CIÊNCIA DA COMPUTAÇÃO

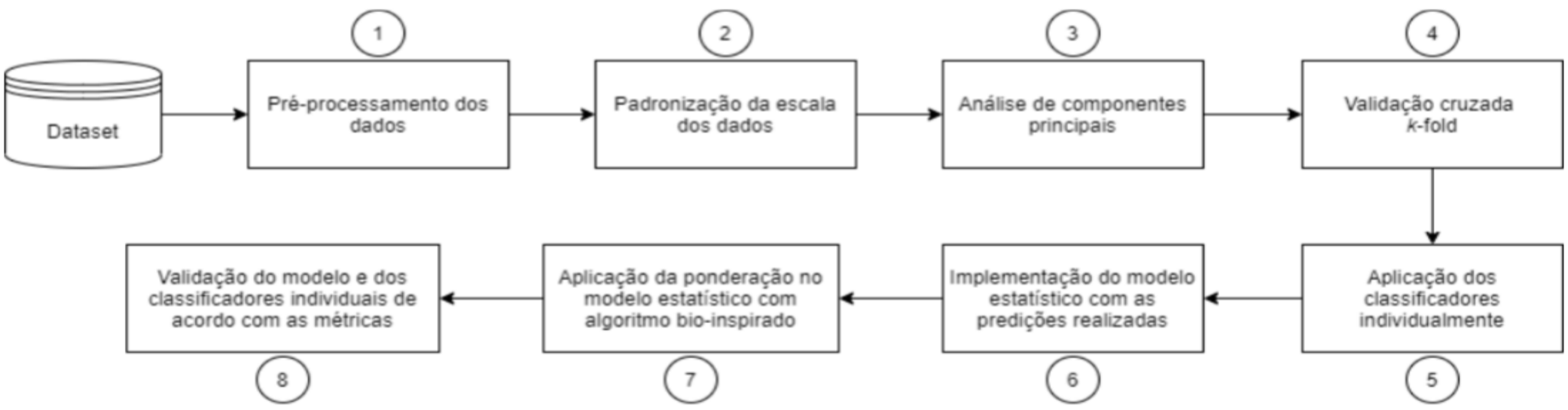
Um modelo bayesiano baseado em algoritmos bio-inspirados para classificação binária

Aluno: Ricardo Morello dos Santos e Thyago Melo dos Santos
(unifrsantos@fei.edu.br, unifthsantos@fei.edu.br)
Orientador: Prof. Paulo Sérgio Rodrigues (psergio@fei.edu.br)

RESUMO

Nos últimos anos, nota-se o crescente aumento na geração de dados digitais, sobretudo por conta da consolidação da internet como meio de comunicação. Proporcionalmente, cresce também a quantidade de algoritmos e metodologias propostas para mineração de dados e identificação de tendências ou padrões, hoje uma tarefa inviável à capacidade analítica humana. No entanto, de acordo com a literatura, estas técnicas apresentam performance diferente quando aplicadas em problemas ou bases de dados diferentes. Assim, este trabalho propõe um modelo bayesiano que agrega a saída de diferentes algoritmos de classificação, ponderando-as de maneira a priorizar o classificador com melhor performance para o problema em questão. Os resultados obtidos mostram que o modelo proposto priorizou os classificadores com maior performance, portanto preservando a saída com maior assertividade, sobretudo na base de dados Telco Customer Churn. Neste caso, a despeito da maior variação nas classificações, o método proposto apresentou estabilidade na classificação. Nas demais bases de dados, quando os classificadores possuem performance similar, o modelo proposto apresentou assertividade também similar aos demais.

Metodologia proposta para resolução de problemas de classificação binária.



Fonte: Os Autores

METODOLOGA

A metodologia proposta neste trabalho é apresentada na figura acima. O pipeline proposto é composto por 8 etapas, e a implementação será feita na linguagem de programação Python. Na etapa 1, pré-processamento dos dados, será feita a transformação de dados categóricos em quantitativos, junto da remoção de campos como identificador dos registros ou número de telefone. Em seguida, na etapa 2, padronização dos dados, será feita a transformação dos dados para o mesmo domínio, para que estes possam ser aplicados no PCA, na etapa 3 . Após a aplicação do PCA, a base de dados será dividida em parcelas para treino e teste dos classificadores, na etapa 4. Por sua vez, os classificadores serão treinados na parcela de treinamento, e será gerado um conjunto de predições para a parcela de teste, na etapa 5. As predições dos classificadores para a parcela de teste serão agregadas no modelo estatístico, que será implementado na etapa 6.

A ponderação das classificações pelo algoritmo bio-inspirado será feito na etapa 7, seguida da avaliação dos resultados na etapa 8.

CONCLUSÃO

Os resultados mostram que o modelo proposto neste trabalho é promissor, priorizando os algoritmos com a melhor performance nos problemas de classificação binária considerados. Da mesma forma, quando os algoritmos de classificação apresentam performance similares, a resposta do modelo iguala-se à dos demais algoritmos. Por outro lado, dentre os classificadores analisados, o Tensorflow foi o que apresentou maior assertividade. Por fim, com relação aos algoritmos bio-inspirados, o PSO obteve performance semelhante ao FA, no entanto, com tempo de execução consideravelmente menor. Os resultados sugerem que o modelo proposto é superior às metodologias do estado-da-arte nas bases de dados e problemas abordados.