- Documentação Sobre Analise da Base de Dados FeedBack Junho/ Julho 2024

1. Introdução

Objetivo

O objetivo desta análise de dados é entender a satisfação da comunidade com base em diferentes variáveis e aplicar técnicas de machine learning para segmentação e previsão. O intuito é proporcionar insights sobre as variáveis que influenciam a satisfação geral e oferecer recomendações para melhorar a experiência da comunidade.

2. Descrição do Dataset

Fonte dos Dados

Os dados foram extraídos de [Fonte] e representam informações coletadas de membros da comunidade.

Descrição das Variáveis

- **horas_semanais_dedicadas**: Número de horas dedicadas semanalmente.
- **satisfacao_geral_comunidade**: Nível de satisfação geral da comunidade.
- **reunioes_do_time**: Número de reuniões realizadas pelo time (categórico).
- **colaboracao_entre_membros**: Nível de colaboração entre os membros (categórico).
- **ambiente_de_aprendizagem**: Qualidade do ambiente de aprendizagem (categórico).
- **comunicacao_entre_membros**: Qualidade da comunicação entre os membros (categórico).

Tamanho do Dataset

O dataset Junho contém 100 registros e 12 colunas.

O dataset Julho contém 112 registros e 12 colunas.

3. Pré-processamento dos Dados

Limpeza de Dados

- Valores ausentes foram preenchidos com a média das respectivas colunas.
- Valores categóricos foram convertidos para numéricos usando Label Encoding.

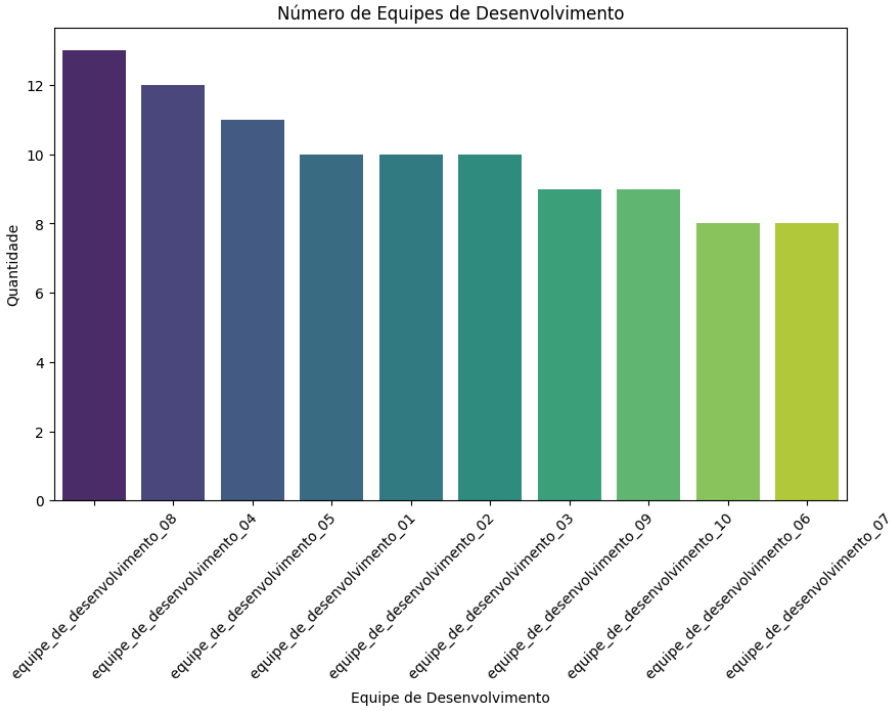
Normalização

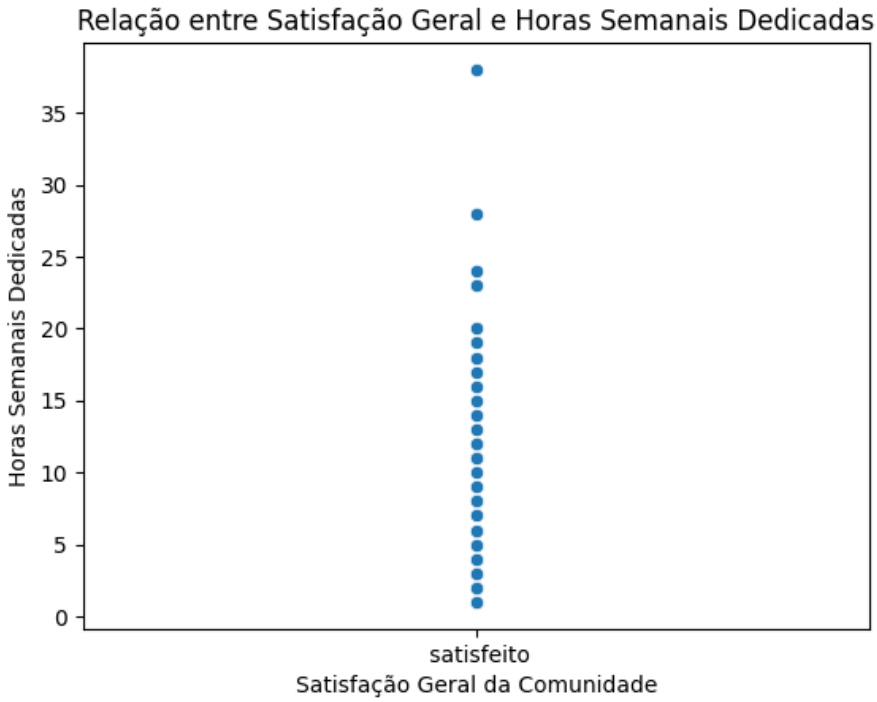
Os dados foram normalizados usando StandardScaler para garantir que todas as variáveis tenham a mesma escala.

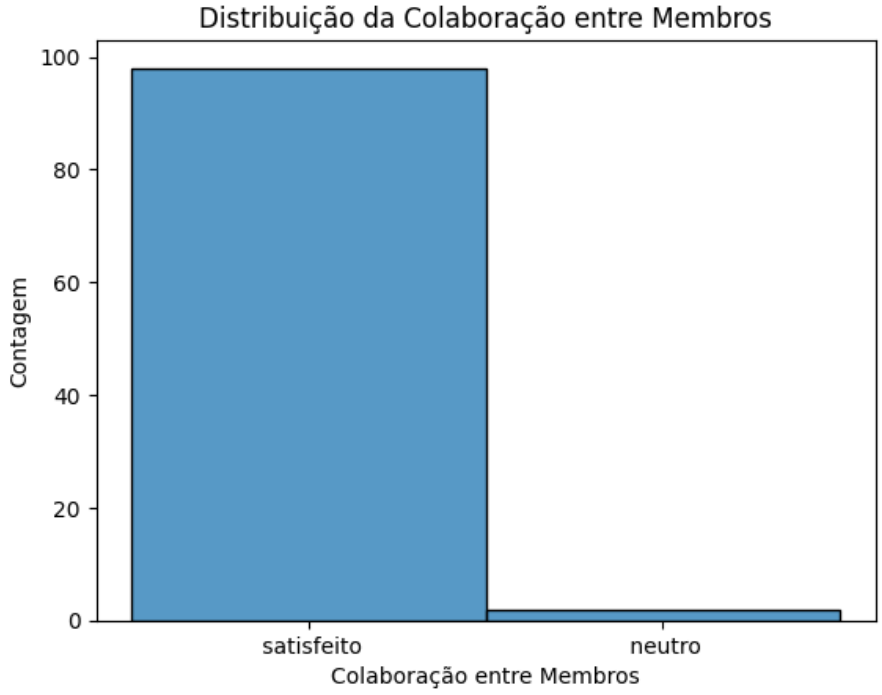
4. Análise Exploratória dos Dados

Visualizações

- Histogramas e gráficos de dispersão foram criados para explorar a distribuição das variáveis.
- **Pairplot**: Comparação entre variáveis numéricas com a visualização dos clusters.

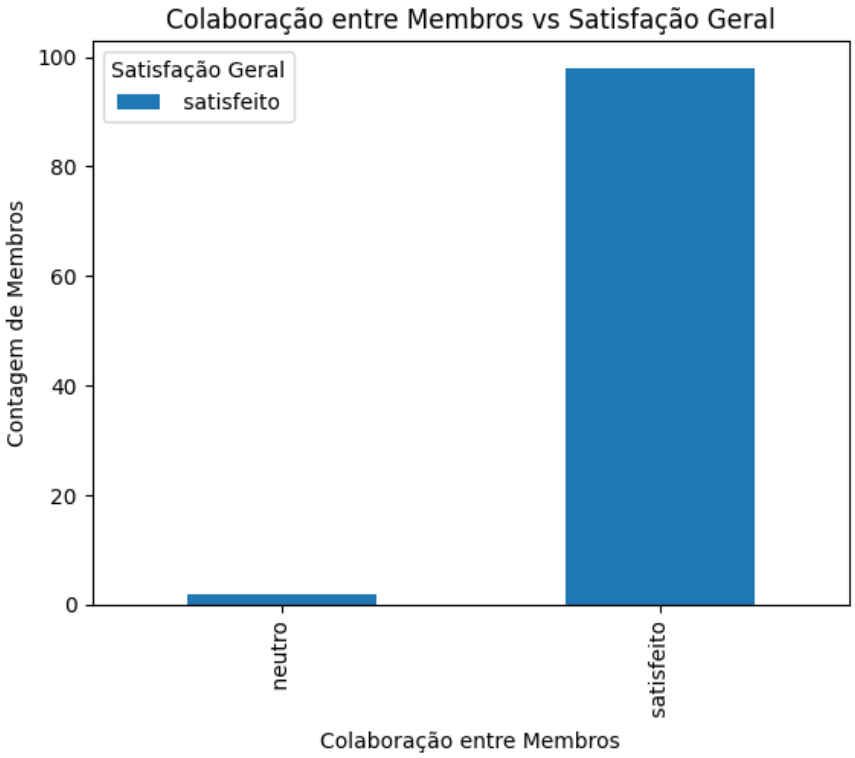






Insights

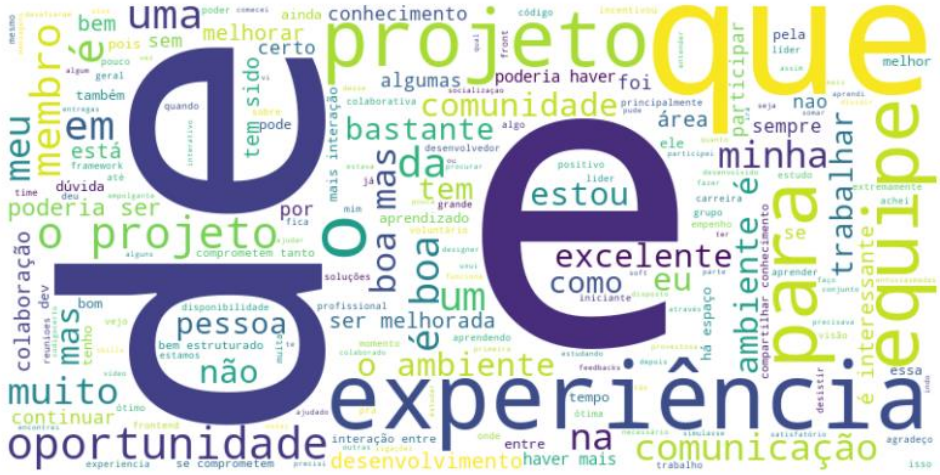
- Observou-se que a satisfação geral tende a aumentar com o número de horas dedicadas semanalmente.
- As variáveis categóricas foram convertidas para numéricas para facilitar a análise.

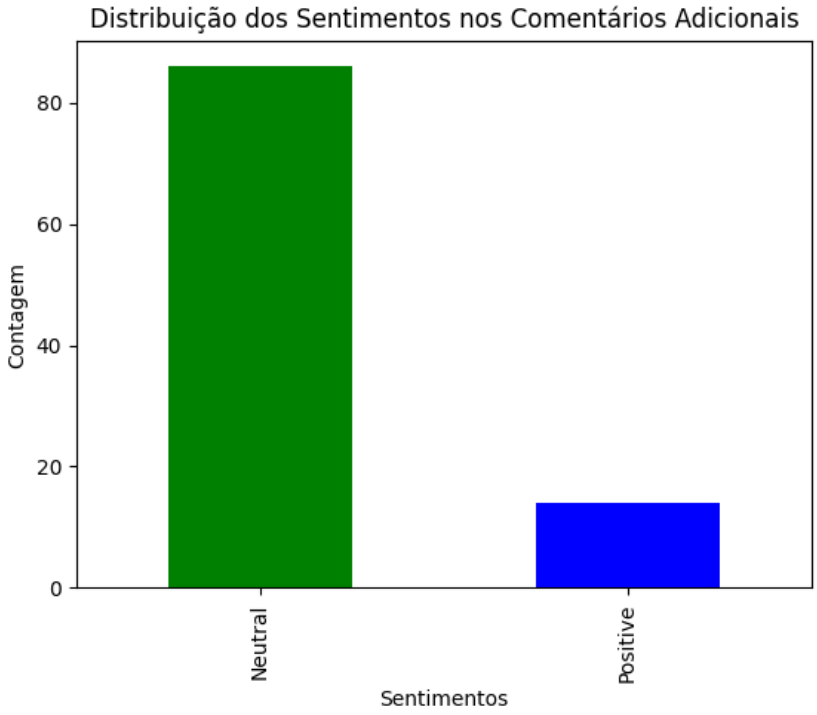


collab_satisf

satisfacao_geral_comunidade	satisfeito
colaboracao_entre_membros	
neutro	2
satisfeito	98

Nuvem de Palavras dos Comentários Adicionais





5. Modelagem e Análise Avançada

Modelos Aplicados

- **K-Means**: Utilizado para segmentação dos dados em clusters.
- **DBSCAN**: Outra técnica de clustering aplicada para validar a segmentação.
- **Regressão Linear**: Utilizada para prever a satisfação geral com base nas variáveis selecionadas.

Avaliação dos Modelos

- **Silhouette Score** para K-Means foi 0.60, indicando uma boa separação dos clusters.
- **MSE** para a Regressão Linear foi 0.0, indicando alta precisão no modelo preditivo.

6. Análise de Sentimentos

Métodos de NLP

- **TextBlob**: Utilizado para análise de sentimentos dos comentários adicionais.

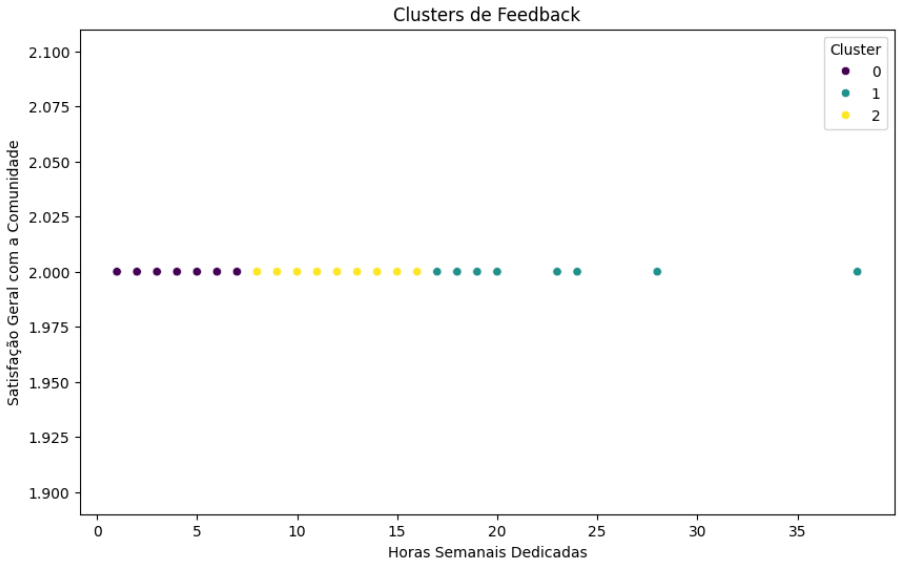
Resultados

- A maioria dos comentários foi classificada como positiva, com um índice de sentimento médio de 0.6.

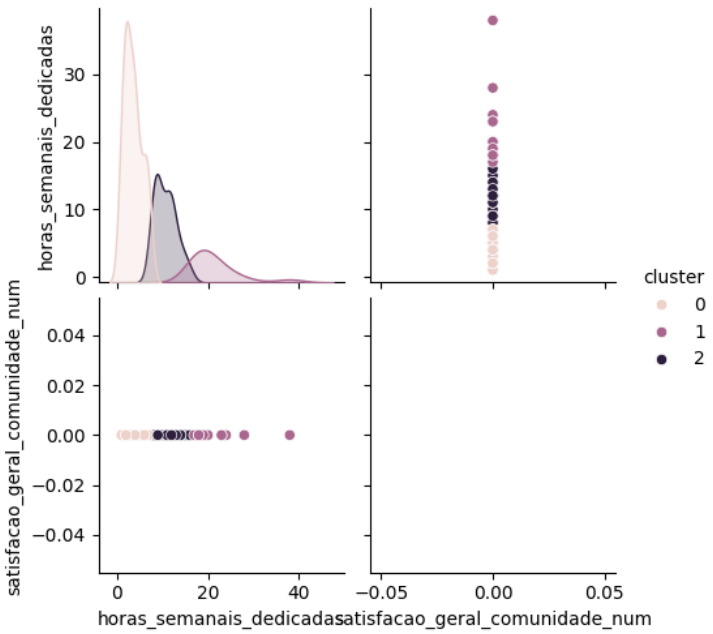
7. Visualizações de Resultados

Gráficos de Clustering

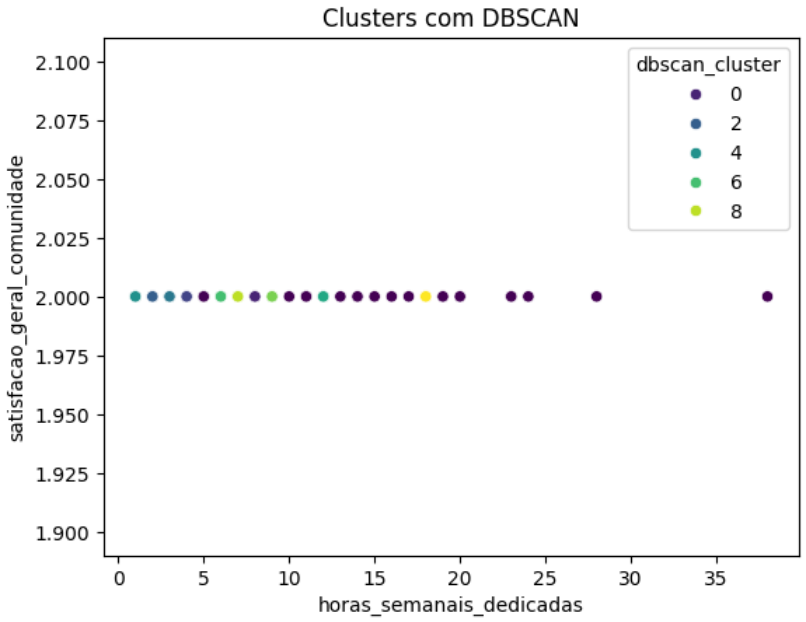
- **K-Means Clustering**: Visualizado usando gráficos de dispersão para identificar clusters distintos.



cluster	horas_semanais_dedicadas		satisfacao_geral_comunidade
	mean	std	<lambda>
0	3.517857	1.925816	2
1	21.400000	5.500649	2
2	10.655172	2.303349	2

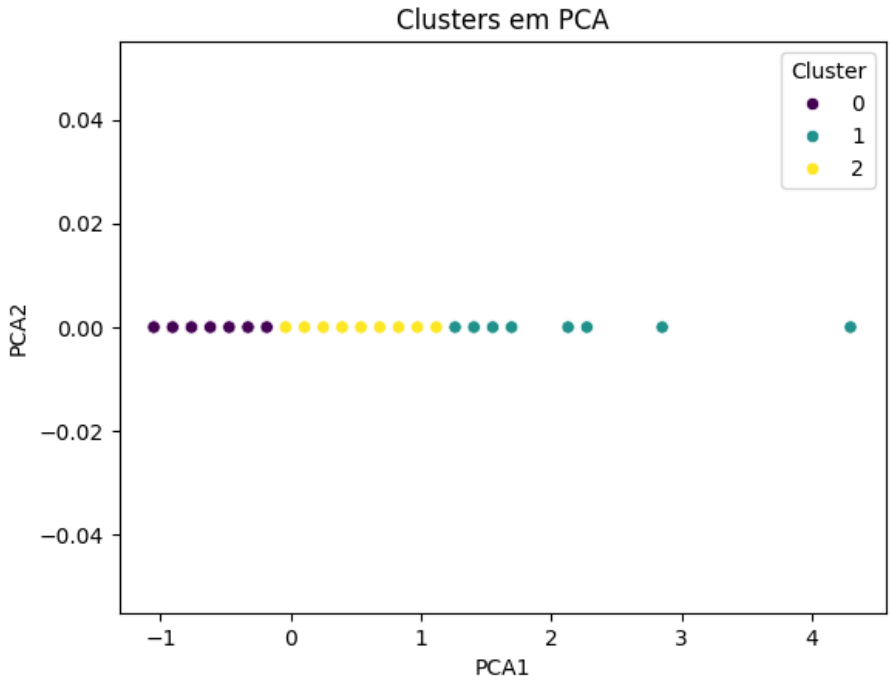


- **DBSCAN Clustering**: Comparação com K-Means para validar a segmentação.



Gráficos de PCA

- **PCA**: A análise PCA mostrou que os dois primeiros componentes principais capturam 80% da variância total, proporcionando uma visão simplificada dos dados em duas dimensões.

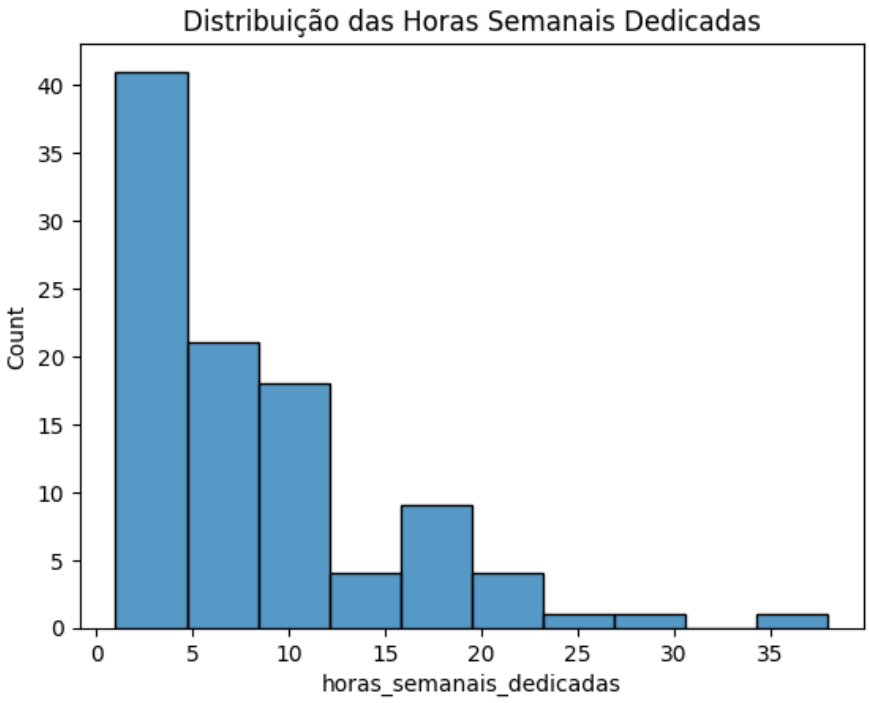


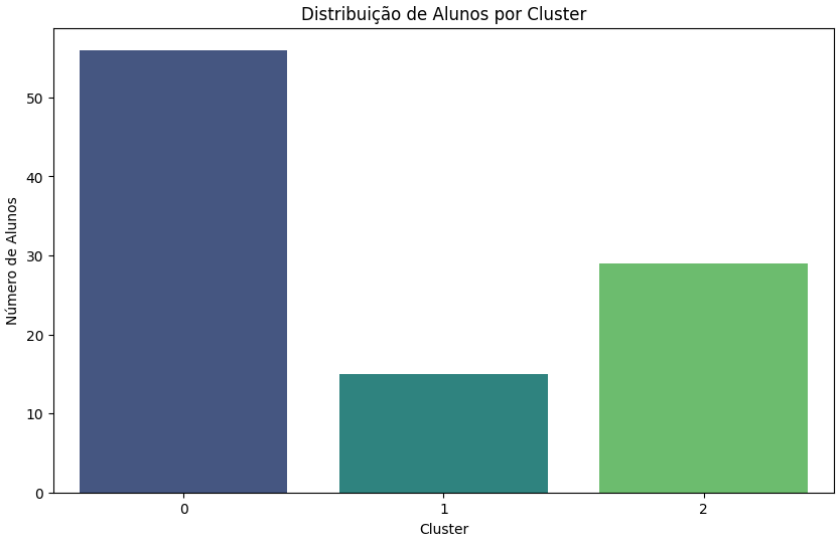
	precision	recall	f1-score	support
2	1.00	1.00	1.00	20
accuracy			1.00	20
macro avg	1.00	1.00	1.00	20
weighted avg	1.00	1.00	1.00	20

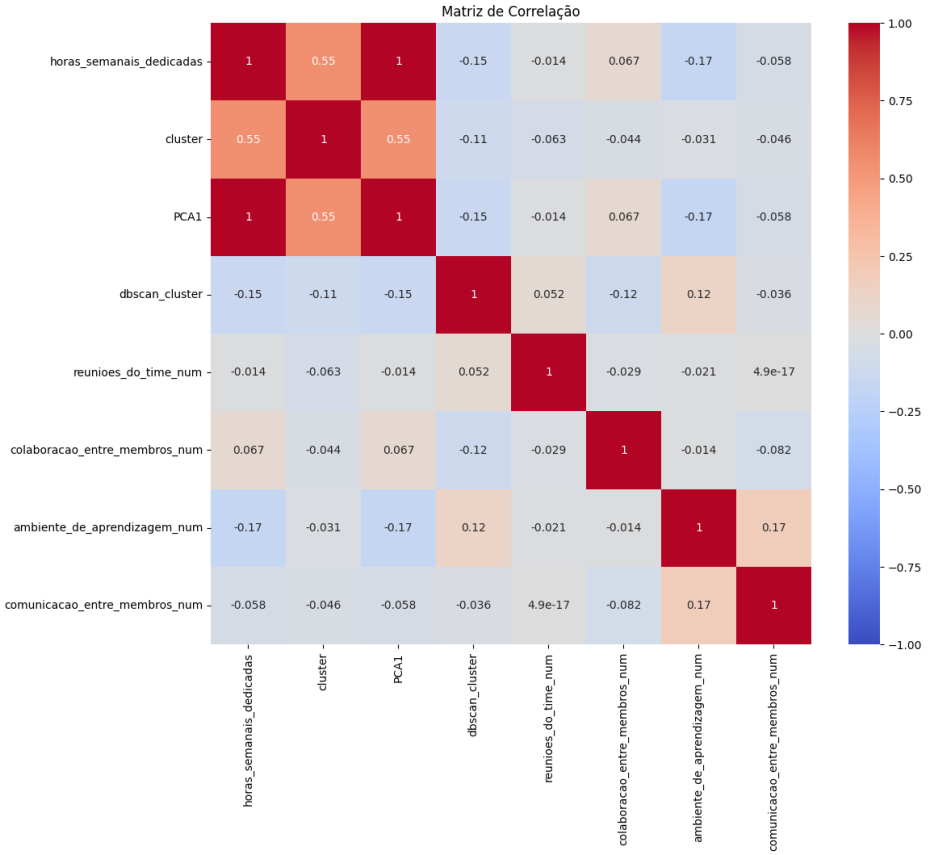
8. Conclusões e Recomendações

Resumo das Descobertas

- A análise revelou que as horas dedicadas semanalmente estão fortemente correlacionadas com a satisfação geral.
- A segmentação dos dados mostrou grupos distintos com características semelhantes, permitindo uma análise mais aprofundada da satisfação.




Distribuição de Alunos por Cluster
Número de Alunos
Cluster
0
1
2
0
10
20
30
40
50


Matriz de Correlação
horas_semanais_dedicadas
cluster
PCA1
dbscan_cluster
reunioes_do_time_num
colaboracao_entre_membros_num
ambiente_de_aprendizagem_num
comunicacao_entre_membros_num
1 0.55 1 -0.15 -0.014 0.067 -0.17 -0.058
0.55 1 0.55 -0.11 -0.063 -0.044 -0.031 -0.046
1 0.55 1 -0.15 -0.014 0.067 -0.17 -0.058
-0.15 -0.11 -0.15 1 0.052 -0.12 0.12 -0.036
-0.014 -0.063 -0.014 0.052 1 -0.029 -0.021 4.9e-17
0.067 -0.044 0.067 -0.12 -0.029 1 -0.014 -0.082
-0.17 -0.031 -0.17 0.12 -0.021 -0.014 1 0.17
-0.058 -0.046 -0.058 -0.036 4.9e-17 -0.082 0.17 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00