Projeto de Aprendizado de Máquina para Problema de Classificação

1. Introdução

O objetivo deste trabalho foi elaborar um projeto de aprendizado de máquina para um problema de classificação utilizando uma base de dados qualquer, de escolha individual. O projeto deveria incluir processos de pré-processamento e seleção de atributos da base de dados, utilização de três algoritmos diferentes de aprendizado de máquina como classificadores e medidas de desempenho, também a critério individual, para comparar o desempenho entre os três classificadores selecionados.

2. Metodologia Proposta

O projeto foi todo desenvolvido na linguagem de programação Python utilizando a biblioteca Scikit-Learn para a implementação dos classificadores e a plataforma Colaboratory do Google como ambiente de desenvolvimento e testes.

2.1. Descrição da Base de Dados

A base de dados escolhida foi a Qualitative_Bankruptcy dataset, disponível no repositório de aprendizado de máquina da Universidade da Califórnia em Irvine. Esta base foi elaborada por Martin. A, Uthayakumar. J e Nadarajan. M no Sri Manakula Vinayagar Engineering College da Universidade de Pondicherry, Índia, em Fevereiro de 2014.

A base de dados contém seis parâmetros ou atributos qualitativos relacionados ao risco de falência de uma empresa mais uma classe nominal indicando se a empresa está ou não em risco de falência. Os atributos e os seus respectivos fatores internos são listados a seguir:

i. Risco da indústria (IR): Políticas governamentais e acordos internacionais, ciclicidade, grau de competição, preços e estabilidade da oferta do mercado, tamanho e crescimento da demanda do mercado, sensibilidade a mudanças nos fatores macroeconômicos, poder competitivo nacional e internacional, ciclo de vida do produto.

ii. Risco de gestão (MR): Capacidade e competência de gestão, estabilidade de gestão, relação entre gerência e proprietário, gestão de recursos humanos, processo de crescimento/desempenho do negócio, planejamento de negócios de curto e longo prazo, realização e viabilidade.

iii. Flexibilidade Financeira (FF): Financiamentos direto e indireto, outros financiamentos.

iv. Credibilidade (CR): Histórico de crédito, confiabilidade das informações, relacionamento com instituições financeiras.

v. Competitividade (CO): Posição de mercado, nível de capacidades essenciais, estratégia diferenciada,

vi. Risco Operacional (OP): Estabilidade e diversidade de compras, estabilidade da transação, eficiência da produção, perspectivas de demanda por produtos e serviços, diversificação de vendas, preço de venda e condição de liquidação, eficácia da rede de vendas.

Para cada atributo uma empresa recebeu, na base de dados, um dentre três valores categóricos que refletem sua condição referente ao respectivo atributo. Esses valores podem ser: “P” para uma condição positiva, “A” para uma condição média e “N” para uma condição negativa. A classe que define se a empresa está ou não sob risco de falência foi definida pela letra “B” para falência e “NB” para não-falência.

No total a base contém 250 observações sem dados faltantes, sendo que 143 são instâncias de “NB” e 107 são “B”. Portanto, trata-se de uma base relativamente balanceada com 57,2% de empresas não falidas e 42,8% de empresas em risco de falência.

2.2. Pré-processamento

A etapa de pré-processamento envolveu a conversão para valores numéricos dos valores originalmente em caracteres contidos na base de dados. Para tal, os valores dos atributos foram mapeados como apresentado abaixo:

Categoria do atributo	Letra correspondente	Valor numérico mapeado
Positivo	P	2
Médio	A	1
Negativo	N	0

Além da conversão dos valores dos atributos, as duas classes foram mapeadas, respectivamente, da seguinte maneira: “B” para 0 e “NB” para 1.

Para seleção de características foi utilizado o coeficiente de Pearson para identificar quais atributos tinham maior correlação com a variável classe. O coeficiente de Pearson é uma medida de correlação linear entre duas variáveis que produz um resultado entre -1 e 1. Onde -1 significa uma correlação linear negativa entre as duas variáveis, 1 significa uma correlação linear positiva e 0 representa a ausência de correlação entre as duas variáveis.

A partir dos resultados foram identificadas correlações lineares positivas entre todos os atributos e a classe, em seguida foram selecionados os atributos que apresentaram coeficiente maior que 0,7 em relação a classe. Os atributos selecionados a partir do corte pelo coeficiente de Pearson foram: Flexibilidade Financeira (FF), Credibilidade (CR) e Competitividade (CO).

2.3. Classificadores

Os três algoritmos de aprendizado de máquina selecionados para os testes de classificação foram: árvores de decisão, máquinas de vetores de suporte e naive Bayes.

Árvores de decisão são algoritmos baseados em uma estrutura de árvore onde cada nó interno realiza um teste de valor de um determinado atributo, cada ramo representa um valor que o atributo pode assumir e as folhas representam uma classe específica relativa a amostra. A ordem de ocorrência dos atributos na árvore deve ser definida pela capacidade do atributo em discriminar a classe. O método escolhido para execução deste trabalho foi o da entropia ou ganho de informação, no qual é considerada a redução da entropia do atributo em relação a classe para definir a posição do atributo na árvore. Atributos com menor entropia e, consequentemente, maior ganho de informação

são posicionados primeiro na árvore, sendo que a raiz da árvore será o atributo com maior ganho de informação inicial de todos.

Máquinas de vetores de suporte utilizam um hiperplano separador para classificar os dados. Dados de treinamento são convertidos para coordenadas de pontos num plano multidimensional. Os pontos mais próximos da linha divisória entre as diferentes classes constituem o vetor de suporte que define o hiperplano separador de classes. O mapeamento dos pontos no plano multidimensional é feito por meio de uma função de kernel. O tipo de função de kernel utilizada neste trabalho foi a função linear.

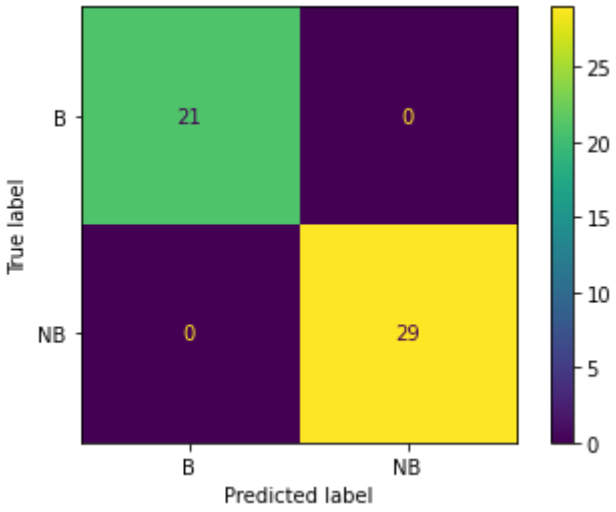
Naive Bayes é um algoritmo utilizado para classificação baseado no teorema de Bayes, que supõe independência condicional entre cada par de atributos dado o valor da classe. A versão do algoritmo utilizada neste trabalho foi a categorical naive Bayes, específica para dados categóricos.

2.4. Avaliação dos Classificadores

A medida de desempenho utilizada para comparar os diferentes algoritmos de classificação testados foi a acurácia. A acurácia é calculada a partir do total de acertos, verdadeiros positivos e verdadeiros negativos, dividido pelo total de observações. É uma medida adequada para bases de dados que não estejam demasiadamente desbalanceadas e representa uma visão geral do desempenho do classificador com base nos resultados obtidos da fase de teste.

3. Resultados e Discussões

Todos os classificadores apresentaram o mesmo resultado, 100% de acertos, para a fase de teste, independentemente da seleção de atributos de maior correlação linear em relação a classe. Portanto, a acurácia calculada para todos os classificadores foi de 1,0. A matriz de confusão derivada dos resultados é apresentada a seguir:



Apesar da classe B (Bankruptcy) ter sido mapeada para o valor “0” na etapa de pré-processamento, seus resultados verdadeiros na matriz de confusão, i.e. empresas da classe B preditas corretamente como tal, devem ser considerados como correspondentes aos resultados verdadeiros positivos, 21 ocorrências nesse caso. Enquanto os resultados da classe NB (Non-Bankruptcy), mapeados para o valor “1”, correspondem aos resulta-

dos verdadeiros negativos, 29 ocorrências. Isto se deve por conta do objetivo primário do classificador ser o de identificar empresas sob risco de falência, consequentemente uma empresa que não estiver sob risco falência será classificada na classe oposta, negativa, àquela que o modelo pretende detectar. As classes B e NB foram mapeadas da forma previamente descrita para corresponder melhor aos valores mapeados dos atributos, onde valores negativos foram mapeados para "0" enquanto médios e positivos para "1" e "2", respectivamente.

4. Conclusões

A partir dos resultados obtidos não foi possível definir qual método de aprendizado de máquina seria o mais adequado para a base de dados utilizada, pois todos os algoritmos testados foram igualmente adequados, apresentando o mesmo desempenho nos resultados, com 100% de acertos. A seleção de características também não apresentou nenhuma influência sobre os resultados. No entanto, podemos dizer que, ao selecionar os atributos de maior correlação, ainda foi possível obter máxima acurácia nos resultados de teste, a um menor custo computacional, utilizando um número reduzido de atributos, com apenas metade dos dados da base de dados original.

5. Links e Referências

Página do repositório UCI contendo informações e download da base de dados. http://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy Acessado: 03/07/2022

Confusion Matrix, Accuracy, Precision, Recall, F1 Score. Autor: Harikrishnan N B. 10/12/2019 https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd Acessado: 03/07/2022

Wikipedia – Confusion Matrix: https://en.wikipedia.org/wiki/Confusion_matrix Acessado: 03/07/2022

sklearn.feature_selection.r_regression:
https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.r_regression.html Acessado: 03/07/2022

Wikipédia – Pearson correlation coefficient:
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient Acessado: 03/07/2022

Scikit-Learn, Naive Bayes:
https://scikit-learn.org/stable/modules/naive_bayes.html Acessado: 03/07/2022