Disciplina: Data Mining (INE5644)
Profa.: Vania Bogorny
Tema: Trabalho Prático Data Mining

UNIVERSIDADE FEDERAL DE SANTA CATARINA

A Kaagle é uma plataforma de competições em Data Mining utilizada por milhares de alunos, professores e pesquisadores do mundo todo. Essa plataforma oferece diversos problemas para explorar, estudar e solucionar. Alguns deles oferecem até valor monetário para aqueles que conseguem a melhor solução. Além disso, a bagagem atingida pelos seus participantes tem feito com que muitas grandes empresas anunciem vagas de emprego através da plataforma Kaagle. Vale a pena ficar atento!

Este trabalho prático trata do problema "Shelter Animal Outcomes", que consiste na predição do destino de cães e gatos em abrigos de animais.

A descrição completa do problema e o conjunto de dados de treino e teste, são encontrados no endereço: https://www.kaggle.com/c/shelter-animal-outcomes. A primeira coisa a ser feita é ler com atenção toda a informação contida no site e observar o conteúdo dos arquivos para poder entender o problema. Este trabalho consiste em 2 partes, descritas a seguir:

Parte 1 - Aplicação da metodologia crisp-DM. Nesta primeira parte deve-se entender o problema, realizando uma leitura detalhada da proposta do problema no link supracitado. Por se tratar de um problema de mineração de dados, onde vocês irão explorar o dataset em busca de padrões, talvez nem toda a informação que você precise para encontrar bons padrões esteja presente nos dados. Assim, observando os arquivos de dados pense em quais outras fontes podem ser utilizadas para enriquecer esse dataset. Esta parte consiste na aplicação de técnicas de preprocessamento, extração de atributos, seleção e transformação dos atributos relevantes, seleção de exemplos (registros), normalização e discretização. Nesta parte será entregue um relatório com a definição do problema de data mining, todos os atributos que serão usados, já transformados de acordo com as técnicas citadas anteriormente, incluindo também: tipo de dado, valores que pode assumir e gráficos de histograma, diagrama de caixas ou outro, relevância dos atributos para o problema. O relatório também deve descrever a técnica de mineração que será usada, e para o caso de classificação, deve apresentar uma primeira descrição de como o conjunto de dados será dividido para treinamento e teste.

Parte 2 - Mineração de Dados e Análise dos Resultados.
Nesta parte será realizada a mineração de dados e a avaliação dos resultados da mineração, utilizando-se as técnicas ensinadas em sala. O número de modelos e tentativas de combinações fica por conta do proponente, no intuito de atingir modelos com a melhor qualidade possível. Os 3 melhores trabalhos e que tenham conseguido o melhor resultado de mineração, e melhores resultados que os três melhores trabalhos do semestre passado, ganharão meio ponto na média final da disciplina.

Uma abordagem para aprimorar os resultados da mineração consiste na criação de novos atributos baseados no resultado de técnicas de *clustering*. Dessa forma, crie novos atributos utilizando as diferentes técnicas de *clustering* apresentadas em sala, e verifique se com isso é possível melhorar a mineração realizada anteriormente.

Finalmente, utilize o conjunto de dados de teste e gere uma saída, conforme pede a descrição oficial do problema no site, faça a submissão da solução no sistema de avaliação da Kaagle e verifique o seu score. Esta etapa deve apresentar um relatório contendo : a descrição da m**ineração de dados, a análise dos resultados, e a relevância e a justificativa dos resultados encontrados**

Atenção. Cada entrega deverá ser feita por meio de: (a) RELATÓRIO impresso em papel do que foi desenvolvido até cada parte; (b) envio pela plataforma Moodle de um arquivo ZIP (bem organizado), contendo os arquivos de seu trabalho: como *scripts*, documentos externos, arquivos de datasets intermediários, modelos construídos, etc. **Entregas de arquivos corrompidos receberão nota Zero, assim o aluno sempre deve conferir os arquivos enviados e não deixar isso para os últimos minutos da entrega.**

A apresentação será através de um vídeo postado no youtube, de no máximo 10 minutos. Cada um dos integrantes do grupo deve apresentar uma parte do trabalho.