CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Detecção de *Outliers* em Finanças de Instituições de Ensino Superior Brasileiras Utilizando Aprendizado de Máquina Não Supervisionado

Nathan C. Freitas, Roberta M. M. Gouveia, Ebony M. Rodrigues, Gabriel Alves, Maria da Conceição M. Batista, Rodrigo Lins Rodrigues

Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

{nathan.freitas, roberta.gouveia, ebony.marquesr, gabriel.alves, maria.cmbatista, rodrigo.linsrodrigues}@ufrpe.br

Abstract. *Educational Data Mining (EDM) is widely used as a tool to gain a better understanding of nuances in the teaching and learning process. In higher education, it is possible to analyze the influence of the financial amounts invested by the institutions, since the presence of related tools can help managers to reduce expenses in their institutions. This work deals with the detection of institutions that present discrepant financial value of expenses (outliers). For this, clusterings are performed in order to create cohesive groups and unsupervised outlier detection methods are applied. The results demonstrate the creation of groups that have outlier institutions with similar profiles. The method provides a means of investigating which institutions are discrepant in terms of their declared expenses, presenting a study on these instances.*

Resumo. *A mineração de dados educacionais (EDM) é amplamente utilizada como uma ferramenta para se obter uma melhor compreensão sobre as nuances existentes no processo de ensino e aprendizagem. Na educação superior, é possível analisar a influência dos valores financeiros investidos pelas instituições, pois a presença de ferramentas relacionadas podem auxiliar os gestores a diminuir gastos nas suas instituições. Este trabalho trata da detecção de instituições que apresentam valores financeiros de despesas discrepantes (*outliers*). Para isso, são realizados agrupamentos visando à criação de grupos coesos e são aplicados métodos não supervisionados de detecção de* outliers*. Os resultados demonstram a criação de grupos que possuem instituições* outliers *com perfis semelhantes. O método proporciona um meio de investigar quais instituições são discrepantes quanto às suas despesas declaradas, apresentando um estudo sobre essas instâncias.*

1. Introdução

A publicação de bases de dados por entes públicos e privados possibilita a realização de estudos em diversos contextos. Compreendendo o emprego de métodos da Estatística e da Ciência da Computação, a Mineração de Dados é uma abordagem comumente usada, uma vez que visa à descoberta de conhecimento por meio da identificação de padrões e tendências relevantes em um conjunto de dados. Em especial, no âmbito da educação,

a Mineração de Dados Educacionais (*Educational Data Mining* – EDM) é utilizada para o aprimoramento de processos de ensino e aprendizagem [Romero and Ventura 2013]. Entre as técnicas amplamente observadas para a mineração de dados, destacam-se os métodos de Aprendizado de Máquina [Costa et al. 2012, Baker et al. 2010].

Considerando as Instituições de Ensino Superior (IES) brasileiras, pode-se usar a Mineração de Dados com o propósito de verificar a influência de determinados fatores no desempenho de discentes. Diante disso, é plausível observar os dados financeiros de IES visando ao entendimento de possíveis diferenças entre as suas despesas e receitas, tendo em mente que a gestão de finanças é relevante para o auxílio à tomada de decisão de uma organização [Freitas et al. 2022]. O investimento realizado pelas IES pode ser objeto de análises sobre o usufruto de ambientes e equipamentos por seus estudantes [Vasconcelos et al. 2021].

Sabe-se que as IES podem ter características diversas que explicam as diferenças de investimento. Entretanto, é possível que instituições com características semelhantes tenham finanças discrepantes. A análise de *outliers* (anomalias ou valores discrepantes) apresenta-se como instrumento para verificar se os recursos estão sendo bem aplicados quando se comparam as variáveis de IES com características semelhantes, o que é de interesse para IES públicas e privadas, pois, no âmbito público, a análise pode alertar sobre possíveis diferenças entre IES públicas de perfis semelhantes, e, no âmbito privado, a análise pode basear a criação de vantagem competitiva, por exemplo.

Este estudo propõe um método de detecção de *outliers* em dados financeiros de IES brasileiras públicas e privadas – em especial, em dados sobre despesas – visando ao aprimoramento das análises que podem ser feitas por gestores educacionais. Para isso, foram aplicados algoritmos de Aprendizado de Máquina Não Supervisionado e técnicas de detecção de *outliers* sobre as bases de dados das edições de 2016 a 2019 do Censo da Educação Superior, que são publicadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). A motivação deste trabalho baseia-se na hipótese de que IES com características semelhantes – como quantidades de discentes, docentes e cursos – podem apresentar despesas discrepantes.

2. Trabalhos Relacionados

As técnicas de mineração de dados educacionais podem ser aplicadas em áreas diversas [Costa et al. 2012, Baker et al. 2010]. No contexto educacional, pode-se construir modelos baseados em dados de estudantes para analisar o seu comportamento, além de fatores sociais e econômicos a fim de melhorar o desempenho [do Nascimento et al. 2021]. Além disso, pode-se tratar de suporte pedagógico com a descoberta e confirmação de teorias científicas educacionais [Costa et al. 2012].

A realização de agrupamentos é apropriada quando se pretende realizar análises de desempenho. No trabalho de [Spanol et al. 2022], propõe-se uma abordagem para auxiliar a análise e a comparação de características de municípios semelhantes com base em indicadores socioeconômicos. A presença de *outliers* demonstrou um menor desempenho para o Coeficiente de Silhueta nos grupos, o que foi necessário tratar. Outra abordagem que utilizou agrupamentos observou avaliações educacionais como o ENADE, em que foram verificadas características relevantes dos grupos, baseando-se no desempenho dos alunos [da Silva Vieira et al. 2022].

As pesquisas relacionadas a análises de despesas públicas se tornaram notórias, utilizando a literatura de detecção de *outliers*. Em uma abordagem univariada, é muito usado o método de quartil, que demonstra bons resultados aplicados a licitações públicas [de Siqueira Gê and Borges 2021], porém é importante levar em consideração os diferentes cenários de instituições, como região geográfica, gestão administrativa e até a distribuição da amostragem dos dados [Freitas et al. 2022]. Com isso, uma análise mais detalhada de *outliers* multivariados pode ser pertinente para considerar as diferentes situações das entidades envolvidas.

Uma análise multivariada de *outliers* leva em consideração uma busca no espaço vetorial sobre as características de uma instância. Essa abordagem foi aplicada a diversas bases por [Goldstein and Uchida 2016], em que foram testados vários classificadores, identificando, assim, as instâncias discrepantes, seus tempos de execução e valores AUC (*Area Under The Curve*). As técnicas podem ser empregadas em diversos contextos, como detecção de fraudes, cibersegurança, *fake news*, transações financeiras, qualidade de dados, entre outros [Goldstein and Uchida 2016][Wang et al. 2019].

No contexto educacional, poucos trabalhos exploraram a aplicação de *outliers* em valores financeiros. O estudo dessas variáveis proporciona uma forma de monitoramento, observando quanto está sendo recebido ou gasto dentro das instituições e que pontos a má gestão desse valores influencia no desempenho da IES e, consequentemente, do aluno. O trabalho [Machado 2022] trata de agrupamento de munícipios com a finalidade de aplicar modelos multivariados de detecção de gastos públicos considerados *outliers* no ensino fundamental. Com um escalonamento dos dados, pode-se identificar possíveis irregularidades existentes nos gastos públicos da educação. O método possui bastantes pontos a serem explorados, como aplicar em outros tipos de bases de dados, busca de acurácia nos métodos não supervisionados, utilização de outras técnicas de detecção, entre outros.

Diante do exposto, o presente trabalho objetiva a aplicação da busca de *outliers* em relação às IES que constam no Censo da Educação Superior. Para isso, utilizam-se agrupamentos de instituições de forma a criar grupos coesos de características, para serem aplicados modelos de detecção de *outliers* multivariados. O método proposto tem o intuito de servir como modelo a ser empregado por gestores de IES a fim de detectar possíveis discrepâncias entre os dados financeiros das instituições.

3. Método

Nesta seção, é apresentado o método empregado neste trabalho, que inicialmente considerou as primeiras etapas do processo de *Knowledge Discovery in Databases* (KDD), com as tarefas de seleção, pré-processamento e transformação de dados [Fayyad et al. 1996]. Após essas etapas, aplicaram-se métodos de agrupamento e de detecção multivariada de *outliers*, abordando as despesas das instituições. Os grupos resultantes foram analisados para a avaliação e interpretação dos resultados.

3.1. Preparação dos Dados

O Censo da Educação Superior, realizado anualmente, caracteriza-se como um instrumento de pesquisa sobre as IES brasileiras. Por meio dele, pode-se obter dados sobre a infraestrutura das IES, despesas e receitas, alunos, cursos e docentes, o que possibilita a análise de variáveis das bases existentes em relação às instituições. A partir dos dados de 2016 a 2019, foram observadas duas abordagens de dados para o experimento:

- **Base reduzida (21 variáveis):** conjunto de dados que contém informações das IES, como categoria administrativa, região, presença de repositório acadêmico, catálogo online, organização acadêmica,etc. Para cada IES foi capturado os valores numéricos da quantidade total de técnicos, cursos, docentes e alunos.
- **Base detalhada (66 variáveis):** possui os mesmos atributos da IES em relação a base reduzida, entretanto os valores numéricos são mais segmentados, como a quantidades de docentes por raça, escolaridade e atuação, cursos por turno e modalidade, técnicos por escolaridade, etc. Em relação a base do aluno, foi levada em consideração apenas a sua quantidade total.

As instituições mantidas pela mantenedora com mais de uma instituição vinculada foram desconsideradas neste estudo, pois seus valores financeiros são replicados, o que não condiz com o valor real da instituição [INEP 2023]. Quanto às variáveis nulas ou inconsistentes, foi utilizado o método de substituição [Hair et al. 2009], que substitui o valor por um conhecido na amostra do atributo ou por medidas de posição (média, mediana e moda). No total, há 7 atributos financeiros de despesas nas IES que foram somados, uma vez que certas instituições possuem os valores distribuídos apenas em algumas das variáveis presentes [Freitas et al. 2022].

As distribuições das variáveis são fatores significativos para a detecção, pois os algoritmos de detecção de *outliers* tendem a ter um melhor desempenho diante de distribuições simétricas. Foi utilizada a *Power Transformation*, que é considerada uma família de funções que aplica transformação monotônica nos dados. Essa transformação é utilizada para estabilizar a variância e tornar a distribuição mais próxima da normal. A Figura 1 demostra o resultado da transformação da variável assimétrica de despesas nas instituições. Após a transformação dos dados quantitativos, foi aplicada a técnica de binarização dos atributos qualitativos por meio do método *One-Hot Encoding*.

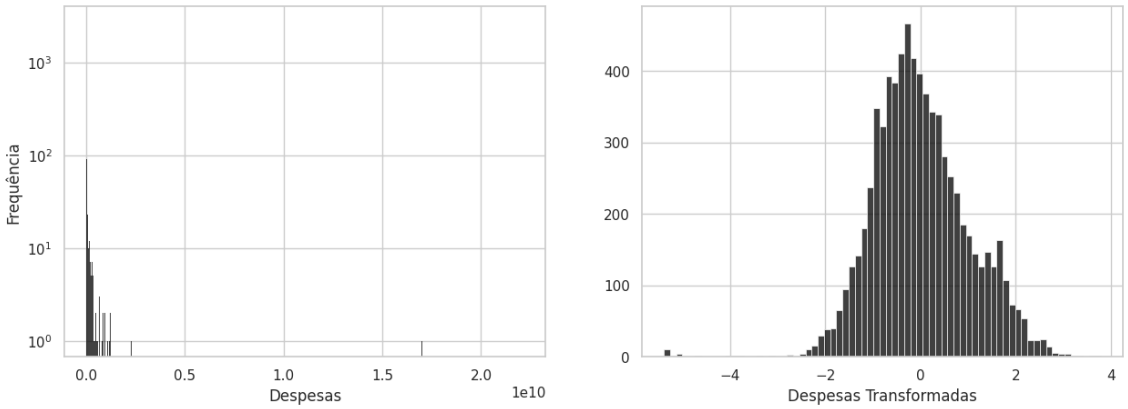

Figura 1. Power transformation em relação às despesas.

3.2. Agrupamento e Métricas

O agrupamento foi realizado considerando todas as variáveis e registros das bases (reduzida e detalhada), que foram devidamente tratados. A realização dessa etapa tem o propósito de criar grupos coesos em diferentes contextos, como a região que a IES está inserida, quantidade de alunos e docentes. Foram utilizados dois métodos de agrupamento:

K-Means: um algoritmo de particionamento em que é necessário definir o número de agrupamentos a serem gerados (k). O objetivo do algoritmo é encontrar centroides, que são definidos como o ponto médio de um determinado grupo; o algoritmo calcula a posição dos centroides de maneira que os grupos possuam características distintas.

Hierárquico Aglomerativo: algoritmo em que inicialmente cada instância é considerada como um grupo individual e gradualmente são formados grupos de forma que, no fim, haverá apenas um único agrupamento. Dendrogramas são bastante utilizados para observação de árvores de relacionamentos.

Para definir o número ideal de grupos em agrupamentos com K-Means, costuma-se utilizar o Método do Cotovelo (*Elbow Method*) [da Silva Vieira et al. 2022]. Procurou-se definir um número ideal de grupos de forma que eles sejam totalmente diferentes entre si e que os seus elementos sejam o mais homogêneos quanto for possível. Para algoritmos hierárquicos, a análise de grupos é realizada graficamente por meio do dendograma.

Este trabalho proporcionou a experimentação de duas bases diferentes que são empregadas, cada uma, em dois algoritmos de agrupamento, contendo, assim, quatro possibilidades de segmentos. Diante disso, foi determinada a escolha de um critério para selecionar o melhor método agrupador, além da base a ser empregada na detecção. Com essa finalidade, tem-se em consideração a criação de agrupamentos que sejam mais distantes (distância inter-*cluster*) entre si e que a distância entre as instâncias dentro de um mesmo agrupamento seja a menor possível (distância intra-*cluster*). Logo, foram empregados os índices de Dunn e Davies-Bouldin, que atendem a esse requerimento.

Os índices possuem intervalos diferentes para lidar com a eficiência do agrupamento. Portanto, foi abordada a estratégia de normalizar os resultados dos índices usando o método *Min-max* para obter a média dos índices dos agrupamentos I (K-Means) e II (Hierárquico Aglomerativo) nas bases reduzida e detalhada. A combinação de base e agrupamento com a maior média das métricas foi escolhida para a detecção de *outliers*.

3.3. Detecção de *Outliers* das Instituições

Ao selecionar a melhor base com seu respectivo algoritmo de agrupamento, realizou-se a detecção de *outliers*, com duas perspectivas a serem verificadas. Na primeira, há a busca de *clusters* (grupos) *outliers*, que podem ser considerados como grupos em que todas as instâncias são atípicas. Para isso, são procurados grupos que estão bastante distantes e com menor número de instâncias, devido ao pressuposto de os valores atípicos aparecerem em número pequeno e disperso em comparação com a quantidade total dos dados. Na segunda abordagem, há a busca de instâncias anômalas, em que, após o agrupamento, são verificadas as instâncias que estão muito distantes em cada grupo.

Para o método de busca de instâncias anômalas, podem ser utilizados algoritmos de detecção multivariada de *outliers*. A biblioteca de Python PyOD proporciona uma variedade de algoritmos para a detecção de *outliers* [Zhao et al. 2019]. Entretanto, diante da literatura existente, foram selecionados para uso neste trabalho os seguintes algoritmos, presentes em estudos no âmbito educacional [Machado 2022]: *Angle-based Outlier Detection, Feature Bagging, Isolation Forest, K-Nearest Neighbors, Local Outlier Factor e One Class SVM*.

Para esses modelos, há um parâmetro que representa a porcentagem de *outliers* na amostra, que, de acordo com a premissa de presença de *outliers* em uma distribuição normal, foi definido como 5%. Devido à presença de vários modelos não supervisionados, considerou-se o grau de anormalidade (grau de qualidade dos *outliers*) de uma instituição, que representa a quantidade de algoritmos que detectam positivo para *outlier*. Quanto maior o grau, maior a probabilidade de se tratar de um verdadeiro *outlier*. Foi determinado, em vista disso, o grau de qualidade maior ou igual a 3, ou seja, que sejam detectados em 3 algoritmos selecionados do total de 6.

4. Resultados e Discussão

Nesta seção, são apresentados os resultados da execução dos agrupamentos K-Means e Aglomerativo, que foram selecionados para serem considerados nos modelos de detecção multivariada de *outliers*. Após isso, são analisadas características dos *outliers* detectados.

Inicialmente buscou-se o número ideal de grupos. O método do cotovelo foi executado nas duas bases (reduzida e detalhada) para encontrar o número ideal (k) de grupos a serem gerados pelo algoritmo K-means. A quantidade ideal para agrupamento com o algoritmo Hierárquico Aglomerativo foi definida com a visualização do dendrograma (Figura 2). As quantidades de grupos podem ser observadas na Tabela 1.

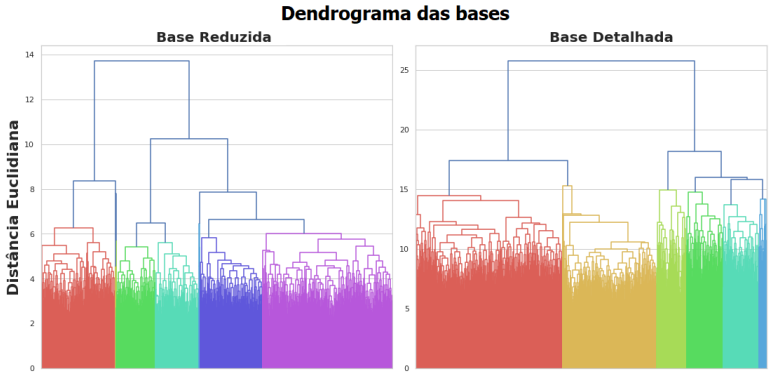

Figura 2. Dendograma de ligação completa aplicado à base reduzida e detalhada dos valores financeiros de despesas das IES.

Tabela 1. Quantidade ideal de grupos.

Base	Algoritmo	Número de grupos
Reduzida	K-means	5
	Aglomerativo	8
Detalhada	K-means	5
	Aglomerativo	6

A quantidade de grupos no método aglomerativo foi definida de acordo com a escolha de um ponto em que a distância euclidiana (distância vertical do gráfico) começou a segmentar de forma muito específica. Como outro critério, considerou-se que o tamanho de um único grupo não fosse superior a 50% da amostra. Para o cálculo de agrupamento, foi empregada a medida de proximidade de ligação completa, que representa o método de cálculo das distâncias dos grupos, que, no caso da base completa, define-se como a maior distância entre dois pontos em cada grupo.

Para selecionar o melhor algoritmo e base de dados, utilizou-se a média normalizada com o método *Min-max* dos índices de Dunn e Davies-Bouldin, que são apresentados na Tabela 2. Percebe-se que a base reduzida com algoritmo Aglomerativo conseguiu grupos mais coesos. Na Figura 3, são apresentados os grupos em relação às despesas e as variáveis numéricas de quantidade de alunos e docentes. A base detalhada teve um desempenho pior nos índices quando comparada à reduzida. A justificativa disso pode ser dada pela incidência de muitos atributos para geração dos grupos, o que pode comprometer o desempenho do modelo [Han et al. 2022].

Tabela 2. Valores e média normalizada das métricas de agrupamento por base de dados.

Algoritmo	Base de dados	Índice Dunn	DBI	Média normalizada
Reduzida	K-means	0.02785	2.64050	0.27335
	Aglomerativo	**0.07997**	**2.34618**	**0.65116**
Detalhada	K-means	0.10101	2.43025	0.64746
	Aglomerativo	0.20024	2.99546	0.50000

Nos gráficos da Figura 3, existem três grupos que possuem poucas instâncias e estão afastados dos demais, havendo pouca sobreposição. Esses grupos podem ser caracterizados como *clusters outliers*. O primeiro, o quinto e o sexto grupos têm, respectivamente, 23, 11 e 12 instâncias.

Figura 3. Grupos do agrupamento Aglomerativo pelos valores numéricos de alunos e docentes em relação às despesas da base reduzida.

O Primeiro Grupo é formado por instituições privadas que possuem despesas com valor muito acima dos gastos comuns das instituições. Algumas podem ser consideradas como possível erro de preenchimento de dados, dado o fato de serem despesas muito altas, principalmente quando comparadas a anos anteriores e posteriores. Um exemplo disso é a Faculdade São Francisco de Juazeiro, que, no ano de 2017, possui uma despesa de R$3.417.569.524,08. Em comparação aos outros anos (2016, 2018 e 2019), esse valor é, em média, 383 vezes maior. Esse grupo se destaca por ter instituições com baixas quantidades de alunos, docentes, cursos e técnicos e com alto valor de despesa.

O Quinto e o Sexto grupos são formados em sua maioria por instituições privadas com o valor de despesa muito baixo, tendo como valor máximo R$22.775,59, que é da instituição extinta Faculdade de Tecnologia Senai Belo Horizonte. A maioria das instituições está extinta, com exceção da Faculdade Einstein, Instituto Infnet Rio de Janeiro e o Instituto de Educação Superior Presidente Kennedy. As instituições extintas possuíam propriedades que, no decorrer dos anos, iriam entrar nesse estado, visto o seu baixo valor de despesas, alunos, docentes, cursos e técnicos. Porém, as instituições não extintas são um alerta a ser levado em consideração, pois seu maior valor é de R$7,00. Assim, pode-se concluir que esses grupos podem ser considerados *outliers*.

Aplicando os algoritmos de detecção de *outliers* nos demais grupos (Segundo, Terceiro, Quarto, Sétimo e Oitavo), têm-se diferentes instâncias consideradas como discrepantes, de acordo com o grau de qualidade de *outliers*, como pode ser visto na Tabela 3. Observando as quantidades da tabela, foi somada a quantidade dos *clusters outliers*, com 46 instâncias. Deve-se levar em consideração que uma instituição pode ser representada por mais de uma instância, apresentando, para tal, anos distintos de observação.

Tabela 3. Quantidade de instâncias *outliers* por grau de anormalidade e porcentagem da quantidade total de instâncias analisadas.

Grau de anormalidade	Quantidade de instâncias	Quantidade de IES	% com a quantidade total de instâncias
1	1224	742	16,44%
2	524	333	7,04%
3	**325**	**211**	**4,36%**
4	217	146	2,91%
5	135	92	1,81%
6	72	49	0,96%

O grau de qualidade de 3 é o que mais se aproxima da porcentagem de 5%, que representa a porcentagem de dados acima de 2 com desvio padrão da média de uma distribuição normal de dados [Morettin and Bussab 2017]. Na Figura 4, pode-se observar a quantidade de *outliers* de acordo com as variáveis quantitativas de alunos e docentes. A partir dela, é válido afirmar que o método utilizado tem o potencial de identificar os *outliers* que estavam às margens dos valores numéricos.

Analisando os resultados, nota-se que uma parte das IES se apresenta expressivamente com uma baixa quantidade de docentes, técnicos e cursos, o que pode ser visto na Figura 4(b) por meio da linha pontilhada vertical. Essas instituições são em parte pertencentes aos *clusters outliers*, porém as que não pertencem a esses grupos, mas que estavam próximas a eles, também foram consideradas *outliers*.

Com as instituições identificadas, é possível realizar uma análise separando a população detectada pelos quartis. As instituições abaixo do primeiro quartil são em sua maioria privadas, representam em parte os *clusters outliers* — o Quinto e o Sexto grupos — e possui o Segundo Grupo (56,79%) majoritário, com uma média de despesa de R$161.377,62. A identificação das instâncias discrepantes nesse grupo pode estar relacionada ao baixo número das variáveis numéricas. Outros grupos detectados foram o Terceiro e o Sétimo, que possuem uma instância detectada.

Os *outliers* identificados entre o primeiro quartil e a mediana são representados

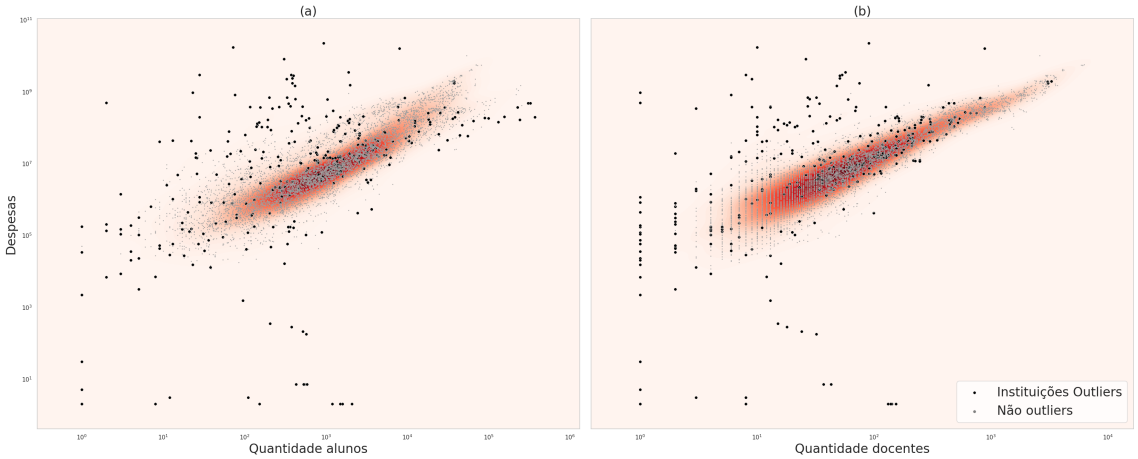

Figura 4. Instâncias detectadas em relação do valor das despesas aos valores numéricos de alunos e docentes.

por três grupos: Segundo, Terceiro (majoritário, com 62,96%) e Sétimo. Suas categorias administrativas têm a predominância das privadas, porém é possível observar algumas instituições públicas federais, estaduais e municipais. Os valores de despesas variam, em média, aproximadamente, de R$2.000.000,00 a R$9.000.000,00. Entretanto, é possível observar uma predominância das regiões Sul, Sudeste e Centro-Oeste dentro desse intervalo, em que os dois últimos possuem uma maior despesas comparadas a outras regiões.

Os detectados entre a mediana e o terceiro quartil foram os que tiveram maior variedade de grupos envolvidos, em que apenas não constam os dois *clusters outliers*, o Quinto e o Sexto grupo. Nesse intervalo, os grupos Segundo, Terceiro e Sétimo são os que apresentaram uma quantidade alta de despesa média, porém a quantidade média de alunos está muito abaixo das demais. Em sua maioria, são instituições privadas, mas foi possível identificar que as instituições estaduais possuem uma quantidade média de alunos muito acima das demais, com 16.199 estudantes.

Por fim, as IES que estão acima do terceiro quartil contêm a maioria das detecções do Primeiro Grupo, que apresenta a maior média de despesa. O Terceiro e o Sétimo grupo possuem um valor alto de despesa em comparação à quantidade média de alunos (abaixo de 1000) e cursos (abaixo de 3). O Quarto grupo apresenta a maior média sem ser considerado um *cluster outlier*, com quase 1 bilhão de reais em despesas em 27 instâncias detectadas. As regiões Nordeste, Sudeste e Sul possuíam o valor médio de despesas acima de 1 bilhão de reais, tendo a região Sul o maior número médio de estudantes, 112.484. A Tabela 4 apresenta um resumo geral em cada quartil analisado.

Diante disso, é possível observar que o método abordado tem capacidade de fornecer análises que possibilitam identificar possíveis relações dos valores financeiros com outras variáveis das instituições no Censo da Educação Superior. A partir disso, podem ser investigadas as instituições que pertencem a cada grupo como *outliers* e a influência de outros valores numéricos, como quantidade de alunos e docentes. Em relação às despesas nessas instâncias, pode-se observar as instituições públicas e analisar os motivos de haver valores discrepantes em IES semelhantes, entre outras análises.

5. Considerações Finais

O presente trabalho demonstrou um método de análise de despesas de IES no Brasil. Foram utilizados algoritmos de agrupamento aplicados à detecção multivariada de *outliers*. Com as análises, foi possível examinar instituições que podem ser consideradas *outliers*, seja por erro de digitação ou devido a características externas, como a extinção da instituição. Com isso, é possível elaborar estudos sobre os motivos que as levaram a serem *outliers*, aplicando medidas que mantenham a instituição nesse estado, caso essa característica seja positiva, ou para mitigar os motivos, caso a característica seja negativa.

Tabela 4. Resumo das características em cada quartil dos *outliers* detectados.

Intervalo Quartil	Agrupamentos presentes (ordinal)	Características de *outliers*
Abaixo de Q1	2°, 3°, 5°, 6°, 7°	• Instituições, em sua maioria, privadas; • Presença dos *clusters outliers*; • Despesas muito abaixo quando comparadas a valores de alunos, docentes, curso;
Entre Q1 e Q2	2°, 3°, 7°	• Em sua maioria, instituições privadas; • Despesas médias das regiões Norte e Nordeste são menores comparadas às regiões mais ao sul;
Entre Q2 e Q3	1°, 2°, 3°, 4°,7°, 8°	• Variedades de grupos; • Grupo 1°, 2° e 3° com despesas mais altas; • Instituição Estadual com número alto de alunos;
Acima de Q4	1°, 3°, 4°, 7°, 8°	• Presença de *cluster outlier*(1°); • O 7° e 3° valores baixos de alunos e cursos; • Regiões com a média de despesas na casa dos bilhões; • Região Sul com alta quantidade média de alunos;

A gestão financeira em uma IES é imprescindível para sua sobrevivência. Uma boa gestão pode ser alcançada tendo uma integração completa de todas as áreas da instituição de forma harmônica [Colombo 2014]. Além disso, a análise externa da instituição também é um dos pilares essenciais para a análise, principalmente para considerar fatores como possíveis crises que possam estar acontecendo no país, mudança de preferência da população sobre cursos superiores, aumento da evasão, entre outros.

O método apresentado não visa medir a acurácia dos algoritmos não supervisionados, mas busca um aumento de probabilidade de identificar verdadeiros positivos por meio da consideração de *outliers* de instituições identificadas em 3 algoritmos distintos. Diante do exposto, é possível realizar trabalhos futuros com vistas a outros métodos de detecção de *outliers*. Executar um método de seleção de atributos na base detalhada seria um meio de evitar a maldição da dimensionalidade, que pode ter causado a criação de grupos com qualidade inferior.

Os dados presentes em bases abertas, como o Censo da Educação Superior, proporcionam uma forma de enxergar as características das IES *outliers* e refletir como atributos quantitativos e qualitativos podem influenciar os gastos da instituição. Para instituições privadas, também deve-se considerar o potencial de aumentar as receitas. No âmbito público, a administração de valores de financeiros de uma IES significa um uso eficiente do dinheiro público e a possibilidade de maiores investimentos em educação, principalmente no que tange à busca por equidade em fatores regionais e sociais. Portanto, a identificação de *outliers* fornece ao gestor uma forma de definir estratégias para melhorar o ensino e para criação de oportunidades de investimento e retorno financeiro.

Referências

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Colombo, S. S. (2014). *Gestão universitária: os caminhos para a excelência*. Penso Editora.

Costa, E., Baker, R. S., Amorim, L., Magalhães, J., and Marinho, T. (2012). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.

da Silva Vieira, A., Bertolini, D., and Schwerz, A. L. (2022). Análise do desempenho no enade dos concluintes de computaçao usando técnica de agrupamento. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 834–845. SBC.

de Siqueira Gê, E. A. and Borges, E. F. (2021). Identificação de outliers em processos de dispensas e inexigibilidades em licitações públicas: um estudo comparativo entre ufrn, ifrn e ufersa nos anos de 2017 e 2018. *Revista Inovar Contábil*, 2(01).

do Nascimento, P. S. C., da Silva Junior, A. S., Schulz, C. L., dos Santos, M. V. R., Maciel, A. M. A., Rodrigues, R. L., do Nascimento, R. R., and Alencar, F. M. R. (2021). Análise dos impactos da gestão do tempo no desempenho acadêmico através da mineração de dados educacionais. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 783–791. SBC.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Freitas, N. C., Gouveia, R. M., de Albuquerque Júnior, G. A., Maria da Conceição, M. B., and Rodrigues, R. L. (2022). A implementation mathematical-computational method for the detection and treatment of financial outliers in higher education. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 740–751. SBC.

Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman editora.

Han, J., Pei, J., and Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

INEP (2023). Resumo técnico do censo da educação superior.

Machado, R. G. (2022). Subsídio às fiscalizações públicas: Identificação dos municípios com gastos discrepantes na educação básica. *Caderno de Finanças Públicas*, 22(01).

Morettin, P. A. and Bussab, W. O. (2017). *Estatística básica*. Saraiva Educação SA.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, 3(1):12–27.

Spanol, M., Oliveira, E., Alves, G., Bittencourt, I. M., Falcao, T. P., and Mello, R. F. (2022). Uso de agrupamento para avaliação de desempenho educacional e apoio à

gestão em áreas de investimento. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 944–955. SBC.

Vasconcelos, J. C., Lima, P. V. P. S., Rocha, L. A., and Khan, A. S. (2021). Infraestrutura escolar e investimentos públicos em educação no brasil: a importância para o desempenho educacional. *Ensaio: Avaliação e Políticas Públicas em Educação*, 29(113):874–898.

Wang, H., Bah, M. J., and Hammad, M. (2019). Progress in outlier detection techniques: A survey. *Ieee Access*, 7:107964–108000.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*.