Classificação do uso de métodos contraceptivos na Indonésia em 1987

1st Caio Rocha Calado *crc@cin.ufpe.br*
2nd Luan Bezerra Coelho *lbc2@cin.ufpe.br*
3rd Luís Felipe Oliveira Costa *lfoc@cin.ufpe.br*
4th Mateus Araújo Neves *man2@cin.ufpe.br*
5th Vinícius Alves Fialho *vaf2@cin.ufpe.br*

***Abstract*—This study proposes an approach to predict the best contraceptive method through statistical analysis of data. Using relevant information about various Indonesian couples from 1987, a predictive model will be developed in Python, incorporating the principles of Naive Bayes classifier and other machine learning techniques. The research aims to provide a valuable tool for identifying the best contraceptive methods based on various influencing factors.**

***Index Terms*—Métodos contraceptivos, Naive Bayes, Modelo Preditor**

I. Introdução

A escolha dos métodos contraceptivos é uma decisão pessoal e complexa que pode ser influenciada por uma ampla gama de fatores socioeconômicos, culturais e religiosos. Dentre esses fatores, a idade, escolaridade, religião, número de filhos e situação profissional do casal desempenham papéis cruciais na determinação do método contraceptivo mais adequado. Cada um desses aspectos molda as percepções e prioridades individuais, afetando diretamente a seleção dos métodos contraceptivos e, por consequência, o planejamento familiar.

Em suma, a escolha dos métodos contraceptivos é uma decisão complexa e multifacetada, refletindo as influências das características individuais, do contexto cultural e das crenças pessoais de cada casal. Entender como esses fatores interagem pode contribuir para o desenvolvimento de políticas de saúde mais efetivas e adequadas às necessidades reais das diferentes populações.

II. Objetivos

Este projeto tem como foco principal prever o método contraceptivo utilizado atualmente por uma pessoa por meio da criação de um classificador ingênuo de Bayes, com base nas suas características, como por exemplo: idade, escolaridade e religião.

III. Justificativa

O desenvolvimento de um classificador para determinar o tipo de anticoncepcional utilizado por mulheres pode trazer benefícios significativos para a saúde reprodutiva e o planejamento familiar. Ao identificar precisamente o método contraceptivo empregado, os profissionais de saúde podem personalizar o tratamento, monitorar possíveis efeitos colaterais e oferecer aconselhamento adequado. Além disso, essa ferramenta pode ser útil em pesquisas de saúde pública, permitindo o desenvolvimento de políticas mais eficazes e promovendo o acesso a métodos contraceptivos seguros e eficientes.

IV. Metodologia

Neste projeto, será abordada a escolha de métodos contraceptivos por mulheres utilizando técnicas de aprendizado de máquina e análise exploratória de dados. Inicialmente, será feita uma análise exploratória para compreender o comportamento das variáveis e suas correlações e distribuições relacionadas ao tema. Adiante, serão aplicadas técnicas de aprendizado de máquina para classificar o uso métodos contraceptivos, considerando informações demográficas e socioeconômicas como parâmetros relevantes. Para isso, será utilizado o algoritmo classificador Naive Bayes. Todo o código para o desenvolvimento deste projeto será implementado na linguagem Python, utilizando o ambiente de desenvolvimento do Google Colaboratory. Para a análise e modelagem de dados, será feito o uso das principais bibliotecas construídas para o *Python*, como *Scipy*, *Numpy*, *Pandas*, *Matplotlib*, *Seaborn* e *Scikit-learn*.

A. Dataset

Para embasar as análises, será utilizado uma base de dados disponível no *UCI Machine Learning Repository*, a qual contém informações sobre 9 tributos, os quais serão selecionados para a análise dos casos.

O *dataset* possui 1473 instâncias e 10 atributos, com a classe inclusa, sendo eles:

1. Idade da esposa: numérico;
2. Grau de educação da esposa: categórico [1, 4], sendo 1 = baixo e 4 = alto;
3. Grau de educação do esposo: categórico [1, 4], sendo 1 = baixo e 4 = alto;
4. Número de filhos: numérico;
5. Religião da esposa: binário (Muçulmana ou não-muçulmana);
6. Ocupação da esposa: binário (Trabalha ou não);
7. Ocupação do esposo: categórico [1, 4];
8. Padrão de vida: categórico [1, 4], sendo 1 = baixo e 4 = alto;
9. Exposição na mídia: binário (Boa ou não);

10. Método contraceptivo utilizado: classe (1 = Não usa, 2 = longo prazo e 3 = curto prazo).

B. Processamento do dataset

Será aplicado um processamento no *dataset* que será dividido em 3 partes:

1. Filtro de Instâncias: Essa é a etapa para remover ou preencher os dados incompletos, mas, como no *dataset* escolhido não há valores faltantes, apenas serão removidas instâncias duplicadas ou irrelevantes para a análise.
2. Seleção de Atributos: Essa é a etapa que devem ser escolhidos os atributos do conjunto de dados que serão utilizados como entradas para os modelos de aprendizado. Esse conjunto de atributos precisa ter uma dimensão razoável, por isso, serão selecionados os atributos relevantes para a análise.
3. Engenharia de atributos: Essa é a etapa de transformação de atributos, de maneira que os tornem prontos para serem entradas dos algoritmos de aprendizagem. Podendo ter técnicas de normalização, one-hot encoding, transformar atributos numéricos em categóricos, vetorização, entre outros.

C. Teorema de Bayes

O Teorema de Bayes é um conceito fundamental na teoria das probabilidades e estatística, o qual determina a probabilidade de um evento acontecer, dado uma informação prévia que pode estar relacionada a este evento. Essa fórmula matemática, atribuída ao matemático e teólogo Thomas Bayes (1702-1761), é amplamente aplicada em diversos campos, incluindo aprendizado de máquina, inteligência artificial, medicina, engenharia e muitos outros.

O teorema de Bayes é representado pela seguinte fórmula:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1}$$

onde:

- $P(A|B)$ é a probabilidade de que o evento A ocorra dado que o evento B ocorreu;
- $P(B|A)$ é a probabilidade de que o evento B ocorra dado que o evento A ocorreu;
- $P(A)$ é a probabilidade a priori de que o evento A ocorra antes de observar qualquer evidência;
- $P(B)$ é a probabilidade a priori de que o evento B ocorra antes de observar qualquer evidência.

O teorema de Bayes permite atualizar as probabilidades iniciais (ou crenças) sobre um evento A com base na nova evidência B que é observada. Essa atualização é especialmente útil quando deve-se fazer inferências ou previsões em situações nas quais se tem informações prévias sobre um evento e, em seguida, obtemos novos dados relevantes.

Uma aplicação comum do Teorema de Bayes é em classificadores probabilísticos, como o já mencionado Classificador *Naive Bayes*. Nesse contexto, o teorema de Bayes é usado para calcular a probabilidade de uma instância pertencer a uma determinada classe com base em suas características (atributos) observadas.

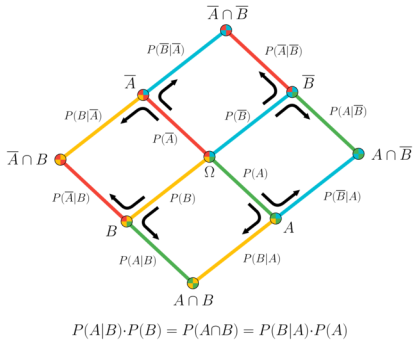

Fig. 1. Teorema de Bayes.

D. Classificador de Bayes

O classificador *Naive Bayes* é um modelo de aprendizagem de máquina que utiliza do Teorema de Bayes para realizar uma classificação probabilística calculada com o teorema. *Naive*, do inglês, significa ”ingênuo”, isso se dá pela suposição que é feita a respeito das variáveis do problema: elas são independentes. Ou seja, os atributos não são correlacionados entre si.

O classificador *Naive Bayes* é amplamente escolhido devido à sua simplicidade, rapidez e facilidade de implementação, garantindo resultados precisos. Em comparação com outros classificadores, seu desempenho é notavelmente superior. Uma de suas grandes vantagens é que funciona bem mesmo com uma pequena quantidade de dados, tornando-o ideal para a classificação de atributos discretos.

Esse método encontra aplicação em várias áreas, incluindo análise de crédito, onde pode ajudar na avaliação do risco de um cliente. Além disso, é útil na análise de texto, permitindo identificar palavras-chave relevantes por meio da frequência de palavras usadas. Isso o torna uma ferramenta valiosa para a classificação de *emails* como *spam*.

Outro campo onde o classificador de *Naive Bayes* se destaca é na área médica, sendo aplicado em diagnósticos. Sua eficiência em lidar com dados discretos o torna uma escolha confiável na busca por falhas em sistemas mecânicos.

Neste projeto, foi utilizado o classificador de Bayes da biblioteca *Scikit-learn* para classificar o método contraceptivo escolhido dado o contexto socioeconômico, filtrando atributos relevantes para a obtenção dessas informações. Dois modelos foram testados durante a pesquisa, o Naive Bayes categórico e o gaussiano:

1) Naive Bayes Gaussiano: O classificador Naive Bayes Gaussiano é um algoritmo de aprendizado de máquina usado para classificar dados em categorias com base em atributos numéricos, assumindo que esses atributos seguem uma distribuição gaussiana (normal). Ele calcula a probabilidade de um dado conjunto de atributos pertencer a cada classe usando o Teorema de Bayes, considerando médias e desvios padrão dos atributos em cada classe. Em seguida, atribui o conjunto

de atributos à classe com a maior probabilidade condicional calculada. Embora o ”Naive” indique uma suposição de independência entre os atributos, o Naive Bayes Gaussiano é eficaz em muitas situações, especialmente quando há muitos atributos, embora essa suposição nem sempre seja verdadeira na prática.

2) Naive Bayes Categórico: O classificador Naive Bayes Categórico é uma variação do algoritmo Naive Bayes que lida com dados categóricos, ou seja, atributos que têm valores discretos em categorias. Ele opera calculando as probabilidades condicionais de um conjunto de atributos pertencer a cada classe com base na frequência de ocorrência de cada categoria nos dados de treinamento. O ”Naive” na nomenclatura refere-se à suposição de independência condicional entre os atributos, o que simplifica o cálculo das probabilidades. Durante a classificação, o modelo multiplica as probabilidades condicionais de cada atributo dado a classe e as probabilidades prévias das classes, selecionando a classe com a maior probabilidade final como a predição.

E. Aplicação

A implementação do classificador Naive Bayes é conduzida utilizando a linguagem Python e o ambiente Google Colab. Também é empregada a biblioteca Scikit-learn, amplamente reconhecida por suas ferramentas em aprendizado de máquina e mineração de dados, incluindo o Naive Bayes.

Para realizar a análise dos dados e as operações matemáticas, são utilizadas as bibliotecas Pandas, NumPy e SciPy, uma vez que elas se integram perfeitamente com o Scikit-learn. O Pandas permite a manipulação e organização dos dados em formato de DataFrame, tornando o pré-processamento uma tarefa mais fluída e organizada. O NumPy é empregado para realizar operações matemáticas e cálculos numéricos, proporcionando uma base sólida para diversas funcionalidades. Além disso, o SciPy complementa o NumPy com funcionalidades adicionais para análise de dados e estatísticas.

Além dessas ferramentas essenciais, é importante mencionar que o projeto se beneficia do uso das bibliotecas Seaborn e Matplotlib. A biblioteca Matplotlib oferece recursos poderosos para criar uma ampla variedade de gráficos e plots, permitindo a representação visual das distribuições e tendências dos dados. Por sua vez, a biblioteca Seaborn enriquece essa visualização com paletas de cores atraentes e funções específicas para gráficos estatísticos, tornando a exploração de dados mais detalhada e informativa.

Ao utilizar todas essas bibliotecas em conjunto, o ambiente de desenvolvimento se torna robusto e eficiente para implementar o classificador Naive Bayes, ao mesmo tempo em que facilita a exploração e visualização dos dados. A combinação dessas ferramentas permite explorar todo o potencial do algoritmo Naive Bayes e obter valiosos insights a partir dos dados analisados.

Em resumo, a utilização das bibliotecas Scikit-learn, Pandas, NumPy, SciPy, Seaborn e Matplotlib, em conjunto com a linguagem Python e o ambiente Google Colab, garantirá uma implementação eficiente e precisa do classificador Naive Bayes, além de proporcionar uma experiência de desenvolvimento mais fluida e facilitada.

O projeto incluirá uma etapa essencial de análise e tratamento dos dados, garantindo a qualidade e relevância das informações utilizadas. Em seguida, o dataset será dividido em duas partes distintas: o conjunto de treinamento e o conjunto de teste. O conjunto de treinamento será utilizado para alimentar o modelo de classificação, permitindo que ele aprenda e se ajuste aos padrões presentes nos dados. Por sua vez, o conjunto de teste será reservado exclusivamente para avaliar a eficiência do classificador Naive Bayes, sendo essencial para medir sua capacidade de generalização em novos dados não vistos durante o treinamento.

Essa abordagem de dividir os dados em treino e teste é fundamental para evitar o sobreajuste (overfitting) do modelo, garantindo que ele não memorize os dados de treinamento, mas, sim, aprenda padrões que possam ser aplicados de forma eficiente em novas amostras. Dessa maneira, é possível obter uma estimativa realista do desempenho do classificador em situações práticas.

Ao avaliar a eficiência do classificador usando o conjunto de teste, poderemos medir sua acurácia, precisão, recall e outras métricas de desempenho. Essas métricas fornecerão detalhes sobre a capacidade do modelo de fazer previsões precisas e auxiliarão na tomada de decisões.

V. Análise exploratória

A análise exploratória desempenha um papel fundamental na implementação de algoritmos preditivos, desdobrando-se como uma fase primordial para compreender profundamente os dados antes de mergulharmos no processo de modelagem. Esta etapa é essencial para garantir a qualidade e a confiabilidade dos dados, além de proporcionar insights valiosos acerca de tendências, padrões e correlações ocultas.

O cerne dessa análise consiste em uma clara e minuciosa investigação da distribuição dos dados, frequentemente auxiliada por meio de representações gráficas, com o intuito de compreender como cada parâmetro pode influenciar a variável de interesse. Dada a natureza diversificada dos conjuntos de dados, diversas abordagens podem ser adotadas na análise exploratória. Em muitos casos, opta-se por uma análise univariada, a qual examina individualmente cada variável. Tal análise nos permite extrair informações cruciais, tais como a média, a mediana, o desvio padrão, os valores mínimo e máximo, bem como a frequência dos valores associados a uma variável específica. Esses dados estatísticos revelam-se essenciais para obtermos uma compreensão profunda dos dados em questão.

No âmbito das variáveis em análise, identificam-se duas categorias distintas: variáveis numéricas e variáveis categóricas, cada qual apresentando suas próprias características específicas. As variáveis numéricas, conforme o próprio termo sugere, são representadas por valores numéricos. Uma classificação adicional relevante entre essas variáveis numéricas é a diferenciação entre variáveis numéricas contínuas e discretas.

No conjunto de dados sob consideração, destacam-se as variáveis numéricas discretas, caracterizadas por valores inteiros. Isso é observado, por exemplo, na idade da esposa e no número de filhos, ambas representadas por valores inteiros do tipo ”int”.

Por outro lado, as variáveis categóricas englobam todas as outras categorias e podem ser binárias, ou seja, apresentarem apenas duas classes distintas, ou contar com três ou até quatro classes diferentes. Essas variáveis categóricas desempenham um papel fundamental na análise, uma vez que capturam informações qualitativas e diferentes categorias que enriquecem a compreensão dos dados, podendo ter um impacto significativo na modelagem e interpretação dos resultados.

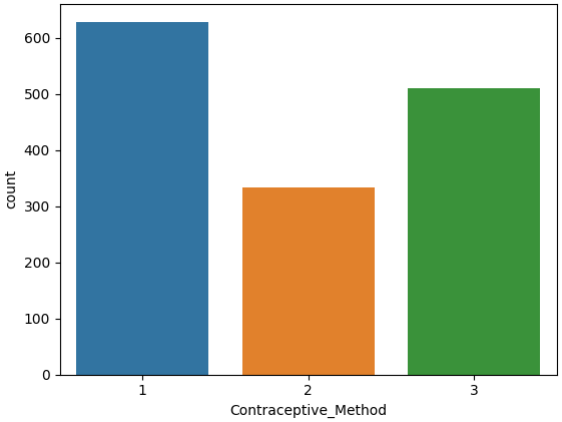

Fig. 2. Distribuição do target.

Realizou-se uma contagem das linhas associadas a cada classe (1, 2 ou 3) da variável target, com o objetivo de avaliar o equilíbrio do dataset. Verificou-se que o dataset já estava balanceado, o que dispensou a necessidade de realizar quaisquer ajustes adicionais de balanceamento.

Com o objetivo de aprofundar a compreensão, foi gerado um gráfico para cada variável, relacionando o número de ocorrências de cada classe do target com o atributo correspondente. Essa análise visa identificar a relevância do atributo e compreender de que forma ele efetivamente influencia o target.

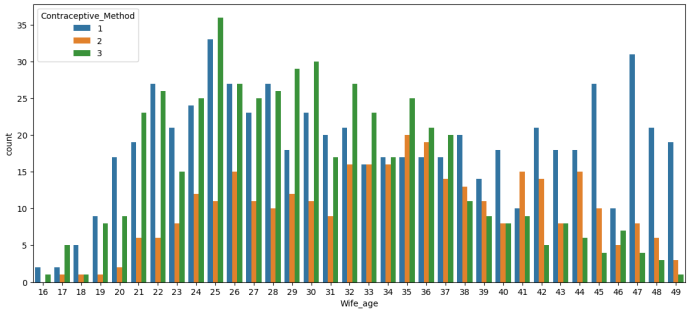

Fig. 3. Distribuição de wife_age.

Conforme evidenciado no gráfico, observa-se que na faixa etária compreendida entre os 24 e os 35 anos, prevalece a utilização de métodos contraceptivos de curto prazo. Posteriormente, a maioria das mulheres opta por não utilizá-los mais.

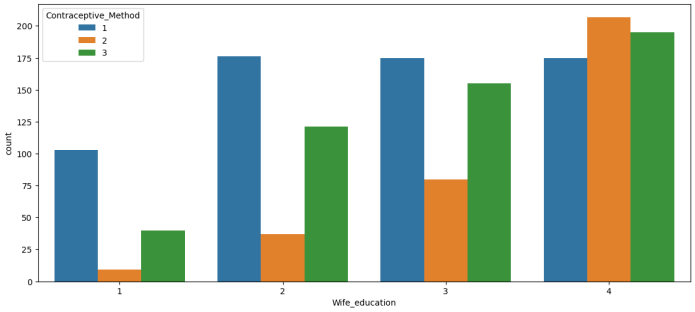

Fig. 4. Distribuição de wife_education.

É evidente que, à medida que o nível de educação aumenta, observa-se um significativo aumento na utilização de métodos contraceptivos, destacando-se a predominância do uso de métodos de longo prazo no mais auto grau.

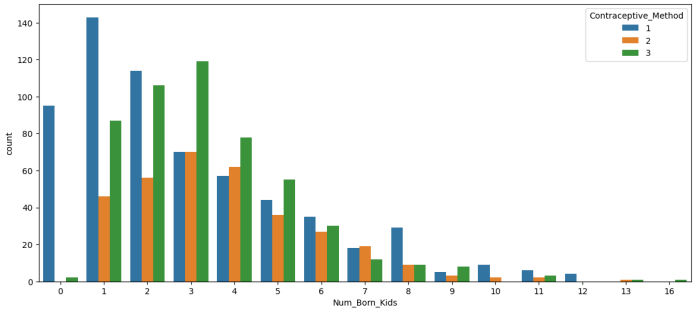

Fig. 5. Distribuição de Num_Born_Kids.

Apenas a partir do quarto filho, nota-se um aumento na utilização de métodos contraceptivos por parte das mulheres, sugerindo a existência de uma cultura que valoriza a maternidade.

Esses são os gráficos para cada variável restante, que possuem menor relevância que as 3 primeiras:

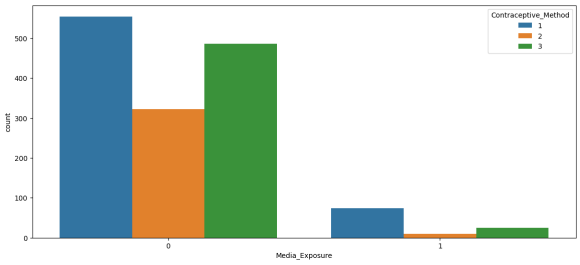

Fig. 6. Distribuição de Media_Exposure.

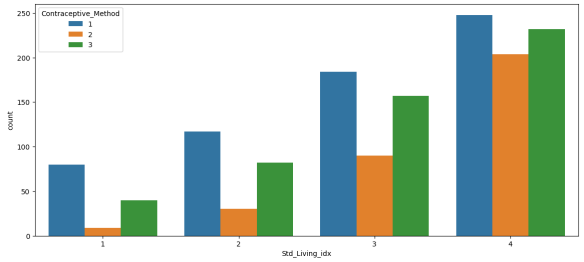

Fig. 7. Distribuição de Std_Living_idx.

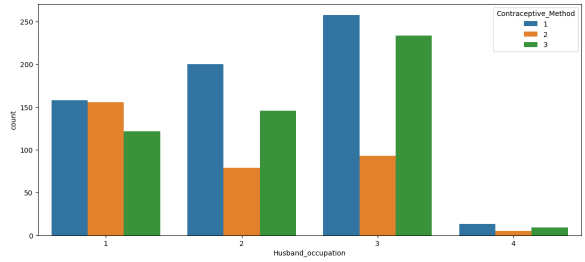

Fig. 8. Distribuição de Husband_occupation.

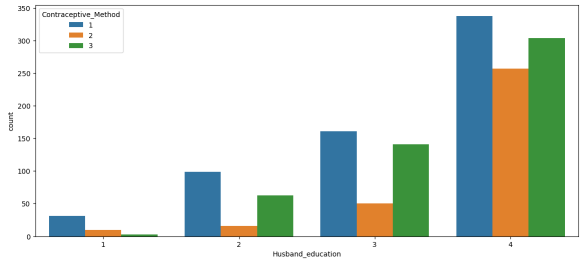

Fig. 9. Distribuição de Husband_education.

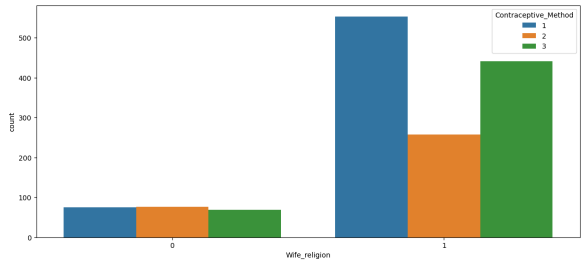

Fig. 10. Distribuição de Wife_religion.

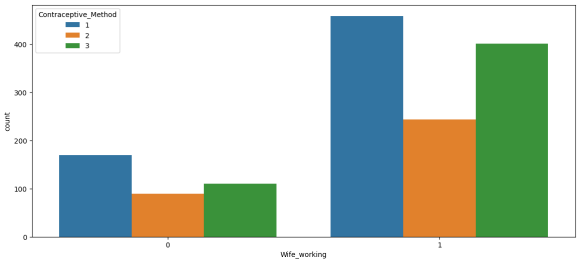

Fig. 11. Distribuição de Wife_working.

Mediante a análise de cada atributo, chegou-se à conclusão de que, embora alguns deles tenham se revelado altamente significativos, ficou evidente a necessidade de empregar um método técnico mais objetivo para identificar as variáveis de maior relevância. Para atender a essa necessidade, optou-se pela aplicação do teste Qui-Quadrado.

O teste Qui-Quadrado, ou teste do qui-quadrado de Pearson, é uma técnica estatística utilizada para avaliar a independência entre variáveis categóricas em uma tabela de contingência. Ele compara as frequências observadas com as frequências esperadas, permitindo determinar se existe uma associação estatisticamente significativa entre as variáveis. Em outras palavras, o teste Qui-Quadrado ajuda a identificar se a presença ou ausência de uma variável está relacionada de forma significativa com a presença ou ausência de outra variável.

O processo envolve o cálculo do valor Qui-Quadrado, que é uma medida da diferença entre as frequências observadas e as frequências esperadas sob a hipótese nula de independência das variáveis. Quanto maior for o valor Qui-Quadrado, mais evidente é a associação entre as variáveis.

No contexto do projeto, o teste Qui-Quadrado foi aplicado para determinar quais variáveis têm uma relação estatisticamente significativa com a variável de interesse, auxiliando na seleção das variáveis mais relevantes para a modelagem. Essa abordagem técnica e objetiva contribuiu para aprimorar a escolha das variáveis a serem incluídas no modelo preditivo, garantindo maior confiabilidade e precisão nas análises.

A implementação foi realizada com auxilio da biblioteca Scikit-Learn, que provê funções para realizar o teste Qui-Quadrado. Com isso, foi concluído que os atributos mais relevantes eram a idade da mulher, o grau de educação da mulher, o número de filhos e a exposição à mídia, em ordem.

Specs	Score
Wife_age	132.680281
Wife_education	45.646138
Num_Born_Kids	45.128054
Media_Exposure	29.235977
Std_Living_idx	18.399817
Husband_occupation	18.136128
Husband_education	9.483911
Wife_religion	3.229343
Wife_working	1.299431

Fig. 12. Teste qui quadrado.

VI. Resultados e discussão

Após a seleção dos parâmetros mais apropriados para um conjunto predefinido, com foco na maximização da precisão, procedemos à utilização da função ”classification_report” disponível na biblioteca Scikit-Learn para a visualização dos resultados.

Primeiramente, o modelo foi testado com as 3 melhores features, segundo o teste do qui quadrado, utilizando diferentes distribuições para treino e teste. Com 30% para teste foi obtida uma acurácia média de 0.542, com 20% foi obtido 0.544 e com 15% foi obtido 0.561. Em seguida, o modelo foi testado

utilizando as 4 melhores features, segundo o teste do qui quadrado, e utilizando 20% do dataset para teste foi obtida uma acurácia média de 0.552. Abaixo temos o resultado de um dos experimentos utilizado para calcular a acurácia média:

```
Accuracy: 0.5728813559322034
Classification Report:
              precision    recall  f1-score   support

           1       0.69      0.64      0.66       132
           2       0.55      0.36      0.43        64
           3       0.47      0.63      0.54        99

    accuracy                           0.57       295
   macro avg       0.57      0.54      0.54       295
weighted avg       0.59      0.57      0.57       295
```

Fig. 13. Avaliação do modelo.

Para compreender melhor os resultados, foi projetada uma Matriz de Confusão, com auxilio da biblioteca Scikit-Learn, para uma visualização gráfica melhor. A matriz de confusão permite que avaliar o desempenho do modelo de machine learning de maneira mais detalhada do que métricas simples, como a acurácia. Além disso ajuda a identificar os tipos de erros que o modelo está cometendo. Podendo ser identificados quantos falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos o modelo está produzindo. Isso é valioso para entender quais classes o modelo tem dificuldade em prever corretamente e onde o modelo está tendo problemas. A matriz de confusão é usada para calcular várias métricas de desempenho, como precisão, recall, F1-score e matriz de confusão balanceada (para classes desequilibradas). Essas métricas são mais informativas do que a acurácia, especialmente quando as classes têm tamanhos diferentes ou quando os custos de falsos positivos e falsos negativos são diferentes.

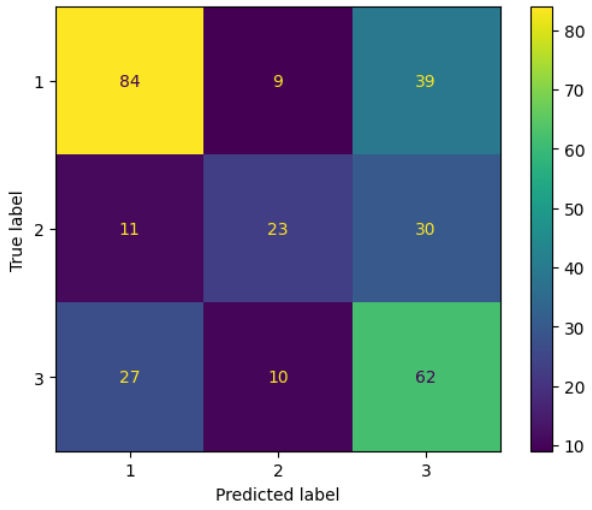

Fig. 14. Avaliação do modelo.

O código pode ser acessado através do link: https://github.com/luanbezerra/Projeto-de-Estatistica.

VII. Conclusão

Após uma análise detalhada, ficou evidente que o modelo utilizado inicialmente não produziu resultados satisfatórios. Esse cenário pode ser atribuído a duas principais causas. Primeiramente, a complexidade inerente aos dados em questão excede a capacidade do modelo atual de lidar eficazmente com ela. Em segundo lugar, foi identificada a ausência de informações críticas que teriam um impacto positivo significativo nos resultados, mas que não foram consideradas no modelo atual.

Uma questão adicional a ser destacada é a possível existência de dependências entre os dados que o modelo atual não leva em consideração. Isso sugere que há relações ocultas e interdependências entre as variáveis do conjunto de dados que não foram adequadamente capturadas pela abordagem atual.

Apesar desses desafios iniciais, nossa pesquisa demonstrou que ainda existe uma promissora linha de investigação. A possibilidade de melhorar a adequação do conjunto de dados com um modelo ainda não testado permanece válida. Isso significa que, ao explorar modelos alternativos e incorporar informações ausentes, há uma oportunidade real de aprimorar nossos resultados e alcançar maior precisão em nossa análise.

References

[1] Contraceptive Method Choice
https://archive.ics.uci.edu/dataset/30/contraceptive+method+choice

[2] Naive Bayes: Como funciona esse algoritmo de classificação
https://blog.somostera.com/data-science/naive-bayes

[3] Machine Learning: Conceitos e Modelos
https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445

[4] What is exploratory data analysis?
https://www.ibm.com/topics/exploratory-data-analysis

[5] Bayes Theorem
https://www.sciencedirect.com/topics/engineering/bayes-theorem

[6] National Indonesia Contraceptive Prevalence Survey 1987
https://dhsprogram.com/pubs/pdf/SR9/SR9.pdf

[7] Feature Selection Techniques in Machine Learning with Python
https://scikit-learn.org/stable/modules/naive_bayes.html

[8] Classificador Naive Bayes na biblioteca Scikit Learn
https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e