**RELATÓRIO AED – AD – MAT3 INTERDISCIPLINAR**

**Logotipo, nome da empresa

Descrição gerada automaticamente**

**Integrantes 2° ano AD:**

**Leonardo de Freitas,**

**Melissa Martins**

**São Paulo**

**2024**

**RESUMO**

1. Este relatório apresenta uma visão detalhada da primeira etapa do desenvolvimento de uma Inteligência Artificial (IA) de classificação, criada no contexto da disciplina de Análise de Dados, ministrada pelo professor Marcelo Modolo. O objetivo principal deste trabalho é documentar todas as etapas concluídas, incluindo a limpeza dos dados, a normalização e as análises realizadas a partir da base de dados selecionada.
2. O projeto interdisciplinar, no qual a IA será aplicada, visa desenvolver um aplicativo voltado para o transporte escolar, com foco em criar um sistema de predição de clientes em potencial. A atividade de Análise Exploratória de Dados (AED) foi essencial para encontrar uma base de dados que estivesse alinhada com o público-alvo do aplicativo, permitindo que, futuramente, essa base seja utilizada em algoritmos de aprendizado de máquina para predizer a probabilidade de uma pessoa tornar-se usuária do serviço.
3. Na sequência, a parte de Análise de Dados (AD) envolveu a implementação de diversos modelos de inteligência artificial, como Decision Tree, Naive Bayes e K-Nearest Neighbors (KNN), para analisar e classificar os dados. Essas análises visam otimizar a predição de clientes em potencial, ajustando os parâmetros e avaliando o desempenho de cada modelo. Além da escolha da base de dados, o trabalho incluiu a aplicação de técnicas de limpeza e preparação dos dados, que servem de base para as análises subsequentes da disciplina Análise de Dados. Se necessário, ajustes serão realizados após o processamento dos dados, visando otimizar o desempenho do modelo preditivo.

­­

**SUMÁRIO**

**DESCRIÇÃO DO APLICATIVO E DO PÚBLICO-ALVO**

ConnectVans é um aplicativo que visa facilitar a busca por transporte escolar/peruas, tornando-os mais acessíveis, de melhor qualidade e ainda contribuindo para a sustentabilidade.

A ideia do aplicativo vem com base na meta 11.2 da ODS 11 na qual diz sobre “Proporcionar o acesso a sistemas de transporte seguros, acessíveis, sustentáveis e a preço acessível para todos." Com este objetivo em mente e ainda vivendo em um ambiente escolar onde a maioria dos alunos vem e vai embora da escola em que estudamos em vans, pensamos em trabalhar nessa meta começando pelo ambiente em que vivemos e convivemos.

**•  Motoristas de transportes escolares**

Visando principalmente os transportadores que levam e trazem os alunos para as escolas, nosso objetivo é oferecer uma otimização de tempo e rotas, além de uma forma de divulgação do seu trabalho, tendo uma organização melhor dos que usufruiriam de seu trabalho.

**• Pais**

Pais terão como cadastrar seus filhos no app para conseguirem as melhores vans, com os preços mais acessíveis. Ficando mais tranquilos sabendo que seus filhos estão em veículos seguros, conduzidos por motoristas treinados e operando em sistemas de transporte bem-organizados.

**BASE DE DADOS**

**Base:** [pesquisa\_responsáveis.xlsx](https://institutogerminare-my.sharepoint.com/personal/melissa_martins_germinare_org_br/Documents/Interdisciplinar/AED%20-%20Interdisciplinar/Bases/pesquisa_responsaveis.xlsx)b

**Descrição:** A base de dados utilizada foi gerada a partir de um formulário destinado aos responsáveis por alunos do ensino fundamental ao ensino médio, com questões relacionadas ao uso de transporte escolar. Foram obtidas 33 respostas no total, sendo que o formulário foi enviado a mais de cinco escolas da região. Para a análise inicial, foi realizada uma consulta com a professora Myrna Kagohara, que autorizou a condução da análise com os dados disponíveis.

**Colunas e Respostas:**

**1 -** Qual é a idade do(s) seu(s) filho(s)?

**6 - 10 anos**

**11 - 14 anos**

**15 - 18 anos**

**2 -** Qual é a distância entre sua casa e a escola?

**Menos de 1 km**

**1 - 4 km**

**5 - 9 km**

**Mais de 10 km**

**3 -** Quanto tempo você leva, em média, para levar seu(s) filho(s) à escola?

**Menos de 10 minutos**

**10 - 20 minutos**

**21 - 30 minutos**

**Mais de 30 minutos**

**4 -** Você possui carro próprio?

**Sim**

**Não**

**5 -** Já fez uso do serviço de um transporte escolar para o(s) seu(s) filho(s)?

**Já e ainda uso**

**Nunca**

**Já, mas não uso mais**

**6 -** Você já teve problemas com transporte escolar?

**Sim**

**Não**

**7 -** Você estaria disposto a mudar o transporte do seu(s) filho(s) para uma van escolar que oferecesse mais segurança e eficiência?

**Sim**

**Não**

**8 -** Você preferiria um serviço de transporte escolar que avise automaticamente sobre atrasos e mudanças de rota?

**Sim**

**Não**

**9 -** Você tem ou já teve dificuldade para encontrar transporte escolar para o seu(s) filho(s)?

**Sim**

**Não**

**10 -** Você tem ou já teve dificuldade para encontrar transporte escolar para o seu(s) filho(s)?

**Sim**

**Não**

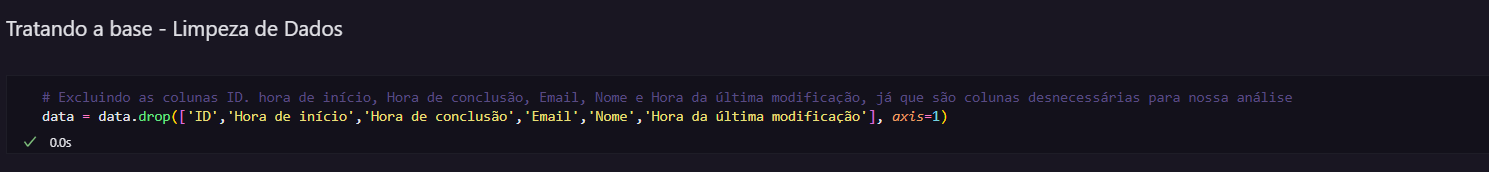
**Às vezes**

**LIMPEZA DOS DADOS**

A base de dados foi gerada por meio de um formulário e, ao ser transferida do Google Forms para uma planilha Excel, colunas como ID, E-mail, Hora de Início (horário em que o respondente iniciou o preenchimento do formulário), Hora de Conclusão, Nome (que estava vazia) e Hora da Última Modificação foram consideradas desnecessárias para a análise, uma vez que não possuíam relação direta com o propósito do aplicativo. Dessa forma, essas colunas foram excluídas logo no início do processo de limpeza dos dados. No que se refere à limpeza de dados para a análise da disciplina de AED, não foram necessários outros ajustes além dessas exclusões, pois as 10 colunas restantes foram essenciais para a criação dos gráficos e análises de correlação.

Interface gráfica do usuário

Descrição gerada automaticamente



**GRÁFICOS E ANÁLISES**

**- Distribuições**

**Interface gráfica do usuário, Aplicativo, Tabela, Excel

Descrição gerada automaticamente**

**Gráfico 1 – Distribuição de Idade dos Filhos:**

**Descrição:** O gráfico apresenta a **distribuição de idades dos filhos em três faixas etárias: 15 a 18 anos, 6 a 10 anos e 11 a 14 anos.** A barra correspondente à faixa etária de 15 a 18 anos é a mais alta, indicando uma quantidade significativa de filhos nessa faixa. As barras para as faixas de 6 a 10 anos e 11 a 14 anos são menores, com a faixa de 6 a 10 anos tendo um número levemente maior de filhos do que a de 11 a 14 anos.

**Análise:** Observa-se uma predominância de filhos mais velhos (adolescentes), o que indica uma forte necessidade de transporte escolar nesse faixa etária. A presença de menos filhos nas faixas de 6 a 10 anos e 11 a 14 anos pode refletir uma diminuição no número de filhos ou uma mudança nas dinâmicas familiares, como um aumento na idade média dos filhos.

**Gráfico 2 – Distribuição de Uso de Serviço de Transporte Escolar:**

**Descrição:** Este gráfico de barras mostra a distribuição de respostas sobre o uso do transporte escolar. A maior parte dos participantes ainda **usa o transporte escolar** ou **usou no passado**, enquanto poucos relataram **nunca terem usado**.

**Análise:** O fato de muitos pais já terem utilizado ou ainda utilizarem o transporte escolar sugere uma familiaridade com o serviço, o que pode facilitar a aceitação de melhorias ou a adesão a um novo sistema de transporte.

**Gráfico 3 – Distribuição de Distância de Casa à Escola:**

**Descrição:** Este gráfico de barras mostra a distribuição da distância da casa à escola. A maioria dos participantes leva **10-20 minutos** para chegar à escola, enquanto grupos menores relataram distâncias **acima de 30 minutos**, **21-30 minutos**, e uma parcela leva **menos de 10 minutos**.

**Análise:** A concentração em uma faixa de 10-20 minutos sugere que uma boa parte dos alunos vive a uma distância moderada da escola, o que pode ser um fator na decisão de usar ou não o transporte escolar.

**Gráfico 4 – Distribuição que Leva para Ir à Escola:**

**Descrição:** Este gráfico de pizza mostra a distribuição de tempos para chegar à escola. A maior parte dos alunos percorre uma distância **superior a 10 km**, seguida por grupos menores que percorrem **5-9 km**, **1-4 km** e **menos de 1 km**.

**Análise:** A maior parcela de alunos percorrendo longas distâncias (acima de 10 km) reforça a importância de um transporte escolar eficiente, já que trajetos longos podem ser mais complicados sem essa opção.

**Gráfico 5 – Distribuição de Problemas com Transporte Escolar:**

**Descrição: O gráfico apresenta as respostas à pergunta "Você já teve problemas com transporte escolar?", com duas barras que representam "Sim" e "Não".** A barra "Não" é mais alta, com mais de 20 respostas, enquanto a barra "Sim" tem aproximadamente 10 respostas.

**Análise:** A maioria dos entrevistados indicou não ter problemas com o transporte escolar, sugerindo uma percepção positiva ou neutra sobre o serviço, enquanto uma minoria relatou dificuldades. A diferença entre as respostas mostra que a experiência com o transporte escolar é majoritariamente favorável.

**Gráfico 6 – Distribuição de preferência de um Transporte mais Seguro e Eficiente:**

**Descrição: O gráfico mostra as respostas à pergunta sobre o uso de um transporte mais seguro e eficiente, com as opções "Sim" e "Não" representadas em duas barras**. A barra "Sim" é consideravelmente mais alta, ultrapassando 20 respostas, enquanto a barra "Não" é mais baixa, com menos de 10 respostas.

**Análise:** Observa-se que a maioria dos entrevistados considera importante ou já adota um transporte mais seguro e eficiente, refletindo uma preferência significativa por melhorias nesse aspecto. A diferença entre as respostas evidencia uma tendência favorável ao uso de transportes que priorizem segurança e eficiência, que é o que nosso aplicativo visa oferecer para os usuários.

**GRÁFICOS E ANÁLISES**

**- Relações entre Colunas**

**Gráfico, Gráfico de barras

Descrição gerada automaticamente**

**Uma imagem contendo Gráfico

Descrição gerada automaticamente**

**Gráfico 7 – Dificuldade em Encontrar Transporte Escolar em Relação à Distância Entre a Casa e a Escola**

**Descrição:** O gráfico exibe a **relação entre a dificuldade em encontrar transporte escolar e a distância** entre a casa e a escola dos alunos. **No eixo horizontal**, temos as categorias de distância: **Mais de 10 km, 5 - 9 km, 1 - 4 km e Menos de 1 km**, enquanto **no eixo vertical** está o número de respostas. **As barras em azul representam aqueles que responderam "Sim" para terem encontrado dificuldade, e as barras em vermelho representam os que responderam "Não"**.

**Análise:** A análise do gráfico revela informações sobre a relação entre a distância de casa à escola e a dificuldade de encontrar transporte escolar. Observa-se que, nas distâncias mais longas (mais de 10 km), a quantidade de respostas afirmativas é consideravelmente maior, sugerindo que quanto maior a distância, maior a dificuldade enfrentada. Já nas distâncias menores, como "Menos de 1 km", a maioria das respostas é negativa, indicando que os pais ou responsáveis têm menos dificuldade em encontrar transporte escolar quando a escola está próxima de casa. Essa tendência pode indicar a necessidade de melhorar as opções de transporte para famílias que vivem mais longe das escolas. ​

**Gráfico 8 – Problemas de Comunicação vs Tempo de Deslocamento**

**Descrição:** O gráfico apresentado é um **histograma que compara problemas de comunicação e horários de transporte, com categorias de tempo de deslocamento.** As categorias são: **Menos de 10 minutos, 10 - 20 minutos, 21 - 30 minutos e Mais de 30 minutos**. O **eixo vertical representa a contagem de respostas**, enquanto o **eixo horizontal mostra as diferentes faixas de tempo**. As barras são coloridas para indicar diferentes respostas à pergunta **"Você já teve problemas com horário e comunicação com o transportador?": azul claro para "Sim", verde claro para "Às vezes" e vermelho para "Não".**

**Análise:** A análise dos dados revela padrões interessantes nas respostas em relação ao tempo de deslocamento. A maioria das respostas "Não" (vermelho) está concentrada nas categorias de tempo de deslocamento mais curto, sugerindo que aqueles que levam menos tempo para transportar seus filhos para a escola tendem a não enfrentar problemas de comunicação. Por outro lado, as respostas "Sim" (azul claro) e "Às vezes" (verde claro) são mais frequentes nas categorias de maior tempo de deslocamento, indicando uma correlação entre um tempo de deslocamento mais longo e a ocorrência de problemas de comunicação. Isso sugere que quanto mais tempo os responsáveis levam para transportar seus filhos, maior é a probabilidade de enfrentarem problemas de comunicação com o transporte escolar. Essa análise pode ser útil para as autoridades escolares ou empresas de transporte ao considerarem melhorias nos serviços.

**CONCLUSÃO**

A análise dos gráficos apresentados revela uma série de insights importantes sobre o uso do transporte escolar, destacando tanto a experiência atual das famílias quanto as áreas que necessitam de melhorias.

**Distribuição de Idade dos Filhos** mostra uma predominância de filhos adolescentes, o que enfatiza a necessidade de um sistema de transporte que atenda eficientemente essa faixa etária. A **distribuição do uso do transporte escolar** indica uma familiaridade com o serviço, sugerindo que os pais estariam abertos a melhorias e novas opções. A **distância entre casa e escola** mostra que a maioria dos alunos leva de 10 a 20 minutos para chegar à escola, um fator que pode influenciar a decisão de usar o transporte escolar.

A análise de **problemas com o transporte escolar** e a **preferência por um serviço mais seguro e eficiente** revela uma percepção majoritariamente positiva do serviço atual, ao mesmo tempo que indica uma demanda por melhorias. O gráfico sobre **dificuldade em encontrar transporte escolar** relaciona a distância com a dificuldade, evidenciando que famílias que vivem mais longe da escola enfrentam maiores desafios, o que reforça a necessidade de soluções mais acessíveis para esses casos.

Por fim, a relação entre **problemas de comunicação e tempo de deslocamento** sugere que quanto maior o tempo de viagem, maior a probabilidade de problemas de comunicação, o que indica uma área clara para intervenção e aprimoramento.

Em conclusão, os dados coletados refletem uma base sólida de satisfação com o transporte escolar atual, mas também destacam oportunidades significativas para melhorias em termos de segurança, eficiência e comunicação, especialmente para famílias que enfrentam desafios devido à distância. A implementação de um sistema de transporte que priorize essas necessidades pode não apenas melhorar a experiência das famílias, mas também contribuir para um ambiente escolar mais seguro e acessível.

**EXPANSÃO DA BASE**

Com a quantidade limitada de dados disponíveis para desenvolver os modelos preditivos, foi necessário implementar uma estratégia de expansão de dados, visando aumentar o volume de informações para o treinamento dos modelos. Inicialmente, partimos dos 33 registros que já havíamos coletado e, a partir deles, realizamos uma expansão que manteve a coesão dos dados reais, permitindo que os padrões fossem reconhecidos pelos algoritmos de maneira eficaz.

**O código de expansão de dados realiza a geração de um conjunto sintético de dados a partir da base original.**

Primeiramente, ele lê a base de dados existente e remove colunas que não são relevantes para a análise, como identificadores e informações pessoais. Em seguida, o código analisa a distribuição das respostas, especificamente focando na preferência dos responsáveis por um serviço de transporte escolar que avise sobre atrasos e mudanças de rota, que foi definida como nossa coluna resposta. Com base nessa distribuição, o código cria um conjunto de dados, garantindo que a quantidade de respostas "Sim" e "Não" fique balanceada, com 150 amostras para cada categoria.

Para isso, uma função é definida para gerar dados sintéticos, onde cada variável é preenchida aleatoriamente com base nas distribuições observadas na base original. Isso inclui informações como a idade dos filhos, a distância entre a casa e a escola, o tempo médio de deslocamento, se os responsáveis possuem carro próprio, o uso anterior de serviços de transporte escolar, a experiência com problemas de transporte, a disposição para mudar para um serviço mais seguro e eficiente, a dificuldade em encontrar transporte escolar e a experiência com horários e comunicação. Após gerar os dados, o código combina os conjuntos de dados que correspondem às respostas "Sim" e "Não" e cria uma nova base expandida contendo 300 registros, com uma distribuição equilibrada entre as duas preferências.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

Finalmente, a base de dados resultante é salva em um arquivo Excel, permitindo que os dados expandidos sejam utilizados posteriormente para modelagem preditiva, contribuindo para a análise de tendências sobre o uso de serviços de transporte escolar.

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Essas estratégias de expansão e balanceamento foram cruciais para viabilizar o desenvolvimento dos modelos preditivos, assegurando que as previsões fossem confiáveis e que o modelo fosse capaz de generalizar bem em diferentes cenários.

**NAIVE BAYES**

**Pré-processamento:**

**Codificação:** Todas as colunas foram codificadas usando LabelEncoder, permitindo a transformação de variáveis categóricas em numéricas.

**Balanceamento de Classes:** O código imprime a proporção das classes para cada coluna, o que ajuda a entender o balanceamento das classes, uma parte crítica na preparação de dados para algoritmos de aprendizado de máquina, especialmente quando se lida com classes desbalanceadas.

**Criação do modelo:**

O modelo foi construído usando o classificador GaussianNB do scikit-learn. Os dados foram divididos em conjuntos de treino (80%) e teste (20%). O código também realiza uma busca em grade (GridSearchCV) para encontrar os melhores hiperparâmetros para o modelo, ajustando o parâmetro var\_smoothing, que é crucial para o desempenho do Naive Bayes.

**- Métricas**

**Acurácia:** **50.0%**

A acurácia é a proporção total de previsões corretas (verdadeiros positivos + verdadeiros negativos) em relação ao total de previsões. O valor de 50% indica que o modelo está fazendo previsões corretas apenas metade das vezes.

**Precisão: 48.28%**

A precisão mede a proporção de verdadeiros positivos em relação ao total de previsões positivas feitas pelo modelo. A precisão de 48.28% significa que, das previsões positivas feitas pelo modelo, menos da metade estavam corretas. Indicando um número considerável de falsos positivos.

**F1-score: 48.28%**

O F1-score é a média harmônica entre a precisão e o recall, sendo uma boa medida quando se busca um equilíbrio entre ambos. O valor de 48.28% indica um desempenho similar em relação à precisão e recall, mas ainda assim sugere que o modelo não está performando bem.

**Recall: 48.28%**

O recall mede a proporção de verdadeiros positivos em relação ao total de positivos reais. O valor de 48.28% sugere que o modelo não está capturando bem os exemplos da classe positiva. O modelo está falhando em identificar quase metade dos casos positivos.

**Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente**

**Verdadeiros Positivos (VP)**: 14 (Classe 1 corretamente prevista)

**Falsos Positivos (FP)**: 15 (Classe 0 incorretamente prevista como Classe 1)

**Falsos Negativos (FN)**: 15 (Classe 1 incorretamente prevista como Classe 0)

**Verdadeiros Negativos (VN)**: 16 (Classe 0 corretamente prevista).

**Interpretação:**

**Verdadeiros Positivos (VP = 14)**: O modelo identificou corretamente 14 instâncias da Classe 1 (a classe positiva). Isso é um bom sinal, pois indica que o modelo é capaz de reconhecer uma quantidade significativa de exemplos da classe de interesse, embora o número ainda possa ser melhorado.

**Falsos Positivos (FP = 15)**: O modelo classificou 15 instâncias da Classe 0 como Classe 1. Isso sugere que o modelo pode estar sendo excessivamente otimista ao prever a Classe 1, o que resulta em um número considerável de previsões incorretas para a classe positiva. Essa situação pode levar a uma diminuição da precisão.

**Falsos Negativos (FN = 15)**: O modelo não conseguiu identificar 15 instâncias da Classe 1, resultando em uma taxa preocupante de exemplos importantes da classe de interesse que não foram reconhecidos. Isso é crítico, especialmente em contextos em que a identificação correta da classe positiva é essencial.

**Verdadeiros Negativos (VN = 16)**: O modelo corretamente classificou 16 instâncias da Classe 0. Isso mostra que o modelo é relativamente eficaz na identificação de exemplos que não pertencem à Classe 1, o que é um ponto positivo para a confiabilidade do modelo ao lidar com a classe negativa.

Em geral, a matriz de confusão indica que o modelo possui um desempenho abaixo do ideal, com um número equilibrado de verdadeiros negativos e positivos, mas enfrenta desafios significativos com falsos positivos e negativos. O aumento na identificação correta de instâncias da Classe 1 e a redução de falsos positivos devem ser focos para melhorias futuras no modelo.

Os resultados obtidos indicam que o modelo de Naive Bayes não está tendo um desempenho satisfatório nos dados fornecidos, com uma acurácia de apenas 50% e um F1-score que sugere um equilíbrio ruim entre precisão e recall. A presença de um número significativo de falsos positivos e negativos indica que o modelo pode não estar capturando bem as características das classes, ou então, que as colunas usadas têm muita aleatoriedade o que dificulta o modelo a encontrar um padrão.

**RANDOM FOREST + NAIVE BAYES**

**Pré-processamento:**

O pré-processamento incluiu a codificação das variáveis categóricas usando LabelEncoder e a remoção de colunas irrelevantes ou com informações excessivas que não contribuiriam para o modelo. Isso é importante para garantir que o modelo se concentre nas características mais relevantes dos dados. O conjunto de dados foi dividido em variáveis independentes (X) e dependentes (y), com a variável de interesse sendo uma questão sobre a preferência por um serviço de transporte escolar. A divisão em conjuntos de treino e teste foi realizada, com 80% dos dados utilizados para treino e 20% para teste.

**Criação do modelo:**

O modelo de Naive Bayes foi inicializado e ajustado aos dados de treino. A busca de hiperparâmetros foi realizada com GridSearchCV para otimizar o parâmetro var\_smoothing, que ajuda a evitar a sobreajuste do modelo.

**- Métricas**

**Acurácia:** **58.33%**

Indica que o modelo acertou aproximadamente 58% das previsões. Embora um valor acima de 50% indique que o modelo tem algum nível de previsão, isso ainda não era suficiente para um modelo confiável em aplicações práticas.

**Precisão: 56.67%**

Mesmo com uma pequena melhoria no valor, a precisão ainda sugere que o modelo pode estar fazendo um número considerável de previsões erradas para a classe positiva.

**F1-score: 57.63%**

Um F1-score de 57.63% mostra que há um equilíbrio moderado entre precisão e recall, mas ainda há espaço para melhoria.

**Recall: 58.62%**

indica que o modelo está identificando uma proporção razoável de exemplos positivos, mas ainda perde um número significativo deles.

**Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente**

**Verdadeiros Positivos (VP)**: 17 (Classe 1 corretamente prevista)

**Falsos Positivos (FP)**: 13 (Classe 0 incorretamente prevista como Classe 1)

**Falsos Negativos (FN)**: 12 (Classe 1 incorretamente prevista como Classe 0)

**Verdadeiros Negativos (VN)**: 18 (Classe 0 corretamente prevista)

**Interpretação:**

**Verdadeiros Positivos (VP = 17)**: O modelo identificou corretamente 17 instâncias da Classe 1 (a classe positiva). Isso é um bom sinal, indicando que o modelo é capaz de reconhecer uma quantidade significativa de exemplos da classe de interesse.

**Falsos Positivos (FP = 13)**: O modelo classificou 13 instâncias da Classe 0 como Classe 1. Isso sugere que o modelo pode ser excessivamente otimista em prever a Classe 1, indicando a necessidade de ajustes para aumentar a precisão.

**Falsos Negativos (FN = 12)**: O modelo não conseguiu identificar 12 instâncias da Classe 1, resultando em um alerta negativo. Isso é um ponto de preocupação, pois indica que o modelo está perdendo exemplos importantes da classe de interesse.

**Verdadeiros Negativos (VN = 18)**: O modelo corretamente classificou 18 instâncias da Classe 0. Isso mostra que o modelo é relativamente eficaz na identificação de exemplos que não pertencem à Classe 1.

Em geral, a matriz de confusão indica que o modelo possui um desempenho razoável, com um bom número de verdadeiros positivos e negativos, mas ainda enfrenta desafios significativos com falsos positivos e negativos, que devem ser abordados para melhorar sua eficácia.

Embora o modelo de Naive Bayes forneça uma base razoável, há uma necessidade clara de melhorar sua performance através de técnicas adicionais e ajustes, garantindo uma melhor identificação dos padrões nos dados.

**ÁRVORE DE DECISÃO**

**Pré-processamento:**

O conjunto de dados original foi carregado a partir de um arquivo Excel. As variáveis independentes (X) foram selecionadas removendo algumas colunas irrelevantes para a predição. Para o pré-processamento, foi utilizado um ColumnTransformer que aplicou OrdinalEncoder para variáveis categóricas e StandardScaler para variáveis numéricas. Após a transformação, o SMOTE foi aplicado para balancear o conjunto de treino.

**Criação do modelo:**

O modelo de árvore de decisão foi implementado utilizando o DecisionTreeClassifier em conjunto com um pipeline para integrar o pré-processamento. A seleção de hiperparâmetros foi otimizada por meio de GridSearchCV com uma busca em uma grade que incluía critérios de divisão, profundidade máxima, e outras características do modelo.

**Métricas –**

**Acurácia: 57.33%**

O modelo apresentou uma acurácia de 57.33%, indicando que ele acerta pouco mais da metade das previsões no conjunto de teste. Esse valor é moderado e pode não ser suficiente para um bom desempenho em aplicações críticas.

**Precisão:** **57.62%**

A precisão ponderada foi de 57.62%, mostrando que o modelo tem um desempenho semelhante à acurácia em termos de evitar falsos positivos, mas ainda comete erros significativos.

**F1-score:** **57.1%**

O F1-score ponderado de 57.1% reflete um equilíbrio entre precisão e recall, mas também confirma que o modelo não é suficientemente robusto para aplicações onde a identificação correta é fundamental.

**Recall: ?**

Embora o valor específico do recall não tenha sido detalhado, é inferido pelo F1-score e pela precisão que o modelo captura uma quantidade moderada de verdadeiros positivos, mas ainda precisa de melhorias.

**Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente**

**Verdadeiros Positivos (VP):** 24 (Classe "Não" corretamente prevista)

**Falsos Positivos (FP):** 19 (Classe "Sim" incorretamente prevista como "Não")

**Falsos Negativos (FN):** 13 (Classe "Não" incorretamente prevista como "Sim")

**Verdadeiros Negativos (VN):** 19 (Classe "Sim" corretamente prevista)

**Interpretação:**

**Verdadeiros Positivos (VP = 24):** O modelo identificou corretamente 24 instâncias da classe "Não", mostrando que ele é eficaz em reconhecer uma quantidade significativa de exemplos desta classe.

**Falsos Positivos (FP = 19):** O modelo classificou incorretamente 19 instâncias da classe "Sim" como "Não", o que indica uma propensão do modelo a classificar erroneamente instâncias da classe positiva como negativa, afetando a precisão.

**Falsos Negativos (FN = 13):** O modelo não conseguiu identificar corretamente 13 instâncias da classe "Não", sinalizando que o modelo está deixando de reconhecer exemplos importantes desta classe, o que afeta o recall.

**Verdadeiros Negativos (VN = 19):** O modelo classificou corretamente 19 instâncias da classe "Sim", evidenciando que ele tem um desempenho razoável em identificar exemplos que realmente pertencem a essa classe.

As métricas sugerem que o modelo de árvore de decisão tem um desempenho moderado, mas não ideal, com todas as métricas em torno de 57%.

Essa análise destaca que, embora o modelo consiga identificar corretamente ambas as classes em várias instâncias, ele ainda comete erros significativos, principalmente na forma de falsos positivos, sugerindo a necessidade de melhorias para aumentar a precisão e o recall.

**CONCLUSÃO FINAL**

Ao analisar os resultados de três modelos de predição diferentes — Naive Bayes, Random Forest + Naive Bayes e Árvore de Decisão — é evidente que todos eles enfrentam desafios significativos, especialmente considerando a qualidade da base de dados utilizada, que não era das melhores.

### Comparação dos Modelos

1. **Naive Bayes**:
   * **Acurácia**: 50%
   * **Precisão**: 48.28%
   * **F1-score**: 48.28%
   * **Recall**: 48.28%
   * O desempenho deste modelo foi insatisfatório, refletindo um equilíbrio ruim entre as métricas. A acurácia indica que o modelo faz previsões corretas apenas metade das vezes. Os falsos positivos e negativos são preocupantes, sugerindo que o modelo não está capturando adequadamente as características das classes, possivelmente devido à aleatoriedade nas colunas utilizadas.
2. **Random Forest + Naive Bayes**:
   * **Acurácia**: 58.33%
   * **Precisão**: 56.67%
   * **F1-score**: 57.63%
   * **Recall**: 58.62%
   * Este modelo apresentou uma leve melhoria em comparação ao Naive Bayes isolado, com um aumento em todas as métricas. No entanto, a precisão ainda é considerada baixa, indicando que, embora o modelo tenha melhorado sua capacidade de identificação de classes positivas, ainda não é confiável para aplicações práticas.
3. **Árvore de Decisão**:
   * **Acurácia**: 57.33%
   * **Precisão**: 57.62%
   * **F1-score**: 57.1%
   * O desempenho geral foi similar ao do modelo Random Forest + Naive Bayes, mas sem a mesma leve melhoria. A árvore de decisão demonstrou um número razoável de verdadeiros positivos, mas também apresentou um alto número de falsos positivos, o que sugere que ajustes são necessários para evitar a sobreposição entre classes.

**Concluíndo**

Embora o uso de diferentes abordagens tenha levado a uma leve melhoria na performance dos modelos, todos apresentaram métricas que indicam uma eficácia abaixo do desejável. A acurácia do Naive Bayes (50%) e o desempenho moderado da Árvore de Decisão (57.33%) e do modelo combinado (58.33%) evidenciam que a identificação correta de padrões nas classes não está sendo realizada de maneira eficaz.

Os resultados sugerem que a qualidade dos dados pode ser um fator crítico para o desempenho dos modelos. Melhorias na coleta de dados e no entendimento das características subjacentes podem permitir a construção de modelos mais robustos, capazes de identificar padrões relevantes e minimizar a ocorrência de falsos positivos e negativos, resultando em um desempenho significativamente melhor.