Treinamento e validação de modelos preditivos com AM
Trabalho 1 – INF01017 – 2022/2

Profa. Mariana Recamonde Mendoza
mrmendoza@inf.ufrgs.br

1 Objetivo

O Trabalho 1 da disciplina consiste no desenvolvimento de modelos preditivos utilizando algoritmos de aprendizado de máquina (AM). O trabalho terá como objetivo permitir que os alunos explorem diferentes algoritmos de aprendizado supervisionado para uma tarefa de classificação de interesse próprio, validando os modelos treinados com uma **implementação própria** da estratégia de *k-fold cross validation* e de métricas de avaliação de modelos.

O trabalho será desenvolvido em **grupos de 2 ou 3 alunos**[1]. Sugere-se o uso da linguagem de programação Python ou R pela ampla disponibilidade de diversos recursos para treinamento de modelos e por permitir fácil integração de funções implementadas pelos grupos.

A seguir são sumarizados os requisitos do trabalho, especificando-se o que deve ser implementado pelos grupos (implementação própria) e o que pode ser aplicado de implementações prontas fornecidas por bibliotecas ou pacotes. Recomenda-se uma leitura atenta destes requisitos e das demais informações fornecidas ao longo do documento.

Cada grupo deve:

- Escolher um conjunto de dados e uma pergunta de pesquisa associada, com foco em problema de **classificação** (binária ou multiclasse). Como sugestão, alguns repositórios que podem ser buscados: PMLB[2], Kaggle[3] e UCI ML Repository[4]. Recomenda-se observar o número de exemplos (i.e., instâncias) por classe, especialmente a classe minoritária, e evitar conjuntos de dados que possuam menos de 10 instâncias em alguma das classes.

- Avaliar a necessidade de pré-processar os dados, como por exemplo, normalizar atributos, realizar imputação de valores, remover atributos ou instâncias com muitos valores faltantes, dentre outros. A necessidade pode ser decorrente das características originais dos dados (isto é, conforme disponibilizados pelo repositório) ou por especificidades dos algoritmos de aprendizado supervisionado escolhidos (ver abaixo).

- Selecionar **ao menos três** algoritmos de aprendizado supervisionado cobertos pelo conteúdo programático da disciplina para treinamento de modelos preditivos focados na pergunta de interesse. Os grupos podem optar por outros algoritmos que não sejam vistos na disciplina **em adição** a estes. Os algoritmos de aprendizado **não** precisam ser implementados pelos grupos: podem ser utilizadas funções prontas disponibilizadas por pacotes ou bibliotecas em linguagens como Python ou R.

- Realizar uma **implementação própria da estratégia de *k-fold cross validation***, que seja genérica o suficiente para permitir o uso de diferentes valores de k (número de folds). A estratégia deve ser aplicada de forma estratificada, isto é, mantendo-se a proporção original de exemplos por classe em todos os folds, independente do número de classes do problema. É

[1] Recomenda-se manter a mesma formação de grupos para todos os trabalhos práticos da disciplina
[2] https://epistasislab.github.io/pmlb/
[3] https://www.kaggle.com/
[4] https://archive.ics.uci.edu/ml/datasets.php

aconselhável que os grupos estudem as funções prontas a serem utilizadas para treinamento dos modelos, a fim de implementarem sua função para validação cruzada de forma compatível com as mesmas (principalmente em relação às estruturas de dados usadas).

- Implementar uma **função própria** para gerar uma matriz de confusão, e a partir dela realizar a quantificação das métricas de acurácia, precisão, *recall* e F1-measure. As métricas também devem ser calculadas por **funções criadas pelos grupos**. Para problemas multiclasse, pode ser utilizada a macro média para sumarizar o desempenho em termos de precisão, recall e F1-measure.

- Aplicar a implementação da estratégia de *k-fold cross validation* e da matriz de confusão e métricas de desempenho ao problema de classificação selecionado, avaliando de forma comparativa o desempenho dos algoritmos de aprendizado supervisionado escolhidos. Os grupos deverão reportar, no mínimo, desempenho médio e desvio padrão obtido através do processo de validação cruzada, discutindo os resultados de forma comparativa entre os algoritmos de aprendizado usados. O uso de gráficos como *boxplot* para avaliar distribuição de desempenho, e outros selecionados a critério do grupo, é desejável e será valorizado na avaliação do trabalho.

- Desenvolver um relatório sobre o trabalho desenvolvido, no qual deverão constar, no mínimo, as seguintes informações:

  - uma definição clara do objetivo do trabalho desenvolvido, ou seja, dos dados utilizados para treinamento dos modelos e da pergunta de pesquisa que o grupo visou investigar com algoritmos de AM. Por exemplo: "*Nosso trabalho tem como objetivo desenvolver um modelo baseado em AM para prever o tipo de vinho (tinto ou branco) a partir de um conjunto de características físico-quimicas. Para tanto, utilizamos o conjunto de dados <breve descrição> disponível em <link para download>.*"
  - uma explicação sobre os dados utilizados, incluindo características como número de instâncias, número de atributos, número de classes, número de instâncias por classe, tipo dos atributos usados (numéricos ou categóricos), se existem valores faltantes, distribuição dos valores dos atributos numéricos, etc.
  - uma menção à linguagem de programação e outros recursos computacionais utilizados no desenvolvimento do trabalho, incluindo pacotes e bibliotecas
  - uma descrição da metodologia usada para desenvolvimento dos modelos preditivos, incluindo a etapa de pré-processamento dos dados, os algoritmos escolhidos, os valores de hiperparâmetros utilizados no treinamento destes algoritmos (e eventuais estratégias adotadas para otimizar estes hiperparâmetros), além de outras técnicas que podem ser aplicadas de forma opcional a fim de aprimorar o desenvolvimento dos modelos, como redução de dimensionalidade e balanceamento de dados[5]
  - na metodologia de trabalho, deve ser incluído o código fonte da implementação feita pelo grupo para o *k-fold cross validation* e para as estratégias de avaliação de modelos (matriz de confusão e métricas de desempenho). O grupo também deverá incluir um pseudocódigo e uma breve descrição de como esta função foi utilizada no pipeline de desenvolvimento de modelos, devendo ficar bem claro como o grupo utilizou os folds gerados para treinar e testar os modelos ao integrar a função implementada com as funções prontas para treinamento de modelos
  - uma descrição e análise dos resultados obtidos. Por análise, entende-se que além de utilizar recursos visuais como gráficos e tabelas para relatar os resultados alcançados com os algoritmos utilizados, os grupos também irão discutir estes achados, discorrendo sobre aspectos como: todos os algoritmos conseguiram resolver a tarefa de classificação abordada? com base nas métricas de desempenho aplicadas, o desempenho dos algoritmos parece satisfatório? como o desempenho variou para estes algoritmos no processo de validação cruzada? algum algoritmo se saiu particularmente melhor, seja por alcançar desempenhos mais altos ou mais consistentes? os algoritmos conseguiram aprender bem para todas as classes do problema, ou demonstraram maior dificuldade com alguma classe

[5]Estes tópicos ainda serão discutidos no escopo da disciplina, mas os grupos ficam livres para pesquisar sobre estes conceitos e empregá-los em seus trabalhos

ou algum tipo de erro (como falsos positivos ou falsos negativos)? A fim de determinar o algoritmo com desempenho mais satisfatório, recomenda-se que os alunos escolham uma métrica alvo a ser otimizada, com base na pergunta de pesquisa abordada. Esta escolha deve ser apresentada no relatório. Entretanto, o relatório deve conter os resultados para as quatro métricas de desempenho solicitadas neste enunciado.

- uma conclusão sobre o trabalho desenvolvido e os resultados alcançados, bem como uma breve discussão de eventuais limitações do trabalho e dos modelos treinados.

Ressalta-se que os alunos podem se basear em material visto na disciplina ou consultado em outras fontes para elaborar a metodologia de treinamento dos modelos, entretanto, a implementação da metodologia de treinamento dos modelos deve ser desenvolvida pelos grupos. Ou seja, não podem ser usados códigos prontos disponibilizados em repositórios como Kaggle ou outros. Se houver quaisquer dúvidas sobre a possibilidade de uso de algum recurso computacional ou software, por favor consultem a professora antes de fazê-lo.

2 Entrega de Resultados: até 03/03/2023

- Os grupos deverão enviar seu código fonte e relatório em PDF pelo Moodle do INF até a data de entrega do trabalho (ver na seção "Critérios de avaliação" a política adotada para entregas com atraso).
- O código fonte pode ser implementado em qualquer linguagem de programação de preferência dos alunos, mas deverá ser enviado com instruções de como rodar o código (um arquivo README, por exemplo). Serão aceitos o envio de código no formato de notebooks, inclusive baseados no uso de ferramenta como Google Colab.
- A simples exportação/impressão de um notebook como PDF não será aceito como relatório. O relatório deve ser um documento devidamente estruturado contendo as informações solicitadas acima.
- O trabalho deverá ser apresentado oralmente pelo grupo, período de aula, em data a ser determinada pela professora

3 Critérios de avaliação

- Pontualidade na entrega do trabalho. Atenção: atrasos na entrega do trabalho serão penalizados proporcionalmente ao tempo de atraso, **sendo descontado 1 (um) ponto por dia de atraso** (o trabalho como um todo vale 10 pontos)
- Corretude na implementação e no uso da estratégia de *k-fold cross-validation* para validação de modelos, bem como das métricas de avaliação solicitadas
- Completude do trabalho e atendimento aos requisitos definidos neste enunciado (ver Seção 1)
- Apresentação dos resultados: qualidade da apresentação e domínio da implementação e resultados, bem como capacidade de arguição acerca dos mesmos
- Qualidade e corretude da metodologia para desenvolvimento dos modelos
- Qualidade do relatório final e da discussão dos resultados, e abrangência em relação aos requisitos de relato definidos no enunciado (páginas 02 e 03)

4 Política de Plágio

Os grupos poderão **apenas** discutir questões de *alto nível* relativas a resolução do problema em questão. **Não** é permitido que os grupos utilizem quaisquer códigos fonte provido por outros grupos, ou encontrados na internet, para a etapa de implementação do *k-fold cross-validation* e das métricas de avaliação de desempenho. Os alunos **poderão** fazer consultas na internet ou em livros **apenas** para estudar o modo de funcionamento destas técnicas e analisar os respectivos **pseudo-códigos**. Toda e qualquer fonte consultada pelo grupo (tanto para estudar os métodos a serem utilizados, quanto para verificar a estruturação da técnica em termos de *pseudo-código*) **precisa obrigatoriamente** ser citada no relatório final.

Qualquer nível de plágio (ou seja, utilização de implementações que não tenham sido 100% desenvolvidas pelo grupo) poderá resultar em **nota zero** no trabalho. Caso a cópia tenha sido feita de outro grupo da disciplina, *todos* os alunos envolvidos (não apenas os que copiaram) serão penalizados. Esta política de avaliação **não** é aberta a debate posterior. Se você tiver quaisquer dúvidas sobre se uma determinada prática pode ou não, ser considerada plágio, ***não*** **assuma** nada: pergunte à professora. Os grupos deverão desenvolver o trabalho **sozinhos**.A professora estará à disposição para sanar dúvidas ao longo do processo - recomendo, no entanto, não deixar as dúvidas para o último momento.