UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

CAIO ALESSANDRO RESNAUER

APLICAÇÃO DE EXPLICABILIDADE DE OUTLIERS PARA COMPREENSÃO DE FATORES QUE INFLUENCIAM NO DESEMPENHO DE INSTITUIÇÕES DE ENSINO SUPERIOR

CURITIBA

2023

CAIO ALESSANDRO RESNAUER

APLICAÇÃO DE EXPLICABILIDADE DE OUTLIERS PARA COMPREENSÃO DE FATORES QUE INFLUENCIAM NO DESEMPENHO DE INSTITUIÇÕES DE ENSINO SUPERIOR

Application of Outlier Explainability for Understanding Factors Influencing the Performance of Higher Education Institutions

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Luiz Celso Gomes Junior

CURITIBA

2023

CAIO ALESSANDRO RESNAUER

APLICAÇÃO DE EXPLICABILIDADE DE OUTLIERS PARA COMPREENSÃO DE FATORES QUE INFLUENCIAM NO DESEMPENHO DE INSTITUIÇÕES DE ENSINO SUPERIOR

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 27/07/2023

Luiz Celso Gomes Junior
Prof. Dr.
Universidade Tecnológica Federal do Paraná

Cesar Augusto Tacla
Prof. Dr.
Universidade Tecnológica Federal do Paraná

João Alberto Fabro
Prof. Dr.
Universidade Tecnológica Federal do Paraná

CURITIBA
2023

AGRADECIMENTOS

Gostaria de começar agradecendo ao meu orientador Prof. Dr. Luiz Celso Gomes Junior, por ter sido meu guia nesta trajetória.

Aos meus amigos que estiveram comigo durante os momentos de trabalho.

À minha família, pelo apoio ao enfrentar este desafio.

Enfim, a todos os que por algum motivo contribuíram para a realização desta monografia.

RESUMO

As observações anômalas em um conjunto de dados, também chamadas de outliers, são de grande relevância em diversos cenários. A detecção de outliers é uma área de pesquisa da mineração de dados que vem sendo estudada e, uma das subáreas relacionada a isto é a de explicabilidade de outliers, que visa encontrar os motivos pelos quais a observação se destaca das demais. A fim de compreender a maturidade destas técnicas, este projeto de pesquisa propõe o estudo do cenário da educação superior brasileira, comparando técnicas tradicionais de análise de dados com as técnicas automáticas de explicabilidade de outliers, utilizando a UTFPR como caso de uso.

Palavras-chave: explicabilidade de anomalias; oam; analise visual; educacao; ensino superior.

ABSTRACT

The abnormal observations in a dataset, also known as outliers, are of great importance in various scenarios. The outlier detection is a research area of data mining that has been studied and, one of the sub-areas related to this is the outlier explanation, wich aims to find the reasons that result into that observation stand out from the others. In order to comprehend the maturity of these techniques, this research project proposes the study of the brasilian higher education, comparing traditional techniques of data analysis with automatic techniques of outlier explanation, using UTFPR as use case.

Keywords: anomaly explainability; oam; visual analysis; education; university degree.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Para um bom desenvolvimento de um país, uma importante tarefa a ser feita é a de analisar e compreender os fatores que influenciam no desempenho de instituições de ensino. De acordo com o Censo da Educação Superior de 2020, existem 2.457 universidades e instituições de ensino superior (IES) ativas no Brasil. Tais instituições podem ser classificadas através de diversos aspectos, como a região em que elas estão localizadas; as áreas de ensino dos cursos ofertados; a quantidade de docentes e de discentes; ou até se são instituições públicas ou privadas. Diante de uma grande pluralidade de pontos que diferenciam as IES, alguns programas públicos são realizados para mapear características destas organizações. O ENADE (Exame Nacional de Desempenho dos Estudantes) é realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) a fim de avaliar como os alunos de cada IES que estão próximos a se formar estão obtendo os conhecimentos e habilidades necessárias para a atuação profissional em cada área, seguindo os conteúdos previstos na diretriz curricular de cada curso. Os dados provenientes do resultado deste exame, assim como diversos outros dados relacionados às IES brasileiras, são de acesso aberto à população e podem ser utilizados para entender pontos que destacam determinadas universidades em relação a outras.

Para identificar padrões e as principais características de um conjunto de dados, é comum se utilizar a Análise Exploratória, que em geral se resume na aplicação de técnicas visuais para a representação do cenário existente no conjunto de dados. É inegável a eficiência deste tipo de técnica, uma vez que é amplamente utilizada desde em ambientes como empresas, que visam buscar formas de aumentar o lucro, até em contextos de pesquisas científicas, onde se busca avaliar hipóteses que expliquem problemas encontrados. Entretanto, existem limitações neste tipo de técnica, devido ao fato de que elas dependem fortemente da ação humana para interpretar cada aspecto a ser analisado. Isto pode se tornar um problema em casos onde o número de variáveis a serem avaliadas é muito grande.

Este trabalho tem como objetivo contrastar análise exploratória tradicional com técnicas automáticas de detecção e explicabilidade de *outliers*. Neste contexto, as técnicas de explicabilidade de *outliers* vêm sendo estudadas em busca de uma forma mais adequada de identificar padrões de anomalias em conjuntos de dados com muitas dimensões. Desta forma, pode-se propor uma hipótese de que ao aplicar técnicas de explicabilidade de *outliers* no cenário da educação superior brasileira, seria possível definir aspectos que tornam, por exemplo, a UTFPR uma instituição que se destaca (positiva ou negativamente) em relação ao cenário atual do ensino superior brasileiro. Busca-se portanto estabelecer a maturidade das técnicas de detecção e explicabilidade para este tipo de tarefa, usando a UTFPR como caso de uso.

1.1 Objetivos

1.1.1 Objetivo geral

Aplicar técnicas de explicabilidade de *outliers* no cenário da educação brasileira, a fim de entender pontos que destacam a UTFPR em relação às demais universidades.

1.1.2 Objetivos específicos

- Identificar fatores de destaque da UTFPR usando técnicas de análise tradicionais (e.g. análise exploratória);
- Escolher e aplicar técnicas de explicabilidade de *outliers* para identificar os fatores de destaque na mesma linha da análise tradicional;
- Compreender e reportar vantagens/desvantagens de utilizar técnicas explicabilidade de *outliers*;
- Identificar e desenvolver pontos de melhorias na biblioteca de explicabilidade de *outliers* implementada por alunos da UTFPR.

1.2 Justificativa

O desenvolvimento de um país depende, em grande parte, da qualidade de suas instituições de ensino. Para garantir um bom desempenho dessas instituições, é crucial analisar e compreender os fatores que influenciam seu funcionamento. Este trabalho se propõe a utilizar técnicas de explicabilidade de anomalias para buscar informações relevantes que possam auxiliar na compreensão de fatores importantes na educação superior, utilizando como caso de uso a UTFPR.

2 REFERENCIAL TEÓRICO E TRABALHOS RELACIONADOS

2.1 Avaliação de desempenho de universidades no Brasil

Para que seja possível o controle nacional sobre a eficiência das instituições de ensino superior em um país, é indispensável o acompanhamento em relação ao desempenho delas. Desde que foi instituída a Constituição da República de 1988, é obrigação do Estado manter um sistema nacional integrado de controle de diversos indicadores vinculados aos órgãos e entidades da administração federal. Desta forma, as Instituições Federais de Ensino Superior (IFES) precisam incluir nos seus relatórios de gestão das contas anuais determinados indicadores de desempenho (NORA, 2014). O ministério da educação, através do INEP, utiliza quatro indicadores de qualidade da educação superior para avaliar o desempenho das IES brasileiras: Conceito ENADE; Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD); Conceito Preliminar de Curso (CPC); e Índice Geral de Cursos (IGC).

O Exame Nacional de Desempenho dos Estudantes (ENADE) é aplicado a alunos dos cursos de graduação das IES brasileiras (tanto federais quanto privadas) que estão prestes a se formar, a fim de avaliar se os conteúdos programáticos dos cursos estão sendo atendidos pelo ensino das instituições. Através do resultado das avaliações, é atribuído a cada curso de cada instituição um conceito expresso em escala contínua e em cinco níveis. O exame é aplicado no Brasil desde 2004, e é obrigatório para que os estudantes que forem convocados se formem. O ENADE é aplicado através de um ciclo avaliativo trienal, que define quais cursos são aplicados em cada um dos três anos de ciclo. O ciclo avaliativo do ENADE também guia a atribuição dos outros indicadores de qualidade fornecidos pelo INEP. O IDD é calculado desde 2014 para os estudantes de graduação que participaram do Exame Nacional do Ensino Médio (ENEM) e do ENADE, a fim de utilizar os resultados dos exames para aproximar o quanto foi agregado de conhecimento ao estudante por parte da IES. O CPC trata-se de um cálculo realizado no ano posterior ao ENADE, que leva em conta o conceito ENADE, o IDD, o perfil dos professores e atributos vinculados à infraestrutura das IES. Os cursos que receberem CPC menor ou igual a 2 recebem obrigatoriamente visitas de avaliadores do INEP a fim de verificar a veracidade do conceito insatisfatório atribuído. Após a visita dos avaliadores, o CPC é então utilizado para atribuição do Conceito de Curso (CC), que é definitivo, e não provisório como o CPC.

O IGC avalia os cursos das IES através das médias dos últimos anos do CPC, a média dos conceitos de avaliação dos programas de pós-graduação stricto sensu atribuídos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a distribuição dos estudantes entre os diferentes níveis de ensino, graduação ou pós-graduação stricto sensu.

Além dos indicadores de desempenho, para entender a situação das IES brasileiras e para que seja possível a comparação em diversos aspectos torna-se necessário uma base unificada contendo informações referentes à infraestrutura das instituições de ensino. Tendo isto em vista, o INEP realiza anualmente o Censo da Educação Superior, que reúne informações

referentes às IES, que inclui vagas abertas, número de matrículas, número de ingressantes e de concluintes e informações sobre os docentes e suas formações. Tal coleta de dados disponibiliza para o governo e para as instituições públicas e privadas um conjunto de dados confiável para que possa ser possível a pesquisa e trabalhos que envolvem projetos de melhoria da educação brasileira.

2.2 Detecção de *outliers*

De forma geral, pode-se definir *outliers* (muitas vezes chamados de anomalias) como observações de um conjunto de dados que, de uma determinada forma, se destoam do restante das informações. Assim, não existe uma definição formal de *outlier* que possa ser utilizada em qualquer situação, mas sim um conceito geral que pode ser definido de forma mais precisa conforme a situação a ser avaliada e ao conjunto de dados.

Em diversas áreas do conhecimento, é de grande importância detectar *outliers*. Muitas vezes, eles podem representar apenas a falha de obtenção de uma informação, mas, frequentemente, podem também trazer informações valiosas e importantes de serem destacadas. Assim, a detecção automática de *outliers* passou a ser uma tarefa estudada e desenvolvida nos últimos anos como uma subárea da mineração de dados (FARIA; COLLI, 2021).

Entre as diversas técnicas existentes de detecção de *outliers*, pode-se classificar em que elas são baseadas: em estatística; em métodos de cálculo de distância; em densidade de *clusters*; em proximidade de *clusters*; e métodos híbridos - que utilizam mais de uma destas técnicas como base (CHANDOLA; BANERJEE; KUMAR, 2009).

2.3 Explicabilidade em Aprendizado de Máquina

As técnicas de aprendizado de máquina vêm sendo utilizadas em diversas áreas, devido à existência de problemas que dificilmente poderiam ser solucionados ao utilizar técnicas convencionais da computação. Um problema decorrente da aplicação de tais técnicas é o de que, muitas vezes, não se pode afirmar que as predições obtidas pelo algoritmo são confiáveis - isto é - não existe uma forma de determinar especificamente quais fatores analisados pelo algoritmo determinaram o resultado.

Neste contexto, estão sendo desenvolvidas as técnicas de explicabilidade de aprendizado de máquina, ou *Interpretable Machine Learning* (IML). Esta área de pesquisa busca definir formas de criar modelos de aprendizado de máquina cujos resultados sejam passíveis de interpretação, ou seja, que permitem a definição de quais recursos do modelo aplicado levaram ao resultado obtido pela aplicação (característica chamada de interpretação local), e que também buscam modelagens que podem ser entendidas como um conjunto lógico (chamada de interpretação global).

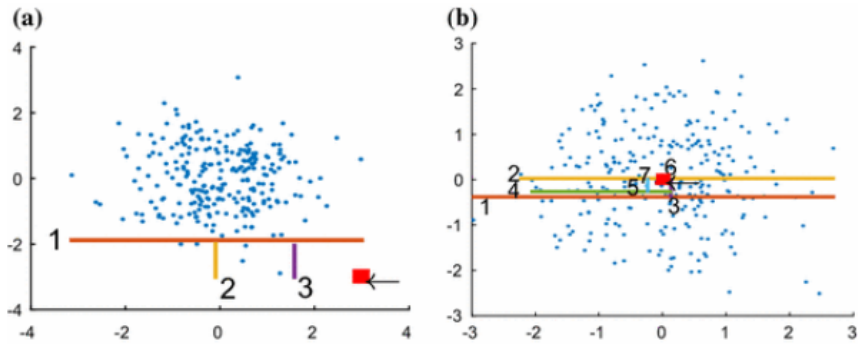
2.3.1 Explicabilidade de *outliers*

Além de detectar, uma outra importante atividade que pode ser desafiadora é a de interpretar o porquê de uma amostra ser considerada um *outlier*. Uma forma de interpretar tal problema seria definir em quais dimensões uma observação se isola das demais em um conjunto de dados. Este problema é chamado de Explicabilidade de *Outliers*. Uma das técnicas desenvolvidas para Explicabilidade de *Outliers* é a chamada *outlier aspect mining* (OAM), técnica que visa pesquisar os subespaços do conjunto de dados a fim de encontrar em qual deles a observação possui uma distância maior em relação ao restante dos dados. É importante destacar que as técnicas de explicabilidade de *outliers* podem ser independentes da técnica utilizada para a detecção de outliers, como é o caso da OAM (SAMARIYA *et al.*, 2020).

Entre as principais técnicas de OAM, a que possui maior quantidade de pesquisas e resultados publicados é a *Score and Search* (SAMARIYA *et al.*, 2020). As abordagens desta técnica dividem a tarefa de OAM em duas etapas: *Score*, que aplica uma função de pontuação a fim de mensurar o quão anômalo o dado é em relação a um subespaço; e *Search*, que gera o conjunto de subespaços os quais serão aplicados a técnica de *Score*, bem como seleciona quais deles mais se destacam.

Um algoritmo importante de *Score and Search* é o *Isolation Path* (IPath), que é derivado da técnica de detecção de outliers *Isolation Forest* (IForest). O IPath é um algoritmo de OAM que utiliza uma métrica de distância que independe da dimensionalidade, o que torna possível a exploração dos subespaços em busca daquele onde a observação é mais facilmente separável das demais (VINH *et al.*, 2016). Pode-se observar na Figura 1 um exemplo de funcionamento do IPath, onde o dado analisado está representado pelo quadrado em vermelho. A ideia é de que um *outlier* pode ser separado do restante dos dados com poucos cortes, como apresentado em (a), enquanto um dado que está cercado por outros precisa de muitos cortes para ser separado (b).

Figura 1 – Fonte: Exemplo de funcionamento do IPath.



Fonte: (SAMARIYA *et al.*, 2020)

Haja visto que poucas implementações relacionadas ao tema estão disponíveis segundo Faria; Colli, 2021, os autores desenvolveram uma biblioteca em *Python* que permite a aplicação de algoritmos de OAM, bem como a análise dos resultados desta aplicação. O algoritmo escolhido para a implementação da biblioteca foi o IPath, devido à baixa complexidade computacional e à presença de pesquisas anteriores validando a técnica.

2.4 Trabalhos Relacionados e Revisão Bibliográfica

2.4.1 Análise de desempenho de universidades brasileiras

É importante definir quais métodos e quais conjuntos de dados podem ser utilizados para que seja possível uma análise adequada das IES brasileiras. Um grupo de indicadores de desempenho disponível para análise é o presente nos relatórios de gestão de contas das IFES. A fim de compreender os resultados da análise dos dados presentes em tais indicadores, Dalla Nora (2014) pesquisou e os correlacionou com o indicador IGC do INEP. Sua conclusão foi de que o relatório de gestão das IFES possui deficiências no conjunto de indicadores, e que dificilmente tais informações podem ser utilizadas a fim de comparar instituições, uma vez que é apresentada uma grande heterogeneidade dos dados entre as IFES e muitas vezes os indicadores não são representados em números, porcentagem ou razão.

2.4.2 Avaliação dos resultados do ENADE

Em relação à avaliação do desempenho das universidades brasileiras, destacam-se os indicadores oferecidos pelo INEP. Tais indicadores levam como principal fator o resultado obtido a partir da aplicação do ENADE para cada curso das IES brasileiras. Assim, ao utilizar tais dados para a avaliação de IES, é preciso entender que tipos de resultados podem ou não serem obtidos e refletir sobre a forma de avaliação oferecida pelo exame. Para que seja possível a argumentação em relação à eficiência ou ineficiência das IES em seus cursos oferecidos, é necessário entender o quão precisa é a metodologia aplicada pelo ENADE.

Com o objetivo de identificar a melhoria na qualidade da educação superior brasileira com a implementação do ENADE, Claudia M. Griboski (2012) analisou os resultados do exame juntamente com os indicadores de qualidade obtidos por visitas *in loco* nas IES. Através da análise das alterações metodológicas que foram realizadas no ENADE desde sua implementação em 2004, a autora conclui que, embora exista um grande potencial de utilização das métricas do ENADE juntamente com os indicadores de qualidade para a melhora da oferta de cursos por parte das IES, de forma geral as informações não são utilizadas em reflexões e análises para a melhora da qualidade de ensino por parte das instituições. Desta forma, destaca-se a

importância de se criar modelos que utilizem de forma ampla os dados obtidos pelo ENADE para a melhora da educação superior brasileira (GRIBOSKI, 2012).

Embora o ENADE consiga avaliar o quão bem os estudantes estão desempenhando em relação aos conteúdos programáticos dos cursos, é importante destacar que existem limites em relação ao desempenho que pode ser avaliado, e este não pode ser isolado de questões sociais, econômicas e culturais dos avaliados. Os autores Júlio C. G. Bertolin, Telmo Marcon (2015), em seu artigo destacam as limitações da avaliação do ENADE e de outros métodos de avaliação de desempenho de IES quando não aliadas à análise socioeconômica e cultural dos estudantes.

2.4.3 Avaliação dos dados socioeconômicos brasileiros

O Programa das Nações Unidas para o Desenvolvimento (PNUD) da Organização das Nações Unidas (ONU), a fim de definir uma medida universal de desenvolvimento econômico e qualidade de vida, criou em 1990 o Índice de Desenvolvimento Humano (IDH). Para estabelecer tal métrica, são levados em conta a Renda Bruta *per capita*, a expectativa de vida, e o nível de escolaridade da população do local a ser avaliado. Este índice é medido em valores que variam de 0 a 1, sendo que valores maiores indicam maior qualidade de vida e maior desenvolvimento econômico.

O IDH geralmente é utilizado a nível nacional, mas também é usado para medir a nível municipal e estadual através do Índice de Desenvolvimento Humano Municipal (IDHM), que se trata de uma adaptação do IDH a nível dos municípios e brasileiros, e que está disponível no Atlas do Desenvolvimento Humano no Brasil[1] - programa realizado em parceria do PNUD com o o Instituto de Pesquisa Econômica Aplicada (IPEA), que, dos anos 1991 a 2010 utilizavam como base o Censo Demográfico, e que a partir de 2012, utiliza como base a Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatistica (IBGE).

Os dados disponibilizados pelo IPEA são de acesso aberto e estão disponíveis no portal do governo brasileiro. O instituto também oferece em sua base de dados índices socioeconômicos de concentração de renda a nível municipal e estadual, sendo eles o GINI, que busca indicar o grau de concentração de renda variando de 0 a 1; e o RDPC - Renda Domiciliar *per capita*. Compreender fatores como estes são de grande importância para a avaliação de desempenho das instituições de ensino (BERTOLIN; MARCON, 2015).

[1] http://www.atlasbrasil.org.br/

3 METODOLOGIA

3.1 Caso de Uso e Análise Exploratória

Para compreender aspectos relacionados ao desempenho das Instituições de Ensino Superior brasileiras, neste trabalho será apresentado como caso de uso a Universidade Tecnológica Federal do Paraná (UTFPR). Para que seja possível a aplicação de algoritmos de explicabilidade de *outliers*, é necessária uma compreensão inicial do conjunto de dados que será utilizado como base para estes algoritmos. Assim, foi realizada uma análise exploratória sobre os dados do Censo da Educação Superior e do ENADE, com foco em destacar os aspectos das universidades que possuem um desempenho superior no exame e, para que isto seja possível, inicialmente é necessário definir de que forma será medido o desempenho.

3.2 Fontes de Dados

As características a serem avaliadas em relação a desempenho serão o conceito médio do ENADE dos cursos e a proporção de concluintes em relação a ingressantes na instituição. O foco destas análises é a de observar as características da UTFPR, de forma que seja possível argumentar em quais aspectos a universidade pode ser considerada um *outlier* quando comparada ao cenário da educação superior brasileira dos anos 2010 a 2021. A origem dos dados e o detalhamento em relação à análise exploratória pode ser encontrado no capítulo 4.

Além disto, como descrito na seção 2.4.2, é importante também vincular as análises de desempenho com os aspectos socioeconômicos e estruturais das instituições. Para que seja possível a avaliação destes aspectos, está disponível também no portal do governo a base de dados do Atlas Socioeconômico Brasileiro, que possui informações como o IDHM de cada município brasileiro, que pode ser vinculado ao município da IES disponível na base do Censo da Educação Superior.

3.3 Aplicações de Técnicas de OAM

A partir da compreensão de tal cenário, passa a ser possível a implementação de técnicas de explicabilidade de *outliers* na base de dados estudada a fim de comparar a eficiência das técnicas tradicionais com as estudadas. Nesta proposta de pesquisa, planeja-se a utilização da biblioteca desenvolvida por Faria; Colli, 2021, que permite a aplicação de técnicas de *Outling Aspect Mining*. Esta implementação foi feita com a linguagem *Python*, e para a manipulação e visualização dos dados foram utilizadas as bibliotecas *pandas, seaborn* e *scipy*.

Uma vantagem de aplicar explicabilidade de outliers para compreender este tipo de cenário é a de poder comparar dados de forma automática em uma grande quantidade de di-

mensões. Desta forma, a partir da realização da análise exploratória dos dados, pode-se definir dados socioeconômicos e dados referentes à infraestrutura das IES para serem utilizados no algoritmo de *Score and Search*. Os dados definidos podem ser encontrados na seção 4 - Análise Exploratória.

A abordagem que será utilizada para a etapa de *Search* é a de *Simple Combination*, que combina todos os conjuntos possíveis de dimensões entre um tamanho mínimo e máximo previamente definidos (FARIA; COLLI, 2021). Já para a etapa de *Score*, a técnica que será utilizada é o IPath, detalhado na seção 2.3.1.

Outra vantagem das aplicações de técnicas de OAM é a automatização do processo de comparação dos dados. Assim, é possível também realizar as análises propostas de forma a separar as áreas de ensino dos cursos oferecidos pelas instituições. Ao fazer isto, torna-se viável a identificação de outliers em relação a áreas de ensino, uma vez que uma instituição pode se destacar apenas em alguma delas e não em todas. Na seção 4, estão listadas as áreas que são definidas na base de dados do ENADE, que podem ser utilizadas nas análises.

Na biblioteca desenvolvida pelos autores citados, existe também a implementação de uma seção focada na visualização dos dados, que pode ser utilizada na interpretação dos resultados. Assim, é possível comparar os resultados das análises para a identificação dos pontos em que a instituição avaliada se destaca e também a interpretação dos motivos pelos quais estes pontos foram considerados outliers.

4 ANÁLISE EXPLORATÓRIA

A fim de compreender a situação da UTFPR nos últimos anos em relação a desempenho quando comparada a outras IES brasileiras, foi feita uma análise exploratória inicial, utilizando-se das técnicas convencionais de análise de dados.

4.1 Conjuntos de Dados

As bases de dados utilizadas na análise exploratória foram os microdados do Censo da Educação Superior (disponível dos anos 1995 a 2021)[1], os microdados do ENADE (disponível dos anos 2004 a 2021) [2], e a base de dados do Atlas do Desenvolvimento Humano do Brasil [3]. As três bases são de acesso aberto e podem ser obtidas no site oficial do Governo do Brasil.

Para que fosse possível a análise adequada das bases de dados, foi preciso realizar tratamentos de dados. Primeiramente, a fim de separar as informações mais recentes, as quais as metodologias do ENADE foram mais atualizadas, foi limitada a extração dos dados a partir de 2010. Além disso, foi criado um campo referente à media do conceito ENADE para os cursos ofertados por cada instituição. No caso da UTFPR, também foi aplicada esta lógica aos diferentes campi da universidade.

Como um dos critérios definidos para a mensuração de desempenho das IES foi a proporção da quantidade de concluintes pela quantidade de ingressantes, foi também criada a coluna "Proporção de Concluintes por Ingressantes", calculada a partir da divisão da coluna "QT_CONC"pela coluna "QT_ING"da base de dados do Censo.

Para que fosse possível o vínculo dos dados socioeconômicos municipais e estaduais do Atlas do Desenvolvimento Humano com as bases do Censo e do ENADE, foi utilizada como chave o campo "Código do Município", que segue a padronização disponibilizada pelo IBGE.

4.2 Análise dos Dados

A análise se iniciou ao observar como a UTFPR desempenhou no último ano quando comparada ao restante das IES na média dos conceitos ENADE. Foi destacado também na histograma o campus UTFPR-CT, o maior campus da universidade.

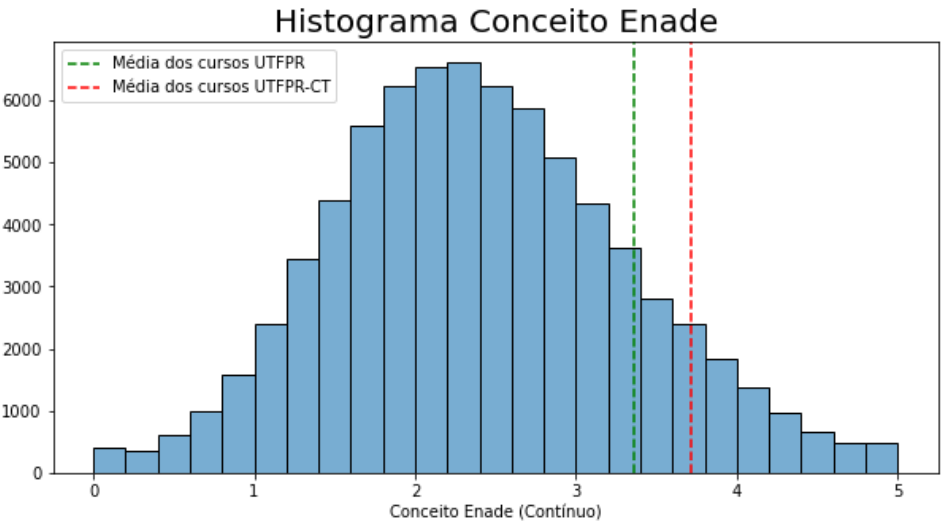
Como pode-se observar na Figura 2, o desempenho da UTFPR é consideravelmente positivo quando comparado ao restante das IES. O campus UTFPR-CT destaca-se ainda mais. Para obter uma métrica para representar o quão acima da média a UTFPR está em relação ao conceito ENADE, foi calculado o z-score da universidade, que resultou em 0.97.

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior

[2] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade

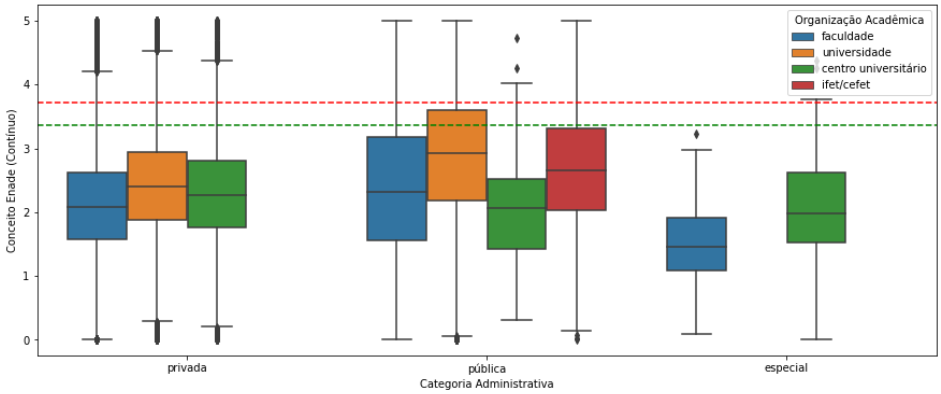
[3] https://dados.gov.br/dataset?tags=%C3%8Dndice+de+desenvolvimento+humano+-+IDH

Figura 2 – Histograma Conceito Enade.



Fonte: Autoria Própria (2023)

Figura 3 – Conceito Enade por tipo de instituição.



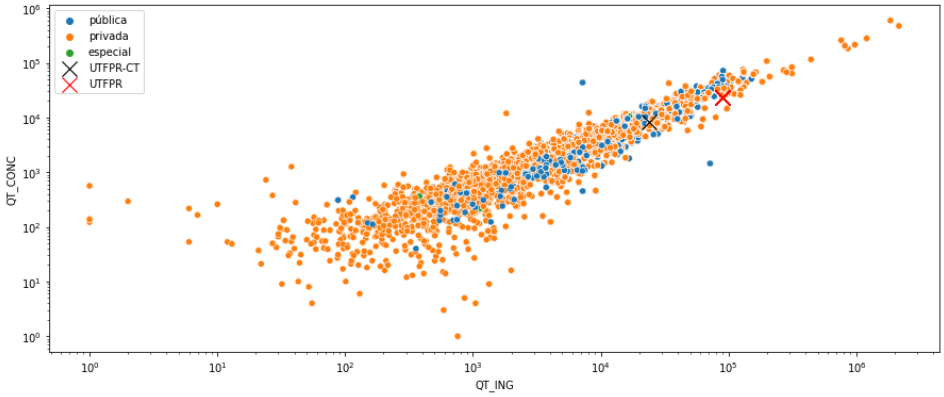
Fonte: Autoria Própria (2023)

Aprofundando um pouco mais a análise, ao separar as instituições de ensino em categoria administrativa (pública/privada/especial) e Organização acadêmica (faculdade/universidade/-centro universitário/ifet-cefet), pode-se observar na Figura 3 que as IES tendem a desempenhar melhor no ENADE quando são universidades públicas. A UTFPR fica acima da mediana apresentada para este tipo de instituição.

Uma outra análise proposta em relação a desempenho foi a de proporção entre concluintes e ingressantes. Assim, foi criada uma coluna 'Proporção de Concluintes por Ingressantes' cujo cálculo é feito através da divisão da coluna "QT_CONC"(quantidade de concluintes) pela coluna "QT_ING"(quantidade de ingressantes). Observa-se no gráfico da Figura 4 a distribuição entre concluintes e ingressantes nas IES brasileiras:

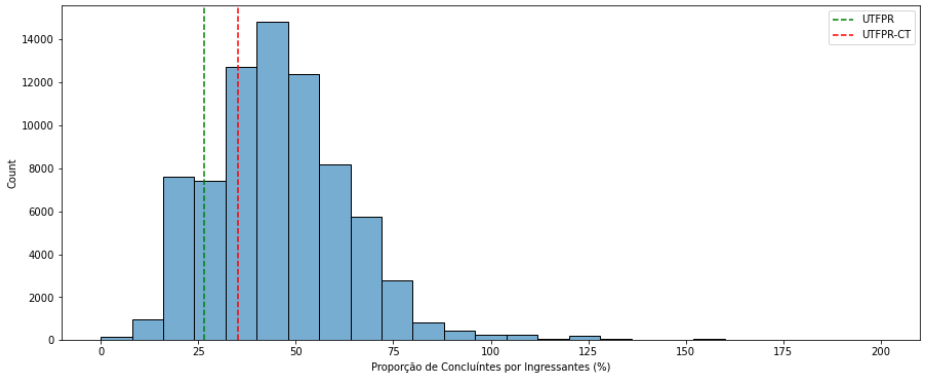
Para entender de forma mais precisa a situação da UTFPR em relação a tal proporção, foi feito um histograma destacando a universidade e o campus UTFPR-CT, disponível na Figura 5.

Figura 4 – Distribuição de concluintes por ingressantes.



Fonte: Autoria Própria (2023)

Figura 5 – Histograma da Proporção de concluintes por ingressantes.



Fonte: Autoria Própria (2023)

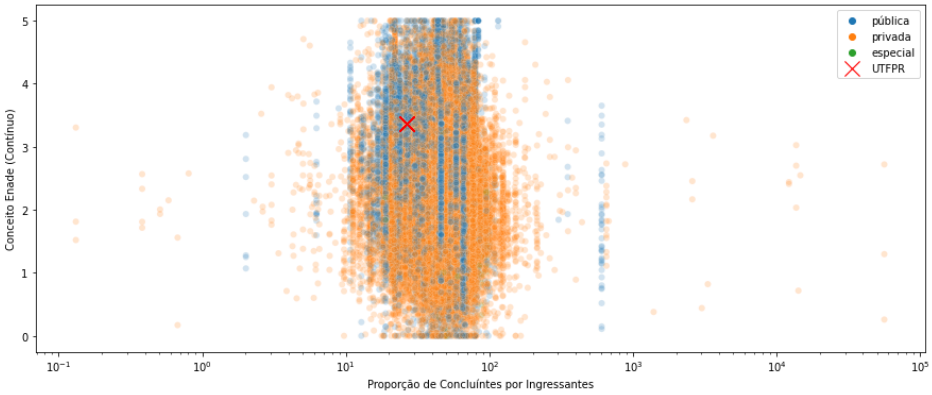
Vemos que, em tal análise, a UTFPR não é de grande destaque, uma vez que fica abaixo da mediana de tal proporção entre as IES brasileiras.

A fim de compreender também se tal análise possui alguma relação com o conceito ENADE, foi feita a distribuição da proporção pela média dos conceitos ENADE, como demonstrado na Figura 6. Porém, não ficou evidente a relação entre tais medidas ao exibir no gráfico de distribuição.

Agora, para verificar as correlações de alguns dados socioeconômicos com o resultado do ENADE, foram levados em conta os dados do Atlas do Desenvolvimento Humano dos municípios onde se localizam as instituições de ensino. Como descrito na seção 2.4.2, é importante que, ao analisar o desempenho de instituições de ensino, levar em conta questões sociais, econômicas e culturais envolvidas. Para aprofundar as análises, os indicadores dos municípios onde se localizam a IES propostos a serem analisados são o IDHM, o GINI e a Renda Per Capita.

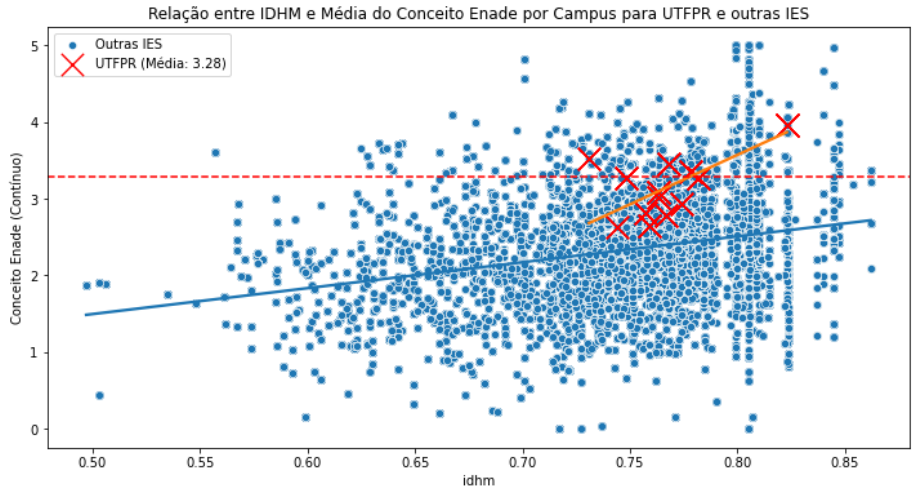
Ao analisar se existe alguma relação entre o desempenho dos cursos da IES no ENADE com IDHM do município da instituição, foi analisado o gráfico de distribuição disponível nas

Figura 6 – Distribuição da proporção de concluintes x Conceito ENADE.



Fonte: Autoria Própria (2023)

Figura 7 – IDHM x Conceito Enade Médio.



Fonte: Autoria Própria (2023)

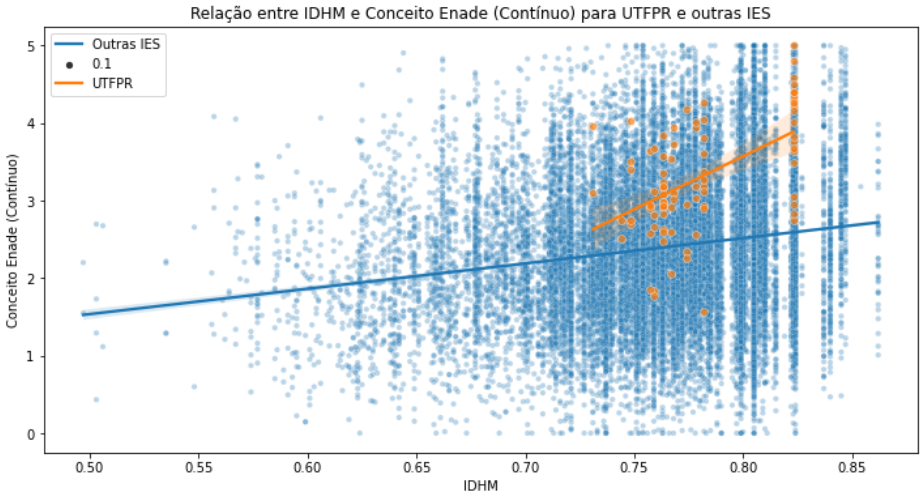
Figuras 7 e 8, as quais trazem, respectivamente: a média do Conceito ENADE de cada campus comparada ao IDHM; e o conceito ENADE de cada curso comparado ao IDHM.

Observando as distribuições, podemos notar que há uma tendência de crescimento do conceito ENADE médio quando o IDHM do município é maior. Em relação a UTFPR, a linha de regressão demonstra que este impacto do IDHM é ainda maior.

Como o campus de Curitiba foi o que mais se destacou nas análises que levam em conta o Conceito ENADE, os dados foram separados em quartis e feita uma nova comparação - a do campus de Curitiba com apenas universidades de cidades com IDHM próximos ao da capital, como apresentado na Figura 9.

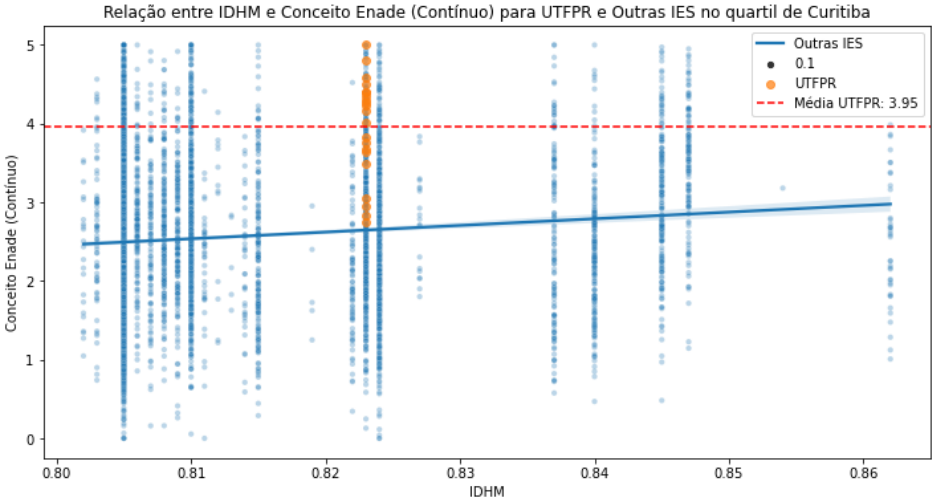
Pode-se verificar no gráfico que todos os cursos o campus de Curitiba ficam acima da linha de regressão das outras IES. Para quantificar o quão acima da média o campus ficou, foi calculado o z-score da UTFPR neste quartil, que resultou em 1.45.

Figura 8 – IDHM x Conceito Enade



Fonte: Autoria Própria (2023)

Figura 9 – IDHM x Conceito Enade (Apenas no quartil de IDHM de Curitiba).



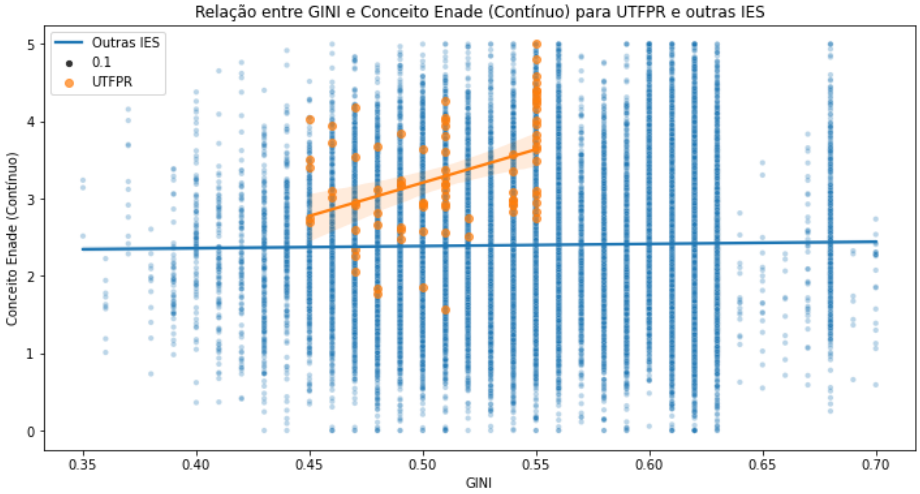
Fonte: Autoria Própria (2023)

As análises feitas sobre o IDHM foram replicadas também aos outros dois indicadores socioeconômicos propostos: GINI e Renda Per Capita. Nas Figuras 10 e 11, estão as distribuições do GINI pelo Conceito ENADE, e nas Figuras 12 e 13, as análises pela Renda Per Capita, sendo que na 11 e na 13 estão selecionados apenas os municípios no mesmo quartil que Curitiba. No quartil de Curitiba em relação ao GINI, o z-score obtido sobre o Conceito ENADE da UTFPR foi de 1.23, e no quartil selecionado na Renda Per Capita, de 1.45.

4.3 Interpretação dos dados

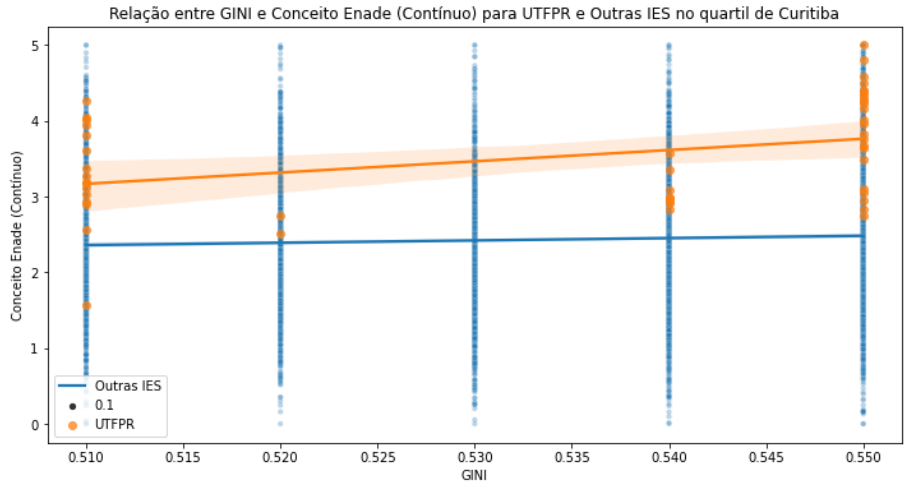
Foi observado nas análises da seção 4.2 que a UTFPR possui um crescimento mais acentuado em comparação com as outras instituições quando se fala na comparação dos da-

Figura 10 – GINI x Conceito Enade



Fonte: Autoria Própria (2023)

Figura 11 – GINI x Conceito Enade (Apenas no quartil de IDHM de Curitiba).



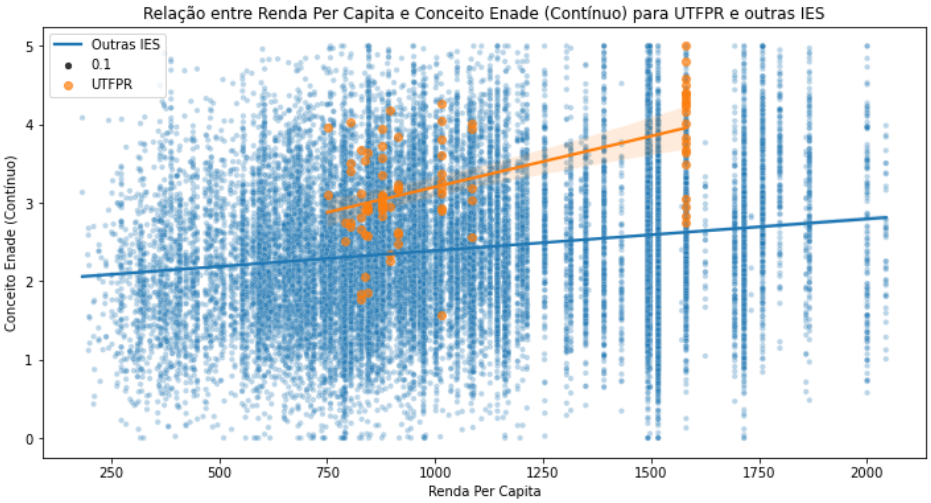
Fonte: Autoria Própria (2023)

dos socioeconômicos ao Conceito ENADE. Os gráficos apresentados indicam que o IDHM e a Renda Per Capita são indicadores em que tem correlação com o Conceito ENADE quando observa-se o conjunto de cursos ofertados pelas IES brasileiras. Porém, o indicador GINI não mostrou correlação aparente nas análises feitas.

Na análise exploratória, constatou-se que a correlação entre IDHM e Renda Per Capita ao ENADE também está presente nos campus da UTFPR, e de forma ainda mais acentuada. Ao calcular o z-score do Conceito ENADE da UTFPR, obteve-se um valor de 0.97, o que indica que a universidade está acima de aproximadamente 83.4% das instituições nesta análise. Isto por si só é um fator de destaque da universidade.

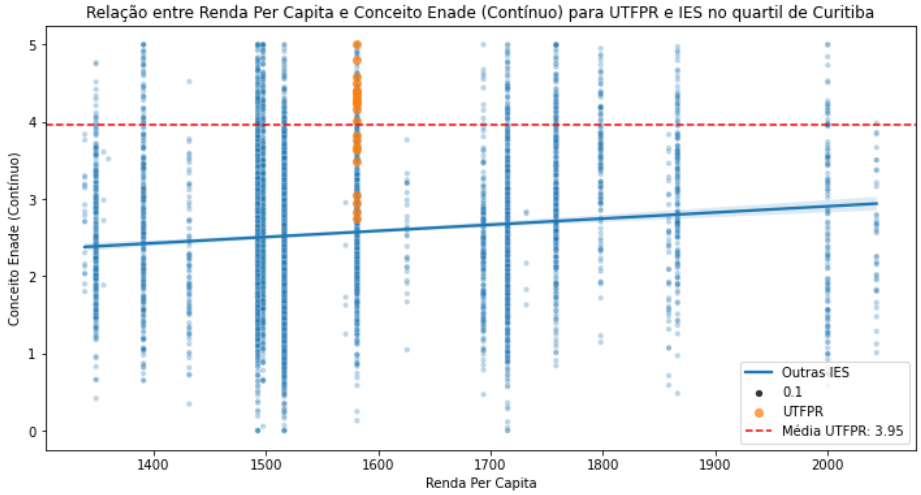
Além disso, o campus de Curitiba foi o que mais se destacou nas análises. Quando comparou-se os cursos ofertados em Curitiba pela UTFPR a outros cursos de cidades com in-

Figura 12 – Renda Per Capita x Conceito Enade.



Fonte: Autoria Própria (2023)

Figura 13 – Renda Per Capita x Conceito Enade (Apenas no quartil de IDHM de Curitiba).



Fonte: Autoria Própria (2023)

dicadores socioeconômicos próximos através da separação dos dados em quartis, foi calculado o z-score de 1.45 tanto na análise pelo IDHM, quanto na de Renda Per Capita, o que significa que, em ambas as análises, o campus de Curitiba fica acima de 92.6% das IES de municípios com indicadores socioeconômicos próximos.

Ao longo da análise exploratória, estes foram os maiores fatores de destaque encontrados na UTFPR em relação a outras IES. Agora, quando for realizada a aplicação das técnicas de OAM, pode-se aproveitar a alta dimensionalidade do algoritmo de *Score and Search* para buscar de forma automática as dimensões em que a UTFPR mais se destaca, e verificar os potenciais ou as limitações dos algoritmos de explicabilidade de outliers em encontrar fatores de destaque para uma observação, quando comparados a técnicas tradicionais de análise exploratória.

5 APLICAÇÃO DA BIBLIOTECA DE OAM

A fim de identificar possíveis vantagens ou desvantagens de utilizar as técnicas de explicabilidade de *outliers*, foram selecionados os principais dados utilizados na análise exploratória para serem aplicados nos algoritmos desenvolvidos na biblioteca de OAM para *Python*. Ao obter os resultados da aplicação da biblioteca, pode-se comparar com as conclusões observadas na análise exploratória, e determinar pontos fortes e fracos de utilizar este tipo de técnica. Além disso, uma proposta deste trabalho é a de observar pontos de melhoria que podem ser implementados na biblioteca a fim de melhorar a utilização em casos como o proposto aqui.

5.1 Conjunto de Dados

As fontes de dados utilizadas nesta etapa do trabalho foram as mesmas descritas na seção 4.1, para que fosse possível uma comparação dos resultados obtidos. As colunas selecionadas para a etapa de *Search* do algoritmo foram:

- **'conceito_enade'**: Conceito Enade;
- **'pro_conc_ing'**: Proporção de alunos concluintes por ingressantes;
- **'idhm'**: Índice de Desenvolvimento Humano Municipal (IDHM);
- **'indice_gini'**: Índice de Gini;
- **'renda_pc'**: Renda per Capita;
- **'proporcao_docentes_dout'**: Proporção de docentes com doutorado pelo total de docentes.

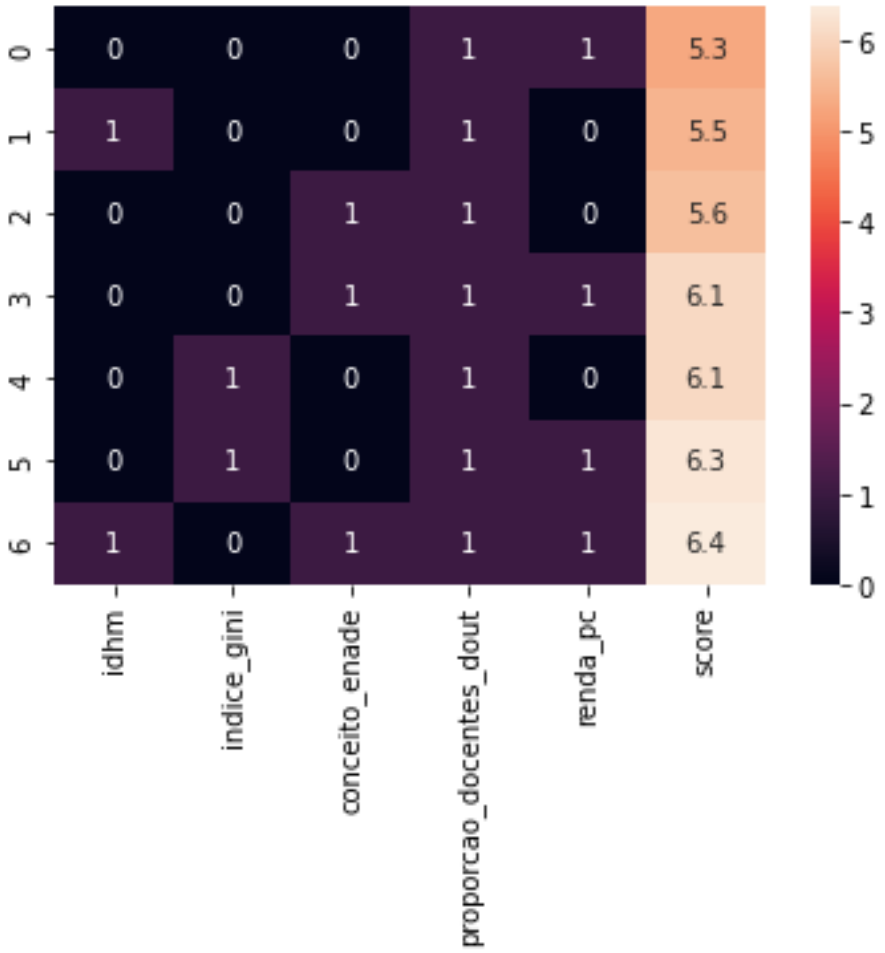
Assim como feito na análise exploratória, as colunas 'conceito_enade' e 'pro_conc_ing' foram as variáveis de qualidade selecionadas.

5.2 Aplicação do algoritmo de *Score and Search*

As análises utilizando a biblioteca foram divididas de maneira similar às feitas na etapa de análise exploratória: inicialmente, avaliou-se a UTFPR em relação às IES brasileiras e, em seguida, o campus de Curitiba (UTFPR-CT) em relação aos outros campi das IES brasileiras.

Para visualizar os resultados do algoritmo de *Score and Search*, foi utilizada a função "visualize_oam_results"disponível no módulo de visualização da biblioteca, que disponibiliza um *heatmap* contendo as dimensões presentes em cada subespaço e o *Score* obtido nele. Nas figuras apresentadas neste trabalho, estão sendo apresentados no *heatmap* os 7 subespaços com melhor *Score* para cada análise.

Figura 14 – Resultado da aplicação da OAM para média da UTFPR vs outras universidades, utilizando Conceito ENADE como variável de qualidade.



Fonte: Autoria Própria (2023)

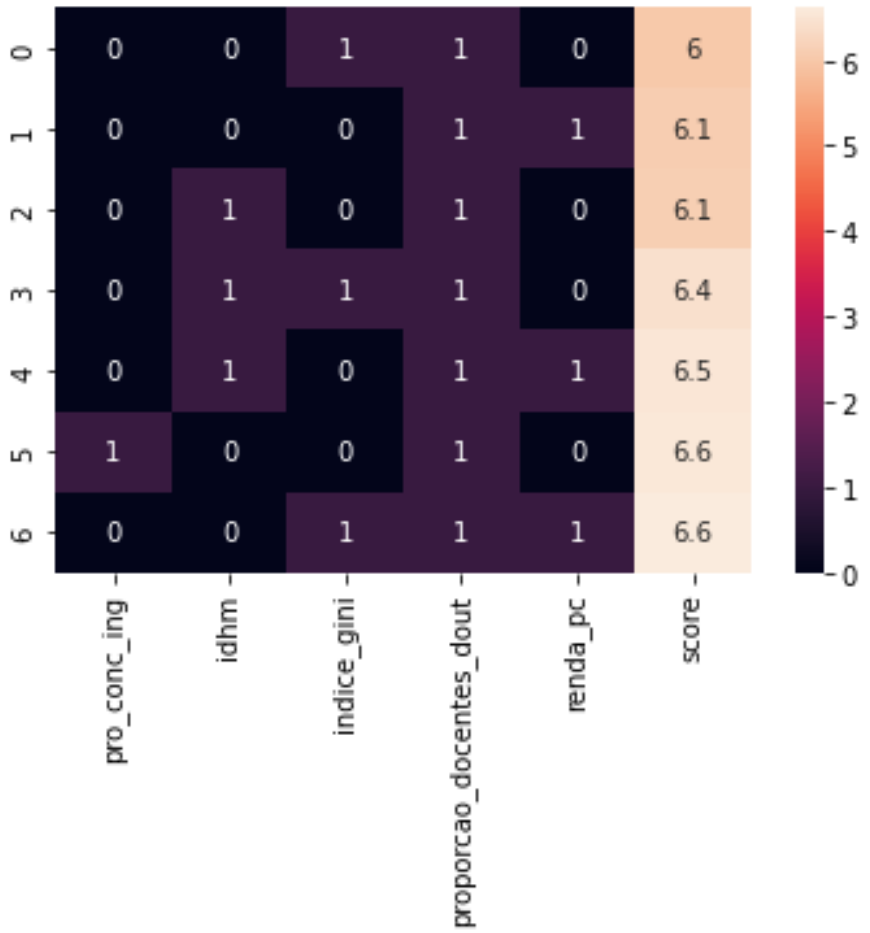
5.2.1 Avaliação da média da UTFPR com a média das outras universidades

Na primeira etapa desta avaliação, o Conceito ENADE foi utilizado como variável de qualidade. O resultado obtido está disponível na Figura 14. O subespaço contendo a variável de qualidade que obteve o melhor resultado foi o que contém a proporção de docentes doutores, com um *Score* de 5.6.

Utilizando a proporção de alunos concluintes por ingressantes como medida de qualidade, obteve-se o resultado exibido na Figura 15. Nesta análise, o subespaço com a variável de qualidade que melhor desempenhou continuou sendo o que contém a proporção de professores com doutorado, com um *Score* de 6.6.

Quando utilizadas as duas variáveis de qualidade simultaneamente no algoritmo, o melhor resultado manteve-se sendo o subespaço contendo a proporção de docentes doutores e o conceito ENADE, com um *Score* de 5.3, como exibido na Figura 16.

Figura 15 – Resultado da aplicação da OAM para média da UTFPR vs outras universidades, utilizando Proporção de Concluintes por Ingressantes como variável de qualidade.



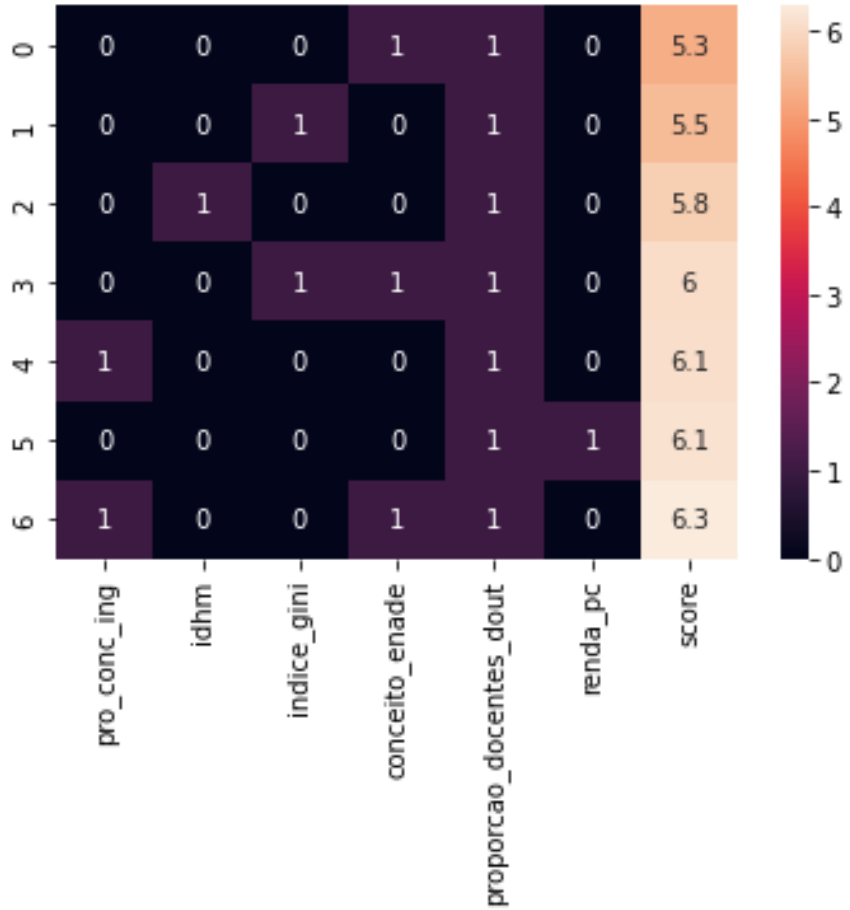
Fonte: Autoria Própria (2023)

Para auxiliar na compreensão dos resultados, foi calculado o z-score de cada uma das variáveis utilizadas no algoritmo para a UTFPR. Estes dados estão disponíveis na Figura 17. Esta maneira de observar os z-scores não está disponível na biblioteca da maneira apresentada na figura, mas foi identificada esta necessidade para compreender melhor os resultados. Neste indicador, não está sendo analisada a relação entre as dimensionalidades, como acontece na OAM, mas serve como uma ferramenta adicional para entender a relevância de cada variável.

5.2.2 Avaliação do campus UTFPR-ct com os campi das outras universidades

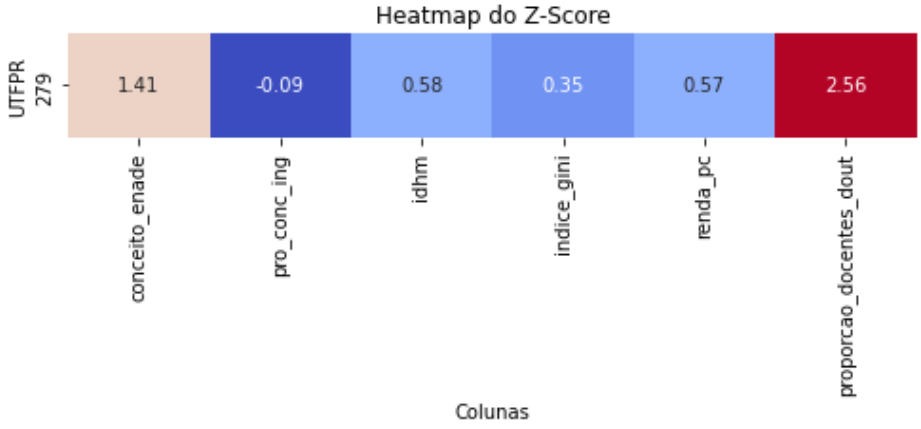
Nesta etapa da análise, foi aplicado o algoritmo de OAM nos campi das IES brasileiras, buscando em quais subespaços o Campus de Curitiba da UTFPR se destaca mais. O *heatmap* da análise utilizando o Conceito ENADE como variável de qualidade está disponível na Figura 18. Neste caso, o subespaço com melhor *Score* foi o de conceito ENADE com a proporção de docentes com doutorado, que obteve um *Score* de 5.3.

Figura 16 – Resultado da aplicação da OAM para média da UTFPR vs outras universidades, utilizando Conceito ENADE e Proporção de Concluintes por Ingressantes como variáveis de qualidade.



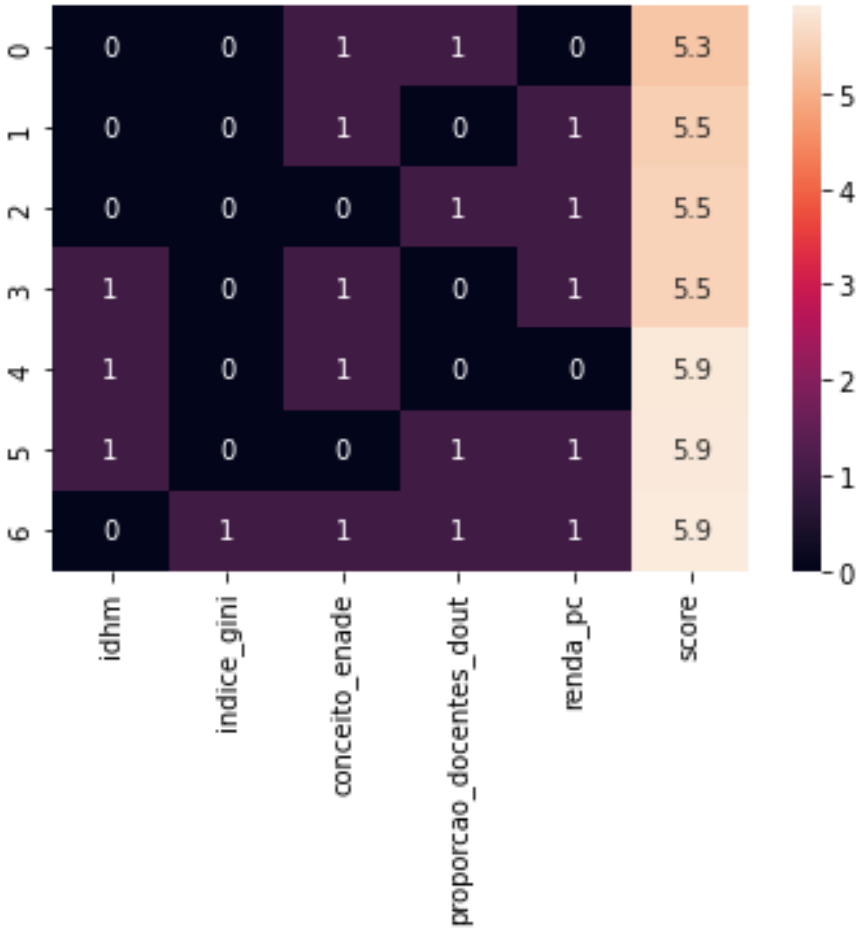
Fonte: Autoria Própria (2023)

Figura 17 – Z-score das variáveis utilizadas para média da UTFPR vs outras universidades.



Fonte: Autoria Própria (2023)

Figura 18 – Resultado da aplicação da OAM para o Campus UTFPR-CT, utilizando Conceito ENADE como variável de qualidade.



Fonte: Autoria Própria (2023)

Na Figura 19, está o resultado do algoritmo para proporção de concluintes por ingressantes como variável de qualidade, cujo subespaço contendo esta variável com melhor *Score* foi o que contem a proporção de docentes com doutorado e a renda per capita, com 6.3.

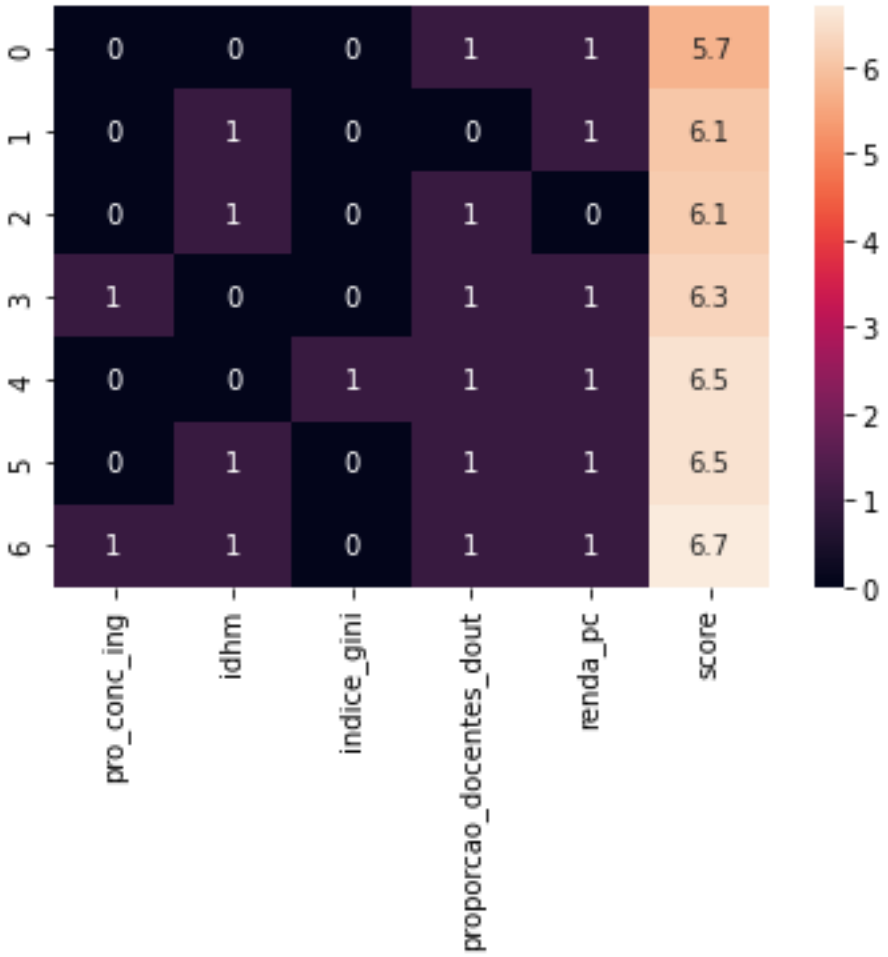
Agora utilizando as duas variáveis nas dimensões do algoritmo, o resultado obtido foi o da Figura 20, com o subespaço do conceito ENADE pela renda per capita em primeiro lugar com um *Score* de 5.5.

Finalmente, foi calculado o z-score do Campus de Curitiba para cada variável, cujo resultado está exibido na Figura 21.

5.3 Interpretação dos Resultados

Ao utilizar os algoritmos de Explicabilidade de *Outliers* da biblioteca de OAM, foi possível entender o conjunto de dados das IES brasileiras a partir de uma perspectiva diferente da observada na análise exploratória.

Figura 19 – Resultado da aplicação da OAM para o Campus UTFPR-CT, utilizando Proporção de Concluintes por Ingressantes como variável de qualidade.

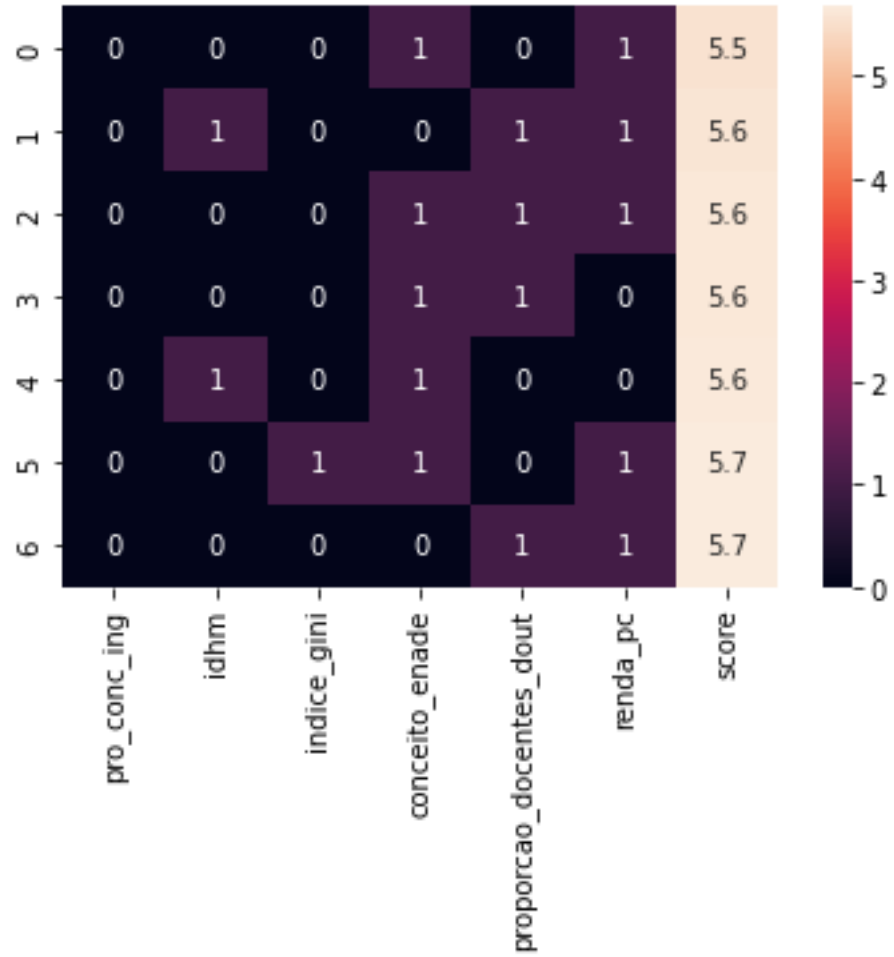


Fonte: Autoria Própria (2023)

Na primeira etapa desta análise, quando foi comparada a UTFPR às médias das outras IES do Brasil, um campo que estava presente nos sete subespaços com melhor *score* de todas as análises feitas na primeira etapa foi a proporção de docentes com doutorado. Como exibido na Figura 17, a universidade possui um *z-score* de 2.56 para este campo, o que indica que esta informação por si só é um grande fator de destaque para a IES, pois indica que a porcentagem de professores doutores, que é de 75.80%, está acima de 99.38% dos dados que foram avaliados.

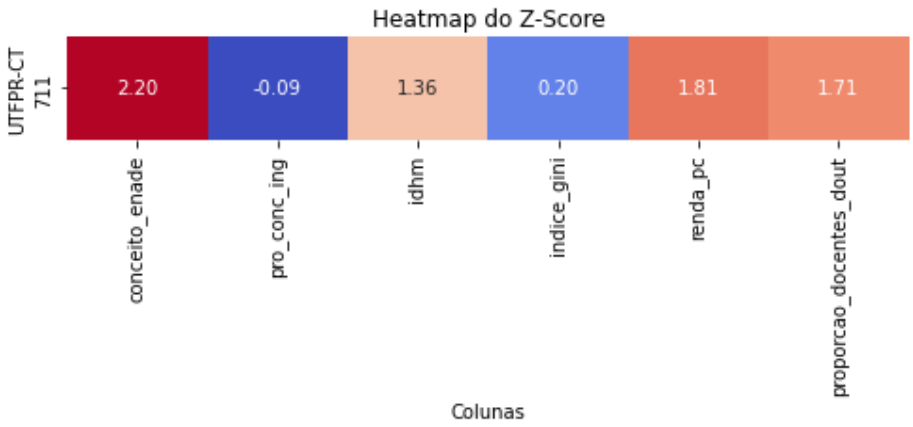
Quando foi utilizado o Conceito ENADE como variável de qualidade, este campo apareceu três vezes entre os sete primeiros subespaços com melhor *score*. Nos três subespaços, a proporção de docentes com doutorado também estava presente; a Renda Per Capita aparece duas vezes; e o IDHM em um subespaço. De fato, é de se esperar que estes sejam valores que influenciem a ter um *score* alto, pois possuem um *z-score* de 0.58 e 0.57 respectivamente. Porém, a maior vantagem de utilizar OAM é a possibilidade de observar as diferentes discrepâncias da observação em espaços multidimensionais. Uma possível interpretação para esta

Figura 20 – Resultado da aplicação da OAM para o Campus UTFPR-CT, utilizando Conceito ENADE e Proporção de Concluintes por Ingressantes como variáveis de qualidade.



Fonte: Autoria Própria (2023)

Figura 21 – Z-score das variáveis utilizadas para UTFPR-CT.



Fonte: Autoria Própria (2023)

análise é de que, de fato o Conceito ENADE da universidade é um destaque para ela, e este destaque não pode ser explicado puramente por ser uma universidade presente em cidades que se possuem um IDHM ou Renda Per Capita altos, uma vez que o *score* dos subespaços que incluem estes campos é muito parecido com o subespaço contendo apenas o Conceito ENADE e proporção de professores doutores e, talvez um fator que explique melhor o alto desempenho da universidade seja a própria proporção de doutores.

A proporção de concluintes por ingressantes, da mesma forma que na análise exploratória, não foi particularmente um fator de destaque para a universidade nas análises com OAM, embora esteja abaixo da média com o z-score de 1.71.

Os resultados obtidos ao utilizar o campus de Curitiba como observação nas análises foram similares aos resultados de quando foi utilizada toda a universidade. Porém, nesta etapa o conceito ENADE se destacou ainda mais, de forma similar ao resultado obtido na análise exploratória. Como apresentado na Figura 20, o subespaço em que UTFPR-CT mais se destacou foi o que contém o Conceito ENADE e a Renda Per Capita, com um *score* de 5.5. Neste caso, uma possível interpretação é de que, como o Conceito ENADE aparece em cinco dos sete subespaços com melhor *score* (mais vezes que a proporção de docentes com doutorado ou a Renda Per Capita), é de fato o principal fator de destaque para o campus, e a proporção de concluintes por ingressantes é uma medida que não é muito conclusiva para tal análise.

6 CONCLUSÃO

Ao utilizar técnicas de OAM no contexto deste trabalho, a principal utilidade foi a de aprofundar as interpretações obtidas na análise exploratória. Os resultados obtidos em ambas as técnicas apontam que o principal fator de destaque da UTFPR é o Conceito ENDADE e, o campo do conjunto de dados utilizado que melhor explica isto é a proporção de professores com doutorado da universidade.

Uma das principais vantagens observadas de utilizar OAM foi a forma automática dos algoritmos ao comparar diversas dimensões, o que pode ser muito útil para entender rapidamente quais campos podem ser mais importantes em um conjunto de dados para um contexto como este. Utilizar apenas a OAM traria resultados inconclusivos e provavelmente não seria a melhor opção. Porém, ao conciliar as duas formas de analisar, é possível obter resultados mais precisos e conseguir argumentos melhores para sustentar os resultados obtidos. Uma sugestão de melhoria observada para a biblioteca é a inclusão dos *heatmaps* dos z-scores das variáveis, como apresentado nas Figuras 17 e 21.

Assim, em trabalhos futuros com um contexto similar, uma boa ideia seria a de incluir análises com os algoritmos de OAM juntamente na etapa inicial da análise exploratória, a fim de mapear diferentes conjuntos de dimensões que expliquem de forma rápida em quais contextos a IES se destaca.

REFERÊNCIAS

BERTOLIN, J. C. G.; MARCON, T. O (des)entendimento de qualidade na educação superior brasileira – das quimeras do provão e do enade à realidade do capital cultural dos estudantes. **Revista da Avaliação da Educação Superior (Campinas)**, 2015.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Comput. Surv.**, v. 41, 07 2009.

FARIA, R.; COLLI, T. **BIBLIOTECA EM PYTHON PARA EXPLICABILIDADE DE ANOMALIAS**. 2021 — UTFPR, 2021.

GRIBOSKI, C. M. O enade como indutor da qualidade da educação superior. **Estudos em Avaliação Educacional**, v. 23, n. 53, p. 178–195, dez. 2012. Disponível em: https://publicacoes.fcc.org.br/eae/article/view/1920.

NORA, R. D. **ANáLISE DA RELAçãO ENTRE OS INDICADORES DE DESEMPENHO DAS UNIVERSIDADES FEDERAIS DA REGIãO SUL DO BRASIL E OS RESULTADOS OBTIDOS NO íNDICE GERAL DE CURSOS (IGC)**. 2014 — UFRGS, 2014.

SAMARIYA, D. *et al.* A new effective and efficient measure for outlying aspect mining. *In*: SPRINGER. **International Conference on Web Information Systems Engineering**. [*S.l.*], 2020. p. 463–474.

VINH, N. X. *et al.* Discovering outlying aspects in large datasets. **Data Mining and Knowledge Discovery**, v. 30, n. 6, p. 1520–1555, Nov 2016. ISSN 1573-756X. Disponível em: https://doi.org/10.1007/s10618-016-0453-2.