CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
2020.1

Aluno: Anselmo de Vasconcelos Cavalcante
Disciplina: Algoritmos 1

Projeto da disciplina

A ideia básica do programa é que ele possa ser utilizado em pesquisas futuras sobre alguma base de dados que eu possa necessitar analisar durante o doutorado. O intuito é que alguns tipos de classificadores (e suas variações), empregados na área de aprendizagem de máquina, sejam aplicados a uma base de dados, utilizando uma aprendizagem do tipo supervisionada, e gerem alguns valores de métricas sobres os dados. O programa deve solicitar ao usuário quais classificadores ele deseja aplicar a base, gerar gráficos para as métricas de **acurácia**, **f1-socre** e **recall**, exibi-los na tela e gerar um relatório em formato PDF contendo as principais informações da base, uma tabela com valores das métricas e os gráficos gerados. Além disso, a tabela deve destacar em verde os melhores resultados para cada uma das métricas e em vermelho os piores. Opcionalmente, o programa desse solicitar ao usuário se deseja informar manualmente os valores de um futuro elemento para que seja realizada a predição (classificação), de acordo com os classificadores treinados. Caso o usuário opte por realizar a predição, as informações do elemento devem constar no relatório e os resultados devem ser exibidos na tabela. O tratamento de erros foi aplicado durante a escolha dos classificadores por parte do usuário, durante a solicitação dos dados do elemento para predição e durante o processo de escrita do relatório PDF no disco.

Submeteu-se o projeto inicial com a ideia de que o programa solicitasse ao usuário o nome da referida base de dados e dos seus campos (características), além disso inicialmente tinha-se a ideia de que o programa também realizasse o pré-processamento dos dados de forma automatizada, porém, com as orientações dos monitores da disciplina, percebeu-se que essas funcionalidades não faziam sentido, uma vez que geralmente as informações da base de dados já se encontram presentes na própria base ou em sua descrição. Além disso, o tratamento de dados pode ser totalmente diferente de uma base para outra. Desta forma, a leitura do nome da base, das suas características e o seu pré-processamento devem ser ajustados manualmente no programa, dentro da função **lerArquivo**, de acordo com as características da base que se deseja analisar. A única exceção é normalização dos dados, procedimento bastante comum, onde para isso criou-se uma função específica chamada **normalizar**.

Vale salientar que o programa inicialmente não deveria possuir a funcionalidade de solicitar ao usuário quais classificadores ele gostaria de aplicar, nem gerar métricas para as métricas f1-score e recall, como também gerar uma tabela no relatório, nem destacar os melhores e piores resultados. Tais funcionalidades foram adicionadas ao longo da criação do programa, sendo algumas delas sugestões dos monitores, que facilitaram a análise do desempenho dos classificadores.

Base de dados utilizada

- Disponível em: https://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.data
- Contém 214 exemplos
- Contem 11 atributos, sendo o primeiro o número de identificação e o último a classe

- Nomes dos atributos: id_number, ri_refractive_index, na_sodium, mg_magnesium, al_aluminium, si_silicon, k_potassium, ca_calcium, ba_barium, fe_iron, type_of_glass
- Nome do atributo da classe: type_of_glass
- Valores da classe

Valor da classe (type_of_glass)	Significado
1	building_windows_float_processed
2	building_windows_non_float_processed
3	vehicle_windows_float_processed
4	vehicle_windows_non_float_processed (none in this database)
5	containers
6	tableware
7	headlamps

- Não contém dados vazios (faltantes)

Bibliotecas externas utilizadas:

- pandas: leitura e manipulação da base de dados.
- sklearn: criação e manipulação dos classificadores. Geração das métricas.
- matplotlib: geração dos gráficos.
- fpdf: geração do relatório em formato PDF.