HELENA PATO MAGALHÃES

RELATÓRIO DO TRABALHO PRÁTICO 3 DE MINERAÇÃO DE DADOS

Classificação

Professor: Dr. Wagner Meira Jr.

1 INTRODUÇÃO

A mortalidade infantil é um importante indicador de saúde e qualidade de vida. Definida como a taxa de óbitos de crianças menores de cinco anos de idade a cada 1000 nascidas vivas, um valor elevado evidencia condições de vida precárias e baixo nível de desenvolvimento. Nesse contexto, a ONU espera que, em 2030, os países acabem com as mortes preveníveis de recém-nascidos e crianças menores de cinco anos de idade.

Dado que grande parte dos óbitos ocorrem em situações de poucos recursos, a cardiotocografia (CTG) é um exame simples e acessível para avaliar a saúde fetal. A análise de seus resultados permite que profissionais da saúde ajam para prevenir a mortalidade infantil e materna.

2 MOTIVAÇÃO

Considerando o contexto apresentado, pretendemos utilizar os dados coletados de 2126 registros de exames CTG, em que a saúde fetal foi categorizada em três classes: normal, suspeita e patológica, para treinar um modelo de classificação que seja capaz de prever o estado de saúde da criança através dos resultados do exame. Dessa forma, contribuímos para a atuação dos profissionais de saúde e para a prevenção da mortalidade infantil.

3 OBJETIVO

Com este trabalho, visamos utilizar técnicas de classificação para prever a saúde fetal, de acordo com os resultados do exame CTG, a fim de identificar quais estão com saúde suspeita ou patológica. Idealmente, o modelo gerado deve ter boa acurácia, tanto no treino quanto na validação, além de ter uma matriz de confusão com poucas classificações errôneas e ser robusto a variações nos dados. Essas condições irão permitir que determinemos se as previsões geradas são interessantes para as análises propostas.

4 METODOLOGIA

Seguimos uma adaptação da metodologia definida no CRISP-DM. Inicialmente, definimos um domínio em que se aplicassem técnicas de agrupamento, então obtivemos dados que o representassem. Depois disso, estudamos o negócio, definindo seus objetivos, avaliando os recursos disponíveis, definindo as metas da mineração de dados e produzindo o plano de projeto.

Então, passamos para o entendimento dos dados. Nessa etapa fizemos uma caracterização das informações obtidas, investigando os campos existentes na tabela, os valores que pode assumir, sua distribuição e correlações entre os atributos. Com isso, definimos a qualidade dos dados. A seguir, preparamos a base para a classificação, de acordo com a necessidade definida na etapa anterior, dimensionamos os dados, reduzimos sua dimensionalidade e tratamos seu desbalanceamento. Na modelagem, testamos diferentes algoritmos e formas de balancear os dados e analisamos as métricas geradas por eles, a fim de escolher o modelo com os melhores resultados.

Avaliamos o modelo final usando dados de teste, medimos as métricas citadas anteriormente, como acurácia, precisão, revocação e F1, o que nos permitiu definir se o modelo foi capaz de generalizar a distribuição dos dados. Dessa forma, chegamos a uma conclusão para a etapa de desenvolvimento. Depois disso, esquematizamos os resultados em um relatório, compilando todas as etapas do processo.

5 DESENVOLVIMENTO

5.1 Trabalhos relacionados

No site Kaggle, é possível ter contato com trabalhos desenvolvidos por outras pessoas em cima da mesma base de dados através de sua página. Em relação à base de dados “Fetal Health Classification”, podemos citar o estudo feito por (ABDELKADER; MAHMOUD, 2022) que fornece uma visão interessante dos dados por diversos ângulos. Nele, os dados são explorados e posteriormente, são usados diversos algoritmos de classificação para prever seus rótulos.

É um trabalho enriquecedor, do qual foi possível aprender técnicas que ajudaram na elaboração deste estudo. Porém, ele deixa a desejar no quesito do tratamento do balanceamento

dos dados. O procedimento usado para que os dados ficassem com proporções mais próximas foi uma simples amostragem aleatória das instâncias de saúde normal em uma menor quantidade. É uma abordagem de undersampling muito simples que não considera a representatividade da amostra.

5.2 Entendimento dos Dados

As seções 5.2 a 6 foram feitas utilizando o Google Collaboratory, portanto, para ver todo o código desenvolvido, acesse o link: <https://colab.research.google.com/drive/1c4FmXNqqVYZ7vw6JpOWSux4hnLK5QCs4?usp=sharing>.

A base utilizada para desenvolver este trabalho foi "Fetal Health Classification", composta por 2126 linhas e 22 colunas, em que cada linha representa um exame CTG e cada coluna contém os resultados gerados por esse exame. Os atributos são todos numéricos, com exceção da classe que rotula cada exame. Parte dos atributos se refere a características do histograma gerado pelo exame.

Não há valores nulos entre os dados. As classes se dividem em 1655 instâncias de saúde normal, 295 suspeita e 176 patológica, portanto observamos que os dados são severamente desbalanceados, o que pode ser visualizado na Figura 1, que ilustra os histogramas das distribuições de cada atributo. Na figura, também podemos notar que muitas das distribuições não seguem uma normal e que cada atributo tem grandezas bastante discrepantes.

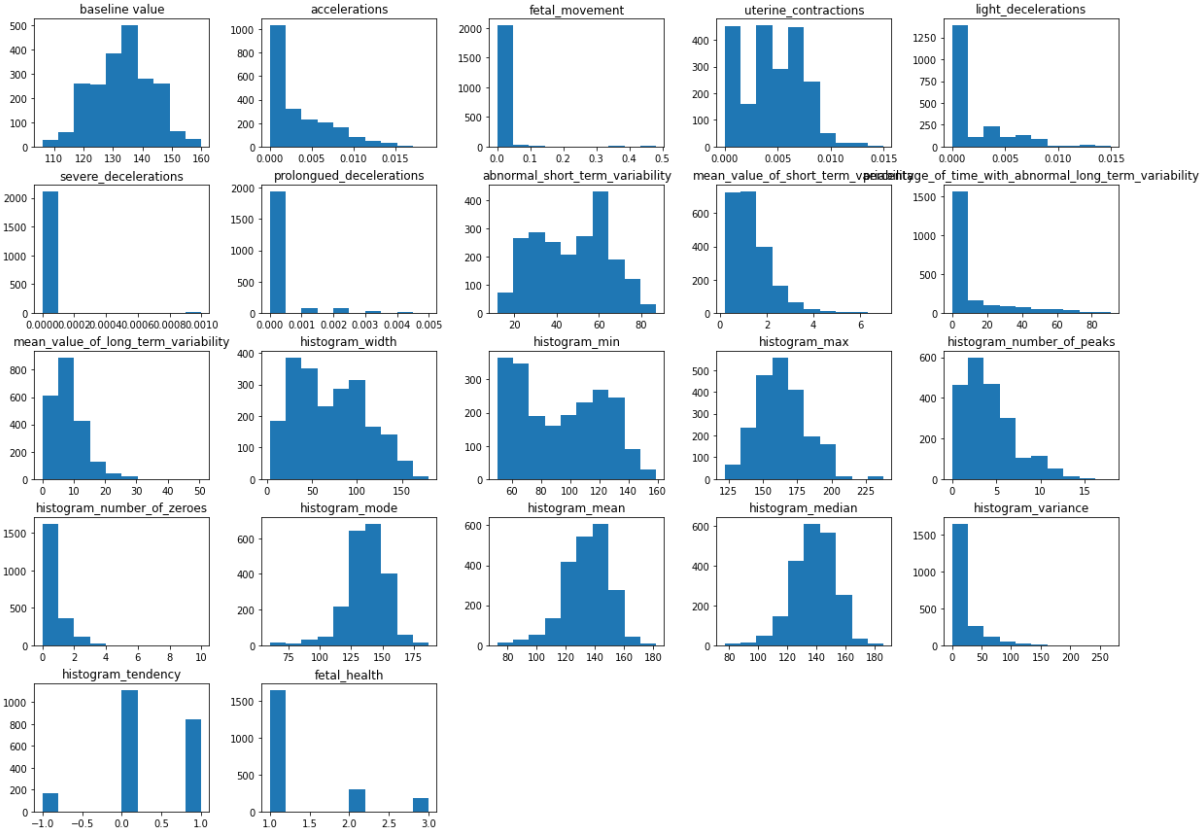


Figura 1 - Histogramas das distribuições dos atributos

Para observar as correlações entre atributos, criamos um gráfico ilustrando esses valores para cada par de colunas dos dados. Através dele, constatamos que não existem muitos atributos correlacionados, e que aqueles que têm um valor mais significativo não fornecem nenhum insight dos dados. Por exemplo, podemos ver uma alta correlação entre a média, mediana e moda do histograma, mas essa relação é óbvia. Além disso outros valores que podem chamar um pouco a atenção se referem a outras medidas do histograma, que não são muito esclarecedoras.

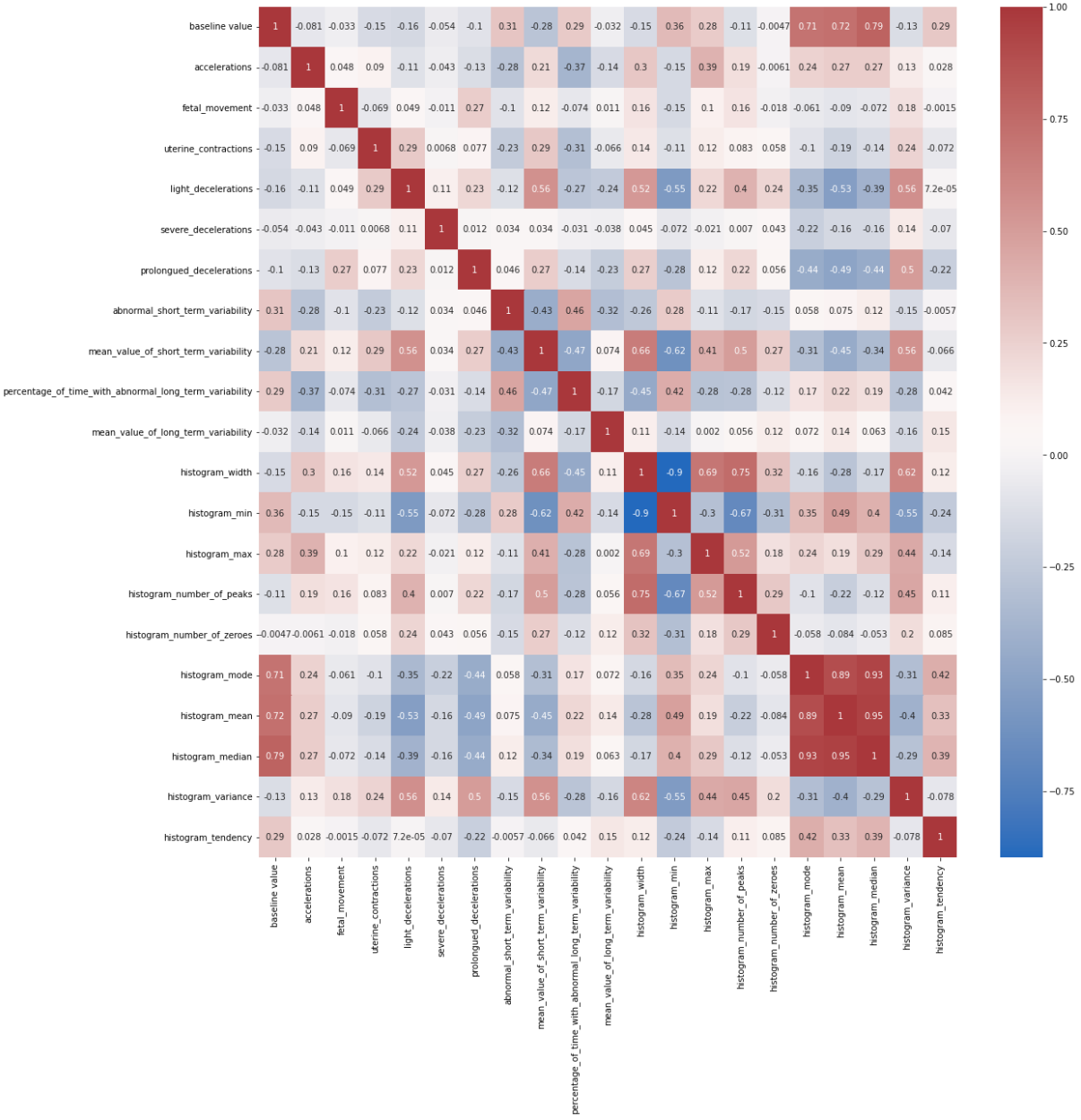


Figura 2 - Correlações entre os atributos

Após essa análise, constatamos que os dados têm uma qualidade interessante para as análises propostas. Fomos capazes de compreender melhor suas características por meio de visualizações. Isso nos levou a concluir que será necessário realizar processamentos nos dados antes que sejam fornecidos a um modelo de classificação.

5.3 Preparação dos Dados

Como os dados eram advindos do site Kaggle, eles já estavam bem formatados. Além disso, não havia valores nulos. Porém, dada a distribuição irregular e as diferenças nas grandezas dos atributos, foi necessário fazer o redimensionamento dos dados. Usamos o método

RobustScaler do ScikitLearn, pois ele faz o redimensionamento dos dados usando estatísticas que são robustas a outliers, ao contrário do método MinMax, por exemplo.

Em seguida, fizemos uma redução de dimensionalidade nos dados, usando a técnica do PCA. Definimos que o algoritmo deveria preservar 99% da variância, e com isso obtivemos 7 componentes principais. Essa é uma redução significativa, em que ficamos com um terço dos atributos iniciais. Com eles, realizamos a transformação da base, a fim de manter somente esses elementos mais relevantes.

Como ilustrado na seção anterior, os dados são desbalanceados, portanto, seria necessário realizar algum processamento para adequá-los. Dessa forma, decidimos fazer um experimento usando diferentes técnicas de balanceamento e comparar os resultados com o uso dos dados originais. Para isso, usamos o método SMOTE para balancear por oversampling. Essa técnica usa interpolação para gerar pontos sintéticos a partir dos dados observados. Já para o undersampling usamos o método ClusterCentroids, que reduz as amostras usando centroides representativos da distribuição. Ambos os métodos provêm da biblioteca ImbalancedLearn.

Após esses processamentos, dividimos os dados em conjuntos de treino, validação e teste nas proporções de 8:1:1. Os dados de treino serão fornecidos ao algoritmo, enquanto os de validação serão usados para comparar os algoritmos e métodos de balanceamento. Já os dados de teste somente serão usados na etapa final de avaliação.

5.4 Modelagem

Para a tarefa de classificação selecionamos dois algoritmos. O primeiro foi a Árvore de Decisão, em que definimos a entropia como métrica para o cálculo do ganho de informação, e não definimos um limite na profundidade da árvore. O segundo foi a Floresta Aleatória, que, apesar de usar árvores, tem uma lógica bastante distinta, por ser um ensemble. Nele também usamos a entropia como métrica, mas dessa vez limitados a profundidade das árvores para 2 níveis e definimos um número de 50 classificadores usados. Após instanciar os modelos, os treinamos com os dados de treino variando os tipos de balanceamento. Para cada variação, calculamos as métricas mostradas na Tabela 1 e desenhamos as matrizes de confusão, na Tabela 2.

Tabela 1 - Métricas em relação ao algoritmo e balanceamento

Algoritmo	**Árvore de Decisão**			**Floresta Aleatória**		
Balanceamento	**Original**	**Oversampling**	**Undersampling**	**Original**	**Oversampling**	**Undersampling**
Profundidade	21	18	13	2	2	2
Acurácia treino	0.999	1.0	1.0	0.812	0.800	0.782
Acurácia validação	0.836	0.895	0.755	0.751	0.806	0.774
Acurácia média da validação cruzada	0.751	0.829	0.758	0.760	0.805	0.716
Precisão	0.845	0.898	0.760	0.659	0.809	0.798
Revocação	0.836	0.895	0.755	0.751	0.806	0.774
F1	0.838	0.895	0.756	0.652	0.807	0.776

Em uma análise geral, podemos observar que a acurácia é mais alta nos dados de treino, o que é intuitivo, mas isso não impede que, em alguns casos, ela não seja elevada nem mesmo nessas condições, como no caso de floresta aleatória com undersampling. Geralmente, o valor cai quando usamos os dados de validação, e diminui mais ainda ao usarmos a validação cruzada. Os modelos que conseguem manter uma acurácia alta na validação cruzada, portanto, são preferíveis.

Outro aspecto que pode ser observado em uma escala geral é a proximidade dos valores de precisão, revocação e F1 de cada modelo. Esse fenômeno indica que o número de falsos positivos e falsos negativos é similar. Ou seja, em cada classe, a quantidade de instâncias erroneamente classificadas é muito próximo.

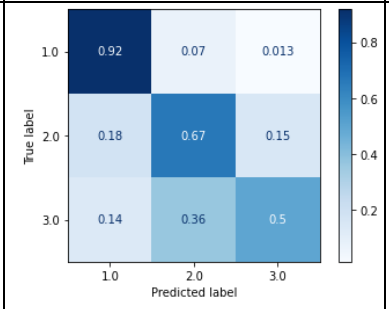
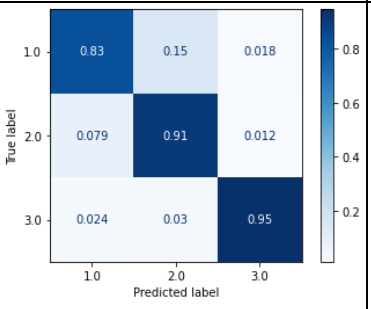
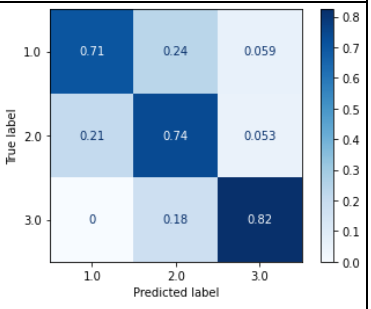
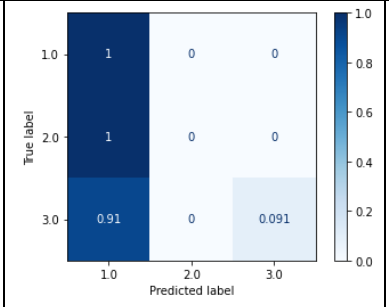
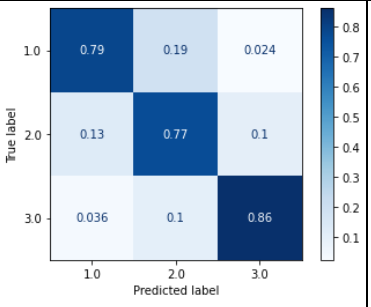
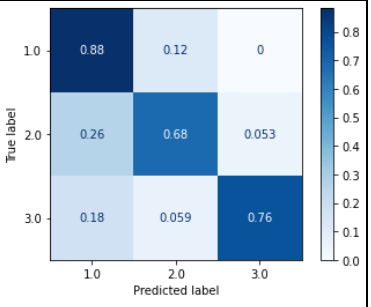
Também podemos notar que o algoritmo de árvore de decisão teve um desempenho melhor do que a floresta aleatória, independentemente do tipo dos dados. Isso é evidenciado pelos maiores valores de acurácia e diagonais mais bem definidas na matriz de confusão. Podemos deduzir então que, nesse caso, o modelo mais simples beneficiou os resultados.

Partindo para análises mais específicas, ao observar os resultados obtidos usando os dados originais, podemos ver que apesar de valores razoáveis de acurácia, ao inspecionar a matriz de confusão podemos ver o fenômeno causado pelo desbalanceamento dos dados. No caso da árvore de decisão, muitas instâncias de saúde patológica foram classificadas como suspeita. Já na floresta aleatória o desbalanceamento fica ainda mais evidente, dado que quase a totalidade das instâncias foi classificada como saúde normal dada, a prevalência dessa classe.

Ao tratar o desbalanceamento usando undersampling, podemos ver uma pequena melhora nesses erros de classificação, na matriz de confusão. Ainda há erros trocando principalmente os rótulos 1 (normal) e 2 (suspeita). A acurácia de ambos os algoritmos usando undersampling também não foi muito boa.

Já o método de oversampling obteve os melhores resultados em ambos os algoritmos. Alcançou as melhores acurácias na validação e manteve um valor alto na validação cruzada. Também gerou os valores mais altos de precisão, revocação e F1, e desenhou uma diagonal muito bem definida na matriz de confusão, com os demais quadrantes quase zerados.

Tabela 2 – Matrizes de confusão de acordo com algoritmo e balanceamento

	Original	Oversampling	Undersampling
Árvore de Decisão			
Floresta Aleatória			

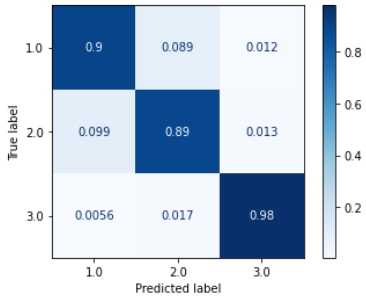
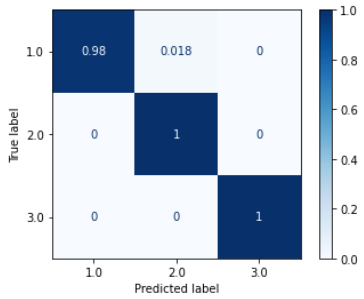
Após essa análise, podemos definir que o modelo que melhor classificou os dados foi a árvore de decisão, usando oversampling. Ele teve melhores valores de acurácia na validação e validação cruzada, melhores valores de precisão, revocação e F1, além de ter uma matriz de confusão com uma diagonal bem definida e poucos erros de classificação em comparação aos demais.

6 RESULTADOS

Com o modelo escolhido na etapa anterior, usamos os dados de teste para medir as mesmas métricas analisadas. Nos dados de teste gerados pelo método de oversampling, obtivemos uma acurácia de 0.924, e usando a validação cruzada, a acurácia caiu para 0.821. Apesar da redução de 0.103, ainda é um valor elevado. Os valores de precisão, revocação e F1 ficaram iguais, em 0.924. A matriz de confusão gerada pode ser vista na Tabela 3.

Testamos, também, o modelo usando os dados originais de teste (sem balanceamento). Neles, o modelo teve um resultado melhor ainda, gerando uma classificação quase perfeita dos dados, como pode ser visto na matriz de confusão, também na Tabela 3. A acurácia foi de 0.986 e a média da acurácia da validação cruzada foi de 0.859. Obtivemos uma precisão de 0.987, revocação e F1 de 0.986. Essa melhora provavelmente se deve ao fato de que os dados originais estão contidos nos dados de oversampling.

Tabela 3 - Matriz de confusão de acordo com o balanceamento nos dados de teste

Oversampling	Original
	

Como os resultados foram bons, mesmo para dados de teste não vistos antes, podemos dizer que o modelo foi capaz de generalizar as características dos dados para fazer sua predição. Um modelo mais simples como a árvore de decisão atingiu resultados satisfatórios em termos de acurácia e de erros de classificação. Ele também possui a vantagem de uma explicabilidade muito alta.

7 CONCLUSÃO

Usando os dados rotulados de exames CTG, fizemos uma exploração de sua distribuição e outras características que nos permitiram definir procedimentos para adequar os dados a um algoritmo de classificação. O desbalanceamento constatado nessa etapa gerou uma oportunidade de fazer experimentos com diferentes tipos de métodos de balanceamento e levou à conclusão de que o método mais adequado neste caso é o oversampling.

O modelo final, apesar de ser uma simples árvore de decisão, foi capaz de atingir bons valores de acurácia e gerar uma matriz de confusão com poucas classificações errôneas. Ele também foi capaz de generalizar bem para dados não vistos antes, mantendo bons resultados nos dados de teste.

Dessa forma o modelo está apto atuar em seu propósito inicial, de ajudar a prever a condição de saúde fetal. Principalmente por causa do baixo número de casos de saúde patológica sendo classificados como normal ou suspeita, dado que os falsos negativos são mais perigosos nesse cenário.

Em conclusão, foi possível realizar um estudo enriquecedor de um problema de classificação, em que todas as etapas de uma tarefa de mineração de dados foram colocadas em prática. Aprendemos muito sobre os modelos usados, as formas de tratar dados desbalanceados e a análise dos resultados sobre diferentes pontos de vista.

8 REFERÊNCIAS

ABDELKADER, Rahma; MAHMOUD, Doaa. **Fetal health classification (full project)**. Kaggle, 18 de nov. de 2022. Disponível em: <https://www.kaggle.com/code/rahmaabdelkader/fetal-health-classification-full-project>. Acesso em: 28 de nov. de 2022.

Ayres de Campos et al. **A program for automated analysis of cardiotocograms.** SisPorto 2.0 (2000) 5:311-318. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/1520-6661(200009/10)9:5%3C311::AID-MFM12%3E3.0.CO;2-9>. Acesso em: 28 de nov. de 2022.

Compare the effect of different scalers on data with outliers. Scikit Learn. Disponível em: <https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py>. Acesso em: 28 de nov. de 2022.

Guia do Usuário CRISP-DM

LARXEL. **Fetal Health Classification**. Kaggle, 11 de out. de 2020. Disponível em: <https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification>. Acesso em: 28 de nov. de 2022.

Over-sampling. Imbalanced Learn. Disponível em: <https://imbalanced-learn.org/stable/over_sampling.html>. Acesso em: 28 de nov. de 2022.

Under-sampling. Imbalanced Learn. Disponível em: <https://imbalanced-learn.org/stable/under_sampling.html>. Acesso em: 28 de nov. de 2022.

ZAKI & MEIRA Jr. **Data Mining and Machine Learning Slides Chapters 18 to 22**. Disponível em: <https://dataminingbook.info/resources/>. Acesso em: 28 de nov. de 2022.