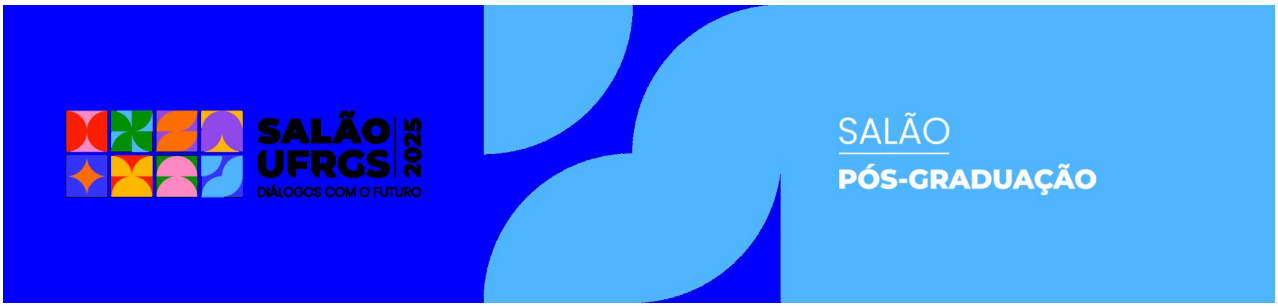


Evento	Salão UFRGS 2025: I SALÃO DA PÓS-GRADUAÇÃO
Ano	2025
Local	Campus do Vale - UFRGS
Título	Aplicação de modelos de machine learning para predição da cor de pele a partir de dados de SNPs
Autor	RAFAEL DIOGO WEIMER
Orientador	MARCIO DORN

RESUMO: A predição de características externamente visíveis (EVCs) a partir de dados genômicos, conhecida como fenotipagem forense por DNA (FDP), tem se destacado como uma ferramenta promissora em investigações forenses. Essa abordagem permite inferir traços fenotípicos como cor de olhos, cabelo e pele a partir de amostras de DNA coletadas em cenas de crime. Dentre os diferentes tipos de marcadores genéticos, os polimorfismos de nucleotídeo único (SNPs) são amplamente utilizados devido à sua abundância e às associações estabelecidas com características humanas, incluindo a pigmentação da pele. Contudo, a maioria dos estudos nessa área é baseada em populações europeias, o que compromete a acurácia das previsões em populações geneticamente diversas. O Brasil, em especial, apresenta grande diversidade genética decorrente de séculos de miscigenação, o que ressalta a necessidade de desenvolver painéis de SNPs adaptados a essa realidade. Nesse contexto, algoritmos de aprendizado de máquina (ML) oferecem uma estrutura robusta para modelar relações entre genótipo e fenótipo. Dado um conjunto de dados rotulados, modelos de ML podem ser treinados para reconhecer padrões e classificar indivíduos com base em informações genéticas. Aplicadas à FDP, esses algoritmos possibilitam a seleção de SNPs informativos e o desenvolvimento de modelos preditivos para diferentes EVCs**.** Nesse contexto, o objetivo deste trabalho é desenvolver modelos de classificação para prever a cor da pele com base em SNPs de indivíduos representativos da população do Rio Grande do Sul (RS), Brasil. Foram utilizados dados genéticos e fenotípicos de 438 indivíduos representativos da população gaúcha. O conjunto inicial incluiu 66 SNPs associados à pigmentação, e a cor da pele de cada indivíduo foi classificada de acordo com a escala de Fitzpatrick em seis categorias: Branco, Pálido, Bege, Marrom Claro, Marrom Médio e Marrom Escuro. Para a construção dos modelos, três abordagens de classificação foram adotadas: (1) classificação multiclasse com as seis categorias da escala Fitzpatrick; (2) classificação com as amostras agrupadas em três categorias: Clara (Branco e Pálido), Intermediária (Bege e Marrom Claro) e Escura (Marrom Médio e Marrom Escuro); (3) classificação com as amostras agrupadas em duas classes: Clara (Branco, Pálido e Bege) e Escura (Marrom Claro, Médio e Escuro). Para cada uma das situações, foram selecionados os 20 SNPs mais informativos utilizando o algoritmo VariantSpark, e três modelos de ML foram treinados: K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Multi-Layer Perceptron (MLP). Os modelos foram avaliados com subconjuntos incrementais de SNPs (de 1 a 20), e em cada situação foi realizada a otimização de hiperparâmetros com validação cruzada utilizando o método Leave-One-Out (LOO). O desempenho de cada modelo foi medido pela métrica F1-score ponderada. De formal geral, os modelos apresentaram melhor desempenho à medida que o número de categorias de cor da pele foi reduzido, sendo a classificação binária a mais eficaz. Na classificação com seis categorias, o SVM obteve os melhores resultados, mesmo com poucos SNPs, embora com maior variabilidade, apresentando F1-score médio de 0,645. Os modelos KNN e MLP mostraram desempenho similar, com f1-score médio de 0,483 e 0,465, respectivamente. Na classificação com três categorias, todos os modelos melhoraram, com o SVM novamente se destacando, seguido por MLP e KNN, com f1-scores médios de 0,833, 0,790 e 0,785, respectivamente. Por fim, na abordagem binária os três algoritmos apresentaram alto desempenho, com destaque para o SVM, que obteve os melhores resultados gerais mesmo com poucos SNPs, obtendo um f1-score médio de 1,0. Apesar dos bons resultados, especialmente na classificação binária, a alta variabilidade em alguns cenários e os F1-scores muito elevados indicam a necessidade de avaliação adicional quanto à possibilidade de overfitting, especialmente nos modelos SVM.

PALAVRAS-CHAVE: Machine Learning; SNPs; Predição de cor de pele.