

Trabalho 2 - Conhecendo os Dados e Pré-processamento

Mineração de Dados

Resumo

1. Introdução
2. Sobre a base
3. Resumo dos tratamentos
4. Reconhecimento das bases
5. Conclusão

Introdução - Informações sobre o trabalho

Deixamos disponível todos os recursos e artefatos necessários e utilizados neste trabalho no repositório do Github, no seguinte link: <https://github.com/usr-bin32/data-mining-ufjf>

A divisão do trabalho foi feita visando que pudéssemos compartilhar as análises um com o outro e ir entendendo em conjunto e tendo ideias juntos do que poderíamos buscar nas bases e também, da mesma forma que foi feita em uma base de dados, fazer na outra o que fosse cabível. O Gabriel ficou responsável de fazer a análise inicial, e o Vinicius dar suporte nessa análise, sendo que o Vinicius ficou mais responsável pela segunda base e pelo relatório final.

Utilizamos os notebooks em Python como base para todo o código e desenvolvimento das análises. E para o relatório visamos deixar bem explicado e com alguns resultados mais relevantes. Portanto, o bruto de tudo que foi feito pode ser encontrado nos notebooks.

Foi utilizado as bases de dados seguintes:

- atitudesEducadores_df.csv
- percepcaosocial_questions.csv
- praticasprofissionais_df.csv
- praticasprofissionais_questions.csv

As bases tiveram intuito de um formulário preenchido por professores, do qual havia o objetivo:

- Curso de sensibilização para a construção de práticas preventivas em contextos escolares na área de álcool e outras drogas
- Ações que envolvam estratégias preventivas eficientes, apontando os componentes eficazes, a relação custo-benefício e a probabilidade de adaptação do programa aos contextos em que as ações de prevenção são operacionalizadas
- Participaram professores, através de um curso online apresentando para tais educadores sobre "Prevenção do uso de drogas". Teve informações sobre pessoas de que utilizaram drogas e em outra base uma observação geral sobre os participantes. Tendo em vista o início e fim das aulas.

As duas bases fazem parte do mesmo curso, cada um a respeito de uma atividade ou parte do curso. Levantando informações sobre o que os profissionais de educação pensam sobre essa questão que é vinda da sociedade.

Conteúdos dos datasets

Base de Atitudes dos educadores


- Escala de atitudes dos educadores da pesquisa do curso online de "Prevenção do Uso de Drogas"
- Base possui 2873 registros utilizáveis, depois de ser tratada

Base de Práticas Profissionais

- Escala de práticas profissionais da pesquisa do curso online de "Prevenção do Uso de Drogas"
- Base possui 3407 registros utilizáveis, depois de ser tratada

Informações a cerca do estudo

- Do estado para educadores de escola pública
- Participaram 4.960 educadores de Minas Gerais, Rio de Janeiro e Paraná
- Online à distância (Moodle)
- Duração de 8 meses
- 5 módulos (4 teóricos e 1 de prática de implementação)
- Quinze Fóruns Virtuais de Discussão do conteúdo específico de cada unidade. Cada Fórum objetivava criar espaço para discussão de uma questão formulada pelo professor do Módulo, abordando os objetivos da unidade.
- Atividades Colaborativas realizadas por educadores pertencentes ao mesmo grupo escolar ao final de cada módulo (quatro ao todo). Cada uma das atividades objetivou a construção de




uma etapa do projeto de prevenção ao uso de álcool e outras drogas a ser implementado no contexto escolar, objetivo primário do curso.

- Questionário avaliativo e conteudista ao final de cada Módulo - respondido por todos educadores. Essa atividade foi realizada três vezes, não sendo aplicada durante o módulo 4 em virtude do maior número de Fóruns de Discussão desta etapa. Estes questionários abordavam questões teóricas chave sobre as temáticas apresentadas no material didático, buscando avaliar o entendimento dos cursistas de pontos essenciais sobre o processo de prevenção do uso de álcool e outras drogas, discutidos em cada Módulo.
- Devido ao fato de o curso ter uma duração considerável, as coletas das informações necessárias aos objetivos almejados no presente artigo foram feitas em momentos distintos.
- Sendo assim, foram realizadas duas coletas de dados para obter os resultados pretendidos. Inicialmente foram coletadas informações sociodemográficas tais como sexo, idade, escolaridade, ocupação e contato prévio com a temática álcool e outras drogas ****(obtidas ao final do Módulo 1 do curso) (n = 3.247)**** e, posteriormente, as informações de avaliação de aspectos do curso e sobre o acesso às ferramentas virtuais de aprendizagem ****(coletadas ao final do Módulo 4) (n = 1.950)****.
- Coleta de Dados. A coleta de dados foi realizada de maneira virtual, através de questionários com questões fechadas para as variáveis sociodemográficas e avaliativas do curso. Os registros de acesso à plataforma foram utilizados para coletar as variáveis de realização das atividades pedagógicas. Os dados foram extraídos utilizando os módulos de questionário oferecidos pela plataforma Moodle.
- No presente estudo, foram investigadas as variáveis abaixo listadas através de questionários de caracterização dos participantes do curso e dos aspectos organizacionais desse, disponibilizados virtualmente.

Sobre o questionário

Questionário sócio-demográfico - aplicado ao final do módulo 1 com o objetivo de caracterizar a amostra a partir das seguintes variáveis: sexo, idade, nível de formação, ocupação, contato prévio com a temática álcool e outras drogas, motivos para participação do curso.

Questionário de avaliação dos processos pedagógicos - aplicado ao final do módulo 4 com o objetivo de avaliar o curso a partir das seguintes questões: adequação do material didático, flexibilidade do prazo de entrega das atividades, avaliação da importância das atividades nos fóruns - interação com outros cursistas, organização do curso, importância do auxílio do tutor. Este questionário foi usado para fins de pesquisa, não sendo, portanto, atribuído a ele notas. As informações nele contidas foram desenvolvidas e analisadas pela equipe de pesquisa do Curso de Prevenção do Uso de Drogas para Educadores de Escolas Públicas.



Registros de atividades - participação dos alunos extraída ao final do curso usando os registros de acessos da plataforma virtual.

Resumo dos tratamentos

A princípio seria feito uma integração entre as bases de dados. Porém, após gerar novos arquivos CSV com quebras de linhas removidas e contendo apenas colunas comuns, foi detectado um número muito pequeno de linhas idênticas entre as duas bases ($n < 300$). O preenchimento de alguns campos não batia entre indivíduos praticamente iguais (por exemplo, uma resposta de um diferia da do outro por uma única letra maiúscula). Portanto, a integração das bases foi descartada.

Vimos que as bases não estavam de uma certa forma, faceis de serem utilizadas, portanto, foi realizado diversos tratamentos, a fim de que conseguíssemos extrair alguns conhecimentos e informação relevante a cerca dos dados.

Abaixo, segue uma relação dos tratamentos realizados:

Remove se tiver linhas totalmente nulas.

Remove linhas/registros que estão "Em progresso", pois estas normalmente só tem a data, sem informações precisas

São de pessoas que iniciaram o preenchimento, mas não finalizaram, ou não enviaram o questionário

Ficando apenas os registros de formulários preenchidos onde consta como "Finalizadas"

Foi removidos outliers do dataset, por exemplo, idades muito altas ou idade abaixo de 18. Julgamos que provavelmente teve erros de digitação e respostas de outros campos de pergunta, que pode ter sido digitado errado por diversos motivos.

ATENÇÃO: interpretação pode ter variado de acordo com o respondente.

Alguns podem ter interpretado como o tempo do último serviço; e outros o total.

A coluna está enviesada (serve apenas para inferência em caso de correlação relevante).

As perguntas não estavam claras

Foram removidos dados invalidos de tempo de serviço, ajustado números e datas que poderiam qualificar um problema para realizar a análise.

Resultados

- Muito mais mulheres do que homens (85.8%). Maioria com mais de 32 anos (Média = 40.7).
 - Seria isso causado pela população de professores nos Estados?
- > O questionário com questões relacionadas à motivação e satisfação de participação no curso foi respondido virtualmente por 1950 educadores-cursistas (41,56% dos alunos ativos durante o período de coleta de dados) durante o período de finalização das atividades do Módulo 4, período este em que 2.429 alunos se encontravam ativos no curso.
 - Quantos profissionais deixaram de fazer o curso na metade?
 - Qual o perfil desses profissionais?
- > Com relação às razões para participação no curso, 29,1% responderam "por ser uma oportunidade de formação continuada", 17,2% "por identificação profissional com o tema", 3,1% "por identificação pessoal com o tema", 0,7% "pelo fato de o curso estar vinculado à Universidade", 0,4% "Por ser um curso à distância" e 0,2% "pelo fato de o curso ser gratuito".
 - Como a motivação de cada um impacta o desempenho?
- > Em relação ao material didático, 99,3% dos participantes consideraram-no muito adequado ou adequado e 0,7% como pouco adequado ou nada adequado. Sobre a flexibilidade dos prazos para realização das atividades, 65,3% avaliaram este como "flexível", 17,4% como "muito flexível", 17,3% como "pouco ou nada flexível".
 - Por que alguns acharam inflexível?
- > Sobre a importância de interação com outros colegas através do ambiente virtual de aprendizagem, 50,3% dos respondentes avaliaram tal interação como "muito importante", 45,1% como "importante" e 4,7% como "pouco ou nada importante"
 - Por que alguns não acharam importante?
- > Com relação à organização do curso de uma forma geral, 50,6% consideraram o curso "organizado", 47,5%, "muito organizado" e 1,9% "desorganizado ou muito desorganizado".
 - Por que alguns consideram desorganizado?
- > Sobre a mediação e auxílio do tutor à distância, 75,1% consideram tal auxílio como "sempre importante", 22,3% disseram que este suporte é necessário "às vezes" e 2,7% consideraram que tal auxílio "não é necessário".
 - Por que alguns consideram desnecessário?
- A figura 2 mostra que a participação tende a cair com o tempo. Mas qual é a distância de tempo exata entre os eventos? Quando eles ocorrem (ex: perto de feriados?)

Perguntas a nível de concordância da base atitudesEducadores_df

- Q01: Usuários de drogas não têm força de vontade.
- Q02: Usuários de drogas têm menor destaque na sociedade.
- Q03: Usuários de drogas não podem ocupar cargos que exigem maior responsabilidade.
- Q04: O uso de drogas representa uma fraqueza de caráter.
- Q05: Usuários de drogas não se preocupam com si mesmos.
- Q06: Usuários de drogas são pessoas moralmente fracas.
- Q07: Usuários de drogas são pessoas sem determinação.
- Q08: Usuários de drogas não querem parar de usá-las.
- Q09: Usuários de drogas raramente prejudicam alguém a não ser a si próprios.
- Q10: A maioria dos usuários de drogas está desempregada.
- Q11: O tratamento raramente ajuda o usuário de drogas.
- Q12: Não se deve ter grandes expectativas na relação com os usuários de drogas.
- Q13: "Quem abusa de drogas pode aprender a diminuir o uso, tendo-o sob controle novamente."
- Q14: Usuários de drogas podem ser ajudados antes de chegarem ao "fundo do poço".
- Q15: As principais causas do uso de drogas é a falta de disciplina e autocontrole.
- Q16: A melhor forma de controlar os usuários de drogas é mantê-los isolados.
- Q17: Existem características que diferenciam os usuários de drogas das pessoas normais.
- Q18: Uma pessoa deve ser hospitalizada assim que apresentar sinais de uso de drogas.
- Q19: A dependência de drogas é uma doença .
- Q20: Os usuários de drogas são pessoas indesejáveis na sociedade.
- Q21: A sociedade não deveria se preocupar em proteger-se dos usuários de drogas.
- Q22: Os usuários de drogas são responsáveis pelos problemas associados ao uso de drogas.
- Q23: Os usuários de drogas devem ser isolados da sociedade.
- Q24: "Uma pessoa seria ingênua em se casar com alguém que tenha sido usuário de drogas, mesmo que estivesse recuperado."
- Q25: As pessoas não gostariam de morar próximo a alguém que tenha sido usuário de drogas.
- Q26: Alguém que tenha um histórico de uso de drogas deve ser impedido de assumir qualquer cargo público.
- Q27: Os usuários de drogas devem ser privados de seus direitos individuais.
- Q28: Usuários de drogas devem ser encorajados a assumir sua responsabilidade por suas atividades diárias.
- Q29: Ninguém tem o direito de excluir os usuários de drogas de sua vizinhança.
- Q30: Os usuários de drogas oferecem mais perigo do que as pessoas imaginam.
- Q31: Os usuários de drogas ainda são ridicularizados.
- Q32: Devem ser gastos mais recursos públicos financeiros no tratamento dos usuários de drogas.
- Q33: A sociedade precisa ser mais tolerante com os usuários de drogas.
- Q34: A sociedade tem a responsabilidade de fornecer o melhor tratamento possível aos usuários de drogas.
- Q35: Os usuários de drogas merecem nossa simpatia.
- Q36: Os usuários de drogas são um peso para a sociedade.
- Q37: Aumentar o investimento nas políticas de drogas é um desperdício de dinheiro público.
- Q38: O numero de serviços de tratamento é suficiente para o número de usuários de drogas .

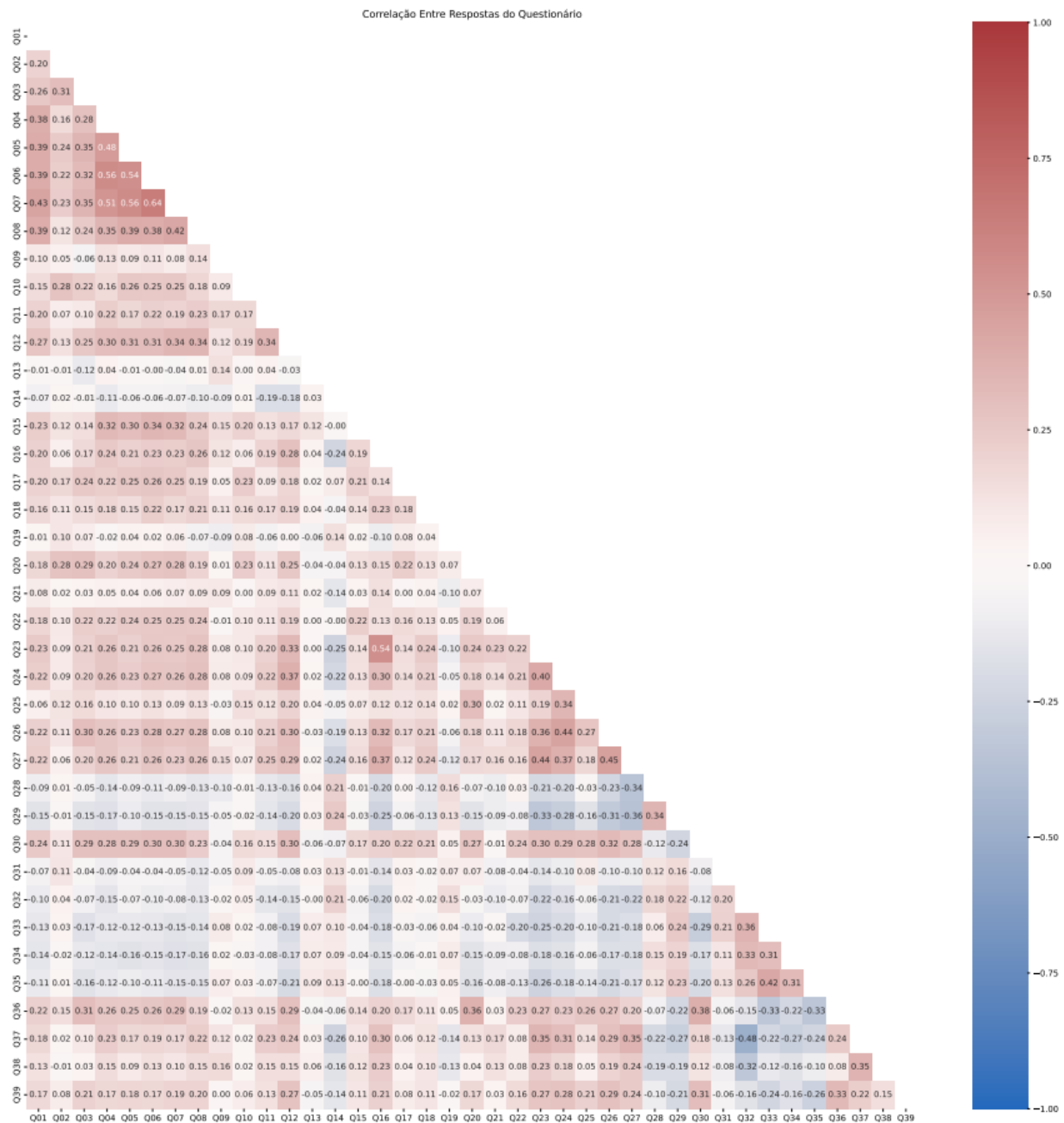
Q39: É melhor evitar alguém que tenha problemas com drogas.

Boxplot geral das perguntas realizadas nos dados da base atitudesEducadores_df



Esse gráfico achamos que ficou bem interessante, perceber o que pode estar na cabeça de uma pessoa quando se fala sobre as pessoas que são usuários de droga. As vezes não é só a força de vontade que falta, as vezes o usuário não consegue mesmo ter ânimo para continuar a vida, ou pior, as vezes não consegue sair da situação. Vemos alguns outliers achando que é melhor evitar essas pessoas, o que é uma péssima ideia, já que evitando você finge que o problema não existe e não ajuda. É interessante ver que a maioria das pessoas concorda que eles precisam de mais empatia para serem ajudados. Ao analisar cada pergunta vemos informações que são muito interessantes para recuperação de conhecimento por parte dos educadores, isso pode vir ainda com preconceitos, com problemas não resolvidos na própria família que podem ter traumatizado, e as vezes até percebermos que algo que achávamos que a resposta seria uma, na verdade na prática acaba sendo outra.


Matriz de Correlação da base atitudesEducadores_df



Vimos como a correlação se destaca em alguns casos, como era de se imaginar que se os usuários de drogas são pessoas moralmente fracas (Q06). eles também acabam sendo pessoas sem determinação (Q07). Vemos muito desse exemplo, que faz sentido a correlação. E alguns casos que realmente não fazem sentido ter muita correlação com nenhuma outra questão que é o caso das perguntas Q28, Q29 e Q31 à Q35, que são até perguntas mais pessoas do que deve ser feito.

Perguntas a nível de concordância da base praticasprofissionais_df

1. Os educadores já estão muito ocupados com suas atividades educacionais para realizarem estratégias de prevenção.
2. Como professor, sou capaz de mediar a construção colaborativa do conhecimento, e posso contribuir para realizar mudanças sociais.
3. Como educador, entendo que é papel da escola orientar seus alunos, almejando a construção colaborativa de sua identidade.
4. Admito ser necessário abordar temas como a promoção de saúde no contexto escolar.
5. A escola é o principal contexto para realização de ações em saúde e educação para os adolescentes.
6. Entendo que trabalhar com a prevenção do uso de álcool seja tão importante quanto a prevenção do uso de outras drogas.
7. Acredito que, ações para reduzir o consumo de álcool e outras drogas entre adolescentes ou atrasar o início de seu uso são positivas.
8. Considero o confronto e o "sermão" como os métodos mais eficazes de abordagem aos usuários de álcool e outras drogas.
9. Conheço o suficiente sobre as causas dos problemas relacionados à álcool e outras drogas para lidar com um aluno usuário de drogas.
10. Oriento meus alunos sobre os padrões de consumo de risco relativos ao uso de álcool e outras drogas.
11. Converso sobre os fatores de risco e proteção para o uso de álcool e outras drogas, porque os adolescentes não têm a capacidade de avaliação dos riscos.
12. Realizo práticas de prevenção aos problemas relacionados ao uso de álcool e outras drogas na escola.
13. Creio que a escola seja capaz de identificar possíveis ações de fortalecimento da rede de apoio ao usuário de álcool e outras drogas.
14. Considero-me preparado para uma avaliação das redes sociais dos adolescentes usuários de álcool e outras drogas.
15. Aceito e trabalho com a repressão ao uso de álcool e outras drogas.
16. Na minha escola, crio ações de parceria com as redes de proteção ao adolescente em situação de vulnerabilidade social.
17. A escola é um ambiente propício para o desenvolvimento de técnicas de prevenção ao uso de álcool e outras drogas.
18. Não existe tempo suficiente durante as aulas para perguntar aos alunos sobre seu consumo de álcool e outras drogas e problemas de saúde relacionados.
19. Procuro participar das estratégias de criação e implementação e de políticas públicas sobre álcool e outras drogas direcionadas à educação.
20. Sou capaz de preparar os meus alunos para se tornarem multiplicadores de ações de prevenção do uso de álcool e outras drogas na escola.
21. Crio espaços de discussões sobre o tema álcool e outras drogas com os alunos.
22. Avalio o histórico familiar de problemas com álcool e outras drogas dos meus alunos.

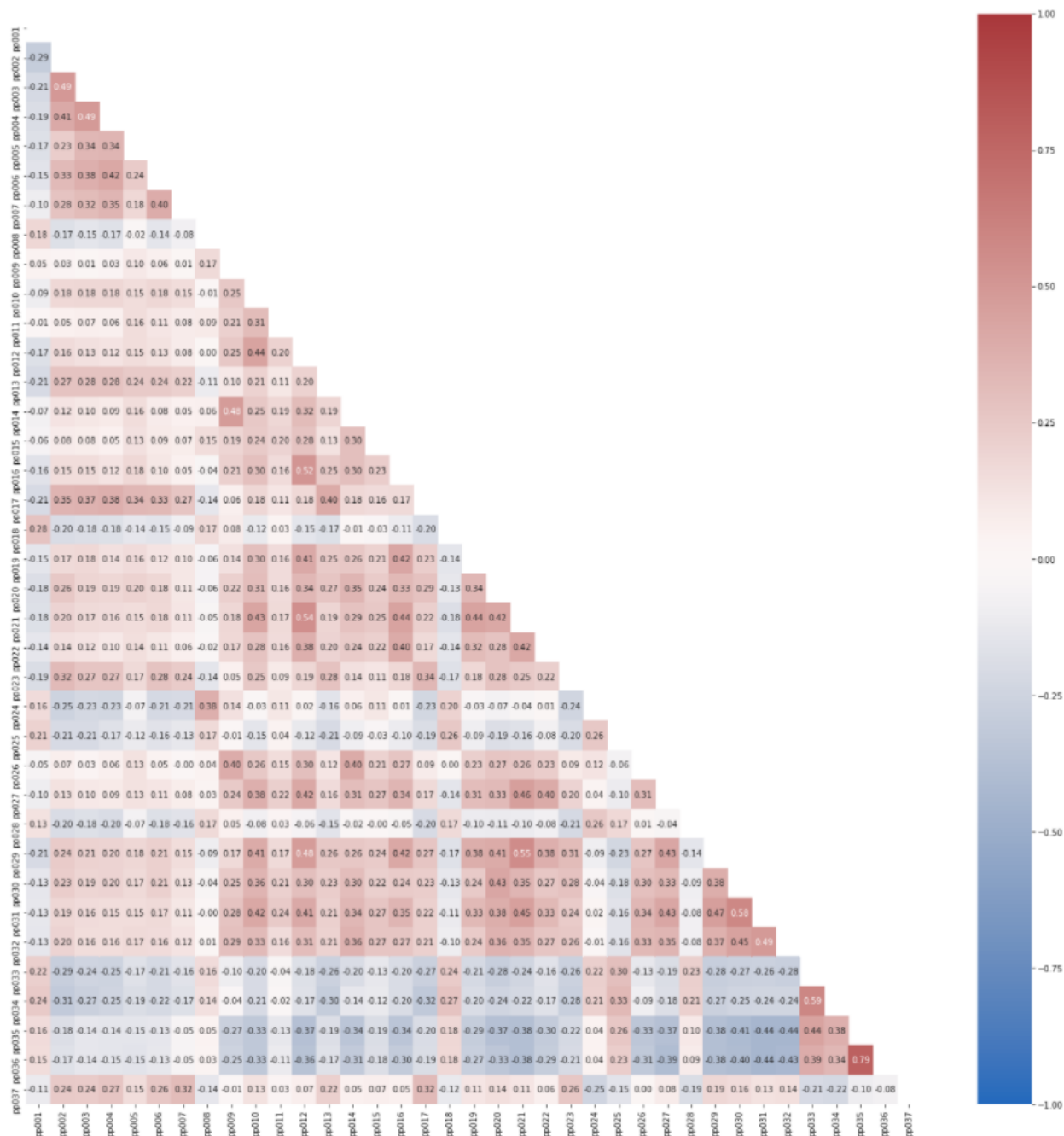
- 
23. Posso manter relações de confiança e apoio com os alunos sem abrir mão de minha autoridade em sala.
 24. Creio que minhas práticas devam ser baseadas no autoritarismo para que previnam o uso de álcool e outras drogas na escola.
 25. Os educadores não têm habilidades interpessoais para conversar e orientar seus alunos sobre o consumo de álcool e outras drogas.
 26. Minha formação acadêmico/ profissional sobre os problemas relacionados à álcool e outras drogas é adequada.
 27. Como parte do meu dia-a-dia na escola, pergunto aos meus alunos sobre seu consumo e sua história de problemas relacionados ao uso de álcool e outras drogas.
 28. A prevenção do uso de álcool e outras drogas na escola não deve fazer parte do projeto político pedagógico.
 29. Eu crio um ambiente propício à interação entre professor-aluno a fim de favorecer as estratégias de prevenção ao uso de álcool e outras drogas.
 30. Posso aconselhar apropriadamente meus alunos sobre o consumo de álcool e outras drogas e seu efeito.
 31. Oriento de forma adequada os alunos a diminuir ou a parar o consumo de álcool e outras drogas.
 32. Eu me sinto seguro para expor minhas preocupações profissionais sobre o padrão de uso de álcool e outras drogas e os riscos associados à saúde dos meus alunos.
 33. Considero que não tenho como ajudar os alunos usuários de álcool e outras drogas.
 34. Em geral, os educadores não podem ajudar seus alunos na redução do consumo de álcool e outras drogas.
 35. Não sei como devo abordar os alunos para auxiliá-los na redução do consumo de álcool e outras drogas.
 36. Não sei o que perguntar aos alunos para obter informações sobre o consumo de álcool e outras drogas.
 37. Suponho que práticas para reduzir o uso de álcool e outras drogas entre adolescentes são benéficas.

Boxplot geral das perguntas realizadas nos dados da base praticasprofissionais_df

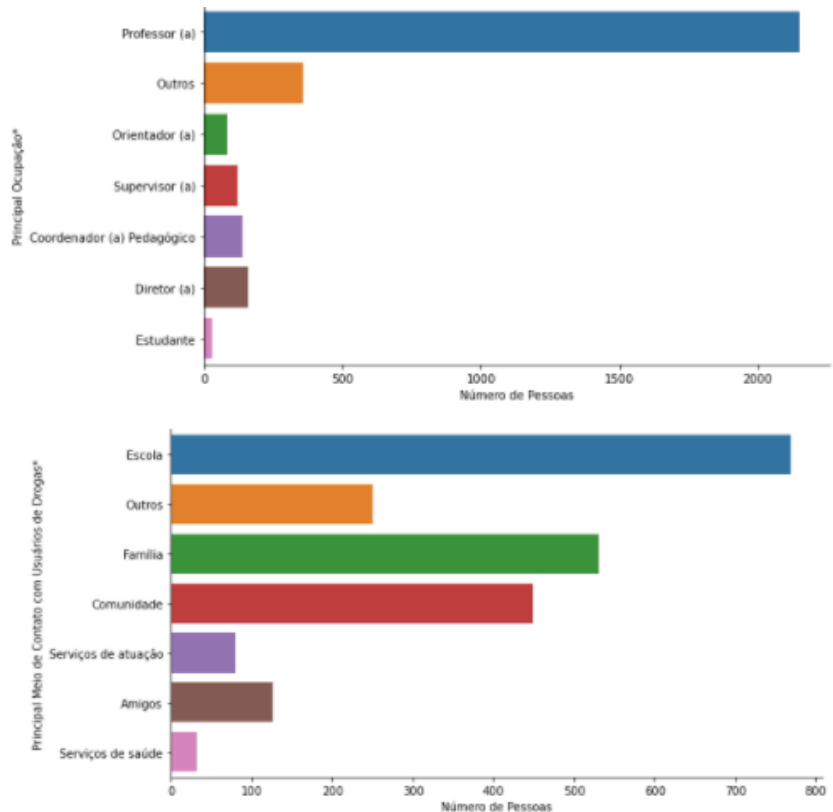
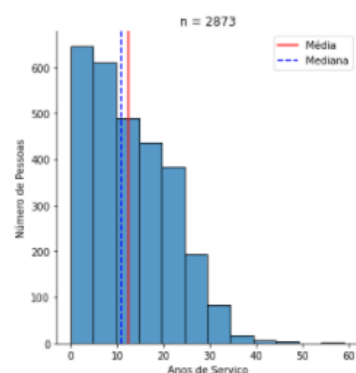
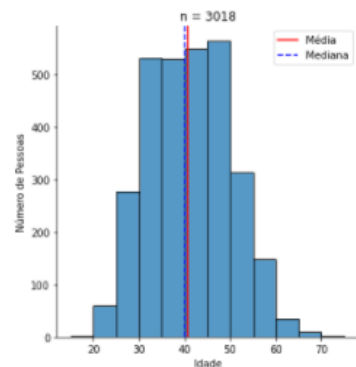


Neste gráfico conseguimos ver um pouco da distribuição e dispersão dos dados nos seus próprios números, a partir do boxplot de cada pergunta temos bem mostrado a realidade dos educadores, e que fazem sentido, e para nós já era o esperado.

Matriz de Correlação da base praticasprofissionais_df



É interessante visualizar como as correlações fazem sentido, como é o caso da correlação 0.79 entre p036 e o35, que mostra que tem uma correlação bastante positiva ao dizer que se uma pessoa não sabe como deve abordar o assunto, ela também não sabe como perguntar. Assim como as correlações negativas que mostra que não é porque não sei que não posso ajudar.




Conclusão

- Existe um viés de sobrevivente nas respostas. Ao final, só temos respostas daqueles que chegaram ao final do curso. Porém, isso também acaba sendo um bom fator, já que conseguimos melhorar a acurácia, com apenas pessoas que estejam interessadas sobre o que está sendo falado.

A partir deste trabalharmos, conseguimos entender de forma prática o quão importante é o trabalho de ser fazer um pré-processamento bem feito, e um entendimento inicial dos dados que são capazes de suprir muitas perguntas que são feitas ao decorrer do estudo. Fazer isso bem feito e pré-ativamente nos dá um ganho muito grande de já trabalhar com dados que conhecemos, sabendo medidas usadas e sua utilidade e utilização no dia a dia de trabalho.

Vimos a partir desses dados que quando eles são preenchidos a mão pode ter uma variedade de possibilidades de respostas, o que muitas vezes pode prejudicar as análises no geral. E por isso é importante que seja feito formulários que ajude o usuário a não precisar digitar coisas nas mãos e já ter opções. E também muito importante fazer esse tratamento dos dados antes de trabalhar com eles.



A partir dos gráficos plotados (outros além dos que estão neste documentos, podemos perceber que realmente em sua maioria temos professores e pessoas que trabalham juntamente com professores, como coordenadores de ensino, supervisores, que estão diretamente envolvidos com a educação nas escolas. Que prezam pela importância do tutor e sabem que o professor bem qualificado pode fazer a diferença no quesito de melhorar a absorção dos alunos e fazer eles aprender e levarem para a vida e para suas famílias. Mostra que quase 100% acha importante a interação entre pares, o que é legal, pois mostra que o dialogo sempre ajuda. E que há pessoas bem variadas, contendo uma distribuição normal de idades, e uma média e mediana de tempo de trabalho de em média 10 anos, isso é muito importante para avaliação, pois mostra que mesmo depois de muito tempo de formado as pessoas ainda buscam se qualificar ainda mais para conseguir dar ainda melhor de si, principalmente em casos tão importantes como esse.