o aluguel. Até o momento em que o apartamento é alugado para outra pessoa. Este é o modelo de processo completo. Se nós zoom em uma parte particular, então você pode ver um pequeno fragmento que é executada pelo, digamos, cerca de trinta anos, os casos. Este caso casas ou apartamentos que se seguiram este caminho sequencial. Sem modelagem você é capaz de ver o que é o processo que está realmente sendo executado. Vamos dar uma olhada em um exemplo mais complicado. Temos vindo a aplicar isso em provavelmente 20 a 30 hospitais diferentes, onde podemos tomar os dados de eventos sobre os pacientes. Portanto, este é um grupo homogêneo de pacientes dentro de um hospital holandês. Podemos olhar para a trilha de auditoria de um único paciente desse paciente, vemos todos os tipos de propriedades, e vemos os eventos que foram executadas. Se tomarmos 600 desses casos. E nós inferir automaticamente um modelo. Vemos este diagrama assustador que você pode ver em um dos lados. Ele está mostrando que um processo de tratamento deste grupo de doentes, é incrivelmente complicado. Não se assustem por este spaghetti como diagrama. Mais tarde, vamos ver todos os tipos de técnicas para simplificar tais diagramas, de modo que você só tem que olhar para as rodovias, ou as peças que você está interessado. No lado direito da tela, você também verá uma graph- modelo baseado, mas agora é não mostrando atividades, mas ele está mostrando departamentos. Então, ao invés de ver a seqüência de uma atividade para a próxima atividade agora vemos como o trabalho está fluindo de um departamento para outro departamento. Então, esses são alguns exemplos práticos que mostram a você como mineração processo pode ser aplicado na prática. Mas não

pára por aqui. Não é suficiente apenas para aprender o modelo, o elemento fundamental é que você pode reproduzir uma realidade em cima de modelos. Se você tiver esses modelos à mão, ou se você automaticamente tem aprendido utilizando descoberta processo é irrelevante. O que é importante é que você tenta reproduzir a realidade em cima do modelo. Assim, na parte superior do slide, você pode ver um traço que aconteceu na realidade. Mais uma vez, este é um exemplo abstrato, mais tarde eu vou mostrar-lhe um exemplo real. Na realidade, nós vimos a sequência de um c d e g. Vamos tentar repetir isso neste modelo. E se fizermos isso, vemos um caminho através do modelo de processo que é de fato possível. Então, nós reproduzir a realidade em cima do modelo e, ao fazer que nós não encontrar quaisquer problemas. Então, nada de surpreendente ainda Vamos agora dar uma olhada em outro caso. Onde vemos a seqüência um c e g. Assim, este pode ser o caminho de um paciente. Este poderia ser o ciclo de vida de um apartamento. Poderia ser alguém aplicar para uma hipoteca. Ele pode ser qualquer coisa. Então, nós olhamos uma seqüência de eventos. Neste caso, na realidade, um primeiro acontece. Isso é possível de acordo com o modelo. Então c acontece, isso também é ainda possível no modelo, mas em seguida, na realidade e aconteceu, e de acordo com o modelo de processo que você só pode tomar uma decisão se também a atividade d está sendo executado é impossível executar d neste momento no tempo . Assim, enquanto repetindo, continuamos, mas nós registramos o fato de que há um problema. Então aqui atividade d, orçamento de verificação, está faltando. E então nós continuamos e no restante do traço, não vemos quaisquer coisas surpreendentes. Assim, podemos ver que os desvios são. Aqui eu mostro-lhe outro exemplo. Se olharmos para este primeiro exemplo de um então nós fazemos ac então vemos que na realidade h aconteceu. Então, houve uma rejeição, enquanto o processo ainda estava no meio. Mais uma vez, fazendo repetição podemos gravar um problema. Fornecer diagnóstico e continuar a nossa replay.

Neste caso, tomamos uma decisão e, finalmente, fazer a aceitação. Então, novamente, nós encontramos um desvio entre a realidade e o modelo. Assim, para verificar se você entendeu estes tipos de coisas. Dê uma olhada neste modelo e os traços de árvores correspondentes. São estes os traços possível neste modelo, ou não? E se eles não são possíveis, onde é que a realidade desviarse desse modelo? Vamos dar uma olhada na resposta. Se dermos uma olhada no primeiro traço. a, b, d, c, e h. Então, podemos ver que isso é impossível. Se tentássemos reproduzi-la, veremos que é impossível fazer as duas coisas b e c. Quando vamos dar uma olhada no segundo traço, encontramos um tipo de problema revertida. Agora não temos uma atividade muito temos uma atividade muito pouco. Atividade d está faltando. Por isso, não se encaixa. Se você tentar repetir o último vestígio você vai ver que você não encontrar quaisquer problemas assim na verdade, é replayable. Então, esses exemplos mostram como você pode aplicar-se. Conformidade técnicas, fazendo a verificação de replay. Mais uma vez, vamos dar uma olhada em um exemplo prático. Este é um fragmento de um modelo de processo maior, que descreve o tratamento de reclamações em um município holandês. Estas são as pessoas que estão reclamando sobre a valorização de sua casa. E por causa disso, eles pensam que eles precisam pagar impostos demais. Vamos dar uma olhada neste processo e nós reproduzir uma realidade em cima desse modelo, vamos encontrar desvios interessantes. Embora o processo global tem uma boa forma física, a relação entre a realidade eo modelo. Eles são muito próximos um do outro. Nós ainda encontrar todo o tipo de desvios interessantes. Assim, por exemplo, aqui a casa foi reavaliado 23 vezes. Embora não era para acontecer. Então, alguma coisa aconteceu na realidade. O que não era possível, de acordo com o modelo. E por repetir realidade no topo do modelo, podemos ver esses tipos de problemas. Repetição não é apenas sobre a conformidade. É também sobre a análise de desempenho. Podemos facilmente reproduzir um evento registrado que tem

marcas de tempo em cima de tal modelo. E cada vez que executar uma atividade. Registrarmos o tempo correspondente, como você pode ver aqui neste slide. Então, nós apenas executar realidade na parte superior do modelo. Gravamos todos os tempos. E no final, sabemos exatamente quanto tempo foi gasto em todas as diferentes partes do processo. E os círculos pretos indicam aqui atrasos entre as atividades individuais. Assim, podemos fazer isso por uma única execução, por um único vestígio. Mas você também pode fazer isso por muitos vestígios. Se você fizer isso por muitos vestígios, você vai ver como muitas vezes uma atividade está sendo executada. Você vai ver quantas vezes um determinado caminho está sendo seguido. Mas você também ver quanto tempo atividades tenham se e você vai ver o que os atrasos são entre essas atividades. Quais são os tempos de espera. Quais são os horários de sincronização. Você pode gravá-los para centenas de milhares ou talvez milhões de casos. E derivar de distribuição de probabilidade a partir deles. Então mais uma vez, vamos dar uma olhada em um exemplo prático. Aqui vamos olhar novamente para um processo onde as pessoas queixam-se da valorização de sua casa. Temos derivado automaticamente um modelo de processo. E esses termos expressos como uma rede de petri. E agora podemos por repetir podemos ver onde os atrasos ocorrem. Então, se você olhar para este slide você pode ver que certos lugares são indicados em roxo. E estes são os lugares onde nós vemos um grande atraso. Portanto, esta é uma indicação de que possíveis problemas de desempenho são. Ele pode ir um passo além. Você pode apontar para dois pontos arbitrários no processo. Você pode ver como muitos casos, neste caso, quantas acusações estão fluindo a partir desta primeira atividade vermelho para a segunda atividade vermelho, e podemos ver quanto tempo levou. E você pode ver aqui as estatísticas correspondentes. Então, por repetir você pode ver exatamente onde os atrasos ocorrem. Você pode começar a investigar o que está causando. Então, para fornecer uma visão geral das coisas que eu tenho vindo a falar, temos um

mundo real onde as coisas estão acontecendo. Temos um sistema de software, isto é, de alguma forma, registrando eventos das coisas que acontecem. As pessoas se opunham à valorização de sua casa. Os doentes em tratamento em um hospital. Então nós temos esses dados do evento. E nós podemos fazer de descoberta, aprendendo automaticamente modelos de processos. Podemos fazer verificação de conformidade. Comparando-se o log de eventos para o modelo de processo e podemos enriquecer, é chamado de reforço aqui, podemos enriquecer o modelo de processo com informações sobre desvios e performance. Por exemplo, a garrafa pescoços que eu acabei de mostrar. Como é que se relacionam com as coisas que eu tenho vindo a falar? Play-out é sobre o uso clássico de modelos de processos que não envolvem quaisquer dados do evento. Jogue-in corresponde à descoberta. Você aprende automaticamente um modelo de processo sem qualquer modelagem de dados de eventos brutos. E você tem de repetição onde você comparar um modelo de processo para um log de eventos para verificar a conformidade, para investigar os problemas de desempenho, etcetera. Então, eu espero que este lhe deu uma boa visão geral dos três tipos diferentes de, mineração processo. Se você quiser encontrar mais informações de alto nível sobre estes três tipos, por favor leia o capítulo um. Muito obrigado por assistir e espero vê-lo em breve.

## **2 - 2 - Lecture 1.2- Different Types of Process Mining (END)**

---

---

## 2 - 3 - Lecture 1.3- How Process Mining Relates to Data Mining

Glad to see you again for this third lecture of the course on Process Mining, Data Science in Action. Today, we will see how process mining relates to data mining. Process mining, as I explained in the last lecture, is the missing link between model based analysis, process model based analysis, and data oriented analysis like data mining. With the goal to answer performance oriented questions and compliance oriented questions. So one can think of process mining as super glue. It's the glue between data and processes. It's the glue between business people and IT people. It's the glue between business intelligence and business process management. It's the glue between performance and compliance. And you can do this at runtime, and at design time. So it's connecting many different things, and that makes it so incredibly valuable. This diagram gives a kind of overview showing that process mining is this world, this discipline connecting business process management to classical data analytics like data mining. And, as we have seen in the last lecture, process mining consists of different types of mining, and here you can see process discovery and conformance checking. Later we will also look at predictive analytics. How you can use process mining to predict things about the future. If you see this diagram, you may ask yourself the question, how about BI, business intelligence? Well if you look at the classical BI tools, they tend to show only spreadsheets, meters, graphs, they try to capture reality in a set of numbers, and that is not sufficient. This is illustrated by Anscombe's Quartet that's the four diagrams that you see here. And let me explain what we are seeing. We are seeing four different data sets, each consisting of 11 elements. They look very different, but if you look at their statistical properties, they are almost the same. So if we look at the mean x coordinate of these 11 elements in each of the four cases the mean is nine, the variance is 11. If we look at the y coordinate, we also can see that all of these four diagrams are exactly the same. Even when we apply linear regression, all of these four data sets, yield exactly the same result.

So, this is illustrating, if a BI tool is just showing numbers, that is often not sufficient. You have to look behind the numbers, because things can be very different. One can think of process mining as finding desire lines. Here you can see a photo that I took at a campus of Tsinghua University in Beijing, and here you see one of my colleagues there walking over a desire line. So this is a trail in a grassy area that is showing what people really do. So, the trail can be seen as the event data, and the sign saying that you should not walk there, that you should use the official route can be seen as the process model. So this is an illustration of what process mining tries to do. You try to uncover the desire lines, what do people really do. And, you can find very surprising things. So for example if you look at this desire line, it is clearly showing that people do not follow the process model. The, the gates that you see here is aimed of keeping cyclists out of this park. But what you can see is that the desire line is showing that people do not follow the things that they should do here. So the gate does not work as it is supposed to work. So using process mining and conformance checking, you can use and show these types of things. Now it's time for a demo that we can see process discovery in action. Let's take a look at an example of process mining in action. What I see here is that I load an event log into the process mining tool ProM, and I apply one of the standard process mining algorithms in ProM. I'm applying the so-called fuzzy miner. And based on raw event data, I automatically learn a process model, as you can see here, and it goes incredibly fast. We can zoom in to the model. We can look at the different activities. And the arcs connecting the activities are showing how cases are flowing through this process model. We can zoom in and zoom out just like when you use Google Maps. So now I look at the same process model at the higher level of abstraction. We see fewer activities because we only show the frequent activities at the highest level. We can also replay reality on such a model. So I'm now loading both the model that I discovered and the event log, and I'm applying a so-called replay algorithm on it. So what you now see

is not a simulation. What you're seeing is that we are replaying reality on top of this model. So all the white dots, refer to real cases. What we can see, you can see the numbers next to the cases, we can follow individual cases. We can see all the events. We can see congestion in the process. We can see frequent paths and infrequent paths. And I think this is a very nice illustration that without any modeling, you can see what is happening in reality, through process mining. So let's take a look at another metaphor showing what process discovery is all about. You can compare process discovery to learning a language based on examples. So here you say see a mother, and the mother is saying sentences to the child. So the mother is saying a b c, and the child is trying to learn the language. So if the mother would always say just the words a b and c. The child would think okay, this language is a b c and that's all that exists. Then if the mother would say an a b d, the child will think oh, at the end there is apparently a choice between c and d. So, this child is very smart, thinking already in terms already of regular expressions. So, this is the model of the language that the child infers. Then mother says another sentence, a d. So, now the child has to think what could this language be about? So it's apparently a and d, or it is a b and then followed by c or d. Now the mother says, a b b c. Another example, trace that we can observe. And now the child makes a jump in learning, and infers that apparently the number of b's is variable, it can be zero or more etc. And so the mother says more words and if it fits into the language, then the child has a good understanding of that language. So, in this example we can see a sentence and compare that to a trace in an event log, and we can think of the language as a process model. So, this is the relationship between understanding a language based on examples from learning a process model just by looking at the examples. What we saw in the last lecture is that next to process discovery we also have conformance checking. In that case, you want to see how does reality deviate from the modeled process. You can compare that to spell checking. So the spell

checker has a model of the language, and you type a piece of text and then it is checked, whether that piece of text fits the language, as it has been formalized. This can be compared to the typical diagnostics that you get when you do conformance checking in the area of process mining. So you see the activities that have happened, but that should not happen, or the other way around. Or you will see activities that were executed too late, or too early, or by the wrong person. This can all be compared by spell checking. So in the remainder of this week, we will focus on the topic of Data Mining. As I have explained, also in the previous lecture, process mining is very different from the classical data mining techniques, but there are many relationships. So in this course, you will also get a basic understanding of what process mining is all about and of course also data mining. So, the growth of the digital universe is driving the fact that many people are using data mining techniques. Initially the term data mining had a very negative connotation. People talked about it as data snooping, fishing, etc. Statisticians did not consider it the proper way to do, but this is now a very mature discipline driven by these huge amounts of data. But data mining is data-centric and not process-centric. So let us take a look at some typical data sets, and then try to think what data mining can do. So, what you see here, is a data set of more than 800 persons that have died at a particular age while having a particular weight and it was recorded whether they were drinking or smoking. So for example, the first row corresponds to a person that died at the age of 44 while being a drinker and a smoker and having a weight of 120 kilos. So this is an example of a data set. And the types of questions that you can ask about such a data set are things like, do people that smoke also drink. When do people get old. What kind of properties do they have in common? The people typically that, that get old. So what is the impact of a certain lifestyle, on the life expectancy of a person? So this is one data set and typical data mining questions. This is another data set. And every row now corresponds to a student at the university. And what we see here

are different columns referring to different courses. We can see the marks that people got for each of these courses. We can see the duration of their studies in months, and we can see whether they passed, failed, or graduated cum laude. So again, if we have such a data set we can ask all kinds of questions. So, are there certain courses that are typically taken together, or do marks of different courses highly correlate? If people fail, what are the typical courses that are leading to such a failure. When and why do people drop out? It's a different type of analysis that you can do based on such a data set. But we can look at all kinds of other data sets. So here you can see the different orders in a cafe. So every row refers to an order, so if you look at the first row, you see a person that has ordered a cappuccino and a muffin. So every row corresponds to an order, and again you can try to learn all kinds of things. So, you can try to find out which are the products which are typically purchased together. You could try to find out are there characteristic groups of customers that typically consume similar things. And all of this can be used to try to promote sales. So these three data sets give all kinds of examples for data mining problems. So if you look at the raw data that we get in each of these examples. Is we, if I get a table, every row in the table corresponds to an instance. And every column in the table refers to a variable, often called attribute or feature or data element. If we look at these variables, these columns, then we can find two main groups. Numerical variables that refer to a number, like for example an age or a weight. And categorical variables that don't have a value that is a number, but these are values taken from a smaller set. So for example, cum laude passed or failed or true and false are examples of categorical variables. If they have an order, they're called ordinal. If not, they are called nominal. So these are the variables, the columns that we see in a data mining problem. So, to check whether you understood what I just explained, take a look at this data set. And my question is, what columns refer to ordinal categorical variables, which to nominal categorical variables, and which ones are numerical

variables? So please think about this for a minute. So the answer to question is that, there are two categorical variables. Referring to the first two columns where the people are a drinker or a smoker, these are not numbers and they are nominal because true and false do not have some natural order. There are two numerical variables, the weight and the age because they refer to a number. There are two types of data mining techniques, generally referred to as supervised learning and unsupervised learning. In the context of supervised learning, we have labeled data. Labeled data means that there is a response variable that labels the instance. And the goal of supervised learning is to learn from the other variables that are called predictor variables, what our response variable is going to be. So instead of response variable, we also talk about dependent variables. And instead of predictor variables we also talked about independent variables. And the goal is to explain the dependent variable in terms of the independent variables. So, classification techniques, like learning decision trees, aim to answer such questions. So what is the class depending on the set of variables that we know. If the response variable is numerical, we typically use regression techniques. And then the goal is to find the function that explains the response variable in terms of these other variables. So, let's take a look at this data set. So, from every instance, it refers to a person, and we know whether the person was drinking or not, smoking or not, and we know their weight. We would like to learn the influence of drinking and smoking on somebody's body weight. If you have that question, and you think of supervised learning, what are the response and predictor variables? I think the answer to this question is relatively easy. In this particular example the response variable is the weight. And we would like to explain the response variable in the predictor variables, whether people are drinking or smoking. Next to supervised learning we also have unsupervised learning, and now the data is unlabeled. In other words, we don't have response variables. And the typical techniques that you will then look at are clustering and pattern

discovery. So you want to find homogeneous groups of for example patients or customers, without aiming to look at a particular response variable. And these are just some examples of the many data mining questions that you can ask. And in the next lectures we will zoom in to some of these questions and I will teach you hands on knowledge to solve these types of problems. There are many data mining tools available, you can see some of them here. In my next lectures, I will use RapidMiner to illustrate these classical data mining techniques. You don't have to install RapidMiner for this course, but if you want, you can do and you can repeat experiments that I will show. There are many differences between process mining and data mining. Let me repeat some of the things. In what way are they common and in what way are they different? They both start from data, but data mining techniques are not process-centric. They look at isolated decisions of the type that I've just shown to you. Topics such as process discovery, conformance checking, bottleneck analysis and all the other things that I showed in the previous lecture, cannot be done using traditional data mining techniques. So you need to have process mining for that. End-to-end process models are crucial. And when you want to discover end-to-end process models, concurrency is important. So when we will talk about process models, we will deal with the topic of concurrency, because it's very important. Process mining assumes a different type of data than the data that we have just seen. We assume that we can see events. These events they have timestamps and they refer to cases. And that is a crucial difference with the data the three data sets that I showed you before. But, process mining and data mining can be combined to answer very advanced questions, so that's very important. If you would like to learn more about data mining and in the next lectures we will discuss some of the techniques in more detail, please read chapter three of the Process Mining book. Thank you for watching, and hope to see you soon.

## **2 - 3 - Palestra 1.3- Como processo de mineração Refere-se a Data Mining**

Fico feliz em vê-lo novamente para esta terceira palestra do curso no processo de mineração, Ciência de Dados em Ação. Hoje, vamos ver como mineração processo refere-se a mineração de dados. Processo de mineração, como expliquei na última palestra, é o elo perdido entre o modelo baseado análise, modelo de processo de análise com base, e análise de dados como a mineração de dados orientado. Com o objetivo de responder a perguntas orientadas desempenho e perguntas orientadas conformidade. Assim, pode-se pensar em mineração processo como super-cola. É a cola entre os dados e processos. É a cola entre os empresários e profissionais de TI. É a cola entre a inteligência de negócios e gerenciamento de processos de negócios. É a cola entre desempenho e compliance. E você pode fazer isso em tempo de execução, e em tempo de design. Então, ele está se conectando muitas coisas diferentes, e que o torna tão incrivelmente valioso. Este diagrama dá uma espécie de visão geral mostra que a mineração processo é este mundo, esta disciplina se conectar a gestão de processos de negócios para análise de dados clássicos como a mineração de dados. E, como vimos na última palestra, mineração processo consiste em diferentes tipos de mineração, e aqui você pode ver descoberta processo e verificação de conformidade. Mais tarde vamos também olhar para análise preditiva. Como você pode usar a mineração processo de prever as coisas sobre o futuro. Se você ver este diagrama, você pode perguntar-se a questão, como sobre BI, business intelligence? Bem, se você olhar para as ferramentas de BI clássicos, eles tendem a mostrar apenas planilhas, medidores, gráficos, eles tentam captar a realidade em um conjunto de números, e isso não é suficiente. Isto é ilustrado pelo Quarteto de Anscombe que é as quatro diagramas que você vê aqui. E deixe-me explicar o que estamos vendo. Estamos vendo quatro conjuntos de dados diferentes, cada uma composta por 11 elementos. Eles

parecem muito diferentes, mas se você olhar para as suas propriedades estatísticas, eles são quase os mesmos. Portanto, se olharmos para os médios x coordenar destes 11 elementos em cada um dos quatro casos, a média é de nove anos, a variação é de 11. Se olharmos para a coordenada y, também podemos ver que todos esses quatro diagramas são exatamente os mesmas. Mesmo quando aplicamos regressão linear, todos estes quatro conjuntos de dados, rendimento exatamente o mesmo resultado. Então, isso está ilustrando, se uma ferramenta de BI é apenas mostrar números, que muitas vezes não é suficiente. Você tem que olhar por trás dos números, porque as coisas podem ser muito diferentes. Pode-se pensar em mineração processo de como encontrar linhas de desejo. Aqui você pode ver uma foto que eu tomei em um campus da Universidade de Tsinghua, em Pequim, e aqui você vê um dos meus colegas de lá andando sobre uma linha de desejo. Portanto, esta é uma trilha em uma área gramada que está mostrando o que as pessoas realmente fazem. Então, a trilha pode ser visto como os dados do evento, eo sinal dizendo que você não deve andar lá, que você deve usar a rota oficial pode ser visto como o modelo de processo. Portanto, esta é uma ilustração do que a mineração processo tenta fazer. Você tenta descobrir as linhas de desejo, o que as pessoas realmente fazem. E, você pode encontrar coisas muito surpreendentes. Por exemplo, se você olhar para esta linha de desejo, é claramente mostrando que as pessoas não seguem o modelo de processo. Os, as portas que você vê aqui é voltado de manter os ciclistas fora deste parque. Mas o que você pode ver é que a linha desejo é mostrar que as pessoas não seguem as coisas que eles deveriam fazer aqui. Assim, o portão não funciona como deveria funcionar. Então, usando mineração processo e verificação de conformidade, você pode usar e mostrar esses tipos de coisas. Agora é hora de fazer uma demonstração de que podemos ver descoberta processo em ação. Vamos dar uma olhada em um exemplo de mineração processo em ação. O que eu vejo aqui é que eu carregar um log de eventos para a ferramenta

ProM mineração processo, e eu aplicar um dos algoritmos de mineração de processo padrão no baile. Estou aplicando o chamado mineiro difusa. E com base em dados do evento matérias, eu aprendo automaticamente um modelo de processo, como você pode ver aqui, e ele vai incrivelmente rápido. Podemos aumentar o zoom para o modelo. Podemos olhar para as diferentes atividades. E os arcos que ligam as atividades estão mostrando como casos estão fluindo através deste modelo de processo. Podemos zoom in e zoom out assim como quando você usa o Google Maps. Então agora eu olhar para o mesmo modelo de processo no nível mais alto de abstração. Vemos menos atividades porque nós só mostrar as atividades frequentes ao mais alto nível. Nós também pode reproduzir a realidade em tal modelo. Então, eu estou agora carregando tanto o modelo que eu descobri e log de eventos, e eu estou aplicando um algoritmo de repetição chamada nele. Então, o que agora você vê não é uma simulação. O que estamos vendo é que estamos repetindo realidade em cima deste modelo. Assim, todos os pontos brancos, referem-se a casos reais. O que podemos ver, você pode ver os números ao lado dos casos, podemos acompanhar os casos individuais. Podemos ver todos os eventos. Podemos ver o congestionamento no processo. Nós podemos ver caminhos freqüentes e caminhos pouco frequentes. E eu acho que isso é uma ilustração muito agradável que, sem qualquer modelagem, você pode ver o que está acontecendo na realidade, através da mineração processo. Então, vamos dar uma olhada em outra metáfora que mostra o que a descoberta processo é tudo. Você pode comparar descoberta processo de aprendizagem de uma língua com base em exemplos. Então aqui você diz ver uma mãe, e a mãe está dizendo frases para a criança. Então, a mãe está dizendo abc, e que a criança está tentando aprender a língua. Então, se a mãe sempre dizia apenas as palavras ab e c. A criança pode pensar bem, esta linguagem é abc e isso é tudo o que existe. Então, se a mãe diria um abd, a criança vai pensar oh, no final há, aparentemente, uma escolha entre c e d.

Então, essa criança é muito inteligente, já pensando em termos já de expressões regulares. Então, este é o modelo da linguagem que a criança deduz. Então a mãe diz outra frase, um d. Então, agora a criança tem que pensar o que esta língua poderia ser? Portanto, é aparentemente a e d, ou é ab e, em seguida, seguido por c ou d. Agora, a mãe diz, a b b c. Outro exemplo, trace que podemos observar. E agora que a criança faz um salto na aprendizagem, e infere que, aparentemente, o número de b do é variável, pode ser zero ou mais etc. E assim a mãe diz mais palavras e se ele se encaixa dentro da linguagem, então a criança tem uma boa compreensão de que a linguagem. Assim, neste exemplo, podemos ver uma frase e que ao comparar um traço em um log de eventos, e podemos pensar a língua como um modelo de processo. Então, essa é a relação entre a compreensão de uma linguagem baseada em exemplos de aprender um modelo de processo só de olhar para os exemplos. O que vimos na última palestra é que ao lado de processo de descoberta também temos verificação de conformidade. Nesse caso, você quer ver como é que a realidade se desviam do processo modelado. Você pode comparar isso a verificação ortográfica. Assim, o corretor ortográfico tem um modelo da língua, e você digitar um pedaço de texto e, em seguida, é verificado, se essa parte do texto se encaixa a língua, uma vez que foi formalizado. Isto pode ser comparado com os diagnósticos típicos que você começa quando você faz verificação de conformidade na área de mineração processo. Então você vê as atividades que aconteceram, mas isso não deve acontecer, ou o contrário. Ou você vai ver as atividades que foram executadas tarde demais, ou muito cedo, ou pela pessoa errada. Isso tudo pode ser comparado pela verificação ortográfica. Assim, no restante desta semana, vamos nos concentrar sobre o tema da Data Mining. Como já expliquei, também na aula anterior, mineração processo é muito diferente das técnicas de mineração de dados clássicos, mas há muitos relacionamentos. Então, neste curso, você também irá receber uma compreensão básica do que a

mineração processo é sobre tudo e, claro, também de mineração de dados. Assim, o crescimento do universo digital está impulsionando o fato de que muitas pessoas estão usando técnicas de mineração de dados. Inicialmente, a mineração de dados prazo teve uma conotação muito negativa. As pessoas falavam sobre isso como espionagem de dados, pesca, etc. Os estatísticos não considerou a maneira correta de fazer, mas isso agora é uma disciplina muito maduro impulsionado por essas enormes quantidades de dados. Mas mineração de dados é centrada em dados e não-centric processo. Por isso, vamos dar uma olhada em alguns conjuntos de dados típicos, e, em seguida, tentar pensar o que a mineração de dados pode fazer. Então, o que você vê aqui, é um conjunto de mais de 800 pessoas que morreram em uma determinada idade, tendo um peso específico de dados e foi gravado se eles estavam bebendo ou fumando. Assim, por exemplo, a primeira linha corresponde a uma pessoa que morreu com a idade de 44 ao mesmo tempo que um bebedor e um fumador e que tem um peso de 120 kg. Portanto, este é um exemplo de um conjunto de dados. E os tipos de perguntas que você pode perguntar sobre um tal conjunto de dados são coisas como, é que as pessoas que fumam também bebem. Quando é que as pessoas envelhecem. Que tipo de propriedades que eles têm em comum? As pessoas normalmente que, que envelhecem. Então, qual é o impacto de um certo estilo de vida, sobre a expectativa de vida de uma pessoa? Portanto, este é um conjunto de dados e questões típicas de mineração de dados. Este é um outro conjunto de dados. E cada linha passou a corresponder a um estudante na universidade. E o que vemos aqui são diferentes colunas referentes a diferentes cursos. Nós podemos ver as marcas que as pessoas tem para cada um desses cursos. Podemos ver a duração de seus estudos em meses, e podemos ver se passou, falhou, ou graduado cum laude. Então, novamente, se temos um tal conjunto de dados que pode pedir a todos os tipos de perguntas. Assim, existem determinados cursos que

normalmente são tomadas em conjunto, ou fazer marcas de diferentes cursos altamente correlacionados? Se as pessoas não conseguem, quais são os pratos típicos que estão levando a um incumprimento. Quando e por que as pessoas desistem? É um tipo diferente de análise que você pode fazer com base em tal conjunto de dados. Mas podemos olhar para todos os tipos de outros conjuntos de dados. Então, aqui você pode ver as diferentes ordens em um café. Assim, cada linha refere-se a uma ordem, por isso, se você olhar para a primeira linha, você vê uma pessoa que encomendou um cappuccino e um muffin. Assim, cada linha corresponde a uma ordem, e, novamente, você pode tentar aprender todos os tipos de coisas. Assim, você pode tentar descobrir quais são os produtos que são normalmente adquiridos em conjunto. Você poderia tentar descobrir o que há grupos característicos de clientes que normalmente consomem coisas semelhantes. E tudo isto pode ser usado para tentar promover vendas. Então, esses três conjuntos de dados dar todos os tipos de exemplos de problemas de mineração de dados. Então, se você olhar para os dados brutos que nós temos em cada um destes exemplos. É que, se eu conseguir uma mesa, cada linha na tabela corresponde a uma instância. E cada coluna na tabela refere-se a uma variável, muitas vezes chamado de atributo ou recurso ou elemento de dados. Se olharmos para essas variáveis, estas colunas, então podemos encontrar dois grupos principais. As variáveis numéricas que se referem a um número, como por exemplo uma idade ou um peso. E as variáveis categóricas que não têm um valor que é um número, mas estes são os valores retirados de um conjunto menor. Assim, por exemplo, cum laude passou ou não ou verdadeiro e falso são exemplos de variáveis categóricas. Se eles têm uma ordem, eles são chamados de ordinal. Se não, eles são chamados nominal. Então, essas são as variáveis, as colunas que vemos em um problema de mineração de dados. Assim, para verificar se você entendeu o que eu acabei de explicar, dê uma olhada neste conjunto de dados. E a minha

pergunta é, o que colunas referem-se a ordinal variáveis categóricas, que para as variáveis categóricas nominais, e quais são as variáveis numéricas? Então, por favor, pense sobre isso por um minuto. Portanto, a resposta à pergunta é que, há duas variáveis categóricas. Referindo-se às duas primeiras colunas onde as pessoas são um bebedor ou um fumante, estes não são os números e eles são nominal, porque o verdadeiro eo falso não tem um pouco de ordem natural. Há duas variáveis numéricas, o peso ea idade porque se referem a um número. Existem dois tipos de técnicas de mineração de dados, geralmente referidos como aprendizado supervisionado e aprendizado não supervisionado. No contexto da aprendizagem supervisionada, que rotulamos de dados. Dados rotulados significa que há uma resposta variável que rotula o exemplo. E o objetivo do aprendizado supervisionado é aprender com as outras variáveis que são chamados de variáveis de previsão, o que a nossa variável de resposta vai ser. Então, ao invés de variável de resposta, também falamos sobre variáveis dependentes. E, em vez de variáveis de previsão também falou sobre as variáveis independentes. E o objetivo é explicar a variável dependente em termos das variáveis independentes. Assim, técnicas de classificação, como aprender árvores de decisão, visam responder a essas perguntas. Então, qual é a classe dependendo do conjunto de variáveis que conhecemos. Se a variável resposta é numérica, que normalmente usam técnicas de regressão. E então o objetivo é encontrar a função que explica a variável resposta em termos dessas outras variáveis. Então, vamos dar uma olhada neste conjunto de dados. Assim, a partir de todos os casos, refere-se a uma pessoa, e nós sabemos que se a pessoa estava bebendo ou não, fumar ou não, e sabemos que o seu peso. Gostaríamos de saber a influência de beber e fumar no peso corporal de alguém. Se você tem essa pergunta, e você acha de aprendizado supervisionado, quais são as variáveis de resposta e de previsão? Acho que a resposta a esta pergunta é relativamente fácil. Neste exemplo particular, a variável de resposta é o peso. E

nós gostaríamos de explicar a variável resposta nas variáveis de previsão, se as pessoas estão bebendo ou fumando. Ao lado de aprendizado supervisionado temos também sem supervisão de aprendizagem, e agora os dados estão sem rótulos. Em outras palavras, não temos variáveis de resposta. E as técnicas típicas que então você vai olhar são agrupamento e descoberta de padrões. Então você quer encontrar grupos homogêneos de, por exemplo, pacientes ou clientes, sem o intuito de olhar para uma variável de resposta particular. E estes são apenas alguns exemplos das muitas questões de mineração de dados que você pode perguntar. E nos próximos palestras vamos ampliar para algumas dessas perguntas e eu vou ensinar-lhe as mãos em conhecimento para resolver esses tipos de problemas. Existem muitas ferramentas de mineração de dados disponíveis, você pode ver alguns deles aqui. Na minha próxima palestra, vou usar RapidMiner para ilustrar estas técnicas de mineração de dados clássicos. Você não tem que instalar RapidMiner para este curso, mas se você quiser, você pode fazer e você pode repetir os experimentos que eu te mostrarei. Há muitas diferenças entre a mineração processo e mineração de dados. Deixe-me repetir algumas das coisas. De que forma eles são comuns e de que forma eles são diferentes? Os dois começam a partir de dados, mas as técnicas de mineração de dados não são centradas no processo. Eles olham para as decisões isoladas do tipo que eu acabei de mostrar a você. Temas como a descoberta de processos, verificação de conformidade, análise gargalo e todas as outras coisas que eu mostrei na palestra anterior, não pode ser feito usando técnicas tradicionais de mineração de dados. Então, você precisa ter o processo de mineração para isso. Modelos de processos end-to-end são cruciais. E quando você quer descobrir modelos de processos end-to-end, a concorrência é importante. Então, quando vamos falar sobre modelos de processo, nós vamos lidar com o tema da concorrência, porque é muito importante. Mineração processo pressupõe um tipo diferente de dados do que

os dados que acabamos de ver. Nós assumimos que podemos ver eventos. Estes eventos têm data e hora e eles se referem a casos. E essa é uma diferença crucial com os dados os dados três conjuntos que eu mostrei antes. Mas, mineração processo e mineração de dados podem ser combinadas para responder a perguntas muito avançados, por isso é muito importante. Se você gostaria de aprender mais sobre a mineração de dados e nas próximas palestras vamos discutir algumas das técnicas mais detalhadamente, por favor leia o capítulo três do livro processo de mineração. Obrigado por assistir, e espero vê-lo em breve.

## **2 - 3 - Lecture 1.3- How Process Mining Relates to Data Mining (END)**

---

---

## 2 - 4 - Lecture 1.4- Learning Decision Trees

Welcome to this lecture of the course on Process Mining: Data Science in Action. Before we focus on process mining, it is good to have a solid understanding of the mainstream data mining techniques. Hence, there will be several lectures on data mining. Today we start by presenting a technique to learn decision trees. This slide shows that process mining is the linking pin between data-oriented analysis, and process-oriented analysis. Before we focus on process discovery techniques, conformance checking, Predictive analytics, and other forms of process mining. We now look at so-called decision tree learning, which is one of many available data mining techniques. So what is a decision tree? In a decision tree we have a number of predictor variables, and based on these predictor variables. We try to predict, what the so-called response variable is. Decision tree learning is a form of supervised learning, because the data is labeled. Labeled using the response variable. In our case, this is categorical data. Let's take a look at some examples. For example, we would like to know what the effect is of lifestyle on how old people get. One can take this as input data, so we know whether people are drinking, smoking, what their weight is. And we know the age at which people have died. Now we try to predict, whether people will die at an older age or at a younger age. So as a response variable, we take the age and we make it discrete. We turn it into a categorical variable, so people that die about 70. Are labeled old, people that die under the age of 70 are labeled young. And the other variables are used as predictor variables. So we would like to predict the response variable in terms of these predictor variables. This is a so called decision tree. It is learned based on the data, and what it tells you, for example, that if people are smoking, they are likely to die young. If people are not smoking, they are not drinking, then they are likely to die old. We can also see numbers in the leaves of such a tree. This is explaining, what was done with the training set used to learn this decision tree. In this case there were 195 smokers. They were all

classified as young, but there were 11 smokers that actually died at an older age so that are classified incorrectly. The other leaves have similar numbers. For example, there were two people that were not drinking, that were not smoking, but still died young. Okay. Let's see whether we can read such a decision tree. Take Mary. She is drinking, but she's not smoking. Her weight is 70 kilos, and she died at the age of 85. So the question is, is Mary classified correctly in this decision tree. The answer is yes. And the red path shows what Mary's path is through this decision tree. So Mary was not smoking, she was drinking, and her weight was below 90, so according to the decision tree she should be classified as old, dying above the age of 70. And indeed she did. Let's take a look at another person, Sue. Is she classified correctly according to this decision tree? Again we follow the path. So Sue is not a smoker and is not a drinker. So it is predicted that she will die at an older age, above 70, but she died at 35. So she's not classified correctly according to this decision tree. Let's take a look at another data set. Suppose that we have data about. The marks that student had for individual courses. If there is a dash in this table it means that the course was actually not taken, and so no result is known. In all the other cases, there's a number between one and ten indicating how well somebody made the exam for that particular course. Then we could be interested in the duration, or we could also be interested in the fact whether people pass, fail, or graduate cum laude. If we are interested in the latter column. Then the response variable is the result, which is cum laude, passed, or failed, and the predictor variables are the grades for all the individual courses. Using such data, we can again learn a decision tree, and here you see an example of it. So. People that had a mark lower than 8 for logic and had a mark of 6 or higher for linear algebra, it is predicted that they have, that they will pass. People that did not make the course on logic, or did not have a grade for it are predicted to fail. So again, the decision tree makes predictions by learning from a larger set of examples. Let's take a look at the third data set. Here we see what

people are buying in a coffee shop, and so people may order cappuccino. latte, espresso. Cafe Americano, tea, and they may eat a muffin or a bagel. As a response variable we take the column muffin, and we again make it discrete into a categorical variable, muffin or no muffin. And so we ignore the numbers in this table. Just whether things are present or not. So this is decision tree that we could learn over this data. So what we see is that people that drink tea according to this decision tree also eat a muffin. People that do not drink a tea, do not drink latte. Are predicted to not order a muffin. So, let's take again a look at some questions. So here you see the check of a visitor of this cafe. And this visitor ordered 2 cappuccinos, 3 lattes, 1 muffin and 2 bagels. Is this particular example classified correctly according to the decision tree? The answer is yes, and here in red, you can see again the path. People that, do not drink tea but drink at least two lattes. They are predicted to also order a muffin. And that is the case in this situation. Let's take a look at another example. A person ordered a bagel, a latte, and a ristretto. The question is, is this person classified correctly? If we follow the path through the decision tree, it is predicted that such a person would order a muffin. But this is actually not the case. And so this would be an instance that would be classified incorrectly by this decision tree. We can use the decision tree to make predictions over unseen instances. So, suppose that we have this check, we know that somebody ordered the bagel, two lattes, and one cappuccino. Then the question is, is it likely that the person also ordered a muffin? Then we can use a decision tree. So, in this situation what would a decision tree tell? The decision tree would tell that this person would indeed order a muffin, and that is the case because the person ordered no tea. At least 2 lattes, and people that fall into this class are predicted to order a muffin. So how does this work? We now have seen what a decision tree is, how we can use it. It can be used for understanding data, for predicting data, but how when can we automatically learn such a decision tree. Well, it is done by splitting nodes to reduce the

variability within every node. So, if we look at this, if we look at 6 persons, and they are all in the same category, 3 are red and 3 are green. Then we have a high entropy. We are very uncertain, whether it should be green or red. The idea is that we split nodes that we are uncertain about into smaller sets, smaller classes that are more homogenous. So, for example, if we split a set of people, into smokers and non-smokers, and we find that the 2 smokers died at a younger age labeled here as red then we reduce the variation within the individual subsets. So, we are still uncertain about people who do not smoke. And we could again split the group of non-smokers into two groups. The people that drink, and the people that do not drink. What we then find is that the variation within the group of people that do not drink. And do not smoke. That all of these people live longer. So, this overview slide shows that we try to go from a bigger class with high entropy, a high degree of uncertainty, to smaller classes where we are more certain about. And, this way we can incrementally build a decision tree. What is crucial for understanding decision trees, is that you have a good understanding of the notion of entropy. So, entropy is the degree of uncertainty. One can also think of entropy as the inverse of compressibility. Or zippability. If there is very little variation within a group, then we can compress the data very much. The goal is now to reduce the entropy in the leaves of the decision tree. And in this way we improve the predictability. Of elements that belong to a particular class. To formalize the notion of entropy, we need to use logarithms. And this is, basic high school math, mathematics. But still it is repeated so that you easily can apply the formulas on the coming slides. For example if we take the logarithm of 2 to the power n, then the result is n. If we take the logarithm of 1 divided by 2 to the power n, we get minus n. Here you can see some examples. So for example, the logarithm of 1 is equal to zero. The logarithm of 1024 is equal to 10. So what is not the formula for entropy? The formula for entropy takes the sum over a set of values. So suppose that we have k possible values. We enumerate

them from 1 to k and then  $\pi_i$  is the probability or in other words, the fraction of elements having this value. So we can estimate this  $\pi_i$ , this probability by taking the number. Of elements that have that particular value by the set of all elements. As we divide  $c_i$  by  $n$ , and then we get the fraction  $\pi_i$ . And then we apply this formula. So we take the sum over  $\pi_i$  times the logarithm of  $\pi_i$ . This looks very complicated, but if one starts to apply it, it will become more clear. So, if we have two groups of people that are represented in a particular class, and of both types of people we have the same amount, then that is the worst case situation. And so we have 3 reds, and 3 green dots. So the entropy if we apply this formula is equal to 1. What this means is that we need to have one bit to encode whether a person belongs to the red labeled class or the green labeled class. If we now split based on the attribute smoker we get more homogeneous groups. Let's see how this is reflected in entropy. So if we compute the entropy of the smokers. There are 2 smokers, and they are both labeled red, meaning that they both die at a younger age. If we now look at the entropy and we fill out a formula, the entropy is equal to zero. We need no information. We need no bits to encode. What people in this class, whether they are dying young or not. Because that they all have the same label. Now we can take a look at the class of all people that do not smoke. 3 persons in the class, live longer. 1 person in this class live shorter. If you now apply the formula, then we get a value of 0.8. And so we can see that the entropy, we do not need to have one bit, we can use slightly less. We can split further based on whether people are drinking or not. So again, we have the class of smokers. Still the entropy is equal to zero. We can take a look at the people that do not smoke but that do drink, and the entropy is equal to 1. Because both groups are equally represented. It is a one-to-one relationship. Finally, we have the people who do not drink and do not smoke, the entropy of the group is equal to zero. So this is the way that we can compute entropy, and here we can see all the numbers that we have just computed. Using this basic formula. We can now take the

weighted average of these 3 decision trees. So if we take a look at the first decision tree there is just one leaf. And in this leaf the entropy is equal to 1. And it is the complete fractions, so 6 divided by 6. So, the overall weighted average of entropy is equal to 1. If we take a look at the second decision tree where we split based on the label smoker, and we take the weighted average of zero and 0.811. Then we can see that the entropy is 0.54. Note that 2 of the 6 persons ended up in the class that had an entropy equal to zero. The rest had an entropy of 0.811. That is why we need to apply the formula in this way. We can split based on the label smoker. Now we have three, leaves. Each of these leaves has a weight, depending on the number of people in that particular leaf. We take the weighted average of the different entropy values, and we see that the resulting entropy is 0.33. So what we can see is that in the first decision tree, the weighted average over all these entropy values was 1, and by splitting, based on the label smoker, it went down from one to 0.54. This means that we had an information gain of 0.46. If we take the tree in the middle, we can split based on the level drinker. And we get an entropy of one-third. So now the information gain is 0.21. So we would like to reduce the entropy. And we would like to split on the labels that maximize the reduction of this entropy, so that maximize information gain. So the idea of the algorithm is now that we continue splitting labels. Trying to maximize information gain, but stop if this is no longer possible. So let's take a look at some questions to see whether you really understand the notion of entropy. Here you see several sets of colored balls. And we would like to compute the entropy of all the individual cells. If you do that then you can compute the overall entropy of all of these 9 different squares. And we can compare that with the overall. Entropy if it would put all the balls in 1 big bin, and then compute the entropy. And so please, try to compute these values that are asked for here. Okay, let's first take a look at the cell in the middle. If we look at the cell in the middle. There are, 16 balls. Two of each color. And there are 8 different colors. If we then

apply the entropy formula, we find that the entropy is equal to three. So, we need to have 3 bits to encode the color of a single ball. In the square in the middle. If we not take a look at the other cells, they all contain 16 balls. But they all have the same color. So the entropy for all these other cells is equal to zero. And we can learn that by just filling out the formula that we have seen before. So, if we now look at the over all entropy, we need to take the weighted average of these 9 different squares and we get an entropy of one-third. That's right, entropy of all cells is zero, the cell in the middle has an entropy of three. There are nine cells, so the resulting weighted average is 0.33. What is the entropy after we take all the different cells and mix them? Then we have 144 balls having 18 different colors. So this is the distribution that we then get, and this corresponds to an entropy of three. So if we now compare these 2 situations, if we have these 9 different cells. Where the cell in the middle has all the colors, and all the other cells have just 1 color, we have a much lower entropy. Then when we have one big cell containing all balls, because then the entropy is equal to 3. So, if we move from the situation with an entropy of 3 to the entropy of one-third, we have an information gain equal to 2.666. So this is what is being used when learning a decision tree. So, you see another example. We start by classifying all people as dying young. Then we split on the attribute smoker. So we go from the attribute that is listed here. To these entropies for the 2 new leaves that we find. Then we can take the weighted average, and compare this weighted average of the overall entropy to the original entropy. And we can see that there is indeed an information gain of 0.107. Although the classification did not change there was information gain. And this is because the group of smokers is now more homogenous than it was before. We can split further. And again, we can look at the original entropy values. This is the weighted average. And compare that to the new situation, where we can basically apply the formulas that we have seen before. So these are the entropy values for the different leaves. We again take the

weighted average. We can compare both, and we see that there is an information gain. Of 0.07. So, this shows you the basic idea of the decision tree algorithm. We start with the root node that has all instances, and then we iteratively go through all the nodes. And see whether we can achieve an information gain. We do that by trying to split on all the possible attributes, and see whether the entropy is indeed reduced. So we select the attribute with the biggest information gain, above a certain threshold. And then we split that node, and we continue doing that until no significant improvements are possible. Then we return the decision tree. So that's the basic idea of learning a decision tree, using the notion of entropy. There are many parameters and variations possible to do this. For example one can define a minimal size of a node before or after splitting, to avoid overfitting, having all kinds of leaves that correspond to for example just a single instance. We can set the threshold on the minimal gain that is needed, and stop splitting if this gain cannot be achieved. We can define a maximal depth of the tree. We can also have, allow for multiple times using a label. Along a certain path. Some decision tree algorithms do not allow it, others do. But rather than using entropy, we can use alternative notions, like the Gini index of diversity. If we have a numerical variable, we can make it. Automatically categorical, using particular types of techniques. And so we split the domain of a numerical variable to make it categorical. We can also do a post pruning of the tree to remove the leaf notes that do not significantly increase the explanatory power of the resulting decision trees. Decision tree learning has many applications also in the field of process mining. So for example, if we take a look at this process model, then we ask ourselves what is driving these decisions? These 2 decision points in this process model, why are certain instances going left and our other instances going right. What is the most likely path of a running case giving its attributes? If we look at the particular case, we would like to predict whether it will be late or rejected. Using decision tree learning on top of process models, we can do that. But

it is crucial to see that these questions require a discovered process, otherwise none of this is possible. So process discovery is necessary before we can use decision tree learning. Today was the first lecture that we start talking about data mining techniques. Please read chapter 3 to learn more about decision tree learning. Thank you for watching this lecture, see you next time.

## 2-4 - Palestra 1.4- Entendendo as Árvores de Decisões

Bem-vindo a esta palestra do curso no processo de mineração: Ciência de Dados em Ação. Antes de focarmos mineração processo, é bom ter uma sólida compreensão das técnicas de mineração de dados convencionais. Assim, haverá várias palestras sobre mineração de dados. Hoje começamos por apresentar uma técnica para aprender árvores de decisão. Este slide mostra que a mineração processo é o pino de ligação entre a análise orientada a dados e análise orientada a processos. Antes de se concentrar em técnicas de descoberta de processos, verificação de conformidade. A análise preditiva, e outras formas de mineração processo. Vamos agora olhar para o chamado aprendizado árvore de decisão, que é uma das muitas técnicas de mineração de dados disponíveis. Então, o que é uma árvore de decisão? Em uma árvore de decisão que têm um número de variáveis de previsão, e baseado nessas variáveis preditoras. Tentamos prever quais o chamado variável de resposta é. Aprendizagem Árvore de decisão é uma forma de aprendizado supervisionado, porque os dados são rotulados. Rotulados com a variável resposta. No nosso caso, este é dados categóricos. Vamos dar uma olhada em alguns exemplos. Por exemplo, nós gostaríamos de saber qual é o efeito do estilo de vida sobre a forma como as pessoas idosas recebem. Pode-se tomar isso como dados de entrada, por isso, saber se as pessoas estão bebendo, fumando, o que o seu peso é. E nós sabemos que a idade em que as pessoas morreram. Agora vamos tentar prever, se as pessoas vão morrer em uma idade mais avançada ou em uma idade mais jovem. Então, como uma variável de resposta, nós tomamos a idade e nós torná-lo discreto. Nós transformá-lo em uma variável categórica, para que as pessoas que morrem cerca de 70. são rotulados de idade, as pessoas que morrem com idade inferior a 70 são rotulados jovem. E as outras variáveis são usadas como variáveis de previsão. Então, nós gostaríamos de prever a variável resposta em termos dessas variáveis de previsão. Este é um chamado de árvore de decisão. É que aprendi com base nos

dados, e que ela lhe diz, por exemplo, que, se as pessoas estão de fumar, eles são propensos a morrer jovem. Se as pessoas não estão fumando, eles não estão a beber, então eles são propensos a morrer de idade. Nós também podemos ver os números nas folhas de uma tal árvore. Este é explicar, o que foi feito com o conjunto de treinamento utilizado para aprender esta árvore de decisão. Neste caso, havia 195 fumantes. Eles foram todos classificados como jovens, mas havia 11 fumantes que realmente morreram em uma idade mais avançada, de modo que são classificados incorretamente. As outras folhas têm números semelhantes. Por exemplo, havia duas pessoas que não foram potável, que não foram fumantes, mas morreu ainda jovem. Ok. Vamos ver se podemos ler tal árvore de decisão. Tome Mary. Ela está bebendo, mas ela não está fumando. Seu peso é de 70 quilos, e ela morreu com a idade de 85. Então a questão é, é Mary classificou corretamente nesta árvore de decisão. A resposta é sim. E o caminho vermelho mostra o caminho de Maria é através desta árvore de decisão. Então, Maria não estava fumando, ela estava bebendo, e seu peso era inferior a 90, isso de acordo com a árvore de decisão que ela deve ser classificada como de idade, morrendo acima da idade de 70. E, de fato ela o fez. Vamos dar uma olhada em outra pessoa, Sue. Será que ela classificou corretamente de acordo com esta árvore de decisão? Mais uma vez, siga o caminho. Então, Sue não é um fumante e não é um bebedor. Assim, prevê-se que ela vai morrer numa idade mais avançada, acima de 70, mas ela morreu aos 35. Então ela não está classificado corretamente de acordo com esta árvore de decisão. Vamos dar uma olhada em outro conjunto de dados. Suponha que temos dados sobre. As marcas que o aluno tinha para cursos individuais. Se houver um traço nesta tabela significa que o curso efectivamente não foi feita, e de modo nenhum resultado é conhecido. Em todos os outros casos, há um número entre um e dez indicando quão bem alguém fez o exame para o referido curso particular. Então nós poderíamos estar interessado na duração, ou

também pode estar interessado no fato de saber se as pessoas passam, falhar, ou pós-graduação cum laude. Se estivermos interessados na última coluna. Então, a variável resposta é o resultado, que é cum laude, passou, ou não, e as variáveis de previsão são as notas para todos os cursos individuais. Usando esses dados, podemos voltar a aprender uma árvore de decisão, e aqui você vê um exemplo disto. So. Pessoas que tiveram uma marca inferior a 8 para a lógica e tinha uma marca de 6 ou superior para a álgebra linear, prevê-se que eles têm, que eles vão passar. As pessoas que não fizeram o curso na lógica, ou não tiveram uma nota para ele são previstos para falhar. Então, novamente, a árvore de decisão faz previsões, aprendendo com um conjunto maior de exemplos. Vamos dar uma olhada no terceiro conjunto de dados. Aqui vemos o que as pessoas estão comprando em uma loja de café, e assim as pessoas podem pedir cappuccino. latte, espresso. Café Americano, chá, e eles podem comer um bolo ou um pão. Como uma variável de resposta tomamos o muffin coluna, e nós novamente torná-lo pontuais em uma variável categórica, muffin ou nenhum muffin. E, assim, ignorar os números nesta tabela. Apenas se as coisas estão presentes ou não. Portanto, esta é árvore de decisão que poderíamos aprender sobre esses dados. Então, o que vemos é que as pessoas que bebem chá de acordo com esta árvore de decisão também comer um muffin. Pessoas que não bebem um chá, não bebo café com leite. Estão previstas para não pedir um muffin. Então, vamos ter novamente uma olhada em algumas questões. Então, aqui você vê o cheque de um visitante deste café. E este visitante pedimos 2 cappuccinos, 3 lattes, um muffin e 2 bagels. É este exemplo específico classificou corretamente de acordo com a árvore de decisão? A resposta é sim, e aqui em vermelho, você pode ver novamente o caminho. Pessoas que, não bebem chá, mas beber pelo menos dois lattes. Eles estão previstos para também pedir um muffin. E esse é o caso dessa situação. Vamos dar uma olhada em outro exemplo. Uma pessoa pediu um bagel, um café com leite e um ristretto. A questão

é, se esta pessoa classificou corretamente? Se seguirmos o caminho através da árvore de decisão, prevê-se que essa pessoa seria encomendar um muffin. Mas isso não é realmente o caso. E assim que esta seria uma instância que seria classificado incorretamente por esta árvore de decisão. Podemos usar a árvore de decisão para fazer previsões sobre instâncias invisíveis. Então, suponhamos que temos essa verificação, sabemos que alguém ordenou o bagel, duas lattes, e um cappuccino. Então, a questão é, é provável que a pessoa também ordenou um muffin? Então, podemos usar uma árvore de decisão. Portanto, nesta situação o que seria uma árvore de decisão dizer? A árvore de decisão diria que essa pessoa seria realmente pedir um muffin, e esse é o caso, porque a pessoa pedidos não chá. Pelo menos 2 lattes, e as pessoas que se enquadram nesta classe estão previstas para encomendar um muffin. Então, como isso funciona? Vimos agora o que uma árvore de decisão é, como podemos usá-lo. Ele pode ser usado para a compreensão de dados, para a previsão de dados, mas como quando podemos aprender automaticamente tal árvore de decisão. Bem, isso é feito por nós de divisão para reduzir a variabilidade dentro de cada nó. Então, se olharmos para isso, se olharmos para 6, pessoas, e eles estão todos na mesma categoria, 3 são vermelhas e 3 são verdes. Então nós temos um alto entropia. Estamos muito incerto, wether que deveria ser verde ou vermelho. A idéia é que nós dividimos nós que estamos incertos sobre em conjuntos menores, classes menores, que são mais homogênea. Assim, por exemplo, se nós dividir um conjunto de pessoas, em fumantes e não-fumantes, e nós achamos que os dois fumantes morreram em uma idade mais jovem marcado aqui como vermelho, depois, reduzir a variação dentro dos subgrupos individuais. Então, ainda estamos incertos sobre as pessoas que não fumam. E nós poderíamos novamente dividir o grupo de não-fumantes em dois grupos. As pessoas que bebem, e as pessoas que não bebem. O que nós, em seguida, encontrar é que a variação dentro do grupo de pessoas que não bebem. E não fumar. Que todas essas

pessoas vivem mais tempo. Então, este slide visão geral mostra que nós tentamos ir de uma classe maior, com alta entropia, um alto grau de incerteza, para turmas mais pequenas onde estamos mais certeza sobre. E, desta forma podemos gradualmente construir uma árvore de decisão. O que é crucial para a compreensão de árvores de decisão, é que você tem uma boa compreensão da noção de entropia. Assim, a entropia é o grau de incerteza. Pode-se também pensar em entropia como o inverso da compressibilidade. Ou zippability. Se houver uma variação muito pequena dentro de um grupo, então podemos comprimir os dados muito. O objectivo agora é para reduzir a entropia das folhas da árvore de decisão. E, desta forma, melhorar a previsibilidade. De elementos que pertencem a uma classe particular. Para formalizar a noção de entropia, precisamos usar logaritmos. E este é, de base matemática do ensino médio, da matemática. Mas ainda é repetida de modo que você pode facilmente aplicar as fórmulas nos próximos slides. Por exemplo, se tomarmos o logaritmo de 2 elevado à potência n, então o resultado é n. Se tomarmos o logaritmo de 1 dividido por 2 elevado à potência n, temos menos n. Aqui você pode ver alguns exemplos. Assim, por exemplo, o logaritmo de 1 é igual a zero. O logaritmo de 1024 é igual a 10. Então, o que não é a fórmula para a entropia? A fórmula para a entropia leva a soma ao longo de um conjunto de valores. Então suponho que temos k valores possíveis. Nós enumerar os de 1 a k e, em seguida, o pi é a probabilidade ou em outras palavras, a fracção de elementos que têm esse valor. Assim, podemos estimar este pi, essa probabilidade, tendo o número. De elementos que têm esse valor determinado pelo conjunto de todos os elementos. Como dividimos  $c_i$  por n, e então nós começamos a pi fração. E, depois, aplicar esta fórmula. Então, tomamos a soma sobre pi vezes o logaritmo do pi. Isso parece muito complicado, mas se alguém começa a aplicá-la, ela se tornará mais clara. Então, se temos dois grupos de pessoas que são representados em uma classe particular, e de ambos os tipos de pessoas que têm a

mesma quantidade, então essa é a pior situação possível. E assim temos 3 vermelhos, e 3 pontos verdes. Assim, a entropia se aplicarmos esta fórmula é igual a 1.0 que isto significa é que precisamos ter um pouco de codificar se uma pessoa pertence à classe marcado vermelho ou verde com a indicação da classe. Se nós agora dividida com base no atributo fumante temos grupos mais homogêneos. Vamos ver como isso se reflete na entropia. Então, se nós calcular a entropia dos fumantes. Há 2 fumantes, e ambos são rotulados vermelho, o que significa que ambos morrem em uma idade mais jovem. Se olharmos agora para a entropia e preencher uma fórmula, a entropia é igual a zero. Não precisamos de informações. Nós não precisamos de bits para codificar. O que as pessoas nesta classe, se eles estão morrendo jovem ou não. Porque que todos eles têm o mesmo rótulo. Agora podemos dar uma olhada na classe de todas as pessoas que não fumam. 3 pessoas na classe, viver mais tempo. 1 pessoa nesta classe viver mais curto. Se você agora aplicar a fórmula, então teremos um valor de 0,8. E assim podemos ver que a entropia, não precisa ter um pouco, podemos usar um pouco menos. Podemos dividir ainda mais com base em saber se as pessoas estão bebendo ou não. Então, novamente, temos a classe de fumantes. Ainda a entropia é igual a zero. Podemos dar uma olhada nas pessoas que não fumam, mas que fazem bebida, e a entropia é igual a 1. Como os dois grupos estão igualmente representados. É uma relação de um-para-um. Por fim, temos as pessoas que não bebem e não fumam, a entropia do grupo é igual a zero. Portanto, este é o caminho que podemos calcular a entropia, e aqui podemos ver todos os números que temos apenas computados. Utilizando esta fórmula de base. Agora podemos tomar a média ponderada destes três árvores de decisão. Então, se vamos dar uma olhada na primeira árvore de decisão não é apenas uma folha. E nesta folha a entropia é igual a 1. E são as frações completas, então 6 dividido por 6. Assim, a média ponderada global de entropia é igual a 1. Se dermos uma olhada na segunda árvore de decisão em que nos

separamos baseado no fumante rótulo, e tomamos a média ponderada dos zero e 0,811. Então, podemos ver que a entropia é 0,54. Note-se que 2 dos 6 pessoas acabaram na classe que teve uma entropia igual a zero. O resto. Teve uma entropia de, 0,811. É por isso que temos de aplicar a fórmula dessa maneira. Podemos dividir com base no fumante rótulo. Agora temos três, sai. Cada uma destas folhas tem um peso, dependendo do número de pessoas em que a folha em particular. Nós tomamos a média ponderada dos diferentes valores de entropia, e vemos que a entropia resultante é 0,33. Então, o que podemos ver é que, na primeira árvore de decisão, a média ponderada ao longo de todos estes valores de entropia foi de 1, e pela divisão, com base no fumante rótulo, ele caiu de uma para 0,54. Isto significa que tivemos um ganho de informação de 0,46. Se tomarmos a árvore no meio, podemos dividir com base no nível bebedor. E temos uma entropia de um terço. Então agora o ganho de informação é de 0,21. Então, gostaríamos de reduzir a entropia. E gostaríamos de dividir nos rótulos que maximizam a redução dessa entropia, de modo que a maximizar o ganho de informação. Assim, a idéia do algoritmo é agora que nós continuamos rótulos de divisão. Na tentativa de maximizar o ganho de informação, mas parar se esta já não é possível. Então, vamos dar uma olhada em algumas perguntas para ver se você realmente entender a noção de entropia. Aqui você pode ver vários conjuntos de bolas coloridas. E gostaríamos de calcular a entropia de todas as células individuais. Se você fizer isso, então você pode calcular a entropia total de todos esses 9 quadrados diferentes. E podemos comparar isso com o global. Entropia se ele iria colocar todas as bolas em uma caixa grande, e depois calcular a entropia. E então, por favor, tente calcular esses valores que são feitas por aqui. Ok, vamos primeiro dar uma olhada no celular no meio. Se olharmos para o celular no meio. Há, 16 bolas. Dois de cada cor. E há 8 cores diferentes. Se, então, aplicar a fórmula da entropia, descobrimos que a entropia é igual a três. Então, a gente precisa ter 3 bits para codificar a cor de

uma única bola. Na praça no meio. Se não tomar uma olhada nas outras células, todos eles contêm 16 bolas. Mas todas elas têm a mesma cor. Assim, a entropia para todas estas outras células é igual a zero. E podemos aprender que, basta preencher a fórmula que temos visto antes. Então, se nós agora olhar para todas as mais de entropia, temos de tomar a média ponderada destes 9 quadrados diferentes e ficamos com uma entropia de um terço. É isso mesmo, a entropia de todas as células é zero, a célula no meio tem uma entropia de três. Existem nove células, de modo que a média ponderada resultante é de 0,33. O que é a entropia depois tomamos todas as diferentes células e misturá-los? Então nós temos 144 bolas com 18 cores diferentes. Portanto, esta é a distribuição que, em seguida, obter, o que corresponde a uma entropia de três. Então, se nós agora comparar essas duas situações, se temos esses nove diferentes células. Quando a célula no meio tem todas as cores, e todas as outras células têm apenas uma cor, que tem uma entropia muito menor. Então, quando nós temos uma grande célula que contém todas as bolas, porque então a entropia é igual a 3. Então, se passar da situação com uma entropia de 3 para a entropia de um terço, temos uma informação ganhar igual a 2,666. Então é isso que está a ser usado quando a aprendizagem de uma árvore de decisão. Então, você vê um outro exemplo. Começamos por classificar todas as pessoas como morrer jovem. Em seguida, dividir o atributo fumante. Então, vamos a partir do atributo que está listado aqui. Para estes entropias para as 2 folhas novas que encontramos. Então, podemos tirar a média ponderada, e comparar essa média ponderada da entropia global para a entropia inicial. E podemos ver que há de fato um ganho de informação de 0,107. Embora a classificação não se alterou, houve ganho de informação. E isso é porque o grupo de fumantes é agora mais homogêneo que era antes. Podemos dividir ainda mais. E mais uma vez, podemos olhar para os valores originais de entropia. Esta é a média ponderada. E que ao comparar a nova situação, em que podemos basicamente aplicar as fórmulas que já vimos antes.

Então esses são os valores de entropia para os diferentes folhas. Voltamos a ter a média ponderada. Podemos comparar os dois, e vemos que há um ganho de informação. De 0,07. Então, isso mostra a idéia básica da árvore de decisão algorithm. We começar com o nó raiz que tem todas as instâncias, e, depois, de forma iterativa passar por todos os nós. E ver se podemos alcançar um ganho de informação. Fazemos isso por tentar dividir em todos os atributos possíveis, e ver se a entropia é realmente reduzida. Por isso, selecione o atributo com o maior ganho de informação, acima de um determinado limite. E, depois, dividir esse nó, e continuamos fazendo isso até que não haja melhorias significativas são possíveis. Então voltamos a árvore de decisão. Então essa é a idéia básica de aprendizagem de uma árvore de decisão, utilizando a noção de entropia. Existem muitos parâmetros e as variações possíveis para fazer isto. Por exemplo, pode definir um tamanho mínimo de um nó, antes ou depois do corte, para evitar superajuste, tendo todos os tipos de folhas, que correspondem, por exemplo, para apenas um único exemplo. Podemos definir o limiar no ganho mínimo que é necessário, e parar a divisão se este ganho não pode ser alcançado. Podemos definir uma profundidade máxima da árvore. Também podemos ter, para permitir que várias vezes usando um rótulo. Ao longo de um determinado caminho. Alguns algoritmos de árvore de decisão não permitem isso, os outros fazem. Mas ao invés de usar a entropia, podemos usar noções alternativas, como o índice de Gini da diversidade. Se temos uma variável numérica, nós podemos fazer isso. Automaticamente categórica, utilizando determinados tipos de técnicas. E por isso, dividir o domínio de uma variável numérica para torná-lo categórico. Nós também podemos fazer um post a poda da árvore para remover as notas de folhas que não aumentam significativamente o poder explicativo das árvores de decisão resultantes. Aprendizagem Árvore de decisão tem muitas aplicações também no campo da mineração processo. Assim, por exemplo, se dermos uma olhada neste modelo de processo, então

nós nos perguntar o que está impulsionando essas decisões? Estes dois pontos de decisão neste modelo de processo, por que certos casos indo para a esquerda e as outras instâncias indo para a direita. O que é o caminho mais provável de um caso de execução dando seus atributos? Se olharmos para o caso particular, gostaríamos de prever se será tarde ou rejeitado. Usando árvore de decisão aprender em cima de modelos de processo, nós podemos fazer isso. Mas é fundamental para ver que estas questões exigem um processo de descoberta, caso contrário, nada disso é possível. Então descoberta processo é necessário, antes de podermos usar a aprendizagem árvore de decisão. Hoje foi a primeira palestra que começamos a falar de técnicas de mineração de dados. Por favor, leia o capítulo 3 para saber mais sobre a aprendizagem de árvore de decisão. Obrigado por assistir esta palestra, vê-lo na próxima vez.

## **2 - 4 - Lecture 1.4- Learning Decision Trees (END)**

---

---

## 2 - 5 - Lecture 1.5- Applying Decision Trees

Welcome to this lecture. Today we will continue with decision trees. The last lecture presented the main concepts. Today, we show more examples and focus on the application of such classification techniques. Here you can see the running example that we have used in the last lecture. Using the notion of entropy, we were splitting notes into smaller notes until they were more homogenous. In such a way that we could understand the data in a better way, and we could make predictions. Today we will look at much more examples to get a better understanding, what the decision tree is, and how it can be used. So let's take a look at this small example, where we have a group of 160 students 100 of these 160 students passed, 60 failed. Suppose that we know the gender of these students. We know whether they are smoking or no. We know whether they are attending lectures. Can we then predict whether students are going to pass or fail based on these attributes? We can do this by building a decision tree and whenever we want to split a node, in this case a root node we want to know what the entropy is. So if we would like to split on the attribute smoker. We are interested to see what the information gain is. On this slide you can see the numbers indicating for the people that smoke and do not smoke. How many people passed and how many people did not pass. So please compute the information gain of this split. The answer can be computed as follows. First we take a look at the root node, we apply the standard formula for entropy, and we find that the entropy is 0.95. Then we split based on the attribute smoker. We get two smaller groups and for each group we compute the entropy. So for one group it is 0.95. For the other group it is also 0.95. So the entropy values in both of the leaf nodes did not go down, compared to the root node. To compute the overall entropy, we need to take the weighted average, and if we do that, we get the entropy for the root node. And of course it is also 0.95. And if we compute the weighted average of the entropy of the two new leaf nodes we find exactly the same value. So there is no information

gain, and actually there was no need to do all of these computations because we could see that the fraction within the two child nodes was exactly the same as it was before, so we did not gain any information. Let us therefore take a look at another attribute. We can split the set of all students into the students that are male and the students that are female. And we now can see that the fractions are different from females compared to males. So the question is, what is the information gain? To compute the answer we need to compare the original entropy of the root node with the weighted average of the entropy of the two new leaf nodes. This is the entropy of one leaf node. This is the entropy of the other leaf node. We can take the weighted average of these two values, and if we do so we can see that there is actually a change in entropy. Unlike based on the label smoker we can see that there is a considerable information gain of 0.2. And so, we are seeing here, if we inspect the decision tree, that females typically have a better study result. So by splitting on this label for the group of females, we know we can better predict what their study results will be. As a last label we look at the attribute attended all lectures. So people that attended all lectures we make a prediction for those and we also make a prediction for the people that did not attend all lectures. So the question is what is the information gain? Again, we can do the same computations. So, we can compare the original entropy of the root node with the weighted average of the entropies of the two child nodes. So, for one child node it is 0.81. For the other child node it's 0. So people that attended all lectures always passed so that's why the entropy within this group is equal to 0. Again we need to compare the weighted averages of the original situation and the situation after splitting and if we do that we get these values. And we can see that there is a considerable information gain by splitting based on this particular attribute. So the information gain is 0.54. If we now compare the three attributes based on which we could split, it is clear that splitting based on whether all lectures were attended provides the best information. So if we build a decision tree, it is

reasonable to start with this particular attribute, and then see whether we need to split other nodes. So if we take a look at the decision tree after one split, attended all lectures, then we find one group which we classify as pass where no further information gain is possible because all the instances are classified correctly. If you take a look, at the other group, the people that did not attend all lectures, we can still look at the other attributes to see whether we can achieve more information gain. All of this can be seen best by simply applying a tool. So in this course we will often use examples taken from RapidMiner. The installation of RapidMiner is optional, but if you would like to play with these ideas using software in the lecture guide you can see how to install it, where to get it, and how to use it. RapidMiner is an integrated extendable environment for machine learning and all kinds of other types of analysis that are focusing on data. RapidMiner also has a so-called marketplace. And there you can also download to so called ProM extension. So in RapidMiner, you can also do process mining using many of the techniques that you will see in later lectures. There are two versions, there are commercial and open-source version and we are using here the open-source version. So, let's take a look at an example. So we take a data set, you can see it's a simple CSV file where we have a column gender, a column age, a column smoker, a column car brand, and a column claim and the latter deserves some explanation. This is a variable indicating whether somebody has claimed insurance, car insurance, in the last year. So for example the first row in this table says there is a, a female customer that has insurance, age 47, is a smoker, drives a Volvo, and she did not claim any insurance in the last year. So if we take such a data set, we are interested in seeing which people are claiming insurance. Can we predict that? So the, we took a data set of 999 customers of an insurance company to learn these types of things. So we would like to know which customers claim insurance by simply using this CSV file. So, the response variable is the column claim. All the other columns correspond to predicted variables. And

we would like to explain the response variable in terms of these predicted variables. So we can feed this to RapidMiner. So we can simply take this CSV file, download it in RapidMiner, then we need to indicate what our role is of all the different columns. And after doing that we have loaded it in the repository of RapidMiner. And now we can apply all kinds of analysis techniques to it. For example, a workflow to build a decision tree. So here we see a sketch of this workflow, in RapidMiner. So we first load the data, then we create a decision tree. Then we apply the decision tree to the data set. And then we look at the resulting performance. So if we take this data set and we apply the workflow just indicated, this is the decision tree that we get. So you can read this decision tree just as you did before. So for example, female drivers don't claim insurance according to this decision tree. Male Alfa Romeo drivers typically claim insurance in the last year as you can see in this decision tree. Male Volvo drivers younger than 25 claim insurance. It's another fact that we can read from this decision tree. So the decision tree predicts for certain groups of customers whether they will claim or not. So what is the quality of this decision tree? We were measuring the performance. And if we, after learning the decision tree, apply it to the same dataset. We can see that of the 513 females, 498 actually did not claim. So less than 3% was wrong. If you look at Alfa Romeo drivers, there were 90 male Alfa Romeo drivers, four did not claim, 86 claimed. The label says they will claim. So less than 5% is wrong. If we look at the group of male BMW drivers, we can see that more than 20% is wrong. So we can measure the quality of this classification using the set that we have used to learn this decision tree. In this table you can see again the data set. But now there is an additional label telling what the predicted class is. So, this is the real class and this the predicted class. So, if we take a look at the smaller fragment of this table one can look at it and try to see, are there any instances that are classified incorrectly here. If you do so you will see that row 11 shows an instance that was classified incorrectly. A male 43 year

old non-smoking Subaru driver was predicted to claim but in reality did not. If you look at row 12 you will find a similar problem. As our BMW driver that was predicted not to claim but actually did claim. So these are examples of where we can see that the decision tree is only making a prediction that will hold for probably the majority of instances, but will not hold for all the instances. You can show this using a so-called confusion matrix. So there were 761 customers that were predicted not to claim and actually did not claim. There were 24 customers that did not claim, but that were predicted to claim. And so the numbers in the green cells, they correspond to things that are good. The numbers in the red cell, they indicate two misclassifications using the decision tree. Let's take a look at a larger data set. Consider an Italian restaurant, where we have 5,000 parties that have a dinner. The menu includes all of these dishes that you see here. But unfortunately of these 5,000 parties, 470 parties had members that got very sick, 313 parties got nauseous, and the remaining parties did not experience any problems. The question is now using decision trees can we predict which people will become sick after eating what. Or if we look at things in hindsight. We would like to understand what are people that got sick or nauseous, what they had in common. What was the dish that caused all of these problems? Again we can formulate a problem in terms of a CSV file. So here you see a data set where all the columns correspond to dishes and drinks. The numbers indicate how many times that particular dish or drink was ordered. And all the row corresponds to a party. And in total we have 5,000 parties. This is the workflow that we have to learn the decision tree problem. Note that all of the parties were labeled in one of these three groups. Very sick, not sick, or nauseous. And then we apply this workflow on this data. So first, we load the data, then we create a decision tree. Then we apply the decision tree to the data set and we measure conformance, exactly as we did before. We set the minimal information gain to 0.1. And so we are using this parameter to see when we need to stop extending the decision tree. If we do

this the RapidMiner returns this decision tree. It's showing that people, that ate this particular pizza and were drinking beer that they typically got very sick. The other parties, according to this decision tree, did not become sick. And so people that did not eat any pizza marinara were okay, that just ate one pizza, they're also still okay. Also the people that were eating this particular pizza and not drinking beer, they also did not get sick. So this node corresponds to parties that ate less than two pizzas marinara, and they did not get sick. This node corresponds to parties that ate multiple pizzas marinara, and that drank beer, and got very sick. These are the people that did not drink beer and also did not get sick. Just like the people that did not eat the pizza. So, the decision tree clearly indicates that the combination of pizzas marinara and beer caused the sickness. What is remarkable if you look at this decision tree is that there were several parties that were classified as nauseous. But it is not, there's no class in this decision tree that is labeled as such. We can see this if we look at the confusion matrix. So in the confusion matrix we can see that 307 of the 313 parties that were nauseous were classified as non sick. The other 6 were classified as very sick. So the decision tree is unable to identify this group. Apparently there is not enough that they have in common to make a proper prediction for this particular class. We can change the information gain and continue to split the tree further. So if we set the information gain to 0.5, this is the decision tree that we get. And you can see that it is becoming very complicated and also seems to over fit. For example, if we look at the node nauseous, that is labeled as nauseous now. You can see that this refers to a very particular group of customers that ate a very specific combination of things. We get an improvement, but we get the improvement only at the cost of severely over fitting the data, a topic that will be addressed also in later lectures. So, we can either classify everybody as not sick. We can have a very detailed decision tree if we set the information gain to even lower values. We get trees that are bigger and bigger. But these bigger trees,

they are clearly overfitting. And if we just predict for everybody that they will not get sick we have a tree that is underfitting. So the idea of these parameters is to balance between these two extremes. It seems that the tree that we showed at the beginning provides a reasonable balance between underfitting and overfitting. It can be used to understand what is happening and it can be used to make predictions or recommendations. And this is exactly how we would like to use these decision trees. So the last two lectures were devoted to decision tree learning. We will look at two additional data mining techniques, but much shorter. That will be association rule learning and clustering. And these will be addressed in the next couple of lectures. As indicated before, chapter three is devoted to these different data mining techniques. Thank you for watching and hope to see you soon.

## 2-5 - Palestra 1.5- Aplicação de Árvores de Decisão

Bem-vindo a esta palestra. Hoje vamos continuar com árvores de decisão. A última palestra apresentou os principais conceitos. Hoje, vamos mostrar mais exemplos e concentrar-se na aplicação de tais técnicas de classificação. Aqui você pode ver o exemplo em execução que temos utilizado na última palestra. Usando a noção de entropia, que estavam se separando notas em notas menores até que eles eram mais homogênea. De tal forma que pudéssemos compreender os dados em uma maneira melhor, e nós poderíamos fazer previsões. Hoje vamos olhar para muito mais exemplos para obter uma melhor compreensão, o que a árvore de decisão é, e como ele pode ser usado. Então, vamos dar uma olhada neste pequeno exemplo, onde temos um grupo de 160 estudantes de 100 destes 160 alunos passaram, 60 falhou. Suponha que nós sabemos o sexo desses alunos. Nós sabemos se eles estão fumando ou não. Nós sabemos se eles estão participando de palestras. Podemos, então, prever se os alunos vão passar ou não com base nesses atributos? Podemos fazer isso através da construção de uma árvore de decisão e sempre que queremos dividir um nó, neste caso um nó raiz, queremos saber o que a entropia é. Então, se nós gostaríamos de dividir sobre o atributo fumante. Estamos interessados para ver o que o ganho de informação é. Neste slide você pode ver os números que indicam para as pessoas que fumam e não fumam. Quantas pessoas passaram e quantas pessoas não passou. Então, por favor calcular o ganho de informação desta divisão. A resposta pode ser calculado como se segue. Primeiro vamos dar uma olhada no nó raiz, aplicamos a fórmula padrão para a entropia, e nós achamos que a entropia é 0,95. Em seguida, dividir com base no atributo fumante. Ficamos com dois pequenos grupos e para cada grupo calculamos a entropia. Assim, para um grupo é 0,95. Para o outro grupo é igualmente de 0,95. Assim, os valores de entropia em ambos os nós de folha não ir para baixo, em comparação com o nó raiz. Para calcular a entropia global, temos de tomar a média

ponderada, e se fizermos isso, temos a entropia para o nó raiz. E é claro que também é 0,95. E se calcular a média ponderada da entropia dos dois novos nós folha encontramos exatamente o mesmo valor. Portanto, não há ganho de informação, e, na verdade, não havia necessidade de fazer todos esses cálculos, pois pudemos ver que a fração dentro dos dois nós filho era exatamente a mesma que era antes, portanto, não obteve qualquer informação. Vamos, portanto, dar uma olhada em outro atributo. Podemos dividir o conjunto de todos os alunos para os alunos que são do sexo masculino e os alunos que são do sexo feminino. E agora podemos ver que as frações são diferentes das mulheres em comparação aos homens. Então a questão é, qual é o ganho de informação? Para calcular a resposta que precisamos comparar a entropia original do nó de raiz com a média ponderada da entropia dos dois novos nós de folha. Esta é a entropia de um nó de folha. Esta é a entropia do outro nó de folha. Podemos tomar a média ponderada destes dois valores, e se fizermos isso, podemos ver que há realmente uma mudança na entropia. Ao contrário com base no fumante rótulo, podemos ver que há um ganho considerável de informações de 0,2. E assim, nós estamos vendo aqui, se nós inspecionar a árvore de decisão, que as fêmeas normalmente têm um melhor resultado do estudo. Então, dividindo neste rótulo para o grupo de fêmeas, sabemos que podemos prever melhor o que seus resultados do estudo serão. Como último rótulo olhamos para o atributo assistiu a todas as aulas. Então, as pessoas que participaram todas as palestras que fazer uma previsão para aqueles e também fazer uma previsão para as pessoas que não compareceram todas as palestras. Portanto, a questão é o que é ganhar a informação? Mais uma vez, nós podemos fazer os mesmos cálculos. Assim, podemos comparar a entropia original do nó de raiz com a média ponderada das entropias dos dois nós filho. Assim, por um nó filho é 0,81. Para o outro nó filho é 0. Então, as pessoas que participaram todas as palestras sempre passou é por isso que a entropia dentro deste

grupo é igual a 0. Novamente precisamos comparar as médias ponderadas de situação original e da situação após a separação e se fizermos que recebemos esses valores. E podemos ver que há um ganho considerável de informações por meio de corte com base nesse atributo particular. Assim, o ganho de informação é de 0,54. Se agora comparar os três atributos com base no que poderíamos dividir, é claro que a divisão com base em se todas as palestras foram atendidos fornece a melhor informação. Então, se vamos construir uma árvore de decisão, é razoável para começar com esse atributo particular, e depois ver se precisamos dividir outros nós. Então, se vamos dar uma olhada na árvore de decisão depois de uma dividida, assistiram a todas as palestras, em seguida, encontramos um grupo que nós classificamos como passe de onde não mais ganho de informação é possível porque todos os casos são classificados corretamente. Se você der uma olhada, no outro grupo, as pessoas que não compareceram todas as palestras, ainda podemos olhar para os outros atributos para ver se podemos conseguir mais informações ganho. Tudo isto pode ser visto melhor, simplesmente aplicando uma ferramenta. Portanto, neste Claro que, muitas vezes, usar exemplos tirados de RapidMiner. A instalação de RapidMiner é opcional, mas se você gostaria de jogar com essas idéias usando software no guia palestra você pode ver como instalá-lo, onde obtê-lo, e como usá-lo. RapidMiner é um ambiente extensível integrada para a aprendizagem de máquina e todos os tipos de outros tipos de análise que estão se concentrando em dados. RapidMiner também tem um chamado mercado. E lá você também pode baixar a chamada extensão de baile. Assim, em RapidMiner, você também pode fazer mineração processo usando muitas das técnicas que você vai ver logo nas suas conferências. Há duas versões, existem versão comercial e de código aberto e nós estamos usando aqui a versão open-source. Então, vamos dar uma olhada em um exemplo. Então, tomamos um conjunto de dados, você pode ver que é um arquivo CSV simples, onde temos um gênero coluna,

uma idade coluna, um fumante de coluna, uma marca de carros coluna, e uma reivindicação coluna e este último merece uma explicação. Esta é uma variável que indica se alguém reclamou de seguros, seguro de carro, no ano passado. Assim, por exemplo, a primeira linha nesta tabela diz que há um, um cliente do sexo feminino que tem seguro, de 47 anos, é um fumante, dirige um Volvo, e ela não solicitou nenhum seguro no ano passado. Então, se tomarmos como um conjunto de dados, estamos interessados em ver que as pessoas estão reclamando de seguros. Podemos prever isso? Assim, o, que levou um conjunto de 999 clientes de uma companhia de seguros de dados para aprender esses tipos de coisas. Por isso, gostaria de saber quais os clientes reivindicar o seguro, basta usar esse arquivo CSV. Assim, a variável resposta é a alegação de coluna. Todas as outras colunas correspondem às variáveis preditivas. E nós gostaríamos de explicar a variável resposta em termos desses valores previstos. Assim, podemos alimentar esta a RapidMiner. Assim, podemos simplesmente pegar este arquivo CSV, baixá-lo em RapidMiner, então precisamos indicar o que o nosso papel é de todas as colunas diferentes. E depois de fazer que ele é carregado no repositório de RapidMiner. E agora podemos aplicar todos os tipos de técnicas de análise a ele. Por exemplo, um fluxo de trabalho para construir uma árvore de decisão. Então aqui nós vemos um esboço desse fluxo de trabalho, em RapidMiner. Por isso, primeiro carregar os dados, então criamos uma árvore de decisão. Em seguida, aplique a árvore de decisão para o conjunto de dados. E então olhamos para o desempenho resultante. Então, se tomarmos este conjunto de dados e nós aplicamos o fluxo de trabalho apenas indicado, esta é a árvore de decisão que temos. Então você pode ler este árvore de decisão, assim como você fez antes. Assim, por exemplo, motoristas do sexo feminino não reivindicar o seguro de acordo com esta árvore de decisão. Masculino motoristas Alfa Romeo tipicamente reivindicar o seguro no ano passado, como você pode ver nesta árvore de decisão. Condutores do sexo masculino com

idade inferior a 25 Volvo reivindicação de seguro. É mais um fato que nós podemos ler a partir desta árvore de decisão. Assim, a árvore de decisão prevê para determinados grupos de clientes se eles vão reclamar ou não. Então, qual é a qualidade desta árvore de decisão? Fomos medir o desempenho. E se, depois de aprender a árvore de decisão, aplicá-lo para o mesmo conjunto de dados. Podemos ver que dos 513 fêmeas, 498, na verdade, não reivindicou. Assim, menos de 3% estava errado. Se você olhar para os motoristas Alfa Romeo, havia 90 homens motoristas Alfa Romeo, quatro não reivindicar, 86 reivindicada. A etiqueta diz que eles vão reclamar. Assim, menos de 5% está errado. Se olharmos para o grupo de pilotos da BMW do sexo masculino, podemos ver que mais de 20% está errado. Assim, podemos medir a qualidade desta classificação usando o conjunto que temos usado para aprender esta árvore de decisão. Nesta tabela você pode ver novamente o conjunto de dados. Mas agora há uma etiqueta adicional dizendo que a classe prevista é. Então, essa é a classe real e esta classe o previsto. Assim, se dermos uma olhada no menor fragmento desta tabela pode-se olhar para ele e tentar ver, existem casos que são classificadas incorretamente aqui. Se você fizer isso você vai ver que a linha 11 mostra uma instância que foi classificada de forma incorreta. Um motorista de 43 anos Subaru não-fumantes de idade do sexo masculino foi previsto para reclamar, mas na realidade não o fez. Se você olhar na linha 12 você vai encontrar um problema semelhante. Como nosso piloto da BMW, que foi predito não reclamar, mas, na verdade, fez reclamação. Então, esses são exemplos de onde podemos ver que a árvore de decisão está apenas fazendo uma previsão de que irá realizar para, provavelmente, a maioria dos casos, mas não vou segurar para todas as instâncias. Você pode mostrar isso usando uma chamada matriz de confusão. Assim, havia 761 clientes que foram preditas não reivindicar e, na verdade, não reivindicou. Havia 24 clientes que não afirmam, mas que foram previstos para reclamar. E assim, os números nas células verdes, eles

correspondem a coisas que são boas. Os números na célula vermelha, eles indicam dois erros de classificação utilizando a árvore de decisão. Vamos dar uma olhada em um conjunto de dados maior. Considere um restaurante italiano, onde temos 5.000 partes que têm um jantar. O menu inclui todos esses pratos que você vê aqui. Mas, infelizmente, destes 5.000 partes, 470 partes tinham membros que ficou muito doente, 313 partes ficou enjoado, e as partes restantes não tenha quaisquer problemas. A questão agora está usando árvores de decisão podemos prever quais as pessoas vão ficar doentes depois de comer o quê. Ou, se olharmos para as coisas em retrospecto. Gostaríamos de entender o que são as pessoas que ficaram doentes ou náuseas, o que eles tinham em comum. Qual foi o prato que causou todos estes problemas? Mais uma vez podemos formular um problema em termos de um arquivo CSV. Então, aqui você vê um conjunto de dados onde todas as colunas correspondem aos pratos e bebidas. Os números indicam quantas vezes esse prato ou uma bebida especial foi encomendado. E toda a linha corresponde a uma festa. E no total, há 5.000 partes. Este é o fluxo de trabalho que nós temos que aprender o problema da árvore de decisão. Note-se que todas as partes foram rotulados num dos três grupos. Muito doente, não doente, ou náuseas. E, então, aplicar esse fluxo de trabalho com esses dados. Então, primeiro, vamos carregar os dados, em seguida, vamos criar uma árvore de decisão. Em seguida, aplique a árvore de decisão para o conjunto de dados e medir a conformidade, exatamente como fizemos antes. Vamos definir o ganho mínimo de informações para 0,1. E por isso estamos usando esse parâmetro para ver quando nós precisamos parar estendendo a árvore de decisão. Se fizermos isso o RapidMiner retorna esta árvore de decisão. Ele está mostrando que as pessoas, que comiam esta pizza especial e estavam bebendo cerveja que eles normalmente ficou muito doente. As outras partes, de acordo com esta árvore de decisão, não adoeceram. E para que as pessoas que não comem qualquer marinara pizza foram bem, que só comia

uma pizza, eles também são ainda bem. Também as pessoas que estavam comendo esta pizza especial e não beber cerveja, eles também não ficar doente. Portanto, este nó corresponde a partes que comiam menos de duas pizzas marinara, e eles não ficar doente. Este nó corresponde a partidos que comeram várias pizzas marinara, e que bebiam cerveja, e ficou muito doente. Estas são as pessoas que não bebem cerveja e também não ficar doente. Assim como as pessoas que não comeram a pizza. Assim, a árvore de decisão indica claramente que a combinação de pizzas marinara e cerveja causador da doença. O que é notável, se você olhar para esta árvore de decisão é que houve vários partidos que foram classificadas como náuseas. Mas não é, não há nenhuma classe nesta árvore de decisão que é rotulado como tal. Podemos ver isso, se olharmos para a matriz de confusão. Assim, na matriz de confusão, podemos ver que 307 das 313 partes que eram náuseas foram classificados como não doente. Os outros 6 foram classificados como muito doente. Assim, a árvore de decisão é incapaz de identificar este grupo. Aparentemente, não há o suficiente para que eles têm em comum para fazer uma previsão adequada para esta classe particular. Nós podemos mudar o ganho de informação e continuar a dividir ainda mais a árvore. Então, se nós definir o ganho de informação para 0.5, esta é a árvore de decisão que temos. E você pode ver que ele está se tornando muito complicado e também parece a mais adequada. Por exemplo, se olharmos para o nó enjoado, que é rotulado como náuseas agora. Você pode ver que esta refere-se a um grupo muito específico de clientes que comeram uma combinação muito específica de coisas. Ficamos com uma melhoria, mas ficamos com a melhora apenas com o custo da severamente ao longo do ajuste dos dados, um tema que será abordado também em palestras posteriores. Então, a gente pode classificar todos como não doente. Podemos ter uma árvore de decisão muito detalhado, se definir o ganho de informação para valores ainda mais baixos. Ficamos com árvores que são cada vez maiores. Mas essas

árvores maiores, eles são claramente overfitting. E se nós apenas prever para todo mundo que não vai ficar doente, temos uma árvore que é underfitting. Assim, a idéia desses parâmetros é equilibrar entre esses dois extremos. Parece que a árvore que mostrou no início proporciona um equilíbrio razoável entre underfitting e superajuste. Ele pode ser usado para entender o que está acontecendo e ele pode ser usado para fazer previsões ou recomendações. E é exatamente assim que nós gostaríamos de usar essas árvores de decisão. Assim, as duas últimas palestras foram dedicados à aprendizagem de árvore de decisão. Vamos olhar para duas técnicas de mineração de dados adicionais, mas muito mais curto. Isso vai ser associação aprendizado de regras e clustering. E estes serão abordados no próximo par de palestras. Como indicado anteriormente, o capítulo três é dedicado a estas técnicas de mineração de dados diferentes. Obrigado por assistir e espero vê-lo em breve.

## **2 - 5 - Lecture 1.5- Applying Decision Trees**

---

---

## 2 - 6 - Lecture 1.6- Association Rule Learning

Glad to see you again. After introducing decision trees as a tool for classification, we now focus on an unsupervised data mining technique, uncovering patterns in data. In this lecture we will learn about frequent item sets and association rules. These will help us to find expected and unexpected patterns. Here you, again see an overview of the different data mining techniques that we are considering in relation to the topic of process mining that we will be exploring later. After talking about decision trees in two lectures, we now move to association rules, and this is an unsupervised learning technique. In other words, we do not have label data. There's no response variable. To explain the topic, it is best to start with an example. One of the earlier applications of association rule mining revealed that people buying beer often also bought diapers. So it's a rule, taking one set of items, implying another set of items. So this is one example of an association rule. Another association rule could be cheese and ham and bread implies butter. So, in other words, customers that buy these three products, typically also buy butter. Or, people that buy oregano, often also buy spaghetti and tomato sauce. So, one set of items implies another set of items. So, all these rules have the form of  $x$  implies  $y$ , where both  $x$  and  $y$  are item sets. Sets of items. Before we talk about how you can compute them, let's first define some quality measures. And we start by explaining the notion of support. The idea of support is to take the number of items, that covers both  $x$  and  $y$  and divided by the total number of instances we have. So we look at the number of instances that has both  $x$  and  $y$ , and we divide that by the total number of instances. It's also always best to explain these concepts by looking at an example. So here we see a table that we have seen before. People ordering cappuccinos americanos, and muffins. And we can take such an example set and automatically learn association rules. We could for example find the rule that people that buy tea and latte, typically also buy muffins. Note that the frequency in the table doesn't matter. We just look, is an item

present or not? So we do not look at quantities, just at presence. So what is support, now, for this particular example? If we compute a support of the rule, tea and latte implies muffin, we compute the number of instances of customers, that bought tea, latte and muffins and divide that by the total number of customers. And if we have a higher value that is better than the lower value because there is more support. The second quality metric is confidence. So if we look at the confidence of a rule, we count the number of instances which covers both  $x$  and  $y$ , and we divide that by the number of instances covering just  $x$ . So in terms of our example, if we would like to know the confidence of the rule tea and latte implies muffin we count the number of customers that ordered tea, latte and muffins and we divide that by the number of customers that just ordered tea and latte and not necessarily muffins. We get another between zero and one, and it's clear that the higher the number, the more confident we can be about the accuracy of such a rule. The third quality metric is the most difficult one. It's called lift. So we look at a fraction of customers that cover both  $x$  and  $y$ , and we divide that by the fraction of customers that covers  $x$  and the fraction of customers that covers  $y$ . So, in terms of our example, and we can write this in the two ways that are indicated here. In terms of our examples this corresponds to taking the number of customers that order tea, latte, and muffins and multiply that by the total number of customers and diving this by the number of customers that order tea and latte and the number of customers that order muffins. If we then look at this fraction and we get a value that is bigger than 1 then  $x$  and  $y$  are positively correlated. They frequently happen together. If they are independent, so they happen independent of each other, then lift will be close to one. If it's lower than one, we get such a number if they are negatively correlated. So how can these three measures be used. They can be used to filter rules a priori. And so to avoid from a computational point of view that we need to look at too many rules at the same time. But they can also be used if we have a large number of rules

to sort them based on the criteria that we find important. These things are very important because typically there is an explosion of association rules. So it's very important to be able to prune the set of rules and to look at the most interesting ones. Typically one is most interested in the rules that have a support that is as high as possible, a confidence close to one. And a lift higher than one, indicating a positive correlation. For example, if we have a rule with the confidence of, let's say, 0.1, it is typically not a very interesting rule, because the confidence is really low. Let's make this clear using an example. So, we take an artificial set of customers, 100 customers that buy diapers and beer. There is just one type of diapers, Pampers. And there are two types of beer Hoegaarden and Dommelsch. And here you can see what these 100 customers bought. We can use such a data set to compute the notions that we have seen before. So let's take a look at four hypothetical association rules. For example, the rule, people that buy Pampers also buy Dommelsch, or people that buy Pampers also buy Hoegaarden, etc. So let's start with a question looking at the first rule. So if we look at the rule, people that bought Pampers typically also bought Dommelsch beer. What is the support, the confidence and the lift? Please take some time to compute these values. Here you can see a computation of these values. So the support is 0.51 because there are 51 customers that bought both Pampers and Dommelsch, and we divide that by the total number of customers. The confidence, is equal to the number of customers that bought Pampers and Dommelsch, divided by the number of customers that bought Pampers, which is equal to 91. So the confidence is 0.56. If we apply the formula for lift, we find that there is a positive correlation. Because we get a value higher than 1, 1.1. Let's now take a look at the other rules. For example, the last rule. People that buy Pampers and Dommelschs typically also buy Hoegaarden. What is the quality of these rules in terms of support, confidence and lift? Well we can do the same computation as what we have done before. It is put in the table here. The first row corresponds to

the to the first rule that we have explored before. If we look at this table then we see that some values are positive and other values are not so positive. For example, the ones that are highlighted now in red, correspond to low values. So, the rule Pampers, implies Hoegaarden, is a bad rule, in terms of support and confidence, and also in terms of lift. And so only 1% of the customers is supporting this rule. The confidence is also very low. So only in 1% of the cases the rule actually holds. And there are many people that buy Pampers, but that do not buy Hoegaarden. The negative, the lift is lower than one, indicating a negative correlation. So let's load this data set into RapidMiner so we can load it into the tool in terms of a CSV file. Load it in the repository and then do the experiment. So here you see a description of the data. Indeed we are looking at 100 customers and we are looking at three different items that people can or did not buy. This is the analysis workflow. So, we start by loading the data, then we do a data transformation. So we convert the numbers to true, false values. Then we compute frequent item sets. And finally, based on these frequent item sets we compute association rules that describe the rules that we are interested in. There are two important parameters here. One parameter is the minimal support. So if you look at an item set. What is the minimal number that this item set should appear in? Confidence is related to the confidence metric that we have explained before. If we use these parameters that were set in the previous slide, then we only find one rule. People that buy Dommelsch also buy Pampers. If we look at the support of this rule as we have computed by hand before the support is 0.51 which is bigger than the threshold. If you look at the confidence of the rule, it's 1 which is also bigger than the threshold that we have set. How to compute this, first of all one could use a brute force approach using these two parameters that you see here. There is minimum support level and there is a minimum confidence level. And we could first compute all the rules and then prune the rules based on these two parameters. So a brute force approach could work like

this, generate all frequent item sets, that have a support that is bigger than this  $\text{minsup}$ , constant that we have chosen. We also look at item sets containing two elements, because we are interested in association rules. And then we partition these frequent item sets into smaller, dejoined sets, and compute the confidence of these. And everything that meets the threshold is then returned as a rule that has been discovered. If you apply this approach, there are two problems. There is a computational problem, that there may be an exponential number of item sets that one needs to consider leading to all kinds of performance problems. Another, perhaps more serious problem, is an interpretation problem. If we find many rules and we return many rules to the user, then the user will be completely confused. So, how to address this problem? Well, there are several techniques to address computation and interpretation problems, and many of them are based on the observation that you can see here. If we have an item set  $X$  that is frequent, then all subsets of that item set are also frequent. If we have an item set that is not frequent, then all supersets are also infrequent, and we not need to consider them. So this can help us to prune the search space and to return the rules that are most interesting. So it can be exploited to improve efficiency and only return the strongest rules. Let's go back to the Italian restaurant that we have described before. Again, we take the data set of 5,000 parties. They can eat these dishes. And the question is, what are products? So, dishes and drinks that are often being combined. Well, if we load the file again in the repository just as we have done before we can now discard the class and, so, we are using unlabeled data. It's an unsupervised method that we are using, we are just finding patterns based on unlabeled data. If we apply this technique of finding association rules on this data set, then, first of all, we need to compute the frequent item sets. We find 153 item sets having a support of at least 0.1. This yields more than 700 association rules if we take a minimal confidence of 0.5. So this explosion of rules can be very confusing to the user. That is why we can increase the

threshold. So we set the threshold of minimal support to 0.3, we set the confidence to 0.9, and if we then again apply the same analysis work flow, we only get 61 association rules. We can for example look at the top three of them. So the first rule says that people that ordered espresso, pizza Siciliana, pizza Romana also ordered vino bianco. We can see the support, confidence, and lift levels here. So it's a positive correlation, the confidence is more than 0.9, and the support is, 32% of all the customers. This is another rule. People that ordered vino rosso, vino bianco, and pizza romano, also ordered espresso and pizza siciliana. Again we have a confidence of 0.9. Finally, we have the third rule in this list. People that ordered both types of wine also ordered pizza siciliana, and the confidence is again above 0.9. And again there's a positive correlation and a support that seems reasonable. So this way we can discover these rules from unlabeled data. Association rules are particular types of patterns. There are other patterns that one can consider. For example one can apply sequence mining techniques. That are essentially the same as association rules but now also look at the ordering of the different events. We can also look at episode mining. It's again similar to association rules but now taking into account a partial order of items. However all of these techniques, do not consider end-to-end process models. For that we need to really use process mining techniques. But often we can use data mining techniques in conjunction with process mining to exploit all the existing techniques like decision trees and association rules in a process oriented manner. Again, in chapter three, you can read more about these basic data mining techniques. Thank you for watching this lecture, see you next time.

## **2-6 - Palestra 1.6- Associação de Regras de Aprendizagem**

Fico feliz em vê-lo novamente. Após a introdução de árvores de decisão como uma ferramenta para a classificação, que agora se concentrar em uma técnica de mineração de dados sem supervisão, descobrindo padrões nos dados. Nesta palestra vamos aprender sobre conjuntos de itens freqüentes e regras de associação. Estes irão ajudar-nos a encontrar padrões esperados e inesperados. Aqui, mais uma vez ver uma visão geral das diferentes técnicas de mineração de dados que estamos considerando em relação ao tema da mineração processo que iremos explorar mais tarde. Depois de falar sobre árvores de decisão em duas palestras, que agora passar para regras de associação, e esta é uma técnica de aprendizado não supervisionado. Em outras palavras, não temos dados da etiqueta. Não há nenhuma variável resposta. Para explicar o tema, o melhor é começar com um exemplo. Uma das aplicações anteriores de mineração de regras de associação revelou que as pessoas que compram cerveja muitas vezes também compraram fraldas. Portanto, é uma regra, tendo um conjunto de itens, o que implica um outro conjunto de itens. Portanto, este é um exemplo de uma regra de associação. Outra regra de associação pode ser queijo e presunto e pão implica manteiga. Então, em outras palavras, os clientes que compram esses três produtos, normalmente também comprar manteiga. Ou, as pessoas que compram orégano, muitas vezes também comprar macarrão e molho de tomate. Assim, um conjunto de itens implica um outro conjunto de itens. Então, todas essas regras têm a forma de  $x$  implica  $y$ , onde  $x$  e  $y$  são conjuntos de itens. Sets de itens. Antes de falarmos sobre como você pode calcular-los, vamos primeiro definir algumas medidas de qualidade. E vamos começar por explicar a noção de apoio. A ideia de apoio é levar o número de itens, que abrange tanto  $x$  e  $y$  e dividido pelo número total de instâncias que temos. Então olhamos para o número de instâncias que tem ambos  $x$  e  $y$ , e nós dividir esse valor pelo número total de casos. Também é sempre melhor para

explicar esses conceitos, olhando para um exemplo. Então, aqui vemos uma tabela que já vimos antes. Pessoas encomendar cappuccinos Americanos, e muffins. E podemos ter um exemplo definido e aprender automaticamente regras de associação. Poderíamos, por exemplo, encontrar a regra de que as pessoas que compram chá e café com leite, também costumam comprar muffins. Note-se que a frequência na tabela não importa. Nós só olhar, é um item presente ou não? Então, nós não olhamos para quantidades, apenas com existência. Então, qual é o apoio, agora, para este exemplo particular? Se computarmos um apoio do Estado, chá e café com leite implica muffin, calculamos o número de casos de clientes, que compraram chá, café com leite e bolos e divida pelo número total de clientes. E se temos um valor mais alto que é melhor do que o menor valor, porque não há mais apoio. A segunda métrica de qualidade é a confiança. Portanto, se olharmos para a confiança de uma regra, contamos o número de instâncias que abrange ambos x e y, e divida pelo número de instâncias que cobrem apenas x. Assim, em termos de nosso exemplo, se nós gostaríamos de saber a confiança da regra e chá latte implica bolinho contamos o número de clientes que ordenou chá, café com leite e bolos e divida pelo número de clientes que apenas pedi chá e latte e não necessariamente muffins. Ficamos com outro entre zero e um, e é claro que quanto maior o número, mais confiantes de que podemos estar prestes a precisão de tal regra. A métrica de terceira qualidade é o mais difícil. É chamado de elevador. Então olhamos para uma fração de clientes que cobrem ambos x e y, e nós dividimos que por fração de clientes que abrange x e a fração de clientes que abrange y. Assim, em termos de nosso exemplo, e podemos escrever isso nas duas formas que são aqui indicados. Em termos de nossos exemplos isto corresponde a tomar o número de clientes que encomendam chá, café com leite, e muffins e multiplique pelo número total de clientes e mergulho isso pelo número de clientes que encomendam chá e café com leite e do número de clientes que a ordem muffins. Se, então, olhar para esta

fracção e nós temos um valor que é maior do que 1, então x e y são positivamente correlacionados. Eles acontecem frequentemente juntos. Se eles são independentes, de modo que eles aconteçam independentes um do outro, em seguida, levante vai estar perto de um. Se for menor do que um, temos um tal número se eles são negativamente correlacionados. Então, como pode ser usado estas três medidas. Eles podem ser usados para filtrar governa a priori. E assim, para evitar a partir de um ponto de vista computacional que temos de olhar para muitas regras ao mesmo tempo. Mas eles também podem ser usados, se temos um grande número de regras para classificá-los com base nos critérios que nós achamos importante. Essas coisas são muito importantes porque normalmente há uma explosão de regras de associação. Portanto, é muito importante ser capaz de remover o conjunto de regras e de olhar para os mais interessantes. Tipicamente, é de maior interesse para as regras que têm um suporte que é tão alta quanto possível, um perto de uma confiança. E um elevador mais elevada do que um, indicando uma correlação positiva. Por exemplo, se temos uma regra com a confiança de, digamos, 0,1, não é tipicamente uma regra muito interessante, porque a confiança é realmente baixo. Vamos deixar isso claro usando um exemplo. Então, tomamos um conjunto artificial de clientes, 100 clientes que compram fraldas e cerveja. Há apenas um tipo de fraldas, Pampers. E há dois tipos de cerveja e Hoegaarden Dommelsch. E aqui você pode ver o que esses 100 clientes compraram. Podemos usar esse conjunto de dados para calcular as noções que temos visto antes. Então, vamos dar uma olhada em quatro regras hipotéticas associação. Por exemplo, a regra, as pessoas que compram Pampers também comprar Dommelsch, ou pessoas que compram Pampers também comprar Hoegaarden, etc. Então, vamos começar com uma pergunta olhando para a primeira regra. Portanto, se olharmos para a regra, as pessoas que compraram Pampers tipicamente também compraram Dommelsch cerveja. Qual é o apoio, a confiança e o elevador? Por favor, tome

algum tempo para calcular esses valores. Aqui você pode ver um cálculo desses valores. Assim, o apoio é de 0,51, porque há 51 clientes que compraram ambos Pampers e Dommelsch, e nós dividir esse valor pelo número total de clientes. A confiança, é igual ao número de clientes que compraram Pampers e Dommelsch, dividido pelo número de clientes que compraram Pampers, que é igual a 91. Assim, a confiança é de 0,56. Se aplicarmos a fórmula para o elevador, descobrimos que existe uma correlação positiva. Porque nós temos um valor superior a 1, 1.1. Vamos agora dar uma olhada nas outras regras. Por exemplo, a última regra. Pessoas que compram Pampers e Dommelschs normalmente também comprar Hoegaarden. Qual é a qualidade dessas regras em termos de apoio, confiança e elevador? Bem que podemos fazer um cálculo idêntico ao que fizemos antes. Ele é colocado na tabela aqui. A primeira linha corresponde ao que a primeira regra que temos explorado antes. Se olharmos para esta tabela, em seguida, vemos que alguns valores são positivos e outros valores não são tão positivos. Por exemplo, aqueles que são realçados agora em vermelho, correspondem a valores baixos. Assim, as Pampers regra, implica Hoegaarden, é uma regra ruim, em termos de apoio e confiança, e também em termos de elevador. E assim, apenas 1% dos clientes está a apoiar esta regra. A confiança também é muito baixa. Assim, apenas em 1% dos casos, a regra seja titular. E há muitas pessoas que compram Pampers, mas que não compram Hoegaarden. O negativo, o elevador é menor do que um, indicando uma correlação negativa. Então, vamos colocar este conjunto em RapidMiner dados para que possamos colocá-la na ferramenta em termos de um arquivo CSV. Carregá-lo no repositório e, em seguida, fazer o experimento. Então aqui você ver uma descrição dos dados. Na verdade, estamos olhando para 100 clientes e estamos a olhar para três itens diferentes que as pessoas podem ou não comprar. Este é o fluxo de trabalho de análise. Então, vamos começar por carregar os dados, em seguida, fazemos uma transformação de dados. Por isso, converter os

números para os verdadeiros, os valores falsos. Em seguida, calculamos conjuntos de itens freqüentes. E, finalmente, com base nesses conjuntos de itens freqüentes calculamos regras de associação que descrevem as regras que estamos interessados. Existem dois parâmetros importantes aqui. Um parâmetro é o suporte mínimo. Então, se você olhar para um conjunto item. Qual é o número mínimo que esse conjunto item deve aparecer em? A confiança está relacionada com a métrica de confiança que já explicado antes. Se usarmos esses parâmetros que foram definidos no slide anterior, então nós apenas encontrar uma regra. Pessoas que compram Dommelsch também comprar Pampers. Se olharmos para o apoio a esta regra que temos calculado pela mão antes do apoio é de 0,51, que é maior do que o limite. Se você olhar para a confiança da regra, é uma que também é maior do que o limite que temos um conjunto. Como calcular isso, em primeiro lugar, pode-se usar uma abordagem de força bruta usando estes dois parâmetros que você vê aqui. Existe nível mínimo de apoio e existe um nível mínimo de confiança. E nós poderíamos primeiro computar todas as regras e, em seguida, podar as regras com base nesses dois parâmetros. Assim, uma abordagem de força bruta poderia trabalhar assim, gerar todos os conjuntos de itens freqüentes, que têm um apoio que é maior do que isso minsup, constante que temos escolhido. Analisamos também conjuntos de itens contendo dois elementos, porque estamos interessados em regras de associação. E, depois, dividir esses conjuntos de itens freqüentes em, conjuntos dejoined menores, e calcular a confiança destes. E tudo o que alcança o limiar é então devolvido como uma regra que foi descoberto. Se você aplicar essa abordagem, há dois problemas. Existe um problema computacional, que pode haver um número exponencial de conjuntos de itens que é necessário considerar que conduz a todos os tipos de problemas de desempenho. Outro problema, talvez mais grave, é um problema de interpretação. Se encontrarmos muitas regras e voltamos muitas regras para o usuário, em seguida, o usuário será completamente

confuso. Então, como resolver este problema? Bem, existem várias técnicas para resolver os problemas de cálculo e interpretação, e muitos deles são baseados na observação de que você pode ver aqui. Se temos um conjunto X item que é freqüente, em seguida, todos os subconjuntos de que conjunto de itens também são freqüentes. Se temos um conjunto que não é freqüente, em seguida, todos os superconjuntos também são freqüentes, e nós não precisamos considerá-los. Então, isso pode nos ajudar a podar o espaço de busca e retornar as regras que são mais interessantes. Por isso, pode ser explorada para melhorar a eficiência e só retornam as regras mais fortes. Vamos voltar para o restaurante italiano que descrevemos antes. Mais uma vez, tomamos o conjunto de 5.000 partes de dados. Eles podem comer esses pratos. E a questão é, quais são os produtos? Assim, pratos e bebidas que são muitas vezes combinadas. Bem, se carregar o arquivo novamente no repositório apenas como temos feito antes que possamos agora descartar a classe e, por isso, estamos usando dados não marcados. É um método não supervisionado que estamos usando, estamos apenas encontrar padrões com base em dados não marcados. Se aplicarmos esta técnica de encontrar regras de associação sobre este conjunto de dados, então, em primeiro lugar, precisamos calcular os conjuntos de itens freqüentes. Encontramos 153 conjuntos de itens que têm um apoio de pelo menos 0,1. Isso gera mais de 700 regras de associação, se tomarmos uma confiança mínima de 0,5. Então, essa explosão de regras pode ser muito confuso para o usuário. É por isso que nós podemos aumentar o limite. Portanto, definir o limite de suporte mínimo a 0,3, montamos a confiança para 0,9, e se, em seguida, novamente aplicar o mesmo fluxo de trabalho de análise, nós só temos 61 regras de associação. Podemos, por exemplo, olhar para os três primeiros deles. Assim, a primeira regra diz que as pessoas que ordenaram espresso, a pizza Siciliana, a pizza Romana também ordenou vino bianco. Podemos ver os níveis de apoio, de confiança, e de elevação aqui. Portanto, é uma correlação positiva,

a confiança é mais do que 0,9, e com o apoio seja, 32% de todos os clientes. Esta é outra regra. Pessoas que ordenou rosso vino, vino bianco, e pizza romano, também ordenou espresso e pizza siciliana. Mais uma vez, temos uma confiança de 0,9. Finalmente, temos a terceira regra nesta lista. Pessoas que ordenaram os dois tipos de vinho também pedimos pizza siciliana, ea confiança está novamente acima de 0,9. E mais uma vez há uma correlação positiva e um suporte que parece razoável. Então, dessa maneira podemos descobrir estas regras a partir de dados não marcados. Regras de associação são determinados tipos de padrões. Existem outros padrões que se pode considerar. Por exemplo pode-se aplicar técnicas de mineração de seqüência. Que são essencialmente os mesmos como regras de associação mas agora olhar também para a ordenação dos diferentes eventos. Nós também podemos olhar para mineração episódio. É novamente semelhante ao de regras de associação, mas agora tendo em conta a ordem parcial de itens. No entanto, todas essas técnicas, não consideram os modelos de processos end-to-end. Para isso, precisamos de realmente usar técnicas de mineração de processo. Mas muitas vezes nós podemos usar técnicas de mineração de dados em conjunto com a mineração processo de explorar todas as técnicas existentes, como árvores de decisão e regras de associação, de forma orientada processo. Mais uma vez, no capítulo três, você pode ler mais sobre essas técnicas básicas de mineração de dados. Obrigado por assistir esta palestra, vê-lo na próxima vez.

## **2 - 6 - Lecture 1.6- Association Rule Learning (END)**

---

---

## 2 - 7 - Lecture 1.7- Cluster Analysis

Welcome to this lecture. Today we will look at another unsupervised data mining technique. We will show how instances can be clustered in homogenous groups using the so called k-means technique. Here again we see the overview of all the different data mining techniques that we are considering. Decision tree learning was an example of a supervised learning technique. Association rules that we discussed in the last lecture is an unsupervised learning techniques just like the clustering approaches we will discuss today. So, what is the basic idea of clustering? We have, instances that have different attributes. Here, we only see two attributes, but it could be hundreds or thousands of attributes. And we would like to group. These instances in homogeneous groups, just as is sketched here in this slide. So, we have a cluster A, a cluster B, and a cluster C of points that are closer related to each other than to the other instances. So, how to compute these clusters? One approach is the so called k-means clustering, where beforehand we need to set the number of clusters. In this case we set  $k$  to 3 indicating that we would like to extract three different clusters. Again, I'm using a 2 dimensional diagram, but this is very misleading. Often, there are hundreds or thousands of dimensions, making clustering much more challenging. So, how does k-means clustering work? We start by, setting so called centroids. These are points in this  $n$  dimensional space. These points could be set randomly, or some other approach could be used like, for example, making it match to one of the instances. After setting these three centroids, what we do is that we decide for every instance to which centroid it is closest. So if we do that we have a blue, a red, and a green centroid. And we start coloring the instances, then this is what we get. So we have assigned all instances to the closest centroid. After coloring the instances, what we can do is that we can look at the instance of a particular color, and compute a new centroid. The centroids that is like the average of all the different points having a particular color. If we do that, then this is the

situation that we get. So for example, if we look at the 3 green dots, then they have a centroids which is like the point in the middle. The same holds for the blue cluster, and the red cluster. After recomputing the centroids, we again start coloring the different instances, again based on the closest centroid. If we do that, then this is the result. So what we can see is that the one instance highlighted here, it is still closest to the blue centroid. After doing this we, again, based on the instances that have a particular color, we again recompute the centroids. And if we do that, then this is the resulting situation. Now take a look at the blue dot closest to the bottom. It is now closest to the red centroid, so in the next iteration it will be added to the red centroid. Now we start recomputing, the middle of all of the centroids, and this is the result that we get. The situation that we see here is a so-called fixpoint. If we now iterate further, nothing will change anymore. Note that approach is non-deterministic if we take a random initialization of the centroids. So, typically the experiment is repeated multiple times. We take the best clustering. That results from all these random trials. This is the way that we can use k-means clustering to find homogenous groups of instances. So the main idea is that instances in a cluster are more similar to each other than those in other clusters. And this has many applications. So for example, we can try to find homogenous groups of customers, of patients, of sessions, of students. Etcetera. After creating these clusters, we get smaller datasets. So for example, now we have partitioned the dataset that is shown here into loyal customers, discount customers, and impuls customers. And now to each of these clusters, we can apply. Other data mining techniques or process mining techniques. For example, we could now build decision trees, based on these individual clusters. Or create association rules for these different clusters. Let's go back to the Italian restaurant that we have used before, but now to simplify things we have restricted the menu to only 6 items. In other ways we just abstract from all the other items, and just look at customers buying these particular dishes and drinks. We can

now apply this mining technique. This is the input that we get, again we have a class label but we can ignore it, because this is an unsupervised method. We would like to find patterns and groups of, customers that belong together without labeling the data before. So, to compute the clusters, we use this, workflow in rapid miner. We load the data sets, we apply k-means clustering, where we set k in this case to 2. And then, at the end, we measure performance. And there are different ways of measuring. The performance of clustering, but they are outside of the scope of this course. If we apply two means clustering to this data set we find two clusters that have approximately the same size. And if we look at the centroids of these two clusters we can clearly see. What these two clusters represent. So, in cluster zero, we can find, typically instances, customers that have bought lasagna, spaghetti carbonara, and drank beer. In the other cluster, we typically find people. That have eaten pizza, pizza margherita or pizza siciliana, and that were drinking wine. So, using clustering, we find two homogeneous groups of customers. Customers that drink beer, and that eat lasagna and or spaghetti. And customers that drink wine and eat pizza. Let's take a look at this in rapid minor and where we try to plot the data to create some additional insights into these clusters and to see that they are actually valid clusters. So here you see a scatter plot, where every dot corresponds to a customer. And we indicate the number of pizza sicilliana that the customer has ordered on the y axis. On the x axis we plot the number of pizzas margherita that people have bought. The color. Correspond to the clustering that was automatically discovered. What we see is that customers in cluster zero, indicated by the blue dots. These are typically not ordering any pizzas. But people in the red cluster, typically, are ordering. Just a pizza margherita, just a pizza siciliana, or multiple pizzas. Here we see another scatter plot. So, on the y axis, we are plotting the number of times that spaghetti carbonara was ordered. On the x axis, we plot the number of times that lasagna was ordered. If you look at the red and the blue dots, they

again correspond to the clusters that we have discovered before. So the blue dots are in cluster zero, and they correspond to customers that typically ordered at least. One time spaghetti carbonara, or one time lasagna, or multiple items of lasagna and spaghetti carbonara. Customers in the red cluster, they typically ordered no pasta. We can also look at the dimension of drinks, and again we can clearly see that the customers in the red cluster are drinking wine, whereas the customer in the blue cluster are drinking beer. So this provides evidence that the clustering we automatically discovered. Actually make sense for this data set. So, k-means is one of several clustering techniques that one can use. One of the drawbacks of k-means is that you need to decide on the value of k up front. Like in the previous example, we set k to 2. If you would like to see whether there are tree clusters that make sense, we need to try the same approach with the k set to three. Other clustering techniques may provide a hierarchy of clusters. And you can see an example here. On the right hand side, you see a so-called dendrogram and it is indicating a hierarchy of clusters. On the other side, you can see the corresponding clusters. So how to read such a diagram, such a hierarchy. A group similar sets of instances in a hierarchical manner. And it can build it in this way in, in an incremental way. If we cut the hierarchy at a particular place, we find as many clusters as the number of lines that we are crossing. So if we take a higher abstraction level, we will only find two clusters. One cluster consisting of a, b, c, and d, and another cluster consisting of the rest of the instances. We can also go lower in the hierarchy, and again if we cut the hierarchy, we find different clusters. So in the example now, we find the cluster consisting of a and b, the cluster consisting of c and d, a cluster containing just e, a cluster containing f and g, a cluster containing h and i, and a cluster containing j. And so we can seamlessly decide on the number of clusters that we would like to have. Clustering can also be used to split event logs. So suppose that we have event logs referring to different types of customers. We could first do clustering based on

the characteristics of the patients or the customers. And then we can automatically build process models for each of the clusters. So this way you can see how process mining and data mining techniques can be combined. This was the last data mining technique that we have discussed, in the next lecture we will focus more on evaluating the quality of data mining results. And then we can switch to more the process-oriented aspects of process mining. Thank you for watching, and I hope to see you soon.

## 2-7 - Palestra 1.7- Análise Cluster

Bem-vindo a esta palestra. Hoje vamos olhar para uma outra técnica de mineração de dados sem supervisão. Vamos mostrar como instâncias podem ser agrupados em grupos homogêneos utilizando a chamada técnica de k-médias. Aqui, novamente, vemos a visão geral de todas as diferentes técnicas de mineração de dados que estamos considerando. Aprendizagem decisão árvore foi um exemplo de uma técnica de aprendizado supervisionado. Regras de associação que discutimos na última palestra é um técnicas de aprendizagem não supervisionada, assim como o agrupamento se aproxima vamos discutir hoje. Então, qual é a idéia básica do agrupamento? Temos, instâncias que têm atributos diferentes. Aqui, vemos apenas dois atributos, mas poderia ser centenas ou milhares de atributos. E nós gostaríamos de grupo. Essas instâncias em grupos homogêneos, assim como é esboçado aqui neste slide. Então, nós temos um conjunto A, um cluster B, e um conjunto C de pontos que estão mais perto uns dos outros do que para as outras instâncias. Então, como calcular esses agrupamentos? Uma abordagem é o chamado k-means clustering, onde de antemão que precisamos para definir o número de clusters. Neste caso vamos definir k a 3 indicando que gostaríamos de extrair três grupos diferentes. Mais uma vez, eu estou usando um diagrama de 2 dimensões, mas isso é muito enganador. Muitas vezes, há centenas ou milhares de dimensões, tornando agrupamento muito mais desafiador. Então, como é k-means trabalho? Começamos, estabelecendo assim chamados centroids. Estes são pontos neste espaço dimensional n. Estes pontos podem ser definidas de forma aleatória, ou algum outro método poderia ser utilizado como, por exemplo, tornando-se corresponder a uma das instâncias. Depois de definir esses três centroids, o que fazemos é que nós decidir para cada instância a que centróide é mais próximo. Então, se fizermos isso, temos um azul, um vermelho e um baricentro verde. E nós começar a colorir as instâncias, então isso é o que temos. Portanto, temos atribuído

todas as instâncias para o centróide mais próximo. Depois de colorir as instâncias, o que podemos fazer é que podemos olhar para o exemplo de uma determinada cor, e calcular um novo centróide. Os centróide que é como a média de todos os diferentes pontos possuindo uma cor particular. Se fizermos isso, então esta é a situação que temos. Assim, por exemplo, se olharmos para os 3 pontos verdes, então eles têm uma centroids que é como o ponto no meio. O mesmo vale para o conjunto azul e vermelho do cluster. Depois de recalcular os centroids, novamente começar a colorir as diferentes instâncias, novamente com base no centróide mais próximo. Se fizermos isso, então este é o resultado. Então, o que podemos ver é que a uma instância destaque aqui, é ainda mais próximo do baricentro azul. Depois de fazer esta que, novamente, com base nos exemplos que têm uma cor particular, novamente recalcular o centróide. E se fizermos isso, então esta é a situação daí resultante. Agora, dê uma olhada no ponto azul mais próximo ao fundo. É agora mais próxima do centroide vermelho, de modo que na iteração seguinte, ele será adicionado ao centróide vermelho. Agora vamos começar recomputing, no meio de todos os centróides, e este é o resultado que nós temos. A situação que vemos aqui é um chamado fixpoint. Se agora interagir mais, nada vai mudar mais. Note-se que abordagem é não-determinístico se dermos uma inicialização aleatória dos centróides. Então, tipicamente, a experiência é repetida várias vezes. Tomamos o melhor clustering. Que resulta de todas essas tentativas aleatórias. Esta é a maneira que nós podemos usar k-means clustering para encontrar grupos homogéneos de instâncias. Assim, a idéia principal é que as instâncias de um cluster são mais semelhantes entre si do que os de outros clusters. E isso tem muitas aplicações. Assim, por exemplo, podemos tentar encontrar grupos homogéneos de clientes, dos doentes, das sessões, de alunos. Etcetera. Depois de criar esses grupos, temos conjuntos de dados menores. Assim, por exemplo, agora temos dividido o conjunto de dados que é mostrado aqui em clientes fiéis, clientes de desconto,

e os clientes Impuls. E agora, para cada um desses grupos, podemos aplicar. Outras técnicas de mineração de dados ou de técnicas de mineração de processo. Por exemplo, agora poderia construir árvores de decisão, com base nesses grupos individuais. Ou criar regras de associação para estes diferentes clusters. Vamos voltar para o restaurante italiano que usamos antes, mas agora para simplificar as coisas que têm restringido o menu para apenas 6 itens. Em outras maneiras que nós apenas abstratas a partir de todos os outros itens, e basta olhar para os clientes que compram esses pratos e bebidas particulares. Agora podemos aplicar esta técnica de mineração. Esta é a entrada que nós temos, mais uma vez temos um rótulo de classe, mas podemos ignorá-lo, porque este é um método não-supervisionada. Gostaríamos de encontrar padrões e grupos de clientes, que devem estar juntos sem rotular os dados antes. Assim, para calcular os clusters, nós usamos isso, o fluxo de trabalho em rápida mineiro. Nós carregar os conjuntos de dados, aplicamos k-means clustering, onde montamos k neste caso a 2. E então, no final, nós medimos o desempenho. E há diferentes formas de medir. O desempenho de clustering, mas estão fora do escopo deste curso. Se aplicarmos dois meios de agrupamento para este conjunto de dados, encontramos dois grupos que têm aproximadamente o mesmo tamanho. E se olharmos para os centroids desses dois grupos, podemos ver claramente. O que esses dois grupos representam. Assim, em conjunto zero, podemos encontrar, normalmente casos, os clientes que compraram lasanha, spaghetti carbonara, e bebeu cerveja. No outro cluster, que normalmente encontramos pessoas. Que ter comido pizza, pizza margherita ou pizza siciliana, e que foram beber vinho. Assim, por meio de agrupamento, encontramos dois grupos homogêneos de clientes. Os clientes que bebem cerveja, e que comem lasanha e ou espaguete. E os clientes que bebem vinho e comer pizza. Vamos dar uma olhada nisso em menor rápida e onde tentamos traçar os dados para criar mais alguns elementos para estes agrupamentos e ver que eles são

realmente aglomerados válidos. Então, aqui você vê um gráfico de dispersão, onde cada ponto corresponde a um cliente. E nós indicamos o número de pizzas siciliana que o cliente encomendou no eixo y. No eixo X marcamos o número de pizzas margherita que as pessoas têm comprado. A cor. Correspondem ao agrupamento que foi descoberto automaticamente. O que vemos é que os clientes em clusters zero, indicados pelos pontos azuis. Estes geralmente não são o pedido de quaisquer pizzas. Mas as pessoas no cluster vermelho, normalmente, está pedindo. Apenas um margherita pizza, apenas uma siciliana pizza, ou múltiplos pizzas. Aqui vemos um outro gráfico de dispersão. Assim, no eixo y, estamos traçando o número de vezes que carbonara spaghetti foi encomendado. No eixo x, marcamos o número de vezes que a lasanha foi encomendado. Se você olhar para o vermelho e os pontos azuis, eles novamente correspondem aos clusters que temos descoberto antes. Assim, os pontos azuis são em aglomerado de zero, e que correspondem aos clientes que tipicamente encomendados pelo menos. Uma vez carbonara spaghetti, ou uma lasanha tempo, ou vários itens de lasanha e espaguete carbonara. Os clientes do cluster vermelho, eles normalmente ordenado não massas. Nós também podemos olhar para a dimensão de bebidas, e mais uma vez, podemos ver claramente que os clientes do cluster vermelho estão bebendo o vinho, enquanto o cliente no conjunto azul estão bebendo cerveja. Portanto, este fornece evidências de que a aglomeração descobrimos automaticamente. Na verdade, faz sentido para este conjunto de dados. Assim, k-médias é uma das várias técnicas de agrupamento que se pode usar. Uma das desvantagens de k-médias é que você precisa decidir sobre o valor de k na frente. Como no exemplo anterior, definimos k para 2. Se você gostaria de ver se existem aglomerados de árvores que fazem sentido, temos de tentar a mesma abordagem com o k definido para árvore. Outras técnicas de agrupamento pode fornecer uma hierarquia de grupos. E você pode ver um exemplo aqui. No lado direito, você

verá uma chamada dendograma e ele está indicando uma hierarquia de clusters. Por outro lado, você pode ver os clusters correspondentes. Assim como ler um tal esquema, tal hierarquia. Um grupo de conjuntos semelhantes de instâncias de uma forma hierárquica. E pode construí-lo desta forma, em, de forma incremental. Se cortarmos a hierarquia em um determinado lugar, encontramos o maior número de agrupamentos como o número de linhas que estamos atravessando. Então, se tomarmos um nível mais alto de abstração, só vamos encontrar dois clusters. Um conjunto que consiste em a, b, c, e d, e um outro conjunto que consiste em o resto dos exemplos. Nós também podemos ir mais baixo na hierarquia, e, novamente, se cortarmos a hierarquia, encontramos diferentes clusters. Assim, no exemplo agora, nós encontramos o conjunto constituído por um e b, o cluster consiste em c e d, um cluster contendo apenas e, um cluster contendo f e g, um cluster contendo H e I, e um cluster contendo j. E assim podemos perfeitamente decidir sobre o número de clusters que gostaríamos de ter. Clustering também pode ser usado para dividir registos de eventos. Então suponho que temos logs de eventos referentes a diferentes tipos de clientes. Nós poderíamos fazer primeiro agrupamento com base nas características dos pacientes ou clientes. E então nós podemos construir automaticamente modelos de processos para cada um dos clusters. Assim, desta forma você pode ver como as técnicas de mineração de mineração e de dados do processo podem ser combinados. Esta foi a técnica de mineração de dados última que temos discutido, na próxima aula, vamos focar mais em avaliar a qualidade dos resultados de mineração de dados. E então nós pode mudar para mais os aspectos orientados para o processo de mineração de processo. Obrigado por assistir, e eu espero vê-lo em breve.

## 2 - 7 - Lecture 1.7- Cluster Analysis (END)

---

---

## 2 - 8 - Lecture 1.8- Evaluating Mining Results

Welcome to the last lecture, of the first week of this course on process mining. In the previous lectures, we learned about various data mining techniques. Today, we discuss ways of measuring the quality of data mining results. We have seen techniques like decision tree learning, association rule learning, clustering, sequence mining and many other techniques. Later we will see techniques for process discovery, conformance checking, for predicting remaining flow times. All of these techniques, produce results of the models, and question is, what is the quality of such models? In the lecture today, we will look at evaluating the quality of data mining results, but many of the ideas will also be applicable to evaluating or at least discussing the results of process mining. The first topic, that we would like to discuss today is the confusion matrix, that we have seen before and the measures based on it. It is best to explain it simply using a set of examples. So this is, a decision tree, to learn which students fail, pass or pass cum laude, based on the course grades that they have. If we create the so-called confusion matrix, we plot the actual class against, the predicted class. So for example, if you look at this matrix, we can see that there were 21 students that passed as the actual class, but the predicted class was that they would have failed. And so this is an incorrect prediction. The green values, indicate the positive results, the red values indicate problems and the higher the red numbers are, the bigger the problem is. We can take another example. So, we were trying to predict, what was causing the fact that certain customers got sick. There were three classes of customers. Customers that were not sick, that were very sick, and a group of customers that was nauseous. The decision tree that we learnt, had problems picking up, patients that are nauseous. And you can see that here, if you take a look at the column indicating the people that are nauseous. If you look at the lower row, you see the people that were actually very sick, and they were all predicted correctly. So all people that were, very sick, were predicted to be

sick. On the other end as I just mentioned, customers, that were nauseous, and there were more than 300 of them, were never predicted to be nauseous. So this way we can see that the confusion matrix gives an indication of the quality of the decision tree. If we take a look at a binary classification, the terms that we can use are true positives. These are the instances, that are actually positive, for example, becoming sick, and are also predicted to be positive. True negatives. Negative instances, predicted to be negative. So, these are, highlighted in green, because they are desirable values. We would like them to be as high as possible. On the other hand, we also have false positives and false negatives. These correspond to instances, that are classified incorrectly. So either negative instances, are predicted to be positive or positive instances are predicted to be negative. If we apply it to the example of the restaurant, then true positives, would correspond to sick customers that were predicted to be sick. If we, for example, look at false negatives, then this corresponds to, sick customers that were predicted to be not sick. So this is indicating the, types of problems that we would like to analyze. We would like to capture these problems in a number. So, the error rate corresponds to the number of false positives and the number of false negatives. So all the problematic cases, divided by the total number of instances. Accuracy is exactly the reverse. We take the number of true positives and true negatives. So all the correct predictions. And we divide that by the total number of instances. Precision and recall are two important metrics, that are often used. So precision corresponds to the number of true positives, divided by the number of true positives plus the number of false positives. Recall, is a number of true positives divided by the number of true positive plus the number of false negatives. We would like precision and recall to be as close to one as possible. The F1 score, is the harmonic mean of precision and recall and we can compute it using the formula shown here. So let's apply this to the following example. So what you see here, is the initial state of a decision tree, before

splitting any of the nodes based on some attributes. And a decision tree, that we discover after splitting based on the attribute smoker and the attribute drinker. We can see the, confusion matrices of both decision trees. So the one consisting of just one, leaf node and the one consisting of three leaf nodes. And we would like to compare them based on precision recall and F1 score. So please compute these values. These values can be computed by just applying the formulas that we have seen before. So if we look at the classification where all the persons are predicted to die young. And nobody is predicted to die old, then precision is 0.63 because 63% of the persons actually died young. Recall is equal to one because we never predict somebody to die at an older age, and F1 score, the harmonic mean of these two values is 0.77. If we look at a more refined decision tree after splitting on these two labels we can see that precision improved. Recall was reduced slightly. And that is due to the two young people that died young, and were predicted to die at an older age. That is the two in the confusion matrix. Recall, is slightly worse than before, but precision is much better, and also the F1-score is much better. So, this is an objective means to compare the two decision trees. So, that was the confusion matrix and metrics like precision and recall. Let's now take a look at cross-validation. Now, to explain why cross-validation is important. It helps to look at the following sentence. Take your ten best friends, and suppose that you would create a decision tree that accurately predicts the length of a friend, based on the person's birth date and eye color. If you have a group of just ten friends it is quite easy to construct such a decision tree. And it will perform perfectly on your ten best friends. But then if you would get another friend, it is very likely that the decision tree will produce an incorrect value. This is the problem of overfitting. So we are overfitting the example set, and by that, we cannot generalize properly. So a definition of overfitting is that the model is too specific for the data set used, and will most likely perform very poorly on new instances. And here you see a rule that could be the rule that would result from overfitting.

So people having a particular birth date and a particular eye color would have a particular length with a very large precision. Underfitting is the other problem. You're not overfitting the data. You're making conclusions that are so general, that they don't say very much and that they don't actually use the data. For example, a rule like, if gender is male, then length is larger than one meter, is a rule that may often be true, but it is severely underfitting, because we, we didn't actually learn from the data that we were analysing. Because of these problems, over-fitting and under-fitting, we typically split the data set into a training set and a test set. Then we apply a learning algorithm that will create a model, for example, a decision tree. Then we take the test set and we try the test set on the model that we have learned based on the training set. And then we measure the performance. So we are testing our model on unseen data. And this is the main idea of cross validation. If you do it like this you're using your data not in an optimal way because you didn't, use all the data to learn as much as possible. That is why you can repeat the principal that is shown here by partitioning your data in  $k$  parts and then use  $k$  minus 1 parts to learn the model and use one of these  $k$  parts for testing the quality of the model. And you can repeat this experiment  $k$  times, so then you are really using all data available to both learn the model and to test the result. There are many possible complications. So one of the complications that we may face if we are evaluating the quality of a model that we have learned is concept drift. Over time, the characteristics of the underlying process may have changed. And, this may, make things very difficult. Another complication with respect to evaluating data is that often we do not have negative examples. For example, if we look at our restaurant, we only know about sick customers that complain afterwards. We do not know anything about that customers that actually did not complain. So often, we have an unbalanced set that we are using to learn without any negative examples. Also, in the context of process mining, this problem is considerable. Because we only see the traces that have

happened not the traces that could not happen. So we are facing similar problems. The focus of this course is clearly on process mining, but we had to learn these basic data mining techniques beforehand. Because we can adopt some of the ideas from data mining. And sometimes in the context of process mining we are also locally using data mining techniques. For example, if we have learned the process model. And in that process model there is a choice. Then after learning the control flow, we may want to use decision tree learning, to see what is influencing the decision in the process, but we can only do that if we have beforehand discovered the process model itself. So in this way process mining extends way beyond classical data mining approaches that do not consider process models at all. So process mining is as a bridge between classical data analysis and classical process analysis. After explaining the core results in data mining, we now move to the more process-oriented part of this course. So the next set of lectures will be devoted to process models and learning these process models. If you would like to read more on data mining take a look at chapter three. Thank you for watching this lecture, see you next time.

## **2-8 - Palestra 1.8- avaliação dos resultados da mineração**

Bem-vindo à última palestra, da primeira semana do curso sobre mineração processo. Nas aulas anteriores, aprendemos sobre as várias técnicas de mineração de dados. Hoje, discutimos formas de medir a qualidade dos resultados de mineração de dados. Vimos técnicas como a aprendizagem de árvore de decisão, a aprendizagem de regras de associação, clustering, mineração sequência e muitas outras técnicas. Mais tarde veremos técnicas para descoberta de processos, verificação de conformidade, para prever restantes tempos de fluxo. Todas estas técnicas, produzir resultados dos modelos, e pergunta é, qual é a qualidade desses modelos? Na palestra de hoje, vamos olhar para avaliar a qualidade dos resultados de mineração de dados, mas muitas das ideias que serão também aplicáveis à avaliação ou, pelo menos, discutir os resultados da mineração de processo. O primeiro tópico, que gostaríamos de discutir hoje é a matriz de confusão, o que temos visto antes e as medidas com base nele. É melhor para explicá-lo simplesmente usando um conjunto de exemplos. Portanto, esta é, uma árvore de decisão, para saber quais os alunos não, passar ou passar cum laude, com base nas notas dos cursos que eles têm. Se nós criamos a chamada matriz de confusão, marcamos a classe real contra, a classe previsto. Assim, por exemplo, se você olhar para esta matriz, podemos ver que havia 21 alunos que passaram enquanto a classe real, mas a classe previu foi que teria falhado. E por isso esta é uma previsão incorreta. Os valores verdes, indicam os resultados positivos, os valores de vermelho indicam problemas e maiores os números vermelhos são, maior é o problema. Nós podemos dar outro exemplo. Então, nós estávamos tentando prever, o que estava causando o fato de que certos clientes ficou doente. Havia três classes de clientes. Os clientes que não estavam doentes, que eram muito doente, e um grupo de clientes que estava enjoada. A árvore de decisão que nós aprendemos, teve problemas Levantar, pacientes que são náuseas. E você pode ver que aqui, se você der

uma olhada na coluna indicando as pessoas que são náuseas. Se você olhar para a linha de baixo, você vê as pessoas que estavam realmente muito doente, e todos foram previu corretamente. Assim, todas as pessoas que foram, muito doente, foi previsto para ser doente. Na outra extremidade, como eu acabei de mencionar, os clientes, que foram náuseas, e havia mais de 300 deles, nunca foram previstas serem náuseas. Assim, desta forma, podemos ver que a matriz de confusão dá uma indicação da qualidade da árvore de decisão. Se dermos uma olhada em uma classificação binária, os termos que podemos usar são verdadeiros positivos. Estes são os exemplos, que são na verdade, positivos, por exemplo, tornam-se doentes, e também são previstos para ser positivo. Verdadeiros negativos. Casos negativos, previsto para ser negativa. Portanto, estes são, destacada em verde, porque são valores desejáveis. Nós gostaríamos que eles sejam o mais alto possível. Por outro lado, temos também de falsos positivos e falsos negativos. Estes correspondem a instâncias, que são classificados incorretamente. Assim, ou casos negativos, estão previstos para haver casos positivos ou negativos são previstos para ser negativo. Se nós aplicá-lo para o exemplo do restaurante, então verdadeiros positivos, corresponderia aos clientes doentes que estavam previstos para ser doente. Se, por exemplo, olhar para os falsos negativos, em seguida, o que corresponde a, clientes doentes que foram previstos como não doente. Então, isso é o que indica os tipos de problemas, que gostaríamos de analisar. Nós gostaríamos de capturar esses problemas em um número. Assim, a taxa de erro corresponde ao número de falsos positivos e os falsos negativos. Assim, todos os casos problemáticos, dividido pelo número total de ocorrências. Precisão é exatamente o inverso. Tomamos o número de verdadeiros positivos e verdadeiros negativos. Assim, todas as previsões corretas. E nós dividir esse valor pelo número total de casos. Precisão e revocação são duas métricas importantes, que são frequentemente utilizados. Assim precisão corresponde ao número de verdadeiros positivos dividido pelo número de

verdadeiros positivos mais o número de falsos positivos. Recall, é um número de verdadeiros positivos dividido pelo número de verdadeiros positivos, mais o número de falsos negativos. Gostaríamos de precisão e recordar a ser tão perto de uma possível. A pontuação F1, é a média harmônica de precisão e recall e podemos calcular-lo usando a fórmula mostrada aqui. Então, vamos aplicar isso a exemplo a seguir. Então, o que você vê aqui, é o estado inicial de uma árvore de decisão, antes da divisão de qualquer um de nós com base em alguns atributos. E uma árvore de decisão, que descobrimos depois de dividir com base no fumante atributo eo atributo bebedor. Podemos ver as matrizes de confusão, de ambas as árvores de decisão. Então, a única constituída por apenas um, e o nó de folha única constituída por três nós de folha. E nós gostaríamos de compará-los com base na recolha de precisão e pontuação F1. Então, por favor calcular esses valores. Esses valores podem ser computados apenas aplicando as fórmulas que já vimos antes. Portanto, se olharmos para a classificação, onde todas as pessoas estão previstas para morrer jovem. E ninguém está previsto para morrer velho, então precisão é de 0,63, pois 63% das pessoas, na verdade, morreu jovem. Lembre-se é igual a um, porque nós nunca prever alguém a morrer em uma idade mais avançada, e F1 marcar, a média harmônica destes dois valores é de 0,77. Se olharmos para a árvore de decisão mais refinado após a separação desses dois rótulos, podemos ver que a precisão melhorada. Lembre foi ligeiramente reduzida. E isso é devido aos dois jovens que morreram jovens, e foram preditied a morrer em uma idade mais avançada. Essa é a dois na matriz de confusão. Lembre-se, é um pouco pior do que antes, mas a precisão é muito melhor, e também o F1-score é muito melhor. Então, este é um objectivo consiste em comparar as duas árvores de decisão. Então, essa foi a matriz de confusão e métricas como precisão e recall. Vamos agora dar uma olhada de validação cruzada. Agora, para explicar por que a validação cruzada é importante. Ele ajuda a olhar para a seguinte

frase. Leve os seus dez melhores amigos, e suponha que você criaria uma árvore de decisão que prevê com precisão o comprimento de um amigo, com base na data de nascimento da pessoa e cor dos olhos. Se você tem um grupo de apenas dez amigos é muito fácil de construir uma tal árvore de decisão. E ele irá executar perfeitamente em seus dez melhores amigos. Mas, então, se você deseja obter um outro amigo, é muito provável que a árvore de decisão irá produzir um valor incorreto. Este é o problema de superajuste. Então, nós estamos overfitting o exemplo set, e por isso, não podemos generalizar corretamente. Assim, uma definição de superajuste é que o modelo é muito específico para o conjunto de dados utilizado, e muito provavelmente irá funcionar muito mal em novas instâncias. E aqui você vê uma regra que poderia ser a regra que resultaria de overfitting. Então, as pessoas que têm uma data de nascimento particular e uma cor de olho em particular teria um comprimento particular com uma grande precisão. Underfitting é outro problema. Você não está overfitting os dados. Você está fazendo conclusões que são tão geral, que eles não dizem muito e que eles não usam os dados. Por exemplo, uma regra como, se o gênero é do sexo masculino, em seguida, o comprimento é maior do que um metro, é uma regra que pode muitas vezes ser verdade, mas é severamente underfitting, porque nós, nós não realmente aprender a partir dos dados que nós éramos analisar. Devido a estes problemas, o excesso de montagem e sub-montagem, dividimos normalmente o conjunto de dados em um conjunto de treinamento e um conjunto de teste. Em seguida, aplicar um algoritmo de aprendizagem que irá criar um modelo, por exemplo, uma árvore de decisão. Em seguida, tomamos o conjunto de teste e tentamos o conjunto de teste do modelo que temos aprendido com base no conjunto de treinamento. E, então, medir o desempenho. Então, estamos testando o nosso modelo em dados invisíveis. E esta é a idéia principal de validação cruzada. Se você fazê-lo assim que você está usando seus dados não de uma forma ideal, pois você não,

use todos os dados para aprender o máximo possível. É por isso que você pode repetir o principal que é mostrado aqui dividindo os dados em  $k$  partes e, em seguida, usar  $k$  menos 1 peças para aprender o modelo e usar uma dessas peças  $k$  para testar a qualidade do modelo. E você pode repetir esta experiência  $k$  vezes, por isso, então você está realmente utilizando todos os dados disponíveis tanto para aprender o modelo e testar o resultado. Há muitas complicações possíveis. Portanto, uma das complicações que podem enfrentar se estamos avaliando a qualidade de um modelo que nós aprendemos é conceito deriva. Ao longo do tempo, as características do processo subjacente pode ter mudado. E, isso pode, fazer as coisas muito difíceis. Outra complicação que diz respeito à avaliação de dados é que muitas vezes não temos exemplos negativos. Por exemplo, se olharmos para o nosso restaurante, só sabemos sobre os clientes doentes que se queixam depois. Nós não sabemos nada sobre o que os clientes que, na verdade, não se queixou. Então, muitas vezes, temos um conjunto desequilibrado que estamos usando para aprender sem quaisquer exemplos negativos. Além disso, no contexto do processo de mineração, este problema é considerável. Porque nós só ver os traços que aconteceram e não os traços que não poderiam acontecer. Então, nós estamos enfrentando problemas semelhantes. O foco deste curso é claramente sobre a mineração processo, mas tivemos que aprender essas técnicas básicas de mineração de dados de antemão. Porque nós podemos adotar algumas das idéias de mineração de dados. E, por vezes, no contexto do processo de mineração também estamos usando técnicas de mineração de dados localmente. Por exemplo, se temos aprendido o modelo de processo. E nesse modelo de processo não é uma escolha. Em seguida, depois de aprender o fluxo de controle, a gente pode querer usar a aprendizagem árvore de decisão, para ver o que está influenciando a decisão no processo, mas só podemos fazer isso se tivermos previamente descoberto o próprio modelo de processo. Assim, deste modo

mineração processo se estende muito além das abordagens clássicas de mineração de dados que não consideram os modelos de processos em tudo. Então mineração processo é como uma ponte entre a análise de dados e análise de processos clássico. Depois de explicar os resultados fundamentais em mineração de dados, que agora passar para a parte mais orientada para o processo deste curso. Então, o próximo conjunto de palestras será dedicado para processar modelos e aprender esses modelos de processo. Se você gostaria de ler mais sobre a mineração de dados dar uma olhada no capítulo três. Obrigado por assistir esta palestra, vê-lo na próxima vez.

## **2 - 8 - Lecture 1.8- Evaluating Mining Results (END)**

---

---