RELATÓRIO TÉCNICO - IFSP

Título: Cursos IFSP

Nome: Carolina dos Anjos Figueiredo

Prontuário: GU3015475

Disciplina: Tópicos Especiais II (TE2D6)

INTRODUÇÃO E JUSTIFICATIVA

Introdução:

O relatório técnico apresenta uma análise dos cursos oferecidos pelo Instituto Federal de Educação, Ciência e Tecnologia de São Paulo e possui informações sobre os cursos oferecidos por todo o Estado de São Paulo, a quantidade de vagas, a carga horária dos cursos, as notas do ENADE e os valores do CPC e CC.

O objetivo principal deste estudo é fornecer insights para compreender o desempenho dos cursos do IFSP, pois através da análise dos dados é possível saber se os cursos estão com boas notas e se ao longo do tempo estão melhorando suas avaliações ou não. Está análise pode servir de base para pesquisas de falhas e pontos a melhorar e visam compreender o motivo destas notas e o que pode ser feito para melhorar.

O desenvolvimento deste relatório surgiu da preocupação com o desempenho dos cursos oferecidos nos municípios de São Paulo, desejando visualizar através das métricas se são boas ofertas, se são cursos de qualidade.

A fonte de dados utilizada neste estudo é o arquivo “Cursos IFSP” disponível no Portal Brasileiro de Dados Abertos, no site dados.gov. O período a ser analisado é o primeiro semestre 2023, com sua última data de atualização em março de 2023. A divulgação transparente dos resultados deste estudo promove a prestação de contas à comunidade acadêmica, aos órgãos reguladores e à sociedade em geral. Isso reforça o compromisso do IFSP com a excelência acadêmica, a transparência e a responsabilidade institucional.

Os dados possuem informações sobre instituições de ensino e cursos, incluindo detalhes administrativos, características dos cursos (como carga horária e modalidade), informações sobre autorizações e reconhecimentos, e dados sobre o funcionamento das instituições. Essas informações são cruciais para entender a oferta educacional e o status das instituições de ensino.

Por meio dessa investigação, busca-se contribuir para a promoção de um ambiente educacional mais dinâmico, inovador e alinhado com as necessidades da sociedade, fortalecendo o papel do IFSP como agente transformador no cenário educacional brasileiro.

***Justificativa*:**

A análise do desempenho dos cursos oferecidos pelo Instituto Federal de São Paulo (IFSP) é de suma importância para compreender a eficácia e qualidade do ensino superior. Este estudo visa contribuir significativamente para o aprimoramento contínuo das práticas educacionais, identificando padrões, tendências e áreas de oportunidade que possam influenciar positivamente o desenvolvimento acadêmico dos estudantes.

Avaliar o desempenho dos cursos no IFSP é crucial para garantir a entrega de uma educação de qualidade. A compreensão dos fatores que impactam diretamente o rendimento acadêmico permitirá a identificação de estratégias eficazes para aprimorar o ensino, o aprendizado e a formação dos estudantes.

FUNDAMENTAÇÃO TEÓRICA

Este estudo concentra-se na avaliação do desempenho dos cursos, utilizando métricas essenciais que fornecem insights valiosos sobre a qualidade educacional oferecida pelas instituições de ensino superior. As principais métricas consideradas são o ENADE, CPC e CC.

O Exame Nacional de Desempenho dos Estudantes (ENADE) desempenha o papel crucial de avaliar o desempenho dos estudantes concluintes nos cursos de graduação. Essa avaliação se concentra não apenas no cumprimento dos conteúdos programáticos delineados nas diretrizes curriculares, mas também na avaliação das competências e habilidades essenciais para a formação geral e profissional.

O CPC, Conceito Preliminar de Curso, desempenha um papel crucial como indicador de qualidade na avaliação dos cursos de graduação. Sua apuração e divulgação ocorrem no ano subsequente à realização do Enade. Esse cálculo é fundamentado na avaliação do desempenho dos estudantes, no valor adicionado pelo processo formativo e em insumos que abrangem as condições de oferta, como corpo docente, infraestrutura e recursos didático-pedagógicos.

Antecedido pelo Conceito Preliminar de Curso (CPC), o Conceito de Curso (CC) é a nota final atribuída durante o processo de avaliação de um curso previsto no Sistema Nacional de Avaliação da Educação Superior (Sinaes).

Para realizar a análise do desempenho, os dados utilizados foram: 'CPC CONTINUO', 'CPC FAIXA', 'CPC ANO', 'VALOR CC', 'ANO CC', 'VALOR ENADE', 'ENADE ANO', ‘NOME DO CURSO’ e ‘MUNICIPIO’.
CPC CONTINUO:

O CPC Contínuo descreve o Conceito Preliminar de Curso (CPC) em uma escala contínua. O CPC é uma métrica que avalia a qualidade de cursos de graduação no Brasil, considerando diversos aspectos como desempenho dos estudantes, corpo docente e infraestrutura.

O CPC Faixa representa o Conceito Preliminar de Curso agrupado em faixas, facilitando a interpretação. As faixas podem indicar diferentes níveis de qualidade, sendo uma forma simplificada de comunicar a avaliação do curso.

O CPC Ano indica o ano ao qual o Conceito Preliminar de Curso (CPC) está associado. É o ano de referência para a avaliação do curso.

O Valor CC refere-se ao valor associado ao Conceito de Curso (CC). O CC é um indicador de qualidade que avalia diversos aspectos do curso, incluindo infraestrutura, corpo docente e projeto pedagógico.

O Ano CC indica o ano de referência para o Conceito de Curso (CC). Este é o ano associado à avaliação específica do CC para um determinado curso.

O Valor ENADE representa o valor associado ao desempenho do curso no Exame Nacional de Desempenho de Estudantes (ENADE), uma avaliação nacional que mede o rendimento dos estudantes em relação aos conteúdos programáticos e habilidades previstas nas diretrizes curriculares.

ENADE Ano indica o ano de realização do Exame Nacional de Desempenho de Estudantes (ENADE), ao qual o valor do ENADE está associado.

Nome do curso refere-se ao nome ou denominação do curso de graduação que está sendo avaliado.

Município indica o município onde está localizada a instituição que oferece o curso. Este dado fornece informações sobre a localização geográfica da instituição de ensino.

Essas variáveis proporcionam uma visão abrangente do desempenho e qualidade dos cursos de graduação, considerando tanto avaliações nacionais quanto métricas específicas de qualidade. A análise dessas variáveis pode ser útil para avaliar e comparar cursos em diferentes localidades e períodos.

DESENVOLVIMENTO (METODOLOGIA E ANÁLISE)

ETL (Extract, Transform, Load)

Extração (Extract):

A fonte escolhida para o desenvolvimento deste estudo foi o dataset "Cursos IFSP" em formato csv, fornecido pelo Portal Brasileiro de Dados Abertos, no site dados.gov. Após a extração, foram utilizadas ferramentas para visualizar as informações, como utilizar os métodos head(), info(), describe(), isnull().sum().

Transformação (Transform):

Inicialmente foram definidas as colunas que vão ser utilizadas, houve uma redução no dataset de 77 colunas para 19 colunas, ficando apenas as informações necessárias para o desenvolvimento, sendo as demais removidas.

Logo em seguida, foram retirados todos os valores nulos do dataset. E também, os dados do tipo float64 foram convertidos para int64, para que os dados sejam do tipo inteiro.

Carregamento (Load):

O dataset com as alterações feitas foram salvos como "cursos.ifsp_FINAL.csv".

KDD (Knowledge Discovery in Databases)

A descoberta de conhecimento em bases de dados (KDD) pode ser definida como o processo de extração de informação a partir de dados registrados numa base de dados

Assim, o processo de KDD utiliza conceitos de base de dados, métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial, dividindo-se nas etapas de seleção, pré-processamento, transformação, DM e avaliação/interpretação.

O KDD será aplicado com o objetivo específico de realizar uma análise abrangente e detalhada do desempenho dos cursos. Este método sistemático envolverá diversas etapas, desde a seleção criteriosa dos conjuntos de dados relacionados aos cursos até a aplicação de algoritmos avançados de mineração de dados.

Coleta de dados: Através do ETL, foi criado o dataset no formato csv, o arquivo possui apenas as colunas necessárias que serão utilizadas e os dados já estão tratados. Este serão os dados utilizados.

Pré-processamento de Dados: Foram separadas do Dataset original as colunas que serão utilizadas no estudo, criando um Dataset com as colunas que são precisas somente. Através do "dropna" foram

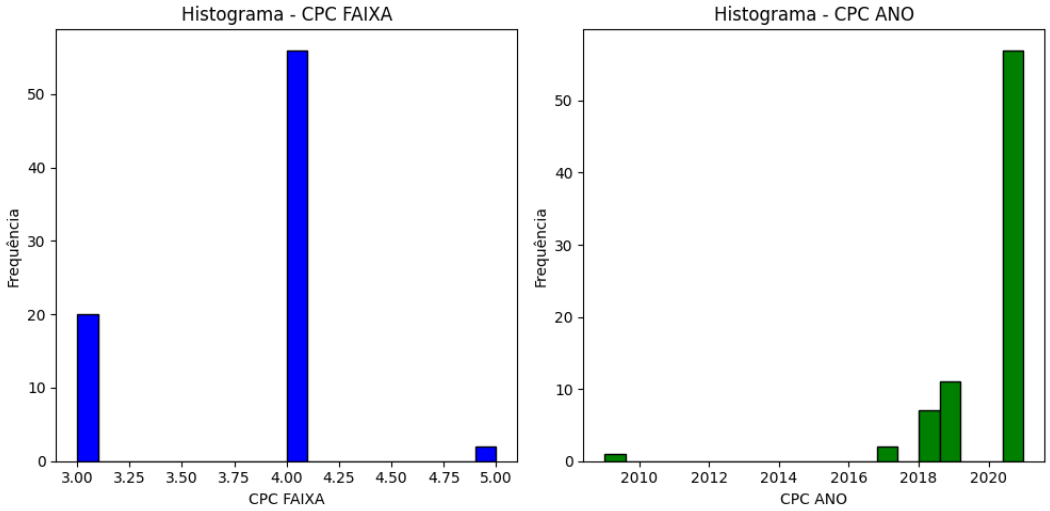
removidos os valores nulos das colunas. Em seguida, foi estabelecido que o código remove as linhas em que pelo menos um valor é igual a zero. E, a conversão do float64 para int64.

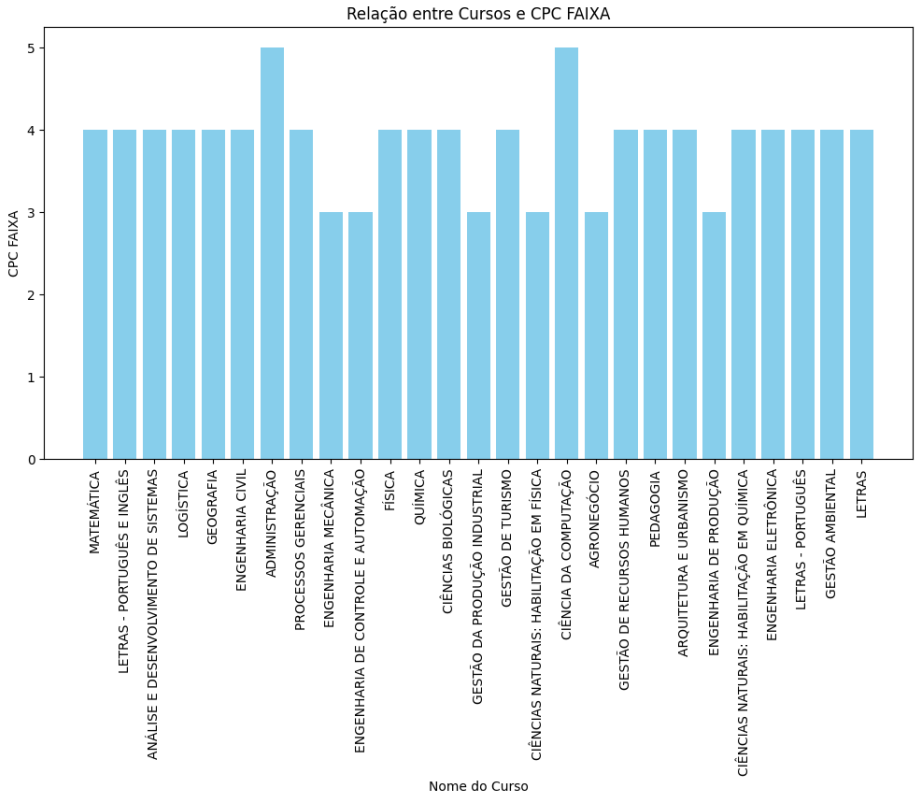
Análise Exploratória de Dados (EDA):

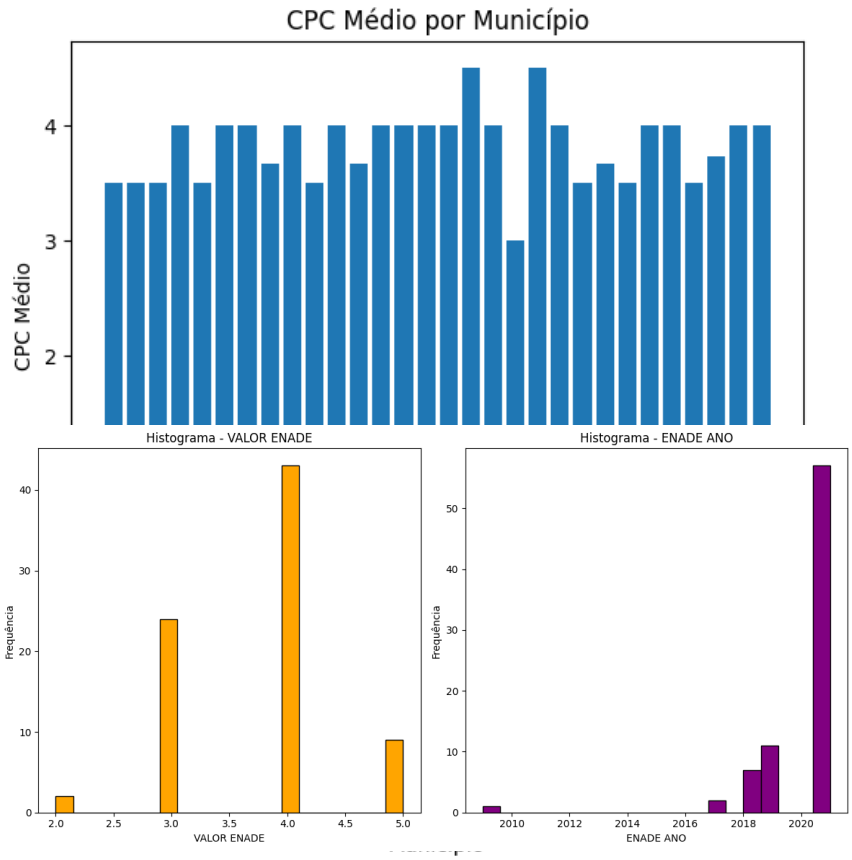
O primeiro gráfico exibe os histogramas apara variáveis numéricas, para ter uma visão geral das variáveis.

Buscando facilitar a visualização dos dados, para avaliar cada um dos valores de desempenho dos cursos, foram separados cada um com seu gráfico. Sendo o primeiro, a análise exploratória sobre o CPC.

CPC (Conceito Preliminar de Curso)





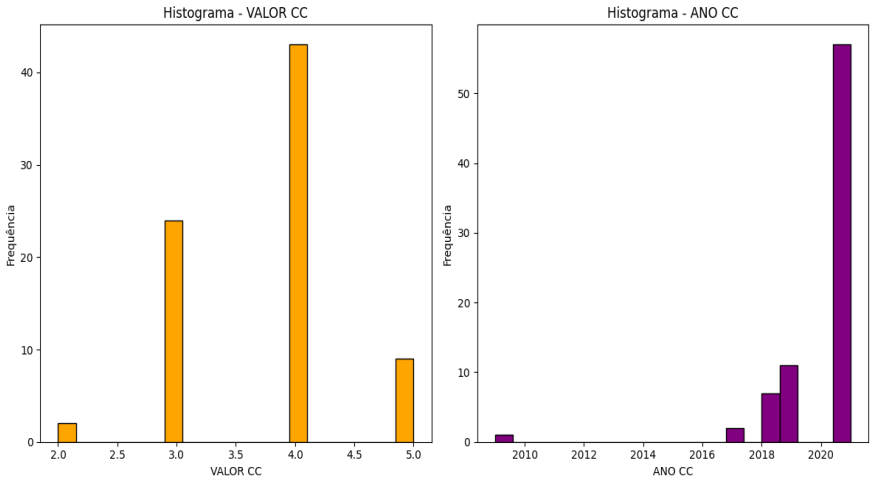


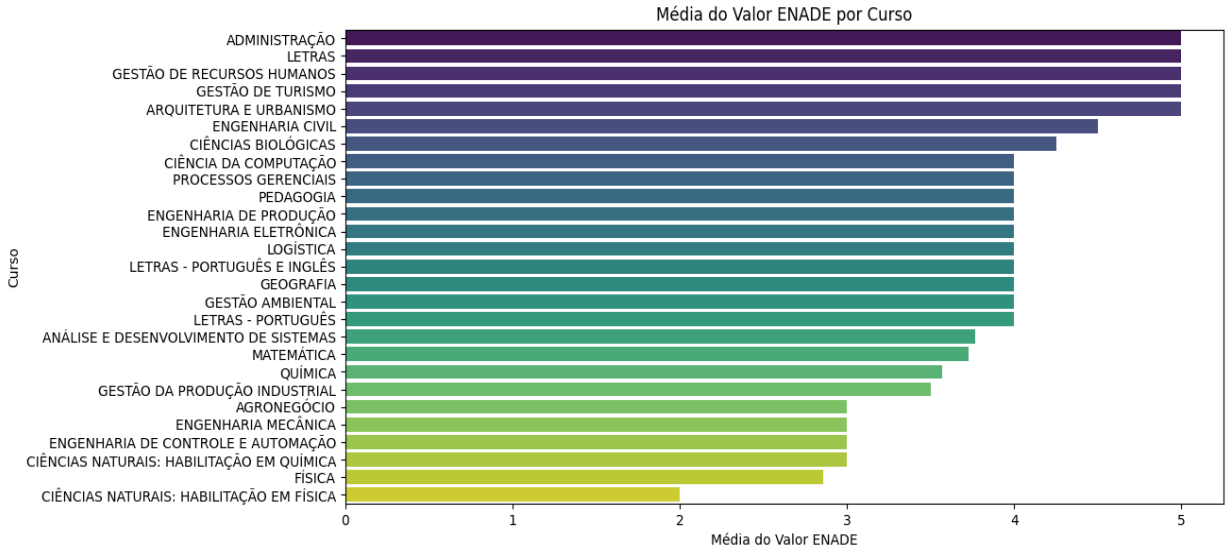

```
[174] contagem = df['CPC FAIXA'].value_counts()
      print(contagem)

4    56
3    20
5     2
Name: CPC FAIXA, dtype: int64
```

Presidente Epitácio e Jacareí foram os municípios com maior valor de média de CPC. Esses dados são referentes aos anos de 2016 até 2022.

É possível analisar que o índice 4.0 foi o valor de mais frequência, de um total de 78 notas, o valor 4.0 ocupou 56 das posições. E que o ano de 2020 até 2022 foram os anos com o maior valor. Administração e Ciência da Computação foram os cursos que tiveram maior valor de avaliação.

ENADE (Exame Nacional de Desempenho dos Estudantes)






```
# Agrupar os dados por município e calcular a média do VALOR ENADE para cada município
resultado_por_municipio = df.groupby('MUNICIPIO')['VALOR ENADE'].mean()

# Exibir o resultado
print(resultado_por_municipio)
```

```
MUNICIPIO
Araraquara              3.500000
Avaré                   3.500000
Barretos                3.500000
Birigui                 3.500000
Boituva                 3.500000
Bragança Paulista       4.000000
Campinas                4.000000
Campos do Jordão        3.666667
Capivari                4.000000
Caraguatatuba           3.750000
Catanduva               3.500000
Cubatão                 3.666667
Guarulhos               4.000000
Hortolândia             4.000000
Itapetininga            3.500000
Itaquaquecetuba         4.000000
Jacareí                 4.500000
Matão                   3.000000
Piracicaba              2.666667
Presidente Epitácio     4.000000
Registro                3.000000
Salto                   4.000000
Sertãozinho             3.666667
Suzano                  4.000000
São Carlos              4.000000
São José dos Campos     4.000000
São João da Boa Vista   3.000000
São Paulo               4.066667
São Roque               4.333333
Votuporanga             3.666667
Name: VALOR ENADE, dtype: float64
```

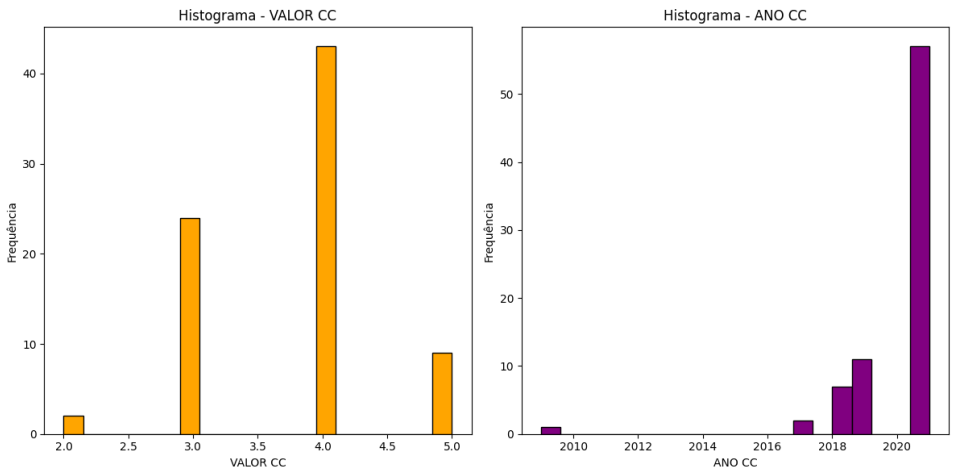
```
contagem = df['VALOR ENADE'].value_counts()
print(contagem)
```

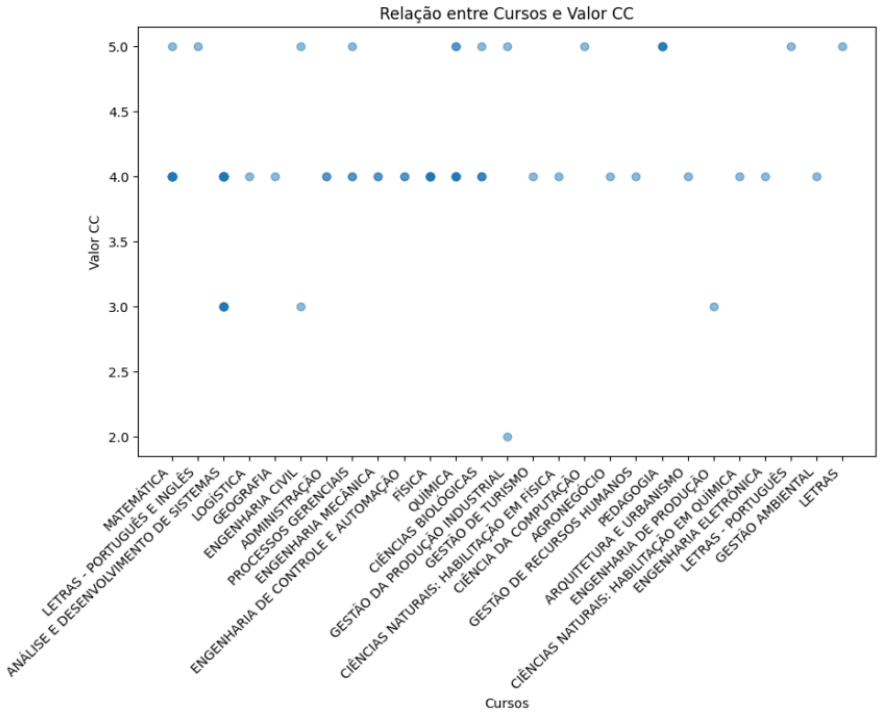
```
4    43
3    24
5     9
2     2
Name: VALOR ENADE, dtype: int64
```

Através destes gráficos é possível analisar que a maior frequência foi de 4.0 e com maior ocorrência dessas notas nos anos de 2020 para cima.

Os cursos com maior data no ENADE foram os de Administração e Letras. De 78 notas, 43 foram com a pontuação 4.0 durantes os anos de 2010 até 2022. As maiores notas foram em Catanduva e São Roque.

CC (Conceito de Curso)



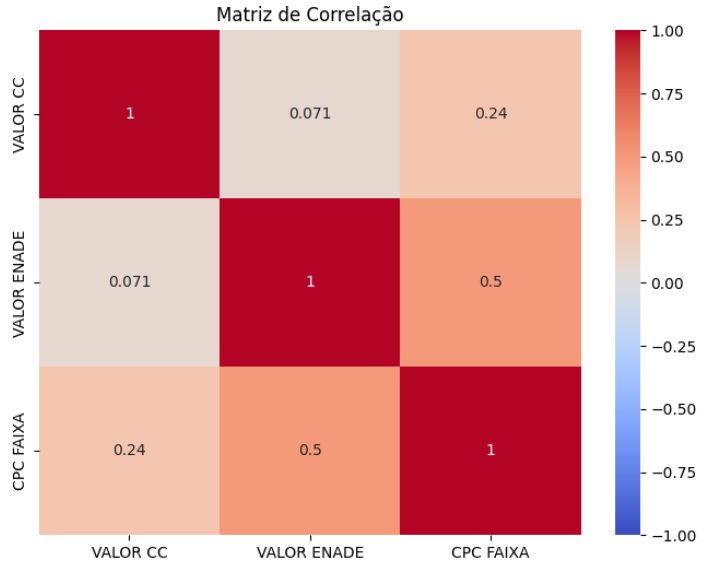



```
cc_por_municipio = df.groupby('MUNICIPIO')['VALOR CC'].mean().reset_index()

# Mostrar o resultado
print(cc_por_municipio)
```

	MUNICIPIO	VALOR CC
0	Araraquara	4.000000
1	Avaré	4.000000
2	Barretos	4.000000
3	Birigui	4.000000
4	Boituva	4.500000
5	Bragança Paulista	3.500000
6	Campinas	4.000000
7	Campos do Jordão	4.333333
8	Capivari	4.000000
9	Caraguatatuba	4.000000
10	Catanduva	4.000000
11	Cubatão	4.000000
12	Guarulhos	3.500000
13	Hortolândia	4.000000
14	Itapetininga	4.000000
15	Itaquaquecetuba	5.000000
16	Jacareí	4.500000
17	Matão	4.000000
18	Piracicaba	4.000000
19	Presidente Epitácio	5.000000
20	Registro	4.000000
21	Salto	2.500000
22	Sertãozinho	4.000000
23	Suzano	4.500000
24	São Carlos	3.500000
25	São José dos Campos	4.500000
26	São João da Boa Vista	4.250000
27	São Paulo	4.000000
28	São Roque	4.333333
29	Votuporanga	4.333333

Através desse gráfico é possível analisar que os cursos com maior valor de CC são Pedagogia e Ciências Biológicas. O valor mais frequente foi de 4.0, com maior frequência de 2010 até 2022. Os municípios com maior valor de CC foram os campus de Itaquaquecetuba e Suzano.

Matriz Correlação com as Três Variáveis.



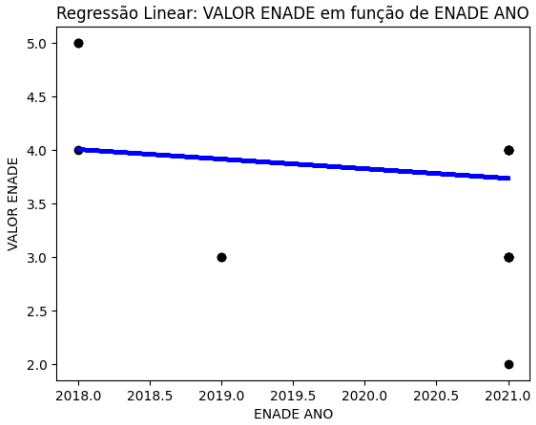
- O VALOR CC tem uma correlação boa com o CPC FAIXA e uma correlação fraca com VALOR ENADE. Portanto, quando o VALOR CC aumenta, pode ser que o CPC FAIXA aumente na mesma proporção, isso acontece porque o CC antecede o CPC.
- O VALOR ENADE tem maior correlação com CPC FAIXA e pouca com VALOR CC.
- O CPC FAIXA tem uma boa correlação com o VALOR ENADE e o VALOR CC

Machine Learning:

Regressão Linear:

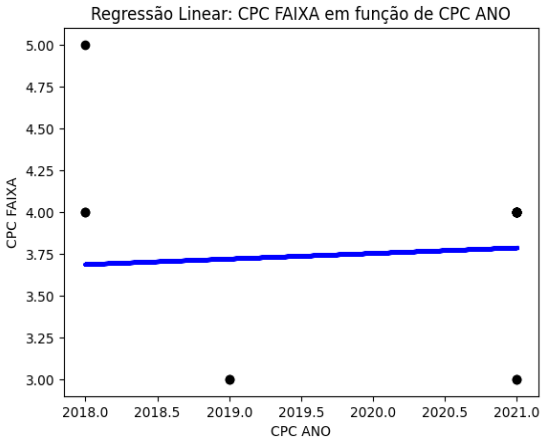
A regressão linear é um algoritmo de machine learning que é usado para prever o valor de uma variável dependente com base em uma ou mais variáveis independentes. A regressão linear simples envolve apenas uma variável independente, enquanto a regressão linear múltipla envolve várias variáveis independentes.

Regressão Linear ENADE:



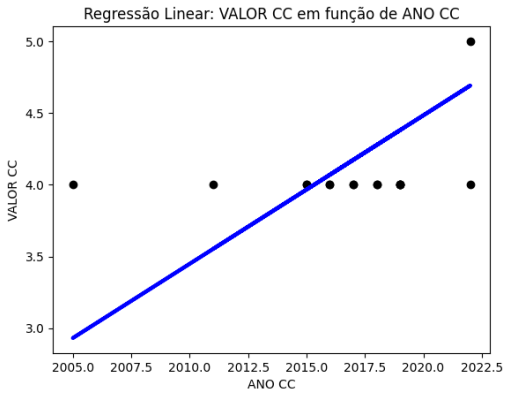
O valor do MSE para VALOR ENADE é 0.6068z. A linha azul representa a regressão linear, indicando a tendência geral dos dados. A linha segue um pouco perto os pontos no gráfico de dispersão, isso sugere uma boa adaptação do modelo aos dados observados. Em resumo, tem um desempenho razoável, com base no MSE e na visualização.

Regressão Linear CPC:



A linha de regressão sugere a relação geral entre 'CPC ANO' e 'CPC FAIXA' com base nos dados de treino.

Regressão Linear CC:



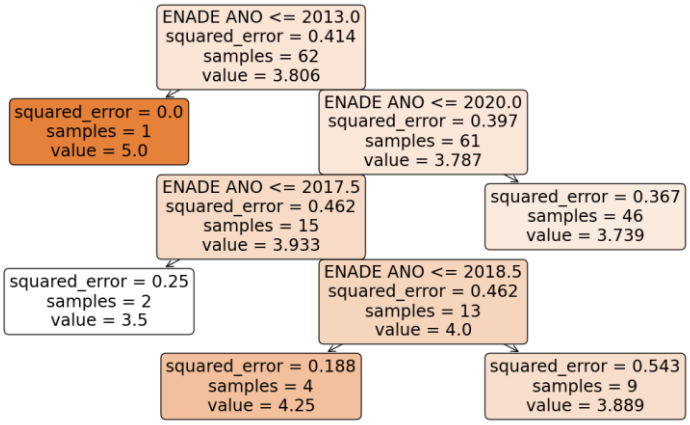
A linha está inclinada para cima, indica uma relação positiva: à medida que o 'ANO CC' aumenta, o 'VALOR CC' também tende a aumentar.
O MSE relativamente baixo sugere que o modelo está ajustando bem a linha de regressão aos dados de teste.

Árvore de Decisão

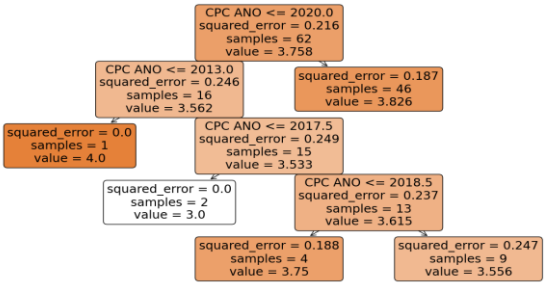
Uma árvore de decisão é um modelo de aprendizado de máquina que é usado para tomar decisões com base em condições e resultados. Pode ser aplicada a problemas de classificação e regressão. A ideia é representar um conjunto de decisões em forma de árvore, onde cada nó interno representa uma decisão baseada em uma característica (ou atributo), cada ramo representa o resultado dessa decisão, e cada folha representa o resultado.

A construção de uma árvore de decisão envolve a seleção iterativa do melhor atributo para dividir os dados, com o objetivo de reduzir a impureza nos nós da árvore.

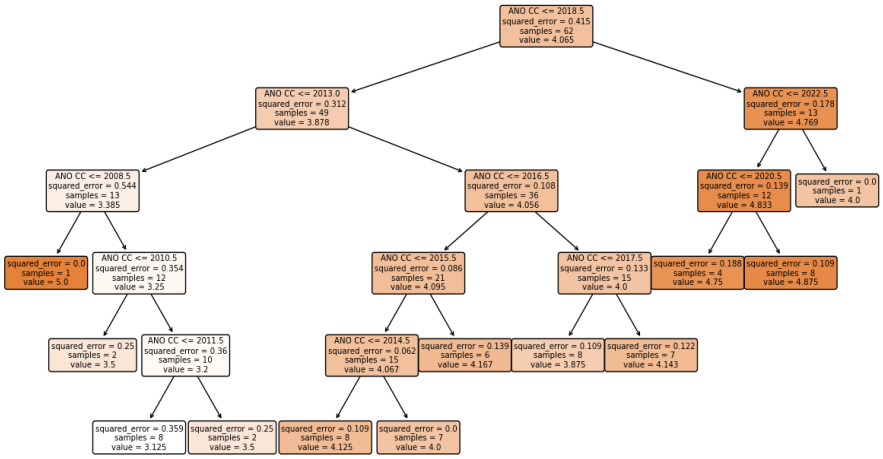
Árvore de Decisão ENADE



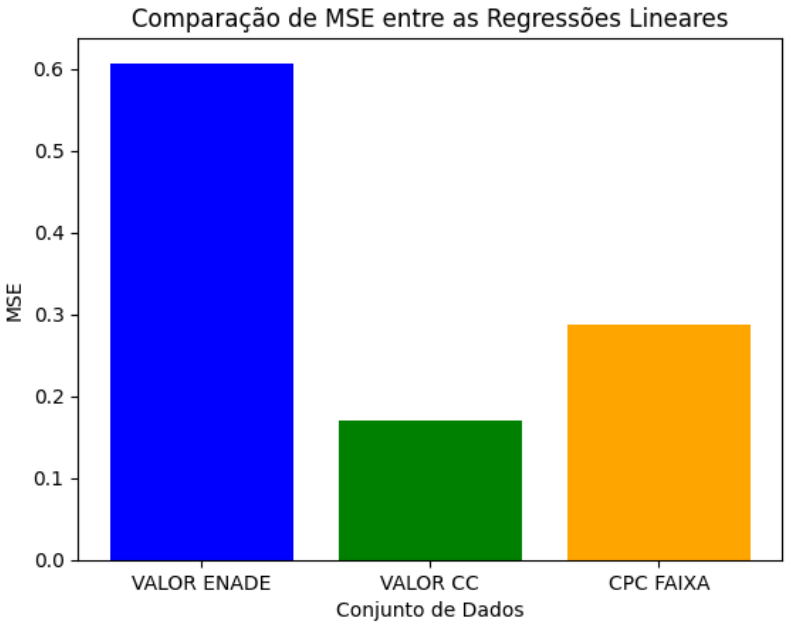
Árvore de Decisão CPC



Árvore de Decisão CC



Comparação de MSE entre as Regressões Lineares



- VALOR ENADE: MSE: 0.6068

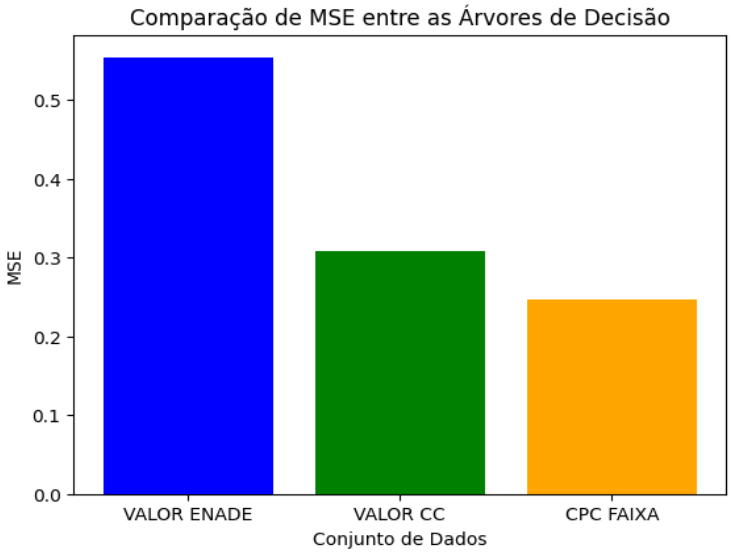
  Este modelo de regressão linear tem um MSE mais alto, indicando que as previsões estão menos precisas em comparação com os outros modelos. Pode ser que a relação entre as variáveis 'ENADE ANO' e 'VALOR ENADE' seja mais complexa ou não linear.
- VALOR CC: MSE: 0.1699

  Este modelo de regressão linear tem um MSE mais baixo, indicando previsões mais precisas em relação ao modelo para 'VALOR ENADE'. Parece ser o modelo mais eficaz entre os três.
- CPC FAIXA: MSE: 0.2880

  O modelo de árvore de decisão para 'CPC FAIXA' tem um MSE intermediário. Isso indica que, em comparação com os modelos de regressão linear, a árvore de decisão tem um desempenho moderado na previsão da variável 'CPC FAIXA'.

O modelo com o menor MSE é considerado o melhor em termos de precisão de previsão nos dados de teste. Neste caso, o modelo de regressão linear para 'VALOR CC' tem o menor MSE (0.1699), sugerindo que é o modelo mais eficaz para prever 'VALOR CC' em comparação com os outros dois modelos.

Comparação de MSE entre as Árvores de Decisão



- MSE para VALOR ENADE: 0.5540

  Este é o MSE associado ao modelo de árvore de decisão para a variável de destino 'VALOR ENADE'.
- MSE para VALOR CC: 0.3078

  Este é o MSE associado ao modelo de árvore de decisão para a variável de destino 'VALOR CC'.

- Um MSE de 0.3078 é inferior ao MSE para 'VALOR ENADE' (0.5540), indicando que o modelo para 'VALOR CC' tem um desempenho relativo melhor.
- MSE para CPC FAIXA: 0.2464
- Este é o MSE associado ao modelo de árvore de decisão para a variável de destino 'CPC FAIXA'.
- Um MSE de 0.2464 é o menor entre os três, indicando que o modelo para 'CPC FAIXA' tem o melhor desempenho relativo entre os três modelos.:

O modelo de árvore de decisão que tem o menor MSE é o que foi construído para a variável 'CPC FAIXA', com um MSE de 0.2464. Em termos de precisão de previsão, o modelo para 'CPC FAIXA' parece ser o mais eficaz entre os modelos avaliados.

CONSIDERAÇÕES FINAIS

O modelo de regressão linear para 'VALOR CC' demonstrou ser mais eficaz em termos de precisão de previsão, apresentando o menor Mean Squared Error (MSE) em comparação com os modelos para as outras variáveis. Isso sugere que a regressão linear é uma abordagem sólida para prever o 'VALOR CC'. O modelo de árvore de decisão para 'CPC FAIXA' obteve o menor MSE entre os modelos avaliados, indicando que é o mais eficaz em termos de precisão de previsão para essa variável específica. Portanto, para entender e prever o 'CPC FAIXA', a árvore de decisão mostrou ser uma escolha mais eficaz. Não foi especificamente mencionado qual modelo obteve o menor MSE para 'VALOR ENADE'. Seria útil avaliar o desempenho dos modelos também para essa variável e verificar se há alguma conclusão específica sobre a precisão de previsão. As conclusões sugerem que diferentes variáveis podem exigir abordagens de modelagem distintas. O desempenho ideal pode depender da natureza específica de cada variável e da relação subjacente com as características disponíveis.

Através da visualização dos gráficos é possível concluir que não há uma discrepância em relação aos valores das três variáveis, possuindo uma média 4.0 ao longo dos anos. Pelo gráfico da análise temporal é possível analisar que a partir de 2010, o ENADE e o CPC iniciaram e que o melhor período para ambos foi de 2017 até 2020. Agora para o valor do CC o auge foi no começo, entre 2005 e 2007, pois só existia esse método avaliador. É possível analisar também que o curso de Administração tem o com melhor valor nas 3 variáveis e que em relação aos municípios, ficam bem divididos entre as variáveis e seus campus.

REFERÊNCIAS BIBLIOGRÁFICAS

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, v. 22, p. 686-690, 2009.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Enade. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade. Acesso em: 27 nov. 2023.

Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - IFSP. (2023). Cursos IFSP. dados.gov.br. https://dados.gov.br/dados/conjuntos-dados/cursos-ifsp

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Conceito Preliminar de Curso (CPC). Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior/conceito-preliminar-de-curso-cpc\>. Acesso em: 27 de novembro de 2023.

OKADA, Hugo Kenji Rodrigues; DAS NEVES, André Ricardo Nascimento; SHITSUKA, Ricardo. Análise de Algoritmos de Indução de Árvores de Decisão. **Research, Society and Development**, v. 8, n. 11, p. e298111473-e298111473, 2019.

Universidade de Brasília (UnB). Conceito de Curso. Disponível em: https://avaliacao.unb.br/conceito-de-curso. Acesso em: 15 de novembro de 2023.