**- 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.

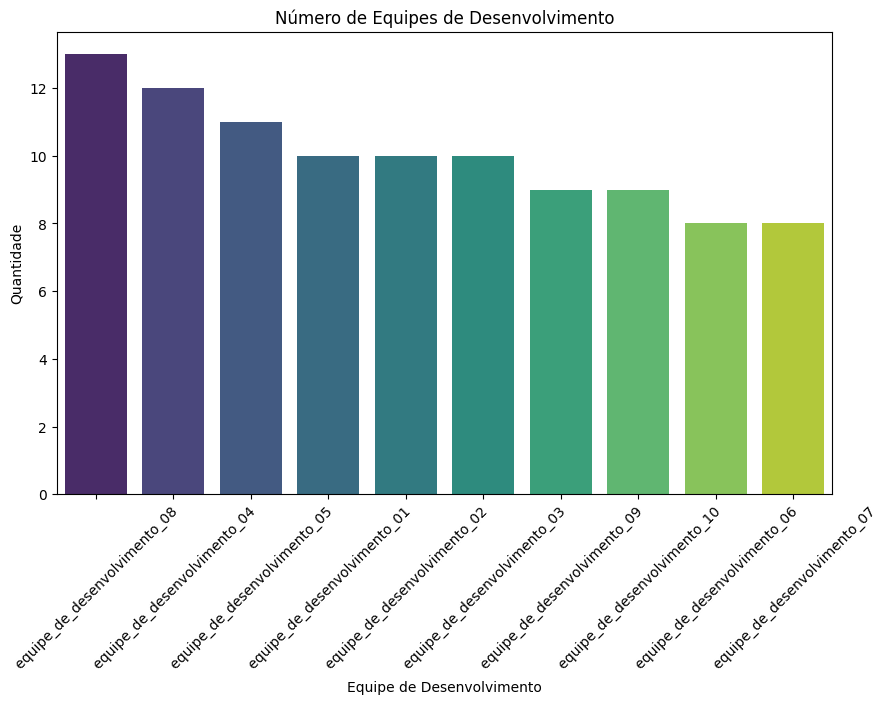
**Normalização**

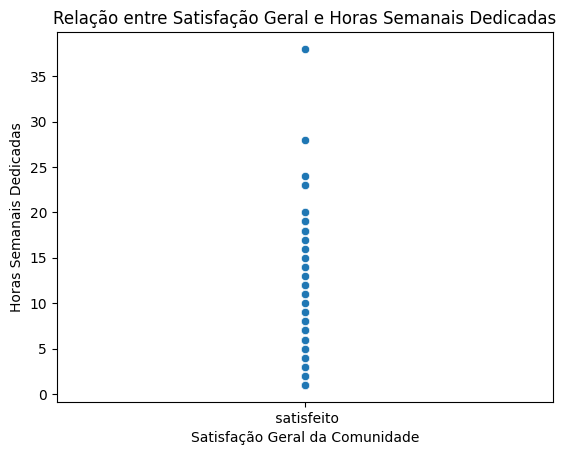
Os dados foram normalizados usando StandardScaler para garantir que todas as variáveis tenham a mesma escala.

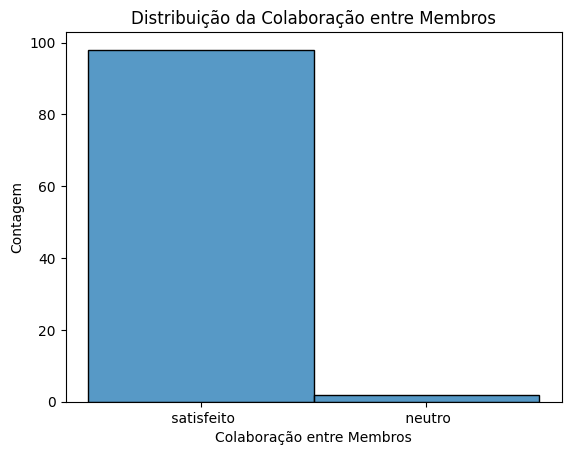
**4. Análise Exploratória dos Dados**

**Visualizações**

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

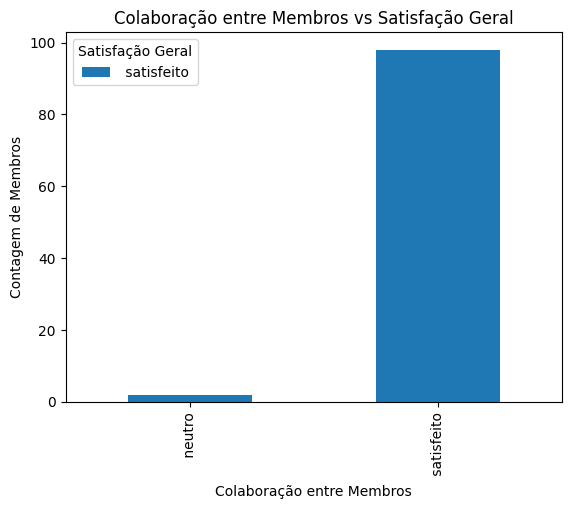


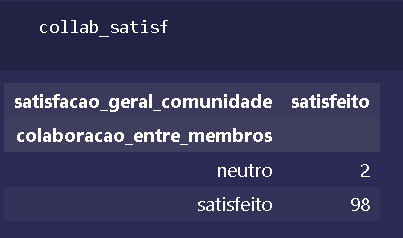


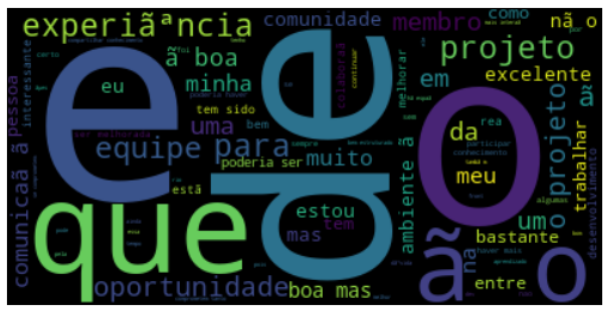


**Insights**

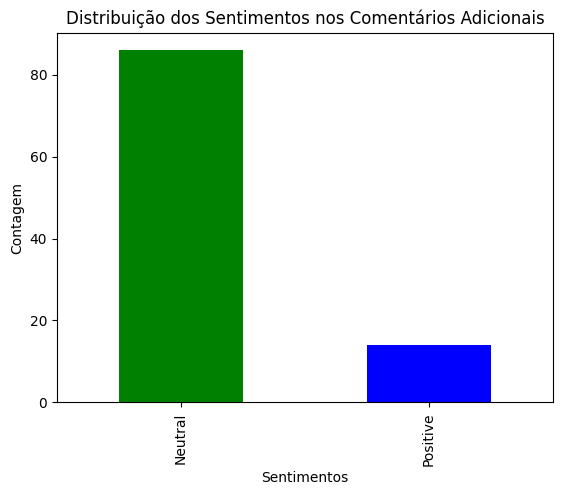
* Observou-se que a satisfação geral tende a aumentar com o número de horas dedicadas semanalmente.
* As variáveis categóricas foram convertidas para numéricas para facilitar a análise.











**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.

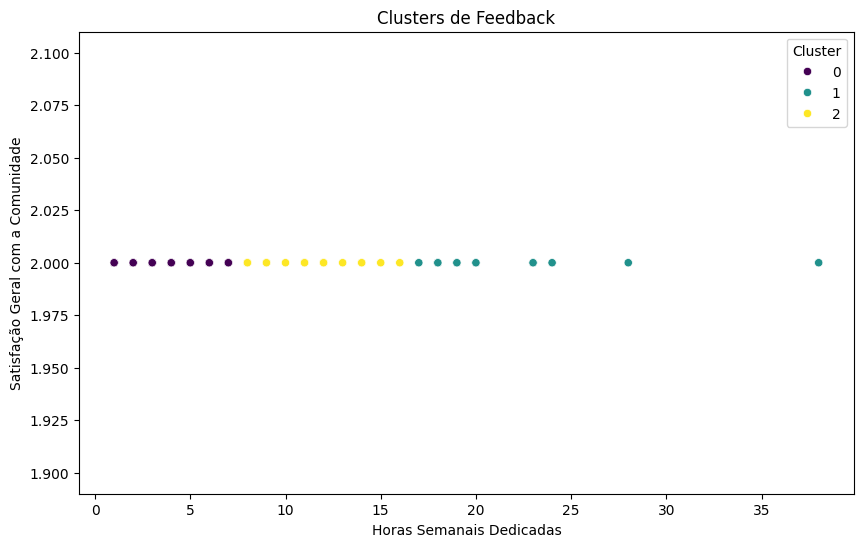
**Resultados**

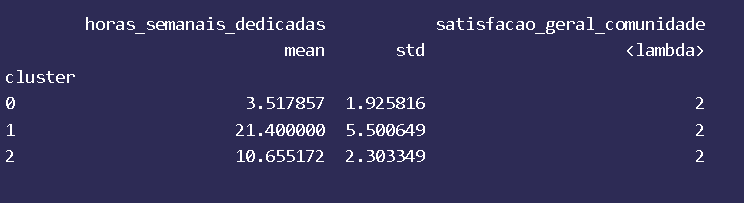
* A maioria dos comentários foi classificada como positiva, com um índice de sentimento médio de 0.6.

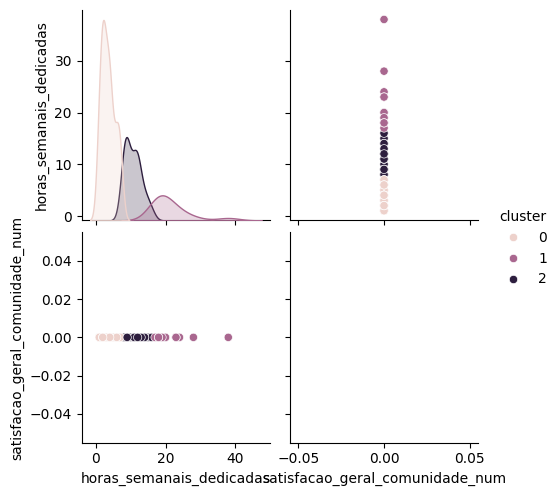
**7. Visualizações de Resultados**

**Gráficos de Clustering**

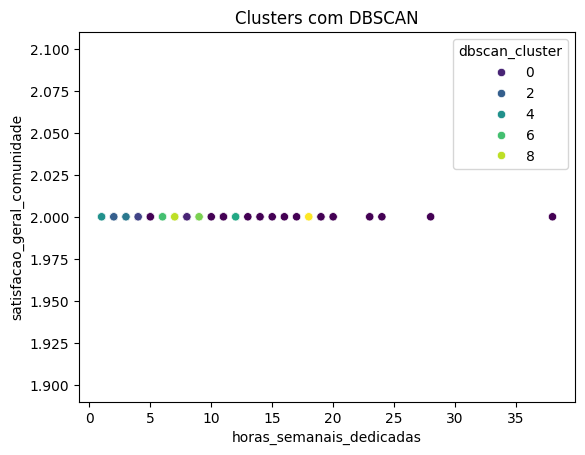
* **K-Means Clustering**: Visualizado usando gráficos de dispersão para identificar clusters distintos.





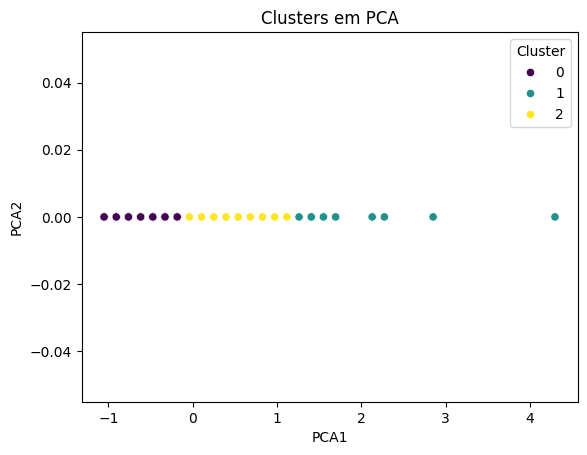


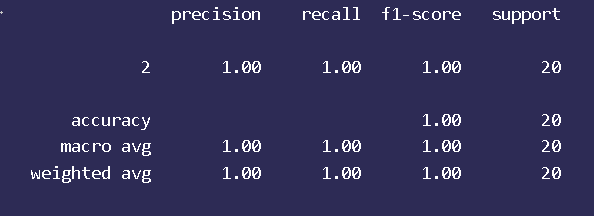
* **DBSCAN Clustering**: Comparação com K-Means para validar a segmentação.



**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.





**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.

