UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CIÊNCIAS DA COMPUTAÇÃO

DANIEL MAHL GREGORINI
DANIEL CASANOVA
MATHEUS BRUSTOLIN

TRABALHO SEGUNDO BIMESTRE

MEDIANEIRA
2024

DANIEL MAHL GREGORINI
DANIEL CASANOVA
MATHEUS BRUSTOLIN

TRABALHO SEGUNDO BIMESTRE

Trabalho do Segundo Bimestre, apresentado à disciplina Sistemas Inteligentes Aplicados, do curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná – UTFPR,. Com o professor Jorge Aikes Junior.

MEDIANEIRA
2024

RESUMO

GREGORINI, Daniel; CASANOVA, Daniel; BRUSTOLIN, Matheus. **Trabalho do Segundo Bimestre de Sistemas Inteligentes Aplicados** – Bacharelado em Ciência da Computação. Universidade Tecnológica Federal do Paraná. Medianeira, 2024**.**

Esse trabalho explora a aplicação de algoritmos de aprendizado de máquina para classificar indivíduos quanto ao consumo de bebidas alcoólicas. Os dados foram extraídos do Serviço Nacional de Seguro de Saúde da Coreia e bibliotecas como Pandas, NumPy e Scikit-learn foram importantes para possibilitar a aplicação de treinamento e pré-processamento. Diversos modelos desenvolveram-se nesse trabalho, incluindo Árvore de Decisão, SVM, k-NN, Naive Bayes, Regressão Linear e Perceptron.

Palavras-chave: Aprendizado de Máquina; Classificação; Pré-processamento de Dados

ABSTRACT

GREGORINI, Daniel; CASANOVA, Daniel; BRUSTOLIN, Matheus. **Second Bimester Project of Applied Intelligent Systems** – Bachelor's in Computer Science. Federal Technological University of Paraná. Medianeira, 2024.

This work explores the application of machine learning algorithms to classify individuals based on their alcohol consumption. The data were extracted from the Korean National Health Insurance Service, and libraries such as Pandas, NumPy, and Scikit-learn were instrumental in facilitating training and preprocessing. Various models were developed in this study, including Decision Tree, SVM, k-NN, Naive Bayes, Linear Regression, and Perceptron.

Keywords: Machine Learning; Classification; Data Preprocessing

SUMÁRIO

1 INTRODUÇÃO

O aprendizado de máquina, uma subárea da inteligência artificial, consiste em métodos que permitem a construção de modelos capazes de identificar padrões e realizar previsões com base em dados históricos. Neste trabalho, exploramos a aplicação de algoritmos de aprendizado de máquina para a classificação de indivíduos quanto ao consumo de bebidas alcoólicas, utilizando um conjunto de dados clínicos coletados do Serviço Nacional de Seguro de Saúde da Coreia.

Para isso, aplicamos diferentes etapas de pré-processamento de dados, como imputação de valores ausentes, codificação de variáveis categóricas e normalização, visando melhorar a qualidade dos dados e aumentar o desempenho dos modelos. Foram utilizados diversos algoritmos de aprendizado supervisionado, incluindo Árvore de Decisão, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Naive Bayes, Regressão Linear e Perceptron.

Cada algoritmo tem uma abordagem diferente na separação e classificação de dados, o que permite identificar qual seria o método mais adequado para o problema proposto. Os resultados foram avaliados com base em métricas de acurácia.

2 MATERIAIS E MÉTODOS

Nesta seção, são descritos os materiais e métodos utilizados para o desenvolvimento do estudo, abrangendo desde as ferramentas e linguagens empregadas até os procedimentos de pré-processamento dos dados e os algoritmos utilizados no treinamento dos modelos. Primeiramente, são apresentadas as ferramentas essenciais para a análise, incluindo bibliotecas da linguagem Python voltadas para manipulação de dados, aprendizado de máquina e visualização. Em seguida, a base de dados utilizada é descrita em detalhes, com ênfase nos atributos analisados e no objetivo do estudo. O pré-processamento dos dados é abordado em etapas que incluem a remoção de valores ausentes, codificação de variáveis categóricas, análise de correlação, normalização, engenharia de atributos e balanceamento das classes, garantindo que os dados estejam adequados para os modelos de aprendizado de máquina. Por fim, são apresentados os algoritmos utilizados no treinamento, detalhando suas principais características e funcionamento.

2.1 Ferramentas utilizadas

Para o desenvolvimento do trabalho, foram utilizadas ferramentas que possibilitam o processamento, análise e visualização dos dados. A linguagem usada foi o Python, destaque-se o uso de suas bibliotecas, como Pandas, usada para estrutura de dados eficientes e conversão. NumPy, usada para operações matemáticas. Joblib, para armazenamento e recuperação de dados no Python, Matplotlib e Seaborn, usadas para a criação de gráficos e plots variados, proporcionando uma visão detalhada dos dados analisados, enquanto a Seaborn, que se baseia na Matplotlib, oferece uma interface de alto nível para a criação de gráficos estatísticos mais elaborados e esteticamente agradáveis, auxiliando na identificação de padrões e tendências..

A biblioteca Scikit-learn, considerada uma das mais robustas e populares para a implementação de algoritmos de machine learning. O Scikit-learn disponibiliza uma vasta gama de algoritmos para tarefas de classificação, regressão, agrupamento e redução de dimensionalidade, além de ferramentas para o pré-processamento dos dados, como normalização, padronização e codificação.

Sua API padronizada e bem documentada facilita a criação de pipelines que encadeiam diversas etapas do processamento, garantindo um fluxo de trabalho limpo, reprodutível e acessível tanto para iniciantes quanto para especialistas na área. Sua utilização foi essencial para o projeto usado em todos os processos de treinamento.

2.2 Descrição da base de dados e tratamentos

A base de dados utilizada neste estudo foi coletada a partir do Serviço Nacional de Seguro de Saúde da Coreia e reúne registros clínicos de indivíduos, abrangendo medições físicas e exames laboratoriais. O conjunto contém uma variedade de atributos quantitativos, como idade (agrupada em intervalos de 5 anos), altura (arredondada para cima a cada 5 cm), peso, pressão arterial sistólica (SBP) e diastólica (DBP), níveis de colesterol total, HDL, LDL, triglicerídeos, glicose em jejum (BLDS), hemoglobina, creatinina e enzimas hepáticas (SGOT_AST, SGOT_ALT, gamma_GTP). Bem como atributos qualitativos, incluindo sexo (masculino ou feminino), avaliação da visão (sight_left e sight_right), audição (hear_left e hear_right) e status de fumante (SMK_stat_type_cd), esses e outros atributos são descritos na Tabela 1.

O objetivo principal deste trabalho é desenvolver modelos de aprendizado de máquina que possam classificar os indivíduos de acordo com o consumo de bebidas alcoólicas, indicado pelo atributo “DRK_YN”. Dessa forma, cada atributo contribui de maneira singular para a compreensão do estado de saúde dos pacientes e para a identificação de possíveis padrões relacionados ao hábito de beber.

Tabela 1 - Todos os atributos da base de dados

Coluna	Descrição
Sex	Sexo
age	Idade arredondada para cima a cada 5 anos
height	Altura arredondada para cima a cada 5 cm
weight	Peso
sight_left	Visão (olho esquerdo)
sight_right	Visão (olho direito)
hear_left	Audição (ouvido esquerdo)

hear_right	Audição (ouvido direito)
SBP	Pressão arterial sistólica
DBP	Pressão arterial diastólica
BLDS	BLDS ou FSG (glicose em jejum)
tot_chole	Colesterol total
HDL_chole	Colesterol HDL
LDL_chole	Colesterol LDL
triglyceride	Triglicerídeos
hemoglobin	Hemoglobina
urine_protein	Proteína na urina
serum_creatinine	Creatinina no sangue
SGOT_AST	SGOT (Glutamato-oxaloacetato transaminase) / AST (Aspartato transaminase)
SGOT_ALT	ALT (Alanina transaminase)
gamma_GTP	y-glutamil transpeptidase
SMK_stat_type_cd	Estado de fumante
DRK_YN	Consumo de álcool (Bebe ou não)

2.3 Pré-processamento e Tratamento dos Dados

Nesta seção, são descritas as etapas aplicadas para preparar os dados antes do treinamento dos modelos. O pré-processamento é essencial para garantir que os modelos de aprendizado de máquina operem de maneira eficiente e produzam previsões mais confiáveis. Para isso, foram realizadas diversas técnicas, incluindo a remoção de valores nulos, codificação de variáveis categóricas, análise de valores ausentes, normalização dos dados, engenharia de atributos e balanceamento das classes. A seguir, detalhamos cada uma dessas etapas.

2.3.1 Codificação de Variáveis Categóricas

As variáveis categóricas foram transformadas em valores numéricos para que os modelos pudessem interpretá-las corretamente. A variável sexo foi convertida, atribuindo 0 para "Male"/"masculino" e 1 para "Female"/"feminino". A variável alvo DRK_YN, que originalmente continha os valores "Y" (bebe) e "N" (não bebe), foi convertida para um formato binário, onde 1 representa consumidores de bebidas alcoólicas e 0 representa não consumidores.

2.3.2 Análise de Valores Faltantes e Imputação

Foram identificados valores ausentes em várias colunas do conjunto de dados. Para os atributos categóricos, como hear_left, hear_right, urine_protein e SMK_stat_type_cd, a imputação foi feita com a moda da distribuição de cada variável, garantindo que os valores mais frequentes fossem usados para preencher os dados ausentes. Para os atributos numéricos, a imputação foi realizada com a média da distribuição de cada variável, minimizando possíveis distorções nos dados.

2.3.3 Análise de Correlação

Foi realizada uma análise de correlação entre as variáveis preditoras e a variável alvo. Esse estudo permitiu identificar atributos com menor influência na previsão, evitando redundâncias e reduzindo o risco de overfitting. Entre as variáveis com menor impacto identificadas estavam hear_left e hear_right, que foram removidas dos modelos para melhorar a eficiência do aprendizado.

2.3.4 Normalização e Padronização

Como os dados apresentavam variáveis com escalas diferentes, aplicamos a padronização usando o método StandardScaler, que transforma os valores das variáveis numéricas para uma distribuição com média zero e desvio padrão unitário. Essa etapa é essencial para modelos que utilizam medidas de distância, como KNN e SVM, garantindo que nenhuma variável domine o cálculo de proximidade entre os dados.

2.3.5 Engenharia de Atributos (Feature Engineering)

Para enriquecer o conjunto de dados e melhorar a capacidade dos modelos, foram criadas novas variáveis derivadas de atributos existentes:

- Pulse Pressure: diferença entre Pressão Sistólica (SBP) e Pressão Diastólica (DBP), usada para avaliar riscos cardiovasculares.

- Cholesterol Ratio: razão entre colesterol total (tot_chole) e colesterol HDL (HDL_chole), útil na avaliação de riscos de doenças cardíacas.

Essas novas features foram incluídas para fornecer mais informações relevantes aos modelos e potencialmente melhorar a discriminação entre os grupos.

2.3.6 Balanceamento dos Dados

Uma análise preliminar revelou um leve desbalanceamento entre as classes da variável alvo, onde aproximadamente 50,01% dos registros pertenciam à classe "não bebe" e 49,99% à classe "bebe". Para evitar que os modelos se tornassem tendenciosos, foi aplicada a técnica de oversampling na classe minoritária, aumentando sua representatividade no conjunto de treino. Esse processo garantiu que os modelos não fossem enviesados para a classe majoritária, permitindo previsões mais equilibradas e generalizáveis.

2.3.7 Divisão dos Dados

Após todas as etapas de pré-processamento, o conjunto de dados foi dividido em 80% para treinamento e 20% para teste, garantindo que os modelos fossem avaliados de forma independente e evitando vazamento de dados entre os conjuntos. Com essa preparação, os dados estavam prontos para a aplicação dos modelos de aprendizado de máquina.

2.4 Algoritmos usados para o treinamento

Nesta seção, são apresentados os algoritmos utilizados no treinamento dos modelos, cada um com abordagens distintas para a classificação. A Árvore de Decisão organiza os dados em uma estrutura hierárquica, enquanto o SVM busca um hiperplano ótimo para separação das classes. O k-NN classifica os dados com base nos vizinhos mais próximos, e os modelos Naive Bayes utilizam probabilidades para inferência. Além disso, a Regressão Linear foi aplicada para classificação por meio de um limiar de decisão, e o Perceptron serviu como base para entender classificadores lineares. A seguir, cada algoritmo é detalhado conforme sua metodologia e aplicabilidade ao problema.

2.4.1 Árvore de Decisão CART

A Árvore de Decisão CART (Classification and Regression Trees) é um modelo de aprendizado supervisionado que utiliza uma estrutura em árvore para tomar decisões com base em critérios que dividem os dados. Esse algoritmo seleciona, a cada nó, a variável e o ponto de corte que melhor separam os dados, utilizando medidas como o índice Gini ou o erro de classificação. Ela é usada em modelos de classificação e regressão

2.4.2 Support Vector Machine

O método que busca do classificador Support Vector Machine (SVM) é encontrar o hiperplano ótimo que separa as classes com a maior margem possível. Em sua versão linear, o SVM procura esse hiperplano diretamente, maximizando a distância entre as classes. Ao utilizar diferentes kernels, o algoritmo transforma os dados para espaços de dimensão superior, o que permite a separação de dados não linearmente separáveis e aumenta a flexibilidade do modelo para lidar com conjuntos de dados complexos.

2.4.3 k-Nearest Neighbors

O k-Nearest Neighbors (k-NN) é um método de classificação que determina a classe de um novo dado com base na maioria das classes dos seus “k” vizinhos mais próximos. Na forma padrão, um valor fixo de k é utilizado, o que define o número de vizinhos considerados. Alterar esse valor pode influenciar a sensibilidade do modelo. Valores menores podem resultar em um ajuste excessivo, enquanto valores maiores podem suavizar a fronteira de decisão. Além disso, a escolha da medida de distância, seja ela Euclidiana, Manhattan ou outra. É importante, pois determina como a proximidade entre os pontos é avaliada.

2.4.4 Gaussian Naive Bayes e Categorical Naive Bayes

Os classificadores Gaussian Naive Bayes e Categorical Naive Bayes utilizam o teorema de Bayes com a suposição entre as variáveis. O Gaussian Naive Bayes é adequado para dados contínuos, assumindo que as características seguem uma distribuição normal, o que facilita o cálculo das probabilidades. Enquanto, o Categorical Naive Bayes é indicado para dados categóricos, onde as variáveis assumem valores discretos.

2.4.5 Regressão Linear

A técnica de estatística de Regressão Linear modela a relação entre uma variável dependente e uma ou mais variáveis independentes através de uma função linear. O objetivo é estimar os coeficientes que minimizam a soma dos erros ao quadrado entre os valores observados e os preditos, o que torna essa abordagem simples e interpretável para compreender a influência de cada variável no resultado.

2.4.5 Perceptron

O Perceptron é um dos algoritmos base para problemas de classificação binária, prestando-se como um classificador na forma de um “padrão” linear. O algoritmo ajusta os pesos das entradas caso a caso para diminuir o erro em inserir a classificação, portanto, é eficiente apenas em representação dos problemas linearmente separáveis. Embora simples, o perceptron é a base para várias mais complexas redes neurais e tem validade técnica e também contribui para dar uma ideia de como o processo de aprendizado supervisionado funciona.

3 RESULTADOS

Neste capítulo são apresentados os resultados dos treinamentos realizados para cada um dos algoritmos testados. A tabela abaixo resume as acurácias obtidas em cada experimento realizado, considerando diferentes abordagens: modelo padrão, remoção de variáveis de baixa importância, aplicação de técnicas de engenharia de atributos (feature engineering - FE) e balanceamento dos dados.

Tabela 2 - Modelos e acurácias.

Modelo	Acurácia
Experimentos sem FE e balanceamento	
Árvore de Decisão (CART)	64,58%
SVM (Kernel RBF)	37,69%
SVM (Kernel Linear)	72,28%
KNN (Padrão, k=5)	68,84%
KNN (k = 10)	70,33%
KNN (Distância Manhattan)	68,95%

Gaussian Naive Bayes	69,14%
Categorical Naive Bayes	69,18%
Regressão Linear (threshold 0,5)	72,06%
Perceptron	69,90%
Árvore de Decisão (após remoção de variáveis)	64,58%
Experimentos com FE e dados balanceados	
Árvore de Decisão (Balanceado + FE)	64,20%
SVM Linear (Balanceado + FE)	72,27%
SVM RBF (Balanceado + FE)	37,02%
KNN Padrão (Balanceado + FE)	67,48%
KNN (k = 10, Balanceado + FE)	69,25%
KNN (Manhattan, Balanceado + FE)	67,60%
Gaussian NB (Balanceado + FE)	69,20%
Categorical NB (Balanceado + FE)	69,18%
Regressão Linear (Balanceado + FE)	72,09%
Perceptron (Balanceado + FE)	70,57%

Modelos Base (Sem FE e Balanceamento)

A Árvore de Decisão (CART) foi utilizada como modelo de referência, apresentando uma acurácia de 64,58%. Devido à sua estrutura hierárquica, pode sofrer com overfitting em conjuntos de dados complexos, limitando sua capacidade de generalização.

O SVM com kernel linear apresentou o melhor desempenho nesta configuração, atingindo 72,28%, sugerindo que as classes possuem uma separabilidade linear bem definida. Em contrapartida, o SVM com kernel RBF obteve um desempenho significativamente inferior (37,69%), indicando que os parâmetros do kernel não foram adequadamente ajustados ou que a transformação não capturou padrões úteis nos dados.

Os modelos baseados em KNN mostraram variações de desempenho conforme a configuração do número de vizinhos. A versão padrão com k=5 obteve 68,84% de acurácia, enquanto aumentar para k=10 resultou em uma leve melhora

para 70,33%, sugerindo que o aumento de vizinhos contribuiu para uma melhor generalização. A mudança da métrica de distância para Manhattan, no entanto, não trouxe impacto expressivo, mantendo a acurácia em 68,95%.

Os algoritmos Naive Bayes, tanto Gaussian quanto Categorical, apresentaram desempenhos estáveis, com 69,14% e 69,18%, respectivamente. Esses resultados indicam que a suposição de independência condicional entre as variáveis não comprometeu a capacidade preditiva dos modelos.

A Regressão Linear, utilizando um threshold de 0,5, demonstrou um desempenho semelhante ao SVM linear (72,06%), reforçando a hipótese de que a separação das classes pode ser bem modelada por um hiperplano linear. O Perceptron apresentou uma performance próxima (69,90%), porém inferior à Regressão Linear e ao SVM linear, possivelmente devido à sua simplicidade e maior sensibilidade a ruídos nos dados.

A remoção das variáveis de baixa importância ('hear_left' e 'hear_right') não teve impacto na performance da Árvore de Decisão, mantendo a acurácia em 64,58%, o que sugere que essas variáveis tinham pouca relevância para a classificação final.

Modelos com Feature Engineering e Dados Balanceados

A introdução de novas variáveis e o balanceamento das classes tiveram impactos distintos nos modelos. A Árvore de Decisão apresentou uma leve queda de desempenho (64,20%), sugerindo que as novas features não contribuíram significativamente para a classificação. O SVM Linear permaneceu estável com 72,27%, evidenciando sua robustez mesmo com a adição de novas características. Já o SVM RBF manteve um desempenho muito baixo (37,02%), reforçando sua inadequação para os dados.

Os modelos KNN mostraram pequenas variações, com a configuração padrão caindo de 68,84% para 67,48%, enquanto a versão com k=10 teve um leve aumento para 69,25%. Isso indica que o balanceamento dos dados teve pouco impacto nessa abordagem. Nos modelos Naive Bayes, as diferenças foram

praticamente inexistentes, com Gaussian NB registrando 69,20% e Categorical NB 69,18%, sugerindo que a distribuição original dos dados já era bem ajustada.

A Regressão Linear manteve sua consistência com 72,09%, consolidando-se como um dos modelos mais eficazes para este conjunto de dados. O Perceptron, por sua vez, obteve uma leve melhora, alcançando 70,57%, possivelmente devido ao balanceamento dos dados, o que pode ter favorecido sua capacidade de generalização.

3.3 Considerações Finais

Os modelos lineares, como o SVM Linear e a Regressão Linear, apresentaram os melhores desempenhos (~72%), sugerindo que as relações entre as variáveis são predominantemente lineares e podem ser bem modeladas sem a necessidade de transformações complexas.

A aplicação de técnicas de balanceamento e engenharia de atributos não trouxe melhorias expressivas para a maioria dos modelos, sugerindo que os dados já possuíam uma estrutura relativamente equilibrada. No entanto, o Perceptron se beneficiou do balanceamento, indicando que, para alguns modelos, a distribuição das classes pode impactar sua capacidade de generalização.

O fraco desempenho do SVM com kernel RBF destaca a importância da escolha do kernel correto e do ajuste fino dos hiperparâmetros para problemas específicos. Dessa forma, os modelos lineares, particularmente o SVM Linear e a Regressão Linear, foram as escolhas mais adequadas para este conjunto de dados.

4 CONCLUSÃO

Neste trabalho, analisamos a aplicação de diferentes algoritmos de aprendizado de máquina para a classificação de indivíduos com base no consumo de bebidas alcoólicas. O estudo envolveu um conjunto abrangente de etapas, desde o pré-processamento dos dados até a avaliação dos modelos treinados, considerando técnicas de balanceamento e engenharia de atributos.

Os resultados obtidos demonstraram que os modelos baseados em classificação linear, como o SVM Linear e a Regressão Linear, alcançaram as melhores acurácias, aproximadamente 72%, sugerindo que as relações entre as variáveis do conjunto de dados podem ser bem representadas por hiperplanos lineares. O Perceptron também apresentou um desempenho próximo, beneficiando-se do balanceamento dos dados, enquanto o SVM com kernel RBF teve um desempenho significativamente inferior, evidenciando a importância da escolha adequada dos hiperparâmetros e do tipo de kernel.

A implementação de técnicas de feature engineering e balanceamento dos dados teve impactos variados. Embora tenha melhorado o desempenho de alguns modelos, como o Perceptron, para outros, como a Árvore de Decisão, a adição de novas variáveis não contribuiu significativamente para uma melhor classificação. Além disso, o balanceamento das classes mostrou-se relevante para evitar tendências nos modelos, garantindo previsões mais justas e generalizáveis.

Dessa forma, concluímos que a escolha do modelo ideal depende fortemente das características do conjunto de dados e do objetivo do problema. Para classificação binária com padrões predominantemente lineares, abordagens como SVM Linear e Regressão Linear são recomendadas. Já para conjuntos de dados mais complexos e não linearmente separáveis, ajustes adicionais nos hiperparâmetros ou a exploração de outras técnicas podem ser necessários.

Como trabalho futuro, sugerimos a investigação de arquiteturas baseadas em redes neurais para melhorar a capacidade preditiva e a generalização dos modelos, bem como a experimentação de técnicas avançadas de seleção de atributos para refinar o impacto das variáveis utilizadas na classificação.