UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E DO DESENVOLVIMENTO

Bruna Damasco de Oliveira

Análise comparativa de modelos *Machine Learning* para a predição de cor de olhos, cabelo e pele em uma amostra da população brasileira

Florianópolis
2022

Bruna Damasco de Oliveira

Análise comparativa de modelos *Machine Learning* para a predição de cor de olhos, cabelo e pele em uma amostra da população brasileira

Dissertação submetida ao Programa de Pós-Graduação em Biologia Celular e do Desenvolvimento da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Mestra em Biologia Celular e do Desenvolvimento

Orientador: Prof. Guilherme de Toledo e Silva, Dr.
Coorientadora: Profa. Elisa C. Winkelmann Duarte, Dra.

Florianópolis

2022

Oliveira, Bruna Damasco de
Análise comparativa de modelos Machine Learning para a predição de cor de olhos, cabelo e pele em uma amostra da população brasileira / Bruna Damasco de Oliveira ; orientador, Guilherme de Toledo-Silva, coorientador, Elisa C. Winkelmann Duarte, 2022.
86 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós Graduação em Biologia Celular e do Desenvolvimento, Florianópolis, 2022.

Inclui referências.

1. Biologia Celular e do Desenvolvimento. 2. Genética Forense. 3. Predição Fenotípica. 4. EVC. I. Toledo-Silva, Guilherme de . II. Duarte, Elisa C. Winkelmann. III. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Biologia Celular e do Desenvolvimento. IV. Título.

Bruna Damasco de Oliveira

Análise comparativa de modelos *Machine Learning* para a predição de cor de olhos, cabelo e pele em uma amostra da população brasileira

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 20 de dezembro de 2022, pela banca examinadora composta pelos seguintes membros:

Prof. Guilherme de Toledo e Silva, Dr.
Universidade Federal de Santa Catarina

Profa. Edna Sadayo Miazato Iwamura, Dra.
Universidade Federal de São Paulo

Prof. Glauber Wagner, Dr.
Universidade Federal de Santa Catarina

Prof. Ricardo Castilho Garcez, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestra em Biologia Celular e do Desenvolvimento.

Documento assinado digitalmente
Evelise Maria Nazari
Data: 24/02/2023 09:03:20-0300
CPF: ***.091.489-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Evelise Maria Nazari, Dra.
Coordenadora do Programa de Pós-Graduação em Biologia Celular e do Desenvolvimento.

Documento assinado digitalmente
Guilherme de Toledo e Silva
Data: 23/02/2023 09:40:36-0300
CPF: ***.522.609-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Guilherme de Toledo e Silva, Dr.
Orientador

Florianópolis, 2022.

RESUMO

A predição fenotípica apresenta-se como uma alternativa para investigações forenses onde a tradicional obtenção de perfis de marcadores do tipo microssatélite não é possível. Tal técnica consiste na análise polimorfismos de nucleotídeo único a fim de predizer as características externamente visíveis (EVC) de um indivíduo, os quais podem ser divididos em traços relacionados ou não à pigmentação de estruturas. Ao longo das duas últimas décadas foram propostos sistemas de predição que correlacionam um conjunto específico de marcadores moleculares com a cor de olhos, cabelos e pele; como por exemplo o HIrisPlex-S (baseado em uma equação de regressão logística multinomial) e o *Snipper* (construído por meio de classificadores Bayesianos). Essas metodologias, contudo, foram formuladas em estudos com populações europeias e geraram resultados conflitantes quando testadas em países com histórico de ampla miscigenação, tal como o Brasil. Levando esse fato em consideração, e utilizando a abordagem de *Machine Learning* (ML) para a resolução de problemas de classificação e *clustering* por meio de modelos matemáticos, o objetivo deste trabalho foi calibrar e aplicar modelos para a predição de cor de olhos, cabelo e pele especificamente em um recorte da população brasileira (composta por 611 indivíduos e 49 marcadores) cedido pelo Laboratório de Imuno-Hematologia e Hematologia Forense da Universidade de São Paulo. O pré-processamento dos dados foi a etapa inicial das análises. Os genótipos foram convertidos em valores numéricos de acordo com os alelos da variante de cada marcador. Indivíduos que continham ao menos uma observação de genótipo NA foram eliminados, assim como SNPs com menos de 1% de variação dentro da amostra. Em seguida, a relação dos marcadores com os fenótipos foi aferida estatisticamente, de forma a prover três grupos de marcadores. Finalmente, os classificadores foram calibrados e aplicados em cada um dos três grupos de acordo com cinco tipos de modelos matemáticos. Seis variantes foram identificadas como não-polimórficas na amostra. Dois marcadores apresentaram resultados inexpressivos nos filtros estatísticos aplicados. Todas as variantes que passaram pelas etapas de triagem estão associadas a pelo menos um dos EVCs analisados. O efeito de variantes do gene HERC2 na cor de olhos, amplamente discutido na literatura, foi corroborado neste estudo. Observou-se também que a definição do tom de pele de um indivíduo parece estar mais diluída entre os vários marcadores estudados. Marcadores dos genes SLC24A5 e SLC45A2 apresentaram bons resultados para a associação com todos os fenótipos. Houve pouca variação na acurácia e sensibilidade dos modelos, independente do conjunto de marcadores e do algoritmo aplicado. Em suma, pode-se afirmar que a metodologia empregada está bem adaptada à amostra utilizada. Salienta-se também a necessidade de que mais estudos sejam realizados na área, principalmente em regiões de alta miscigenação, a fim de estabelecer um sistema de predição que contemple as particularidades genéticas de diferentes populações.

Palavras-chave: Genética Forense; Predição Fenotípica; EVC.

ABSTRACT

Phenotype prediction has emerged as an alternative in forensic investigations where the traditional microsatellite profiling is not possible. This technique consists in the analysis of single nucleotide polymorphisms (SNP) in order to predict an individual's externally visible characteristics (EVC), which can be divided into pigmentation traits and non-pigmentation traits. Over the course of the past two decades prediction systems correlating an specific set of molecular markers and eye, hair and skin color have been proposed; such as the HIrisPlex-S model (based on multinomial logistic regression) and Snipper (built on Bayesian classifiers). These methodologies, however, were established in studies with European populations and have yielded conflicting results when tested in countries with a history of high admixture, like Brazil. Considering that, and with the aid of Machine Learning approaches aimed for the resolution of classification and clustering problems, the goal of this study was to calibrate and apply models for the prediction of eye, hair and skin color in a sample of the brazilian population (composed of 611 individuals and 49 markers) provided by the Laboratório de Imuno-Hematologia e Hematologia Forense of the University of São Paulo. Data preprocessing was the first step of the analysis. Genotypes were converted into numeric values considering the variant allele of each marker. Individuals that had at least one missing observation were eliminated, as well as SNPs with less than 1% of variation in the sample. Next, the association between markers and phenotypes was statistically determined with the intention of separating three groups of makers. Lastly, the classifiers were calibrated and applied in each of the three groups under different mathematical models. Six SNPs were identified as non-polymorphic in the sample. Two markers have yielded poor results in the statistical filters applied. All of the variants that have passed the triage stage are associated with at least one of the EVCs analyzed. The effect of SNPs of the HERC2 gene in eye color, amply discussed in the literature, have been corroborated in this study. It was also observed that the definition of skin tone seems to be diluted in the many studied variants. Markers from the genes SLC24A5 and SLC45A2 have been associated with all the phenotypes. There was little variation in accuracy and sensibility of the models, regardless of the marker subset or the algorithm applied. In conclusion, the employed methodology is well adapted to the analyzed sample. It is also worth mentioning the necessity of further studies in the area, especially in regions of high admixture, with the intent of establishing a prediction system that contemplates the genetic particularities of different populations.

Keywords: Forensic Genetics; Phenotype Prediction; EVC.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC	*Area Under the Curve*
CODIS	*Combined DNA Index System*
CORR	Marcadores Submetidos ao Teste de Correlação Linear de Pearson
DCT	Enzima TYRP2
DQ	Dopaquinona
DT	Árvore de Decisão
EVC	*Externally Visible Characteristic* (característica externamente visível)
FP	Falsos Positivos
FPR	Taxa de Verdadeiros Negativos
FN	Falsos Negativos
LR	Regressão Logística
MI	Marcadores Submetidos ao Teste de Informação Mútua
ML	*Machine Learning* (aprendizado de máquina)
MSH	Hormônio Estimulante de Melanócitos
MULTICOL	Marcadores Submetidos ao Teste de Multicolinearidade
NB	*Naive Bayes*
NDIS	*National Index System*
NDNAD	*UK's National DNA Database*
NGS	Sequenciamento de Nova Geração
RF	*Random Forest*
RIBPG	Rede Integrada de Bancos de Perfis Genéticos
ROC	*Receiver Operating Characteristic*
SNP	*Single Nucleotide Polymorphisms* (polimorfismos de nucleotídeo único)
STR	*Short Tandem Repeats* (microssatélites)
SVM	*Support Vector Machines*
TPR	Taxa de Verdadeiros Positivos
VIF	*Variation Inflation Factor*
VN	Verdadeiros Negativos
VNTR	*Variable Number of Tandem Repeats* (minissatélites)
VP	Verdadeiros Positivos

SUMÁRIO

1. INTRODUÇÃO

1.1 GENÉTICA FORENSE

Butler (2015) divide a história do desenvolvimento da Genética Forense em quatro fases cruciais: exploração, estabilização, crescimento e sofisticação. A primeira delas iniciou-se em meados dos anos 1980, com a publicação da técnica de DNA *fingerprinting*, a qual postula que a identificação de um indivíduo pode ser feita a partir da análise de regiões altamente polimórficas conhecidas como minissatélites, ou *Variable Number of Tandem Repeats* (VNTR) (GILLS; JEFFREYS; WERRETT, 1985; BUTLER, 2015). Deste ponto em diante, a etapa inicial é marcada pela elaboração dos primeiros protocolos voltados ao contexto forense envolvendo ferramentas conhecidas na Genética, como a Reação em Cadeia da Polimerase (PCR) e, em menor escala, as enzimas de restrição (BUTLER, 2015).

A fase de estabilização configura o período de intensificação dos estudos com microssatélites (*Short Tandem Repeats* ou STR), associados tanto aos cromossomos autossômicos quanto ao cromossomo Y (BUTLER, 2015). Além de serem sequências mais curtas, os STRs geralmente apresentam tamanhos similares, de modo a tornar possível a análise de múltiplos desses marcadores ao mesmo tempo (PINHEIRO, 2015). Levando em consideração esse princípio, diferentes estudos foram publicados ao longo da década de 1990 relacionando conjuntos específicos de microssatélites à identificação humana (KIMPTON et al., 1994; SPARKES et al., 1996). A padronização de metodologias que utilizam diferentes regiões polimórficas, ou *multiplex*, propiciou a criação de importantes bancos de dados, como o *National Index System* (NDIS) nos Estados Unidos (WERRETT, 1997), e a Rede Integrada de Bancos de Perfis Genéticos (RIBPG) no Brasil (SILVA JUNIOR et al., 2020).

O novo milênio foi marcado pelo rápido crescimento da Genética Forense (BUTLER, 2015). O DNA mitocondrial entrou em evidência, impulsionado pela vantagem de possuir inúmeras cópias em uma única célula, e por estar associado à linhagem matrilinear de um indivíduo (PARSON et al., 2014; ZAVALA et al., 2019). Novos estudos buscaram empregar polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms* ou SNP) no contexto de investigações criminais, resultando no desenvolvimento da técnica de DNA preditor de fenótipo (*DNA Phenotyping*) (KAYSER, 2015; CHAITANYA et al., 2018) e na expansão da área da Genealogia Forense (WICKENHEISER, 2019).

Finalmente, Butler (2015) projetou em sua quarta fase um momento de sofisticação de diversos conceitos envolvidos na área. Ainda segundo o autor, tal etapa implica em um

aumento na sensibilidade e especificidade de técnicas conhecidas, assim como no avanço das ciências ômicas dentro do contexto forense.

1.2 MARCADORES GENÉTICOS DE INTERESSE FORENSE

Dentre os diferentes tipos de variações genéticas inerentes ao DNA humano, duas classes de polimorfismos se destacam por sua aplicabilidade forense: microssatélites e polimorfismos de nucleotídeo único (BUTLER, 2015; PINHEIRO, 2015; TURCHETTO-ZOLET et al., 2017; ZANELLA et al., 2017).

1.2.1 STR

Os microssatélites constituem regiões do genoma compostas por grupos de dois a sete nucleotídeos repetidos em padrões curtos, como mostra a Figura 1 (NUSSBAUM; MCINNES; WILLARD, 2016). A população humana está sujeita a uma grande diversidade em relação ao número de repetições dentro de um mesmo *locus*, de modo que regiões STR constituem marcadores multialélicos com alto grau de polimorfismo. Tal característica aumenta o poder discriminatório, permitindo assim que a identificação do indivíduo, tanto em contexto forense quanto em circunstâncias nas quais se faz necessário estabelecer graus de parentesco, possa ser realizada a partir de análises comparativas dessas variações (PINHEIRO, 2015; NUSSBAUM; MCINNES; WILLARD, 2016; ZANELLA et al., 2017).

O perfil genético de um indivíduo é obtido através da amplificação (por meio de técnica de PCR) de regiões microssatélites. Os fragmentos previamente amplificados são analisados de acordo com seus diferentes tamanhos (associados diretamente ao número de repetições) e, posteriormente, comparados com amostras de referência, quando possível (NIMS et al., 2010; LINACRE; TEMPLETON, 2014; UDOGADI et al., 2020). Apesar do grande potencial de variabilidade, Zanella e colaboradores (2017) afirmam que as regiões imediatas aos STRs são bem conservadas dentro de indivíduos de uma mesma espécie. Logo, é possível desenhar sequências iniciadoras (ou *primers*) que funcionam para todos os possíveis alelos de um locus em seres humanos, facilitando assim a realização da PCR.

Além de serem utilizados para fins comparativos em casos específicos, os perfis gerados podem ser também incluídos em bancos de dados. Dentre os mais conhecidos, temos o NDIS associado ao *Combined DNA Index System* (CODIS) nos Estados Unidos, o qual armazena perfis individuais compostos por 20 marcadores STR (BUTLER; WILLIS, 2020), o *UK's National DNA Database* (NDNAD) estabelecido em 1995 (AMANKWAA;

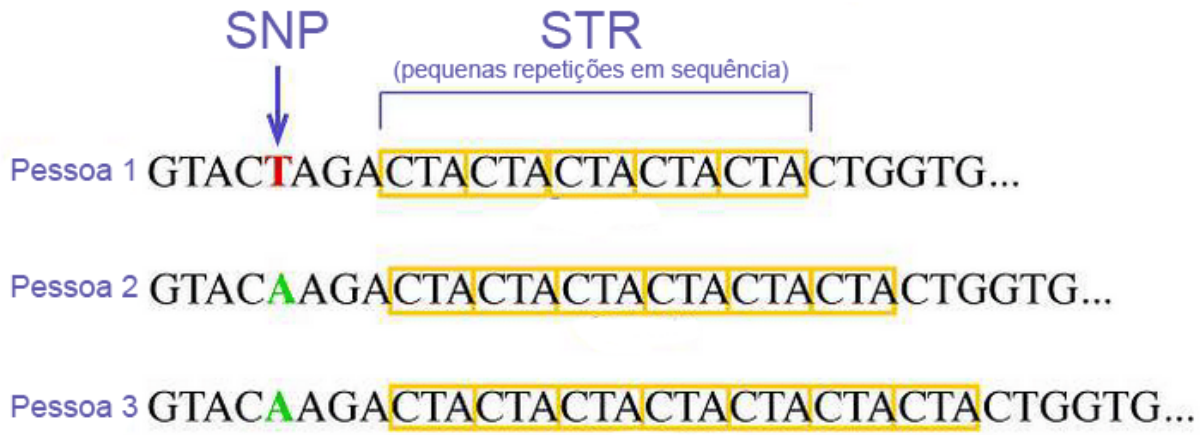
MCCARTNEY, 2019) e o *INTERPOL'S DNA Database* que armazena cerca de 250 mil perfis submetidos por investigadores de 84 países (INTERPOL, 2022).

1.2.2 SNP

Polimorfismos de nucleotídeo único têm origem em mutações pontuais dentro de um genoma. Esse tipo de variação pode ser classificado em dois grupos: transições e transversões. No primeiro caso, um nucleotídeo é substituído por outro do mesmo tipo, por exemplo, uma base purina (A) entra no lugar de outra base purina (G). Enquanto que na segunda categoria, um nucleotídeo é substituído por outro de diferente classe, ou seja, uma base purina (A ou G) entra no lugar de uma pirimidina (T ou C), e vice-versa. Os SNPs são caracterizados em sua maioria como bialélicos, tendo em vista que, apesar de qualquer uma das quatro bases poder ocupar um mesmo locus, nas amostragens em geral observa-se apenas duas possibilidades ocorrendo em diferentes frequências, como representado na Figura 1 (NUSSBAUM; MCINNES; WILLARD, 2016; ALBERTS et al., 2017). O principal critério de subdivisão dos SNP envolve seu posicionamento no genoma. Um maior número de polimorfismos é encontrado em áreas não-codificantes. Entretanto, a presença dessas variações em éxons é amplamente conhecida. Caso um SNP de região codificante não altere a proteína resultante, ele é classificado como sinônimo. É possível ainda que variações causem alterações de aminoácido no produto proteico, logo, tal SNP é considerado não-sinônimo. Finalmente, há também a possibilidade da formação de um códon de parada em posição incorreta, ou a modificação de um sítio de *splicing* devido a uma mutação pontual. Tais polimorfismos podem ainda se localizar em regiões regulatórias, de modo a afetar a expressão do gene e, eventualmente, impactar no fenótipo final (NUSSBAUM; MCINNES; WILLARD, 2016; TURCHETTO-ZOLET et al., 2017).

O estudo de marcadores bialélicos em contexto forense se beneficiou da evolução do Sequenciamento de Nova Geração (NGS) e dos avanços da bioinformática (YANG; XIE; YAN, 2014; TURCHETTO-ZOLET et al., 2017). Ensaios de genotipagem de SNP resultaram na criação de *multiplexes* que podem ser utilizados tanto na área da genealogia (PHILLIPS, 2018) quanto em casos nos quais se busca inferir sobre os traços fenotípicos de um indivíduo (WALSH et al., 2011; WALSH et al., 2013; CHAITANYA et al., 2018).

Figura 1. Representação dos dois tipos de polimorfismos de maior interesse forense: SNP e STR.



Fonte: Rashid, Othman e Zainudin (2017).

1.3 DNA PREDITOR DE FENÓTIPO

A fenotipagem forense tem como objetivo a predição de características externamente visíveis (*Externally Visible Characteristic* ou EVC) a partir de um conjunto de SNP informativos. As EVCs podem ser divididas em duas categorias distintas: traços relacionados ou não à pigmentação de estruturas (KAYSER, 2015; VIRMOND et al., 2016; SCHNEIDER; PRAINSACK; KAYSER, 2019).

Como exemplos de atributos não relacionados aos processos de pigmentação, pode-se citar a estatura, morfologia facial, e a estrutura capilar. A dinâmica dos múltiplos genes envolvidos nesses processos ainda não está bem elucidada, de modo que, até o momento, não existem metodologias capazes de fazer predições acerca de grande parte das características agrupadas nessa categoria (KAYSER, 2015; SCHNEIDER; PRAINSACK; KAYSER, 2019). Segundo Kayser (2015), a idade cronológica de um indivíduo também pode ser incluída no conjunto de EVCs independentes da pigmentação, pois pode ser inferida visualmente até certo ponto. A estimativa de idade por técnicas moleculares é bem estabelecida, em especial a análise dos padrões de metilação das ilhas CpG dispersas ao longo do genoma (KAYSER, 2015; HONG et al., 2017; SCHNEIDER; PRAINSACK; KAYSER, 2019; MARANO; FRIDMAN, 2019).

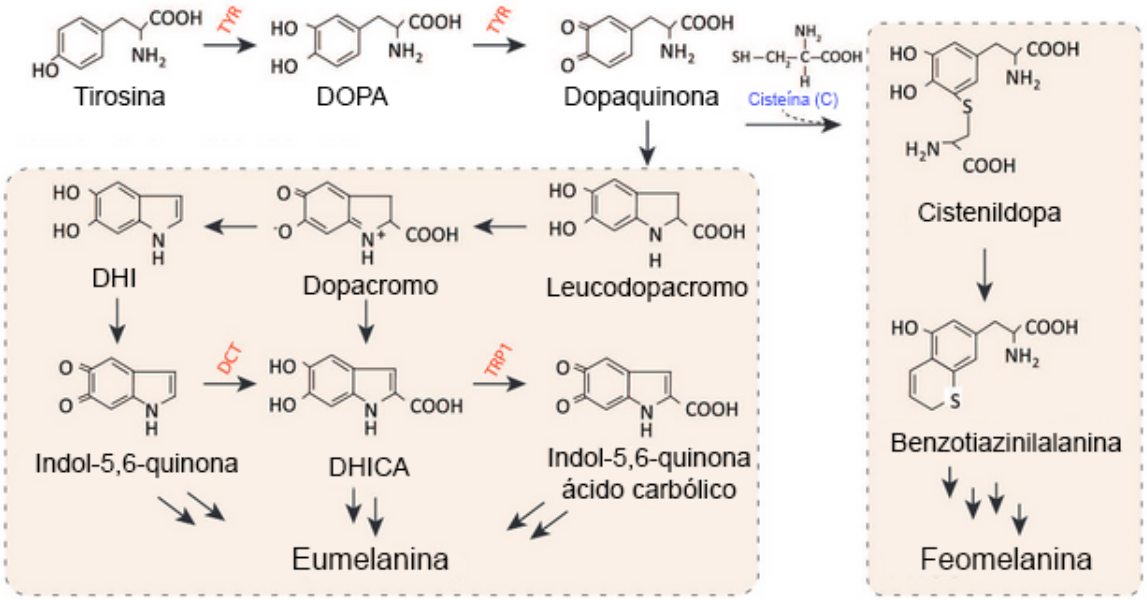
Em contrapartida, a pigmentação de estruturas tende a ser mais estudada dentro da Genética Forense. A determinação da cor de olhos, cabelos e do tom de pele em seres humanos está associada a um número menor de alelos. Logo, dentro do conceito de DNA preditor de fenótipo, a maioria dos autores busca estudar esses três pontos (KAYSER, 2015).

1.3.1 Pigmentação humana

A melanina é a principal molécula envolvida na pigmentação de estruturas como olhos, cabelo e pele. Tal polímero é produzido e armazenado em organelas específicas denominadas melanossomas, as quais são encontradas exclusivamente em melanócitos. Essas células, por sua vez, são originadas dos melanoblastos que migram das cristas neurais para a região basal da epiderme, folículos capilares, íris e orelha interna (STURM; LARSSON, 2009; STURM; DUFFY, 2012; VIDEIRA; MOURA; MAGINA, 2013; VIRMOND et al., 2016; WAKAMATSU; ZIPPIN; ITO, 2021).

A síntese de melanina é conhecida como melanogênese, e pode resultar em dois tipos de pigmento: a eumelanina, mais escura e insolúvel em água, e a feomelanina, em tons avermelhados e ou/amarelados e em geral mais solúvel (VIDEIRA; MOURA; MAGINA, 2013; VIRMOND et al., 2016; WAKAMATSU; ZIPPIN; ITO, 2021). O processo de melanogênese tem início com a oxidação da tirosina pela enzima tirosinase presente nos melanócitos, formando assim a Dopaquinona (DQ). No caso da eumelanina, a DQ sofre uma série de reações, gerando moléculas de Leucodopacromo e os intermediários DHI e DHICA, os quais ficam sujeitos à ação das enzimas TYRP1 e TYRP2 (também conhecida como DCT) antes de originar o pigmento. A biossíntese da feomelanina envolve a adição de cisteína à DQ, formando compostos intermediários, como a Benzotiazinilalanina, que vão sendo oxidados ao longo de uma cadeia de reações (STURM; DUFFY, 2012; VIDEIRA; MOURA; MAGINA, 2013; HIDA et al., 2020, WAKAMATSU; ZIPPIN; ITO, 2021). Vale ressaltar que a melanogênese é um processo misto, ou seja, o pigmento final produzido nos melanócitos possui tanto eumelanina quanto feomelanina em diferentes proporções (ITO, 2003; WAKAMATSU; ZIPPIN; ITO, 2021). A sequência de reações da melanogênese está representada na Figura 2.

Figura 2. Biossíntese de eumelanina e feomelanina.



Fonte: Horrell, Boulanger e D'orazio (2016).

Além de produzir a melanina, as células especializadas presentes na epiderme também são responsáveis pelo transporte da mesma para queratinócitos adjacentes (DISCHIA et al., 2015; MAROÑAS et al., 2015; WAKAMATSU; ZIPPIN; ITO, 2021). Estima-se que um único melanócito epidermal esteja em contato com até 36 queratinócitos (SLOMINSKI et al., 2004; RACHMIN et al., 2020), os quais eventualmente migram para as camadas mais superficiais da pele após receberem melanossomas cheios de pigmento (HUNT et al., 1995; RACHMIN et al., 2020). Tons de pele mais escuros estão associados a melanossomas maiores, bem pigmentados, com grande quantidade de eumelanina e pouca feomelanina presente, enquanto que tons de pele mais claros provém de melanossomas menores, menos pigmentados, que se amontoam em pequenos grupos e contém uma proporção menos díspar entre eumelanina e feomelanina (ALALUF et al., 2002; BARSH et al., 2003; MAROÑAS et al., 2015).

Apesar de também envolver o transporte de melanina para queratinócitos próximos (SLOMINSKI et al., 2005; LEERUNYAKUL; SUCHONWANIT, 2020; WAKAMATSU; ZIPPIN; ITO, 2021), a pigmentação de folículos capilares tem o diferencial de não ser um processo contínuo. Pêlos, em geral, possuem um ciclo de nascimento e regressão caracterizado por três etapas bem distintas, sendo que a melanina tem um papel ativo somente no período de crescimento, denominada como a fase anagênica (SLOMINSKI; PAUS, 1993; LEERUNYAKUL; SUCHONWANIT, 2020). Cabelos pretos têm origem em folículos com

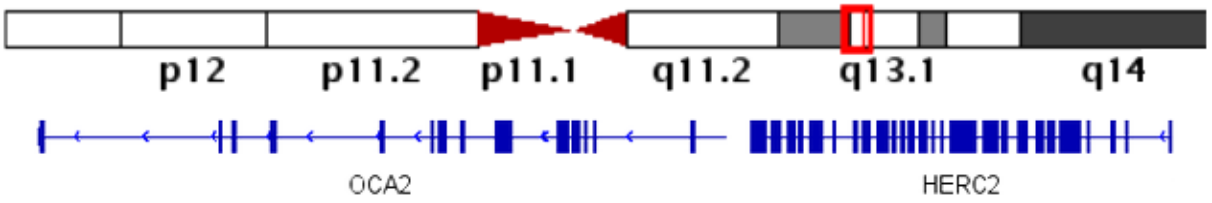
melanossomas grandes e com alta concentração de eumelanina. Já cabelos castanhos estão associados a melanossomas elipsóides e com quantidade de eumelanina levemente menor. Cabelos loiros são formados quando a quantidade de eumelanina e feomelanina são similares. E, finalmente, cabelos ruivos surgem a partir de concentrações baixas de ambos os pigmentos (MAROÑAS et al., 2015; LEERUNYAKUL; SUCHONWANIT, 2020).

No processo de pigmentação de olhos, não há transferência de melanossomas para outras células. Os próprios melanócitos dispersos ao longo da camada mais anterior da íris são os grandes responsáveis pela coloração dessa estrutura (IMESCH; WALLOW; ALBERT, 1997; STURM, 2004; STURM; LARSSON, 2009; MAROÑAS et al., 2015). As variações no tamanho, distribuição e concentração de pigmentos dentro dos melanossomas estão diretamente associadas aos possíveis fenótipos, assim como a capacidade de absorção de luz por parte das moléculas de melanina. Olhos castanhos se originam a partir de grandes quantidades de melanina absorvendo bastante luz. Em contrapartida, olhos azuis são formados quando há menos melanina presente nas células e, consequentemente, menos luz é absorvida. Tonalidades de verde e mel ocorrem na presença de níveis intermediários de melanina. A presença de um anel mais escuro em torno da pupila pode também ser observado em alguns casos, independente da cor do olho em si (IMESCH; WALLOW; ALBERT, 1997; STURM; LARSSON, 2009; MAROÑAS et al., 2015).

1.3.2 Genes envolvidos na pigmentação humana

A melanogênese é um processo complexo e que conta com a atuação direta ou indireta de uma gama de genes. Ao longo das etapas iniciais da biossíntese desse pigmento, temos a atuação do gene OCA2, o qual sintetiza uma proteína-p constituinte da membrana de vesículas transportadoras de tirosina, o aminoácido precursor de melanina. A expressão de OCA2 está fortemente associada a HERC2, seu vizinho no cromossomo 15, como representado na Figura 3. O sítio promotor de OCA2 localiza-se no íntron 86 do gene HERC2 e determinadas variantes deste último podem acarretar em falhas na ligação de fatores de transcrição à região promotora de OCA2, de modo a reduzir sua expressão e, consequentemente, a produção de eumelanina (STURM; LARSSON, 2009; DORGALELEH et al., 2020; SAFRAN et al., 2021).

Figura 3. Localização e constituição dos íntrons (blocos azuis) dos genes OCA2 e HERC2, vizinhos na porção q13.1 do cromossomo 15.



Fonte: Thorvaldsdottir, Robinson e Mesirow (2012).

O gene MC1R codifica o receptor transmembrana do hormônio estimulante de melanócitos (MSH). A ligação dessa molécula com seu receptor também serve como ativador da produção de eumelanina. Quando determinadas mutações ocorrem neste gene, a proteína final adquire uma conformação que impede a sua ligação com o MSH, causando assim um aumento na produção de feomelanina. Além de provocar diversidade na pigmentação de estruturas de cabelos e pele, as variações na produção dos dois tipos de melanina a partir de receptores não-funcionais são consideradas como fatores de risco para diferentes tipos de câncer de pele. A eumelanina atua como uma forma de proteção frente à exposição aos raios UV, propriedade que não é inerente à feomelanina. Com isso, tais raios têm maior chance de desencadear mutações carcinogênicas (ZANNA et al., 2008; DORGALELEH et al., 2020; SAFRAN et al., 2021).

A tirosinase, codificada pelo gene TYR, atua como principal catalisadora das reações iniciais da conversão da tirosina em DQ e compostos intermediários (SAFRAN et al., 2021), um dos moduladores de tal enzima é o produto de TYRP1 (GHANEM; FABRICE, 2011; DORGALELEH et al., 2020). As moléculas codificadas pelo gene ASIP atuam diretamente sobre o produto de MC1R, de maneira a estimular a produção de feomelanina (MARONAS et al., 2015; SAFRAN et al., 2021), enquanto que o KITLG sintetiza o ligante para um receptor responsável pela regulação da quantidade de melanócitos e a distribuição do pigmento na epiderme ao longo do desenvolvimento (PICARDO; CARDINALI, 2011; MAROÑAS et al., 2015). O gene BNC2 também atua na melanogênese, produzindo uma proteína dedo de zinco relacionada principalmente à diversidade de tons de pele (SAFRAN et al., 2021).

É possível que variações na pigmentação de olhos, cabelo e pele possam também ser explicadas em parte pelo efeito pleiotrópico de HERC2 e OCA2 em outros genes (POŚPIECH et al., 2011; DORGALELEH et al., 2020). Os integrantes da família de carreadores de soluto, como SLC24A4, SLC24A5 e SLC45A2 estão sujeitos a tal efeito, assim como o gene IRF4, o qual codifica um dos fatores de regulação de interferons (POŚPIECH et al., 2011;

DORGALELEH et al., 2020; SAFRAN et al., 2021). Estudos de fenotipagem detectaram também a associação dos genes EXOC2, PIGU (WALSH et al., 2013), ANKRD11, RALY e DEF8 (CHAITANYA et al., 2018) com a melanogênese na pele e cabelo.

1.3.3 Modelos preditivos

Diferentes estudos propõem modelos preditivos a fim de estimar EVCs relacionadas à pigmentação de estruturas a partir de uma amostra de referência. Um dos primeiros modelos de destaque constitui o sistema *multiplex* de genotipagem denominado IrisPlex, o qual indica a probabilidade da cor da íris de um indivíduo corresponder a azul, castanho ou intermediário a partir de uma expressão de regressão logística. Variantes do tipo SNP de seis genes foram selecionados para compor o IrisPlex: HERC2 (rs12913832), OCA2 (rs1800407), SLC24A4 (rs12896399), SLC45A2 (rs16891982), TYR (rs1393350) e IRF4 (rs12203592). A publicação original aponta a eficácia considerável deste método na distinção entre olhos azuis e castanhos. Além disso, uma outra vantagem é a alta sensibilidade do sistema, possibilitando a análise de DNA degradado e de pequenos fragmentos (WALSH et al., 2011).

Em 2013 o modelo foi atualizado com o intuito de englobar também a pigmentação capilar. O HIrisPlex é composto por 24 variantes no total, contemplando os seis parâmetros de sua versão anterior, além de 17 novos SNP e uma alteração do tipo InDel. A matemática probabilística por trás desse sistema é semelhante à primeira publicação, com leves ajustes para que seja possível classificar a cor de cabelos em preto, castanho, loiro ou ruivo. As novas inclusões no *multiplex* foram: MC1R (Y152OCH, N29insA, rs1805006, rs11547464, rs1805007, rs1805008, rs1805009, rs1805005, rs2228479, rs1110400 e rs885479), TYR (rs1042602), EXOC2 (rs4959270), SLC45A2 (rs28777), TYRP1 (rs683), SLC24A4 (rs2402130), KITLG (rs12821256), PIGU/ASIP (rs2378249). O polimorfismo do gene SLC45A2 (rs16891982) e os SNP relativos aos genes HERC2, OCA2 e IRF4 são compartilhados entre a predição da cor de olhos e cabelos (WALSH et al., 2013).

A última atualização do sistema incluiu também a predição do tom de pele, utilizando métodos similares aos descritos anteriormente para categorizar essa EVC de acordo com a classificação de Fitzpatrick: muito pálida (I), pálida (II), intermediárias (III e IV), morena escura (V) e negra (VI) (WARD et al., 2017; CHAITANYA et al., 2018). O HIrisPlex-S é composto pelas 24 variantes de sua versão anterior e pela adição de 17 outros SNP: ANKRD11 (rs3114908), BNC2 (rs10756819), SLC24A4 (rs17128291), HERC2 (rs2238289, rs1129038, rs6497292 e rs1667394), TYR (rs1126809), OCA2 (rs1470608, rs1800414, rs12441727 e rs1545397), SLC24A5 (rs1426654), ASIP (rs6119471), RALY (rs6059655),

MC1R (rs3212355) e DEF8 (rs8051733). A predição do tom de pele envolve tanto os novos SNP quanto 19 variantes presentes no HIrisPlex. A publicação do novo sistema foi acompanhada pelo lançamento de uma *webtool* capaz de fazer a estimativa dos três parâmetros desejados a partir da entrada de dados relativos aos genótipos dos marcadores analisados (CHAITANYA et al., 2018).

Entretanto, a aplicação da tríade de *multiplex* que compõem o HIrisPlex-S em diferentes populações evidenciou as limitações de seu uso. Em primeiro lugar, classificações do tipo intermediária (incluindo cabelo castanho claro/loiro escuro) possuem menor probabilidade de estarem corretas. Ademais, a taxa de predições corretas tende a ser menor e são obtidos mais resultados inconclusivos em populações que possuem um grau de miscigenação maior do que a amostra de referência para a elaboração do sistema (DEMBINSKI; PICARD, 2014; DARIO et al., 2015; SALVORO et al., 2019).

Outros modelos surgiram com o mesmo objetivo, utilizando variantes e metodologias diferentes para inferir sobre EVCs relacionados à pigmentação de estruturas. O método conhecido como *Snipper* utiliza classificadores bayesianos em suas predições e possui também uma *webtool* para análises gratuitas (RUIZ et al., 2013). Hart e colaboradores (2013) aplicaram testes de proporção binomial para obter suas predições, enquanto que Allwood e colaboradores (2013) empregaram uma árvore de decisão em suas estimativas.

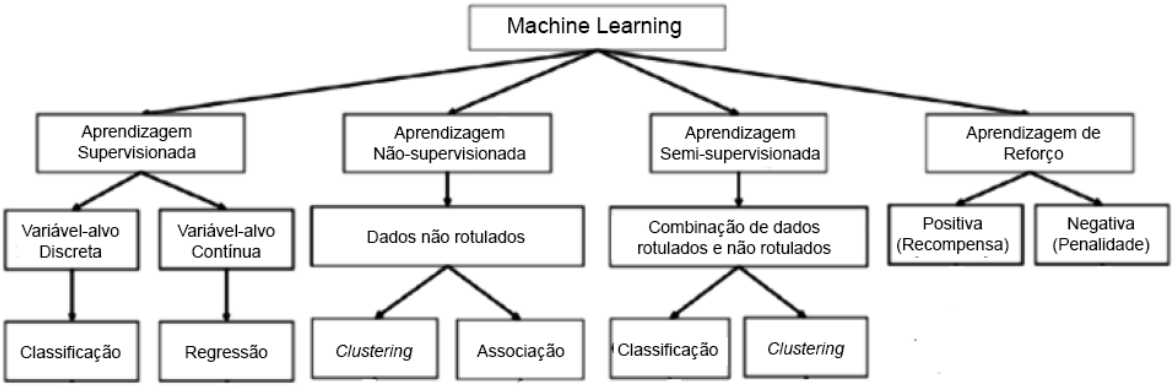
1.4 *MACHINE LEARNING*

O conceito de *Machine Learning* (ML) envolve o aprendizado a partir da experiência dentro do contexto computacional, ou seja, algoritmos que buscam padrões para seleção e predição de variáveis desejadas ou o aperfeiçoamento de um determinado programa de acordo com a sua aplicação ao longo do tempo. Existem pelo menos quatro diferentes tipos de métodos que permitem essa aprendizagem: supervisionada, não-supervisionada, semi-supervisionada e de reforço (HAN; KAMBER; PEI, 2012; JORDAN; MITCHELL, 2015; SARKER, 2021).

As três primeiras abordagens estão relacionadas à natureza dos dados analisados. Se tais dados estiverem previamente rotulados ou classificados (cada *input* possui um único *output*), o algoritmo aplicado funcionará a partir da aprendizagem supervisionada. Caso rótulos não estejam presentes, busca-se traçar parâmetros e criar classificações finais através das propriedades dos dados analisados com algoritmos de aprendizagem não-supervisionada (HAN; KAMBER; PEI, 2012; JORDAN; MITCHELL, 2015; SARKER, 2021). É possível ainda que a resolução de um problema específico envolva dados rotulados e não-rotulados, de

forma a caracterizar a aprendizagem semi-supervisionada (HAN; KAMBER; PEI, 2012; SARKER, 2021). Já a aprendizagem de reforço está associada a um processo de reconhecimento de padrões mais independente e automatizado por meio de sistemas de recompensa ou penalidade, e se faz presente em softwares mais complexos, como os de robótica, por exemplo. (KAELBLING; LITTMAN; MOORE, 1996; SARKER, 2021). As diferentes abordagens de aprendizagem e suas respectivas subdivisões estão representadas na Figura 4.

Figura 4. Abordagens de aprendizagem em ML e suas respectivas subdivisões.



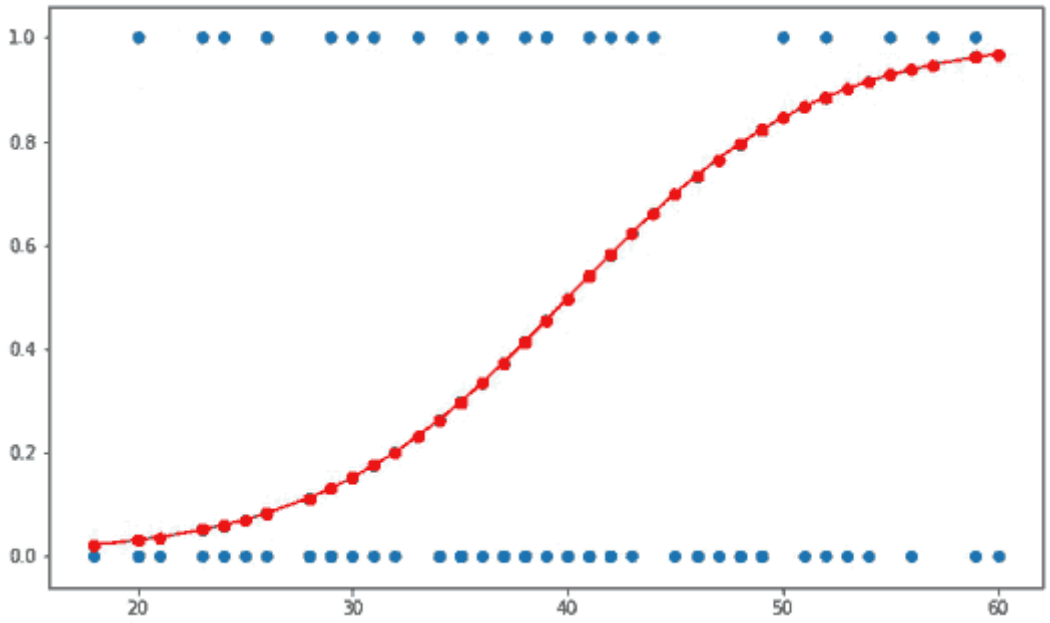
Fonte: Sarker (2021).

O presente trabalho foca na aprendizagem supervisionada que, por sua vez, pode ser dividida em algoritmos de classificação e regressão de acordo com o tipo de variável que se busca predizer (HAN; KAMBER; PEI, 2012; JORDAN; MITCHELL, 2015; SARKER, 2021). Algoritmos de classificação envolvem a predição de variáveis tidas como discretas, ou seja, que podem assumir valores específicos dentro de um conjunto determinado. Por outro lado, algoritmos de regressão admitem variáveis contínuas, as quais podem assumir qualquer valor dentro de um intervalo específico (HAN; KAMBER; PEI, 2012; KALIYADAN; KULKARNI, 2019). Os modelos aplicados na análise em questão são de classificação, tendo em vista que os dados utilizados são do tipo discretos.

Comumente aplicadas em problemas de classificação, expressões de Regressão Logística (*Logistic Regression* - LR) calculam a probabilidade de um determinado conjunto de dados pertencer às possíveis categorias a partir de uma função sigmóide (CESSIE; VAN HOUWELINGEN, 1992; SARKER, 2021) (Figura 5). As árvores de Decisão (*Decision Trees* - DT) utilizam a estrutura de nós e galhos para realizar predições; cada nó define um atributo

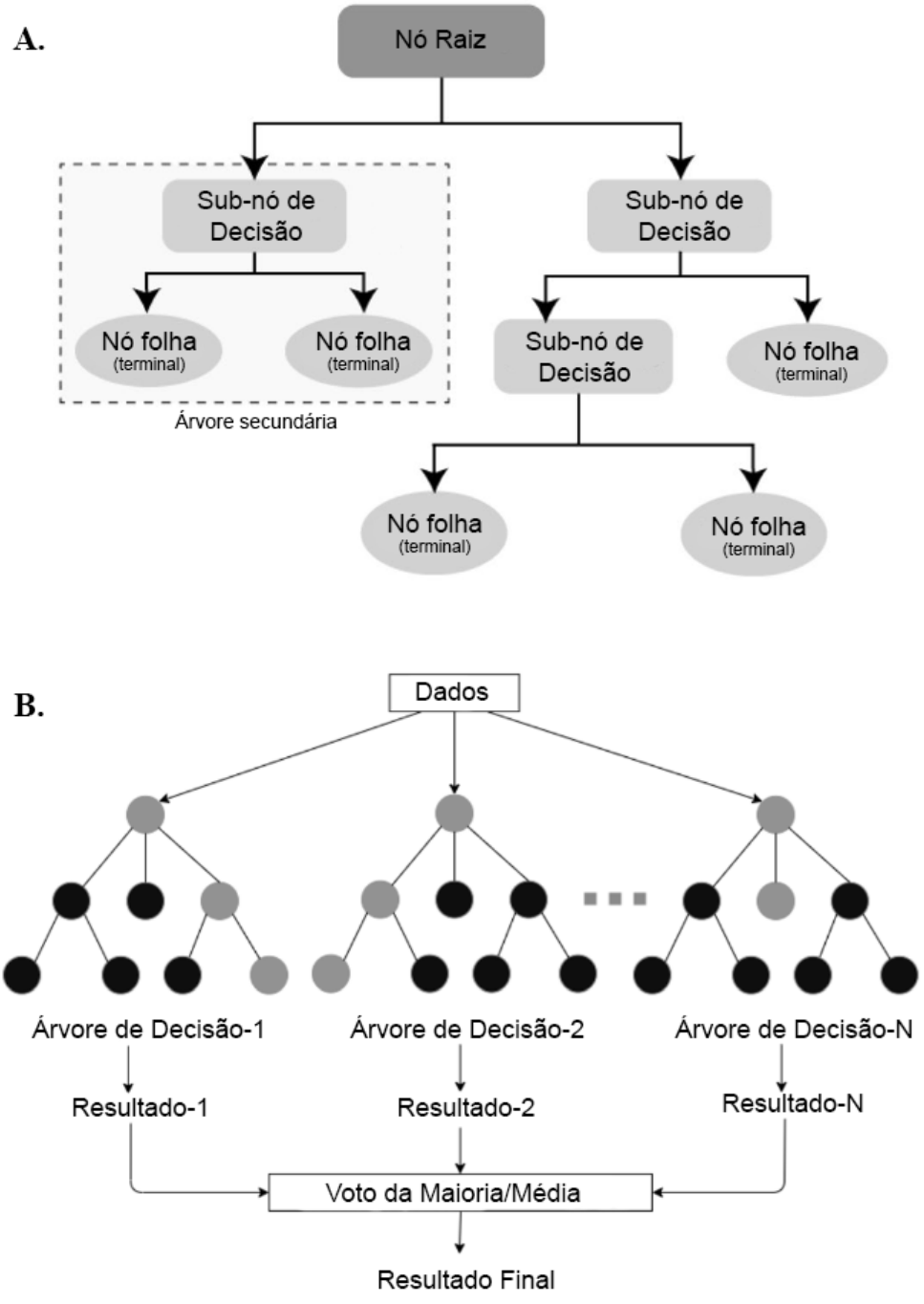
e os galhos que se originam a partir desse estão relacionados a características específicas, podendo levar a um resultado final ou a formação de outro nó (SALZBERG, 1994, HAN; KAMBER; PEI, 2012; SARKER, 2021) (Figura 6). Os classificadores do tipo *Random Forest* (RF) funcionam a partir de múltiplas árvores de decisão em paralelo, em que cada uma gera uma classificação e o resultado final é a categoria com maior frequência dentro desse conjunto (HAN; KAMBER; PEI, 2012; BREIMAN, 2001; SARKER, 2021) (Figura 6). Os algoritmos do tipo *Naive Bayes* (NB) utilizam conceitos de probabilidade e o teorema de Bayes para fazer predições levando em consideração a ideia de independência entre cada par de características (HAN; KAMBER; PEI, 2012; WEBB, 2016; SARKER, 2021) (Figura 7). A construção de um hiper-plano onde os dados são expressos através de pontos agrupados também é uma abordagem que pode ser empregada em classificações, e para isso se recorre aos algoritmos do tipo *Support Vector Machines* (SVM) (HAN; KAMBER; PEI, 2012; PEDREGOSA et al., 2012; SARKER, 2021) (Figura 8).

Figura 5. Esquema mostrando o funcionamento dos algoritmos de Regressão Logística. A curva vermelha representa a função sigmóide do classificador. Pontos acima da curva são incluídos em um categoria, pontos abaixo da curva são incluídos em outra.



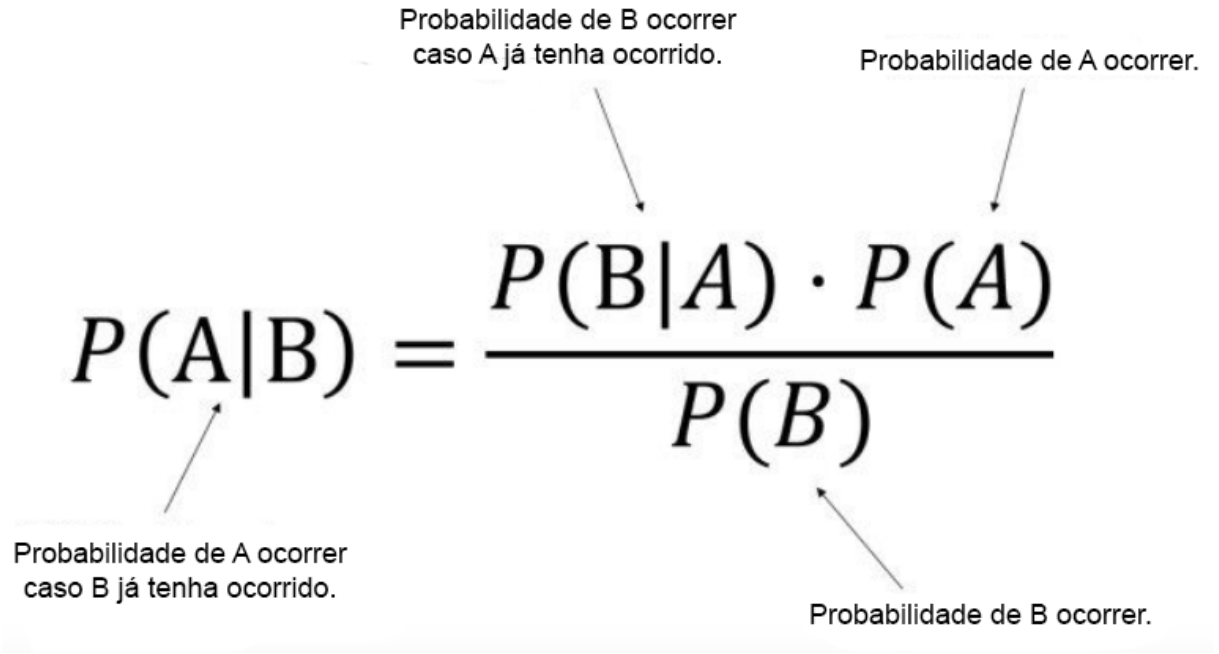
Fonte: Menon (2022).

Figura 6. Esquema mostrando o funcionamento dos algoritmos de Árvore de Decisão (A) e de *Random Forest* (B).



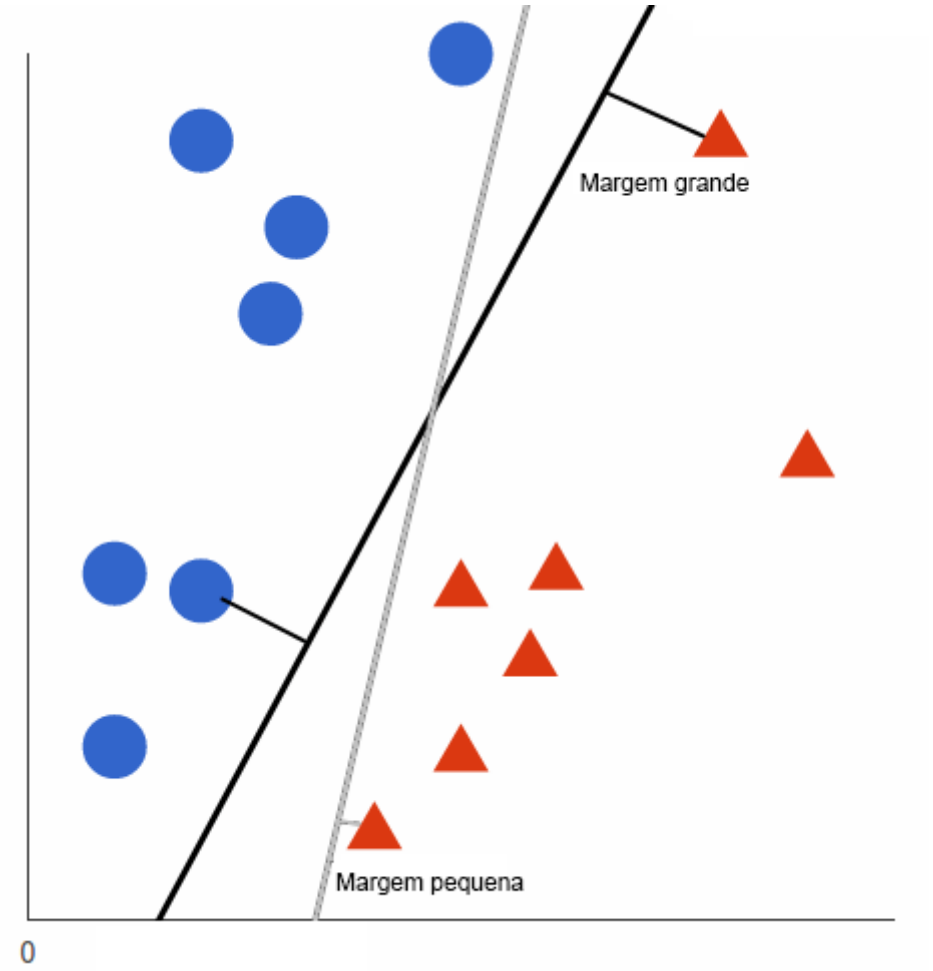
Fonte: Sarker (2021).

Figura 7. Fórmula matemática do classificador *Naive Bayes* envolvendo o cálculo com probabilidades e a independência de cada característica.



Fonte: Chauhan (2022).

Figura 8. Esquema mostrando o funcionamento do algoritmo *Support Vector Machine*. A linha mais escura representa o melhor hiperplano para a separação dos dados em duas categorias (círculo azul e triângulo vermelho), com uma margem grande entre cada ponto e a divisão. Já a linha clara representa um hiperplano menos eficiente para a separação dos dados e com margem menor entre os pontos e a divisão.



Fonte: Le (2022).

Dentro do contexto forense, a predição e o reconhecimento de padrões a partir das diferentes metodologias de ML podem ser empregados na detecção de associações criminosas e eventuais atos ilícitos, tais como fraudes, perpetuados em meios digitais (QADIR; VAROL, 2020). Além disso, os ramos da genética (WALSH et al., 2011; ALLWOOD et al., 2013; HART et al., 2013; WALSH et al., 2013; RUIZ et al., 2013; CHAITANYA et al., 2018) e de análise de imagens (NOWROOZI et al., 2021) e alimentos (RANBIR et al., 2022) podem se beneficiar da aplicação dos algoritmos anteriormente mencionados.

2. JUSTIFICATIVA

Os sistemas preditivos de maior destaque na área forense foram elaborados com base em dados obtidos de diferentes coortes dentro da população europeia (WALSH et al., 2011; WALSH et al., 2013; RUIZ et al., 2013). Em versões mais recentes do HIrisPlex-S, o *multiplex* proposto foi também testado em indivíduos com variadas origens biogeográficas, a partir de dados anteriormente coletados por outros projetos de genotipagem humana, sob a suposição de que todos possuíam os mesmos fenótipos. Ainda que as conclusões previamente reportadas sugiram que tal modelo possa ser extrapolado para outras populações globais (CHAITANYA et al., 2018), estudos posteriores, visando a aplicabilidade dessa metodologia em países como Portugal e Itália, alcançaram resultados conflitantes (DARIO et al., 2015; SALVORO et al., 2019).

O mesmo ocorreu quando o HIrisPlex-S foi testado na população estadunidense, a qual notoriamente apresenta uma maior taxa de miscigenação em comparação com a Europa (DEMBINSKI; PICARD, 2014). Por certo, quando o assunto é o continente americano, em particular a região sócio-cultural conhecida como América Latina, sabe-se que processos históricos levaram a uma constituição genética populacional tida como tripla, envolvendo populações ameríndias, européias e africanas (RUIZ-LINARES et al., 2014). De fato, Palmal e colaboradores (2021), após testarem o modelo em uma amostra que contempla 6500 indivíduos originários do México, Colômbia, Peru, Chile e Brasil, sugerem que o sistema HIrisPlex-S deve ser aplicado com ressalvas dentro de coortes latinoamericanas. Um estudo restrito à população brasileira, realizado por Carratto e colaboradores (2019), também encontrou resultados que corroboram as falhas previamente reportadas na predição das três EVCs.

Dessa forma, faz-se necessário buscar métodos alternativos aos já publicados na área da fenotipagem forense, com o intuito de construir um modelo que se adeque melhor às características populacionais da América Latina. Assim sendo, a hipótese do presente trabalho é que, com o auxílio de ferramentas de ML, seja possível estabelecer um preditor baseado em uma amostra da população brasileira, visando um melhor desempenho frente ao sistema HIrisPlex-S.

3. OBJETIVOS

3.2 OBJETIVOS GERAL

Aplicar modelos do tipo *machine learning* para a predição de características externamente visíveis associadas à pigmentação de estruturas (olhos, cabelo e pele) a partir de um conjunto de SNPs em uma amostra da população brasileira.

3.2 OBJETIVOS ESPECÍFICOS

- Verificar a frequência de 49 marcadores relacionados à pigmentação de estruturas;
- Selecionar os marcadores de maior importância na definição de cor de olhos, cabelo e tom de pele na amostra;
- Aplicar modelos preditivos com cinco algoritmos diferentes a partir dos marcadores selecionados e calibrar seus hiperparâmetros;
- Avaliar e comparar os modelos criados a partir de sua acurácia e sensibilidade.

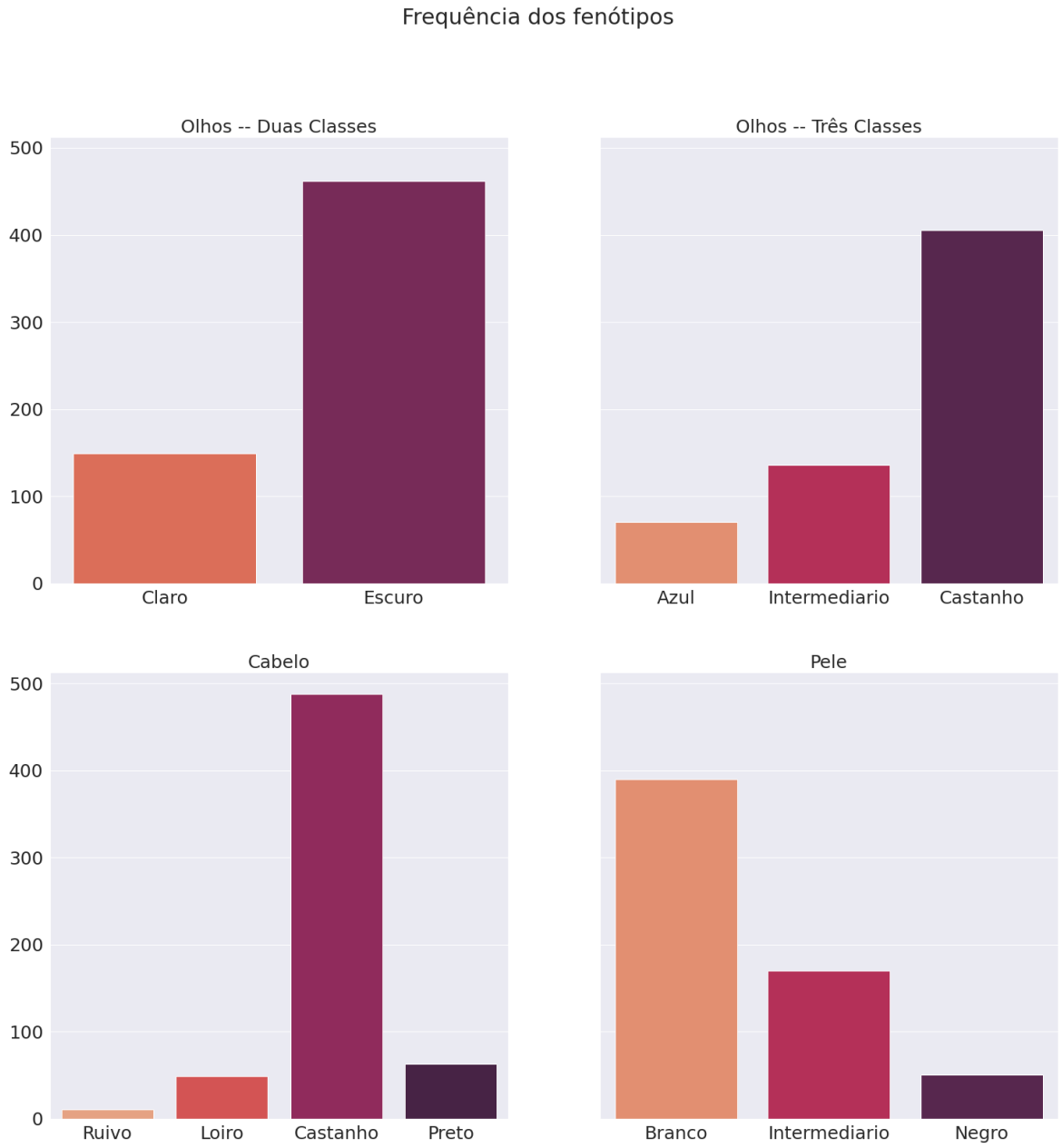
4. METODOLOGIA

4.1 AMOSTRA

Os modelos preditivos foram aplicados a partir de um recorte da população brasileira, cujos dados foram cedidos pelo Laboratório de Imuno-Hematologia e Hematologia Forense, do Depto de Medicina Legal, Bioética, Medicina do Trabalho e Medicina Física e Reabilitação, da Faculdade de Medicina da Universidade de São Paulo. A amostra contém os resultados da genotipagem de 49 marcadores em 611 indivíduos.

Entre as variantes presentes, 41 delas compõem o sistema HIrisPlex-S (Apêndice A) enquanto que outros oito SNPs (Apêndice B) foram sequenciados por terem sido associados à pigmentação de estruturas na população brasileira em estudos prévios (DURSO et al., 2014; LIMA; GONÇALVES; FRIDMAN, 2015; ANDRADE et al., 2017). Além disso, a amostra contém também dados sobre o sexo biológico (398 mulheres e 213 homens) e os fenótipos de cor de olho, cabelo e tom de pele dos participantes. A quantidade de indivíduos alocados em cada uma das categorias de fenótipos encontradas na amostra está representada na Figura 9.

Figura 9. Frequência dos fenótipos identificados na amostra analisada: cor de olhos (dividido em classificações de duas e três classes), cabelo e tom de pele.



Fonte: Autor (2022).

4.2 PRÉ-PROCESSAMENTO E TRIAGEM DOS MARCADORES

O pré-processamento da amostra e a seleção de marcadores para os modelos preditivos se deram a partir de pacotes desenvolvidos na linguagem *Python* (versão 3.9.7) especificamente para a manipulação e visualização de dados. Inicialmente, verificou-se a frequência alélica de todos os 49 marcadores, com o intuito de identificar a taxa de variação desses dentro do recorte populacional em questão. Marcadores monomórficos, ou seja, com

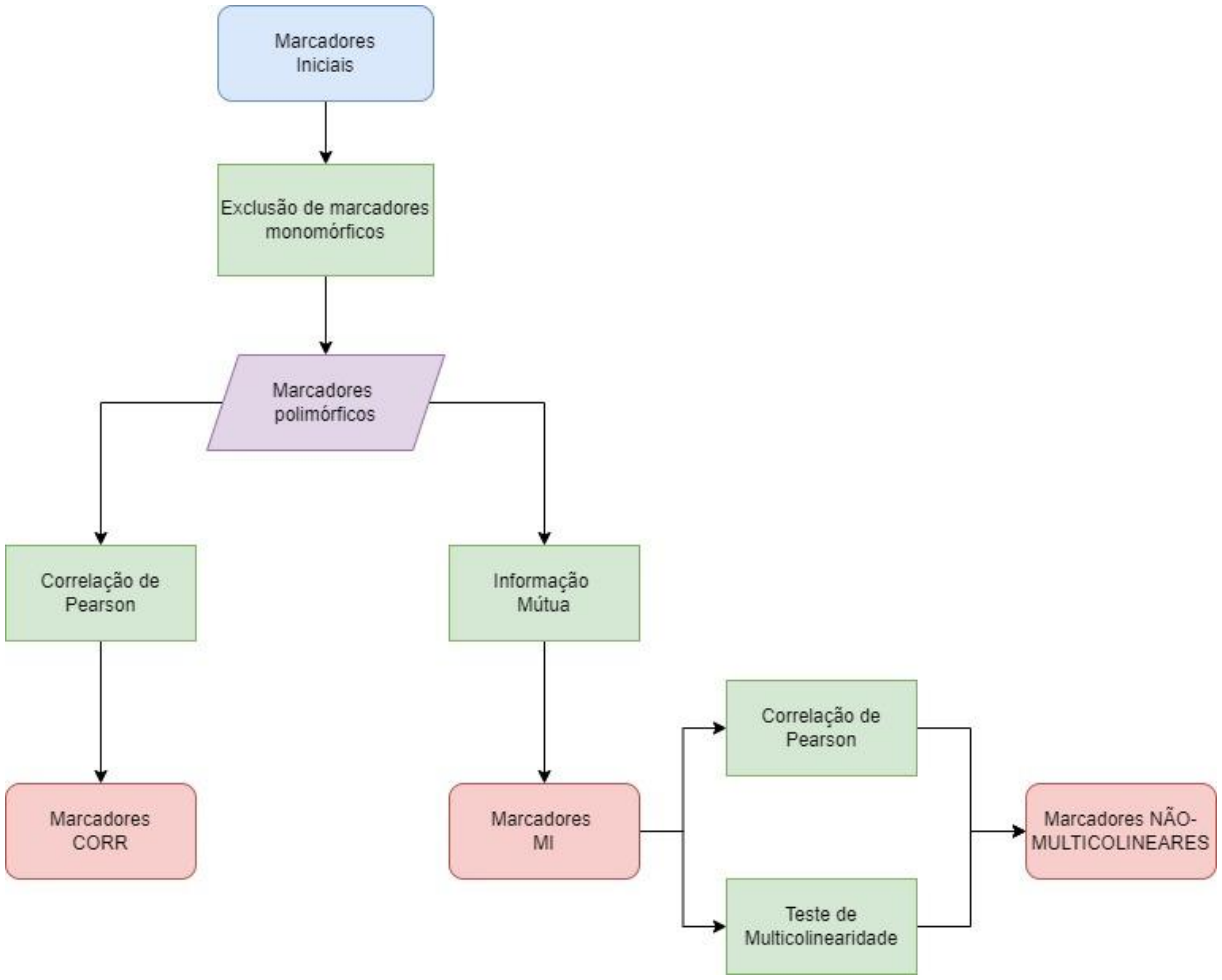
taxa de variação menor do que 1%, foram excluídos das análises posteriores (NUSSBAUM; MCINNES; WILLARD, 2016; ALBERTS et al., 2017).

Partindo desse ponto, uma nova triagem resultou em três subgrupos distintos de marcadores selecionados por meio de diferentes metodologias estatísticas. Dois desses conjuntos foram originados a partir da verificação entre as variáveis independentes e o fenótipo, utilizando os testes de correlação linear de Pearson (ZOU; TUNCALI; SILVERMAN, 2003; SCHOBER; BOER; SCHWARTE, 2018) e de Informação Mútua (ZHOU; WANG; ZHU, 2021). No primeiro caso, removeram-se marcadores sem correlação com os quatro fenótipos, ou seja, cujo valor variou entre 0,1 e -0,1 (SCHOBER; BOER; SCHWARTE, 2018). Para o segundo subgrupo, levou-se em consideração que o teste de Informação Mútua pode gerar um valor entre zero e um, e que quanto mais distante do primeiro, maior é a relação entre variáveis (ZHOU; WANG; ZHU, 2021). Sendo assim, para o desenvolvimento deste trabalho, foram excluídas as variantes em que esse valor correspondeu a menos de 1% de relação (0.01).

Finalmente, o terceiro conjunto de marcadores foi formado a partir de um passo extra com o subgrupo de SNPs obtidos a partir do teste de Informação Mútua. Tal conjunto foi triado novamente de acordo com a multicolinearidade, conceito associado ao grau de relação dos marcadores entre si (NASIR et al., 2020). A presença de duas variáveis altamente relacionadas dentro de um algoritmo de ML pode ser redundante (NASIR et al., 2020) ao ponto de interferir no processo de aprendizagem do modelo. Para verificar isso, utilizou-se a função *Variation Inflation Factor* (VIF) presente na biblioteca *statsmodels* (versão 0.14.0) e um novo teste de Pearson entre pares de variantes. A partir de tais análises foi possível verificar quais marcadores possuíam alto grau de relação com outros (VIF superior a 10) e também excluir uma das variantes de um par com valor elevado de correlação entre si (superior a 0,85 ou inferior a -0,85) (ZOU; TUNCALI; SILVERMAN, 2003; SCHOBER; BOER; SCHWARTE, 2018).

O fluxograma presente na Figura 10 esquematiza o processo de triagem de marcadores e a separação dos três subgrupos anteriormente mencionados.

Figura 10. Fluxograma representando a triagem de marcadores.



Fonte: Autor (2022).

Tendo em vista a natureza categórica dos dados, os genótipos foram convertidos em valores numéricos dependendo da quantidade de alelos de variante presentes (0, 1 ou 2). Por exemplo, para o SNP HERC2 rs12913832, cujo alelo variante é G, a conversão resultaria nas seguintes alterações: 0 (AA), 1, (AG) e 2 (GG). De maneira similar, os fenótipos foram também convertidos em valores numéricos crescentes considerando as características que surgiram mais recentemente dentro do processo evolutivo humano e o espectro de possibilidades. Sendo assim, o Quadro 1 abaixo apresenta os valores convencionados para as EVCs analisadas.

Quadro 1. Conversão dos fenótipos a valores numéricos.

Fenótipo	0	1	2	3
Olhos (2 categorias)	Escuro	Claro	--	--
Olhos (3 categorias)	Castanho	Intermediário	Azul	--
Cabelo	Preto	Castanho	Loiro	Ruivo
Pele	Negro	Intermediário	Branco	--

Fonte: Autor (2022).

O último passo da etapa de pré-processamento consistiu na exclusão de todos os indivíduos que tiveram pelo menos um dos marcadores com problemas na genotipagem em etapas anteriores à elaboração do presente trabalho, uma vez que a inclusão de dados "*missing*" ou faltantes nos algoritmos não é viável.

4.3 APLICAÇÃO DOS MODELOS

Cinco algoritmos foram selecionados para a testagem dentro da amostra analisada a partir dos grupos de marcadores previamente selecionados. Três desses classificadores foram escolhidos por terem sido abordagens previamente utilizadas em estudos de fenotipagem de traços relacionados à pigmentação de estruturas; os modelos englobados pelo HIrisPlex-S utilizam Regressões Logísticas (WALSH et al., 2011; WALSH et al., 2013; CHAITANYA et al., 2018), enquanto que o *Snipper* é baseado em um algoritmo do tipo Naive Bayes (RUIZ et al., 2013), e, finalmente, Allwood e colaboradores (2013) fizeram predições a partir de uma Árvore de Decisão. Outros dois classificadores, *Random Forest* e *Support Vector Machine*, foram escolhidos por serem abordagens clássicas para resolução de problemas de classificação por meio de ML (SARKER, 2021).

Considerando os métodos escolhidos, realizou-se a calibragem dos hiperparâmetros visando a máxima acurácia possível com os dados disponíveis. Cada um dos cinco algoritmos faz predições a partir de funções estatísticas particulares. A maneira como os modelos aprendem os padrões presentes na amostra está relacionado a tais funções, de modo que é possível controlar e ajustar os detalhes (hiperparâmetros) envolvidos nesse processo de aprendizagem. Com isso em mente, utilizou-se a função *GASearchCV*, presente no módulo *sklearn-genetic* (versão 0.5.1), a qual associa uma metodologia computacional inspirada nos processos de seleção natural (KATOCH; CHAUHAN; KUMAR, 2020) a diversas validações cruzadas a fim de testar conjuntos de hiperparâmetros até encontrar a melhor combinação dos

mesmos, levando em conta a acurácia dos algoritmos. Tais análises resultaram em uma tabela, semelhante à Tabela 1, com valores para cada um dos modelos e dos fenótipos estudados.

Tabela 1. Exemplo representativo dos valores (acurácia e parâmetros associados) gerados a partir da calibragem de hiperparâmetros com a função *GASearchCV*.

	model	best_score	best_params
0	svm	0.862319	{'C': 2, 'kernel': 'rbf', 'gamma': 'scale', 's...
1	random_forest	0.867150	{'bootstrap': False, 'max_depth': 12, 'max_lea...
2	logistic_regression	0.806763	{'C': 3, 'solver': 'newton-cg', 'warm_start': ...
3	decision_tree	0.869565	{'criterion': 'gini', 'splitter': 'random', 'm...
4	multinomial_naive_bayes	0.828502	{'alpha': 26, 'fit_prior': False}

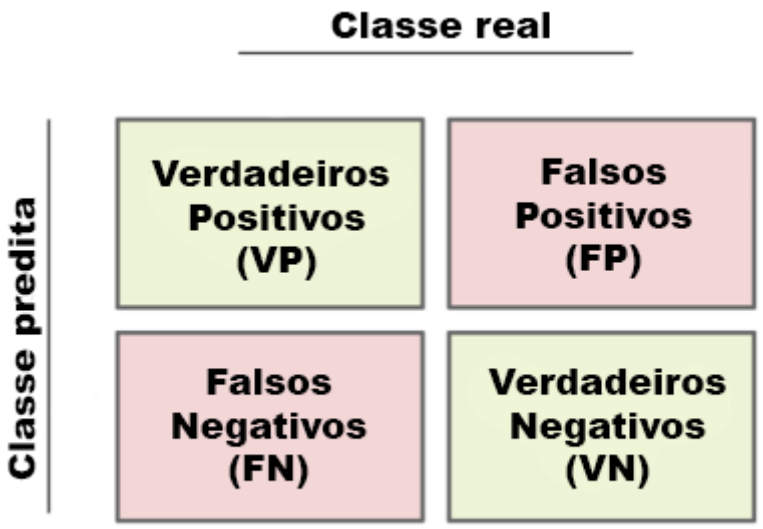
Fonte: Autor (2022).

Com a definição dos hiperparâmetros referentes aos cinco modelos, os dados foram alocados de maneira randômica em dois grupos: treino (80%) e teste (20%). Em seguida, foram criados os algoritmos de classificação para cada um quatro fenótipos analisados de maneira individual e isolada. Para tal, foram utilizados os pacotes *Pycaret* (versão 2.3.6) e *scikit-learn* (versão 1.0.2).

4.4 AVALIAÇÃO DO DESEMPENHO DOS MODELOS

O desempenho dos classificadores foi avaliado sob duas principais perspectivas: acurácia e a sensibilidade dos modelos. O primeiro dos pontos foi calculado de forma automática pelos pacotes supracitados de acordo com a taxa de predições corretas realizadas por cada modelo. Além disso, foram também criadas matrizes de confusão, as quais esquematizam de forma didática o número de verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN) obtidos a partir das predições do algoritmo (LUQUE et al., 2019), com o intuito de verificar a acurácia de cada uma das categorias, como evidenciado na Figura 11.

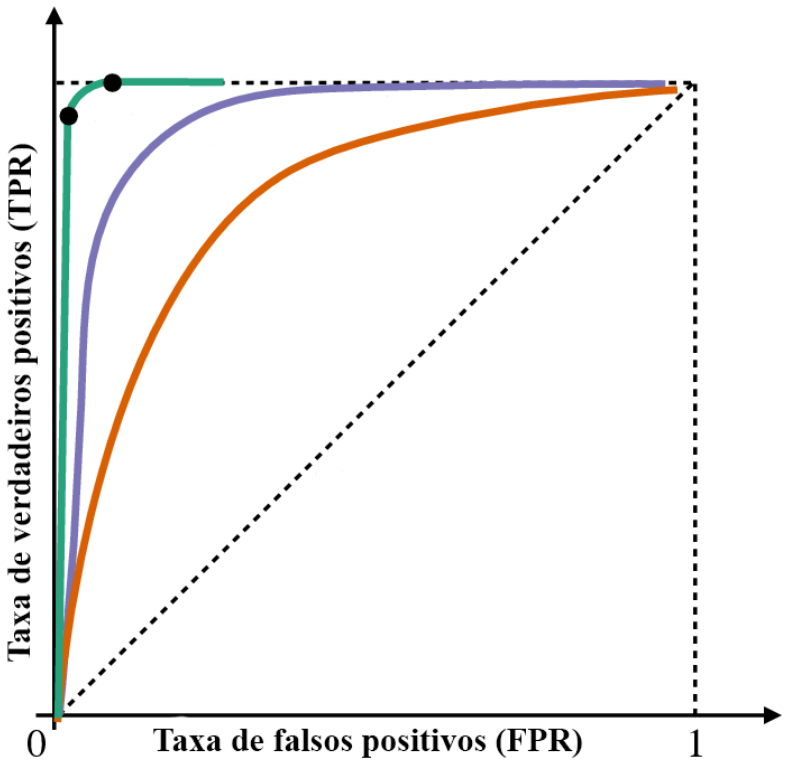
Figura 11. Exemplo de matriz de confusão.



Fonte: Bittrich *et al*. (2019).

A sensibilidade dos algoritmos de classificação foi avaliada através das curvas de característica de operação do receptor (*receiver operating characteristic* ou ROC), também geradas na etapa de análise dos resultados. Quando plotadas em um gráfico, as curvas ROC são representações da relação entre a taxa de verdadeiros positivos (*true positive rate* ou TPR) e a taxa de verdadeiros negativos (*true negative rate* ou FPR) calculada ao longo de todos os pontos de dados de uma amostra. A eficácia de predição do modelo em questão pode ser inferida de acordo com a proximidade de sua respectiva curva ROC ao canto superior e esquerdo do gráfico, o chamado índice de Younden, que representa o balanço ideal entre sensibilidade e especificidade. Outra forma de se avaliar um preditor seria pelo cálculo da área sob a curva (*area under the curve* ou AUC), que, por sua vez, pode variar entre 0,5 e 1, sendo esse último o valor de um classificador perfeitamente adaptado aos dados imputados (HOO; CANDLISH; TEARE, 2017). A Figura 12 esquematiza as características das curvas ROC de maneira ilustrativa.

Figura 12. Exemplos de curvas ROC para classificadores: mais acurado (verde), de acurácia intermediária (roxo) e menos acurado (laranja). A curva tracejada indica a linha de base para o algoritmo em questão.



Fonte: Gwirtz *et al*. (2020).

5. RESULTADOS

5.1 PRÉ-PROCESSAMENTO E TRIAGEM DOS MARCADORES

A verificação das frequências genotípicas dentro da população resultou na exclusão de seis marcadores com menos de 1% de variação dentro da amostra, ou seja, monomórficos: OCA2 rs1800416 (100% AA), SLC24A5 rs16960620 (99,32% AA), MC1R rs312262906 (100% AA), MC1R rs11547464 (99,49% GG), MC1R rs1805006 (99,32% CC) e MC1R rs201326893 (100% CC).

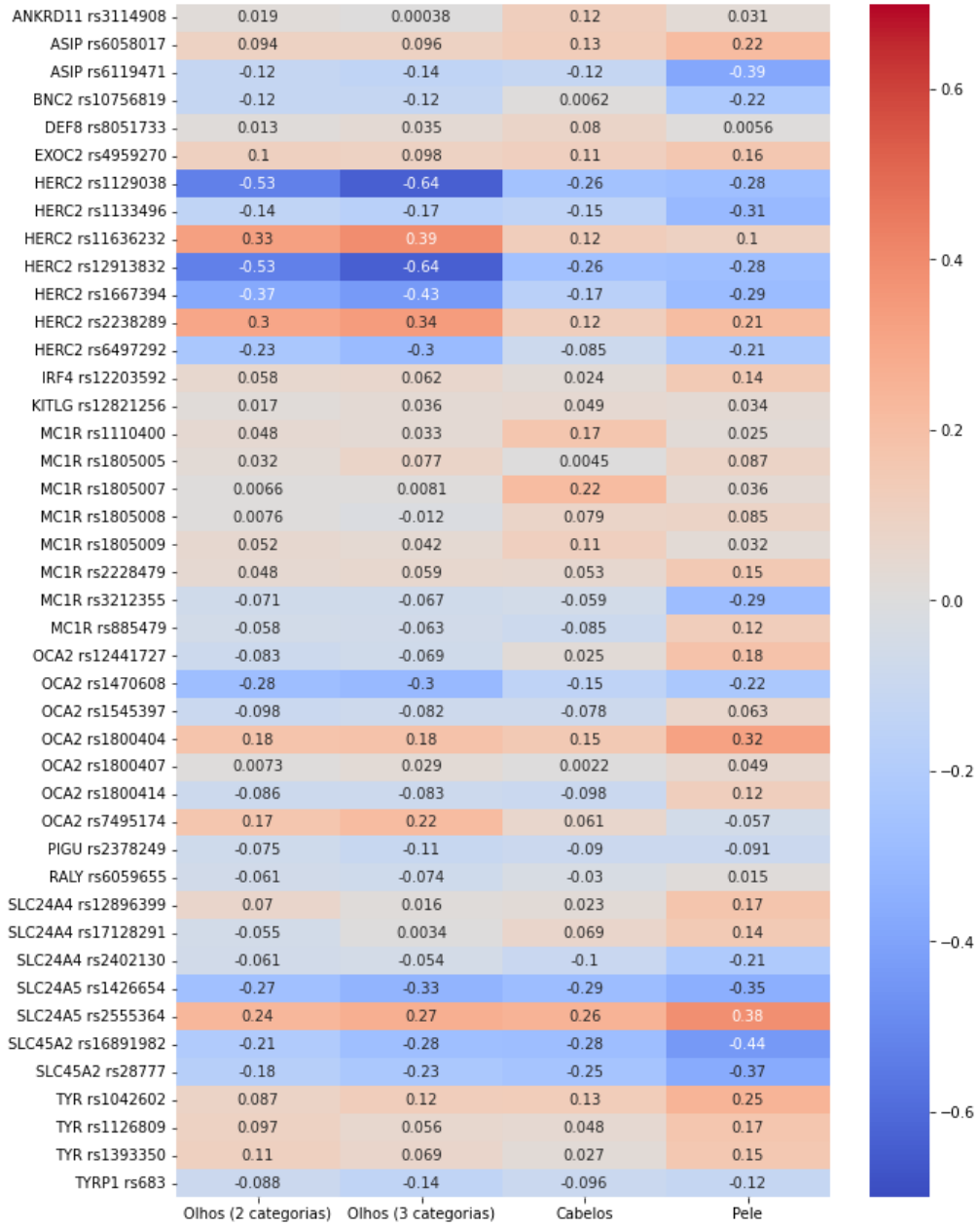
5.1.1 Marcadores submetidos ao teste de correlação linear de Pearson (CORR)

O conjunto de SNPs polimórficos da amostra foi submetido a testes de correlação linear de Pearson, a fim de quantificar a relação entre os mesmos e os quatro fenótipos analisados. O resultado de tais testes pode ser visualizado por meio de gráficos de mapa de calor, representados na Figura 13. A partir dos índices de correlação calculados, foram removidas desse subgrupo de marcadores sete variantes cujo valor calculado estava no intervalo entre 0,1 e -0,1 (MC1R rs1805005, MC1R rs1805008, OCA2 rs1545397, OCA2 rs1800407, KITLG rs12821256, RALY rs6059655, DEF8 rs8051733). Ao total, a amostra final nesse caso contou com 36 marcadores e 488 indivíduos.

5.1.2 Marcadores submetidos ao teste de Informação Mútua (MI)

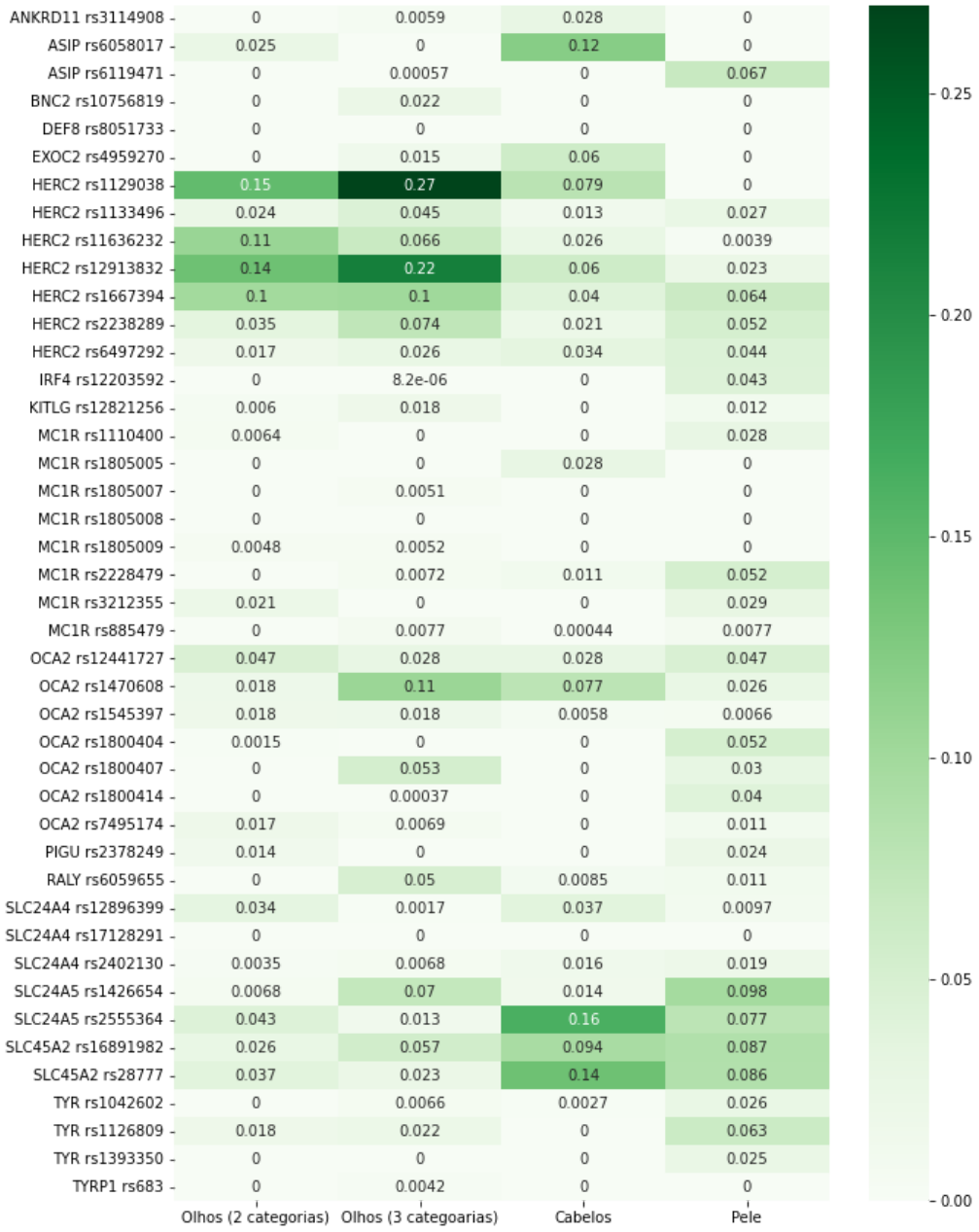
Os marcadores polimórficos na população foram também submetidos a teste de informação mútua, com o intuito de identificar as relações dos mesmos com os quatro fenótipos analisados. O resultado de tais testes pode ser visualizado na Figura 14. Sete SNPs apresentaram valores inferiores a 0.01 neste teste, de forma a serem removidos ao final dessas análises (MC1R rs885479, TYRP1 rs683, DEF8 rs8051733, SLC24A4 rs17128291, MC1R rs1805009, MC1R rs1805008 e MC1R rs1805007). A amostra final contou com 36 marcadores e 494 indivíduos.

Figura 13. Mapa de calor com o resultado dos testes de correlação linear de Pearson entre os SNPs analisados e os quatro fenótipos. Sete marcadores com valores presentes entre o intervalo 0,1 e -0,1 foram excluídos da amostra final (MC1R rs1805005, MC1R rs1805008, OCA2 rs1545397, OCA2 rs1800407, KITLG rs12821256, RALY rs6059655, DEF8 rs8051733).



Fonte: Autor (2022).

Figura 14. Mapa de calor com o resultado dos testes Informação Mútua entre os SNPs analisados e os quatro fenótipos Sete marcadores com valores inferiores à 0.01 foram excluídos da amostra final (MC1R rs885479, TYRP1 rs683, DEF8 rs8051733, SLC24A4 rs17128291, MC1R rs1805009, MC1R rs1805008 e MC1R rs1805007).



Fonte: Autor (2022).

5.1.3 Marcadores submetidos ao teste de Multicolinearidade (MULTICOL)

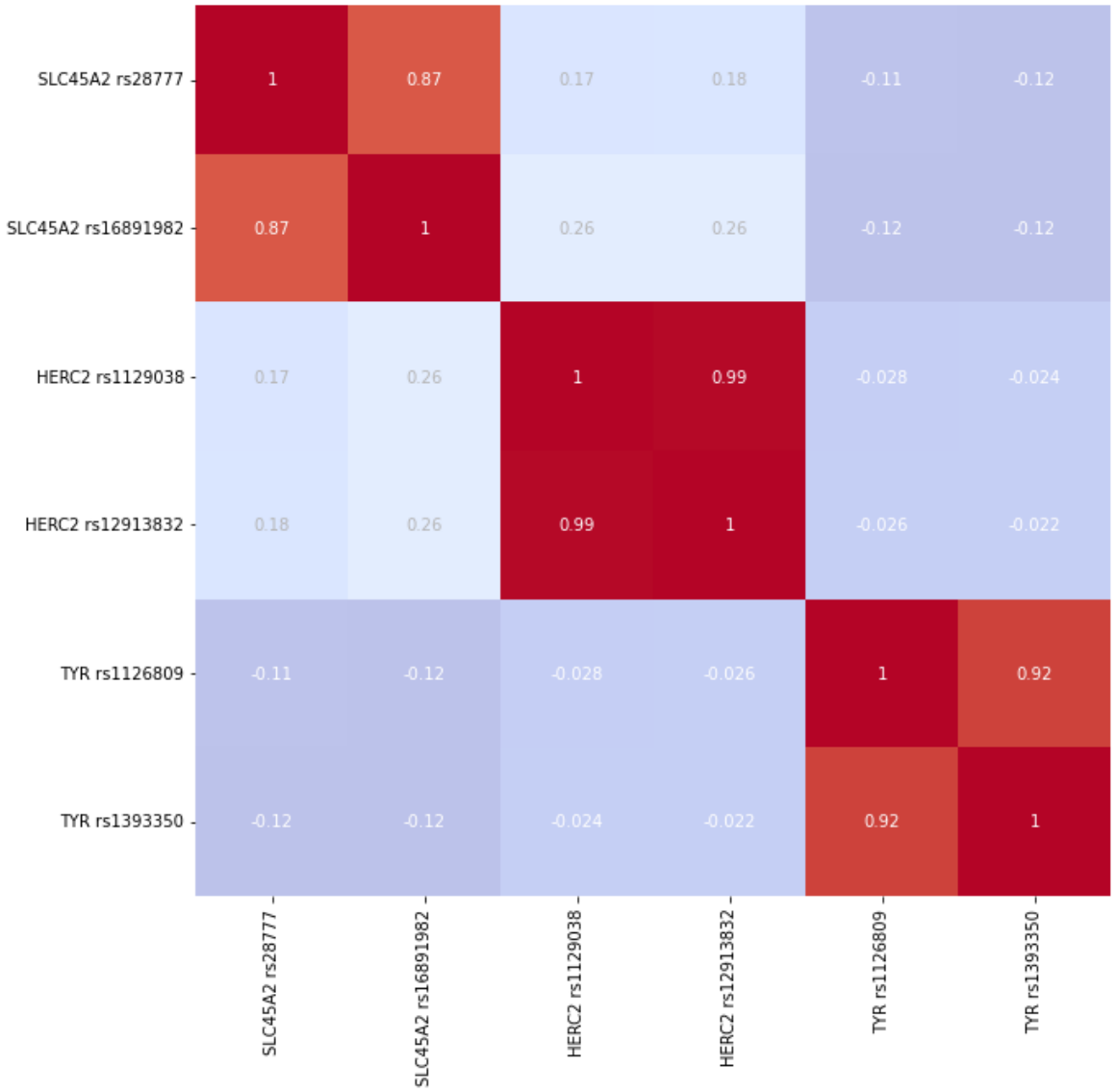
O subconjunto de marcadores selecionados a partir do teste de informação mútua foi submetido a análises de multicolinearidade. Os resultados dos cálculos de VIF podem ser visualizados de maneira individual no Quadro 2. Valores superiores a 10 já podem indicar a presença de multicolinearidade nas variantes analisadas (ZOU; TUNCALI; SILVERMAN, 2003; SCHOBER; BOER; SCHWARTE, 2018). As interações entre pares de SNPs também foram avaliadas, sendo que as relações que geraram valores maiores que 0,85 ou menores que -0,85 neste teste foram consideradas fortes. A Figura 15 apresenta os três casos onde se observou correlação forte entre os marcadores (o mapa de calor completo com as interações de todas as variantes encontra-se no Apêndice C). Ao final dessa etapa, três marcadores foram removidos (SLC45A2 rs28777, TYR rs1126809, HERC2 rs1129038), sendo um de cada par com alto valor de correlação. A amostra final contou com 33 marcadores e 496 indivíduos.

Quadro 2. Resultados do teste VIF para cada um dos marcadores do grupo MULTICOL.

Marcador	VIF	Marcador	VIF
HERC2 rs12913832	332.70	TYR rs1042602	2.48
HERC2 rs1129038	325.99	SLC24A4 rs12896399	2.18
SLC24A5 rs2555364	18.32	SLC24A4 rs2402130	2.16
OCA2 rs7495174	16.32	ASIP rs6119471	2.02
ASIP rs6058017	15.13	OCA2 rs1545397	1.96
HERC2 rs2238289	13.53	HERC2 rs11636232	1.93
SLC45A2 rs16891982	10.48	ANKRD11 rs3114908	1.89
TYR rs1393350	9.30	PIGU rs2378249	1.57
HERC2 rs1667394	9.13	OCA2 rs1800407	1.55
TYR rs1126809	9.03	OCA2 rs1800414	1.52
SLC45A2 rs28777	8.69	HERC2 rs1133496	1.50
OCA2 rs1800404	6.98	IRF4 rs12203592	1.38
OCA2 rs1470608	5.28	MC1R rs1805005	1.35
HERC2 rs6497292	4.26	RALY rs6059655	1.32
OCA2 rs12441727	3.01	MC1R rs2228479	1.24
BNC2 rs10756819	2.79	MC1R rs3212355	1.17
EXOC2 rs4959270	2.72	KITLG rs12821256	1.14
SLC24A5 rs1426654	2.71	MC1R rs1110400	1.01

Fonte: Autor (2022).

Figura 15. Mapa de calor com os maiores resultados do teste de correlação linear de Pearson entre pares de marcadores. Um marcador de cada um dos pares com correlação forte (resultados maiores que 0.85 e menores que -0.85) foram excluídos da amostra final (SLC45A2 rs28777, TYR rs1126809, HERC2 rs1129038).



Fonte: Autor (2022).

Ao longo das etapas de pré-processamento e triagem dos marcadores, as 49 variantes iniciais passaram por diferentes testes a fim de selecionar as variáveis de maior importância em relação aos quatro fenótipos. Primeiramente, foram excluídos seis marcadores que não apresentavam variação dentro da amostra. Em seguida, filtros estatísticos foram aplicados às variantes polimórficas, originando em três subgrupos com conjuntos diferentes de SNPs: CORR (36 marcadores), MI (36 marcadores) e MULTICOL (33 marcadores). Dentre os resultados aqui apresentados, destaca-se a relação entre os genes HERC2 e OCA2 com a cor

de olhos e dos genes SLC24A4, SLC24A5, SLC45A2 e ASIP com os fenótipos de cor de cabelos e pele.

5.2 MODELOS PREDITIVOS

Ao todo, 60 modelos foram aplicados englobando cinco algoritmos (LR, NB, DT, RF e SVM), quatro fenótipos analisados e três conjuntos distintos de marcadores (CORR, MI e MULTICOL). O resultado da calibragem dos hiperparâmetros para cada um dos classificadores encontra-se no Apêndice D.

5.2.1 Modelos CORR

Os algoritmos de maior e menor acurácia aplicados para as predições de cor de olhos (em 2 e 3 categorias), cabelo e tom de pele a partir do grupo de marcadores CORR estão presentes na Tabela 2.

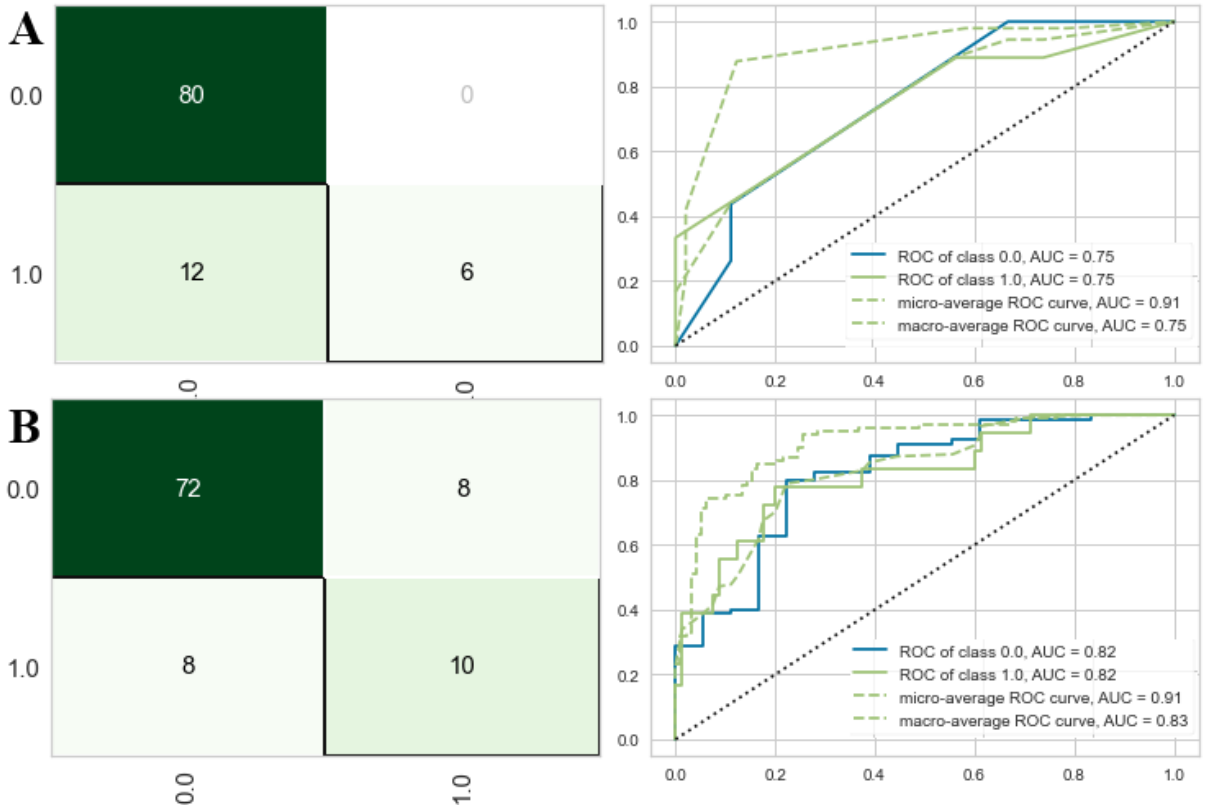
Tabela 2. Maior e menor acurácia dos modelos CORR para os quatro fenótipos considerando os algoritmos de Árvore de Decisão (DT), *Random Forest* (RF), *Support Vector Machines* (SVM), Regressão Logística (LR) e *Naive Bayes* (NB).

	Maior Acurácia	**Menor Acurácia**
Olhos (2 categorias)	DT (86,47%)	LR (80,32%)
Olhos (3 categorias)	RF (75,41%)	NB (72,13%)
Cabelo	RF (80,12%)	LR (76,84%)
Pele	NB (73,97%)	DT (69,87%)

Fonte: Autor (2022).

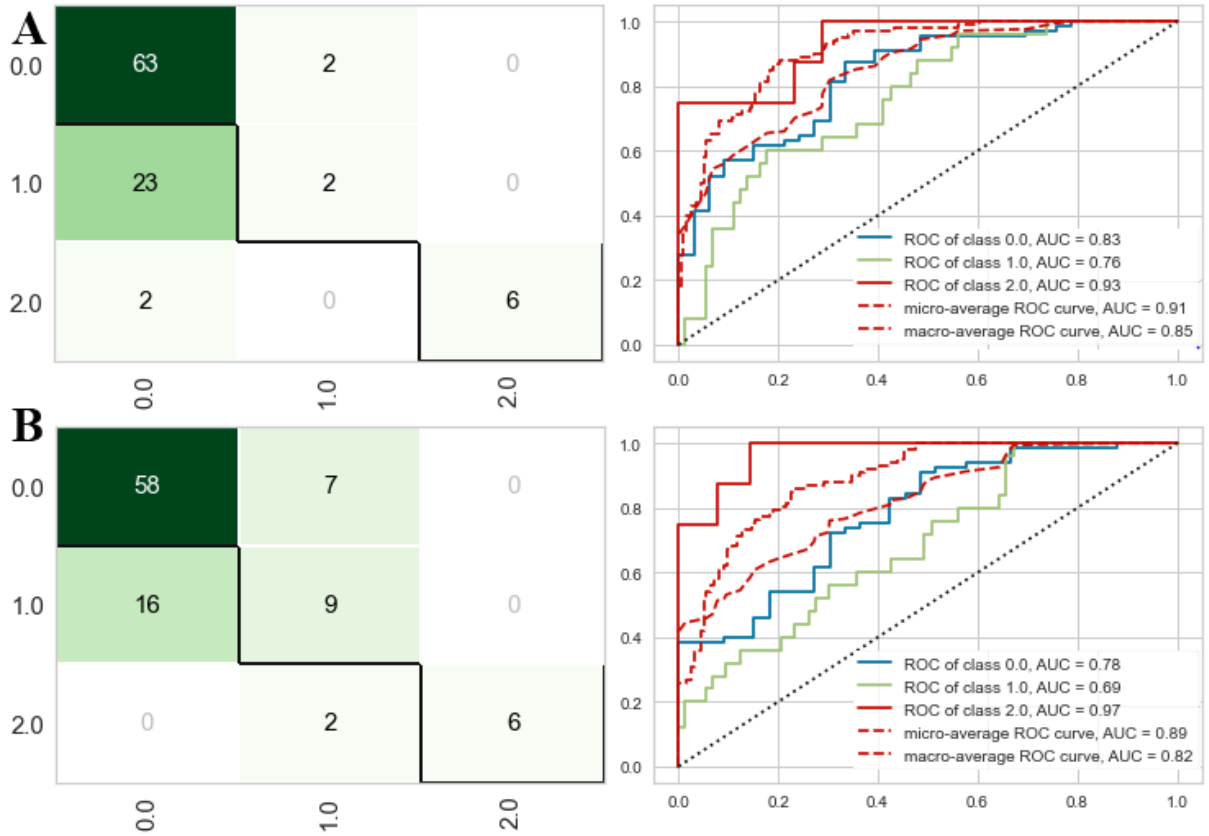
A performance de tais modelos foi também analisada de acordo com a matriz de confusão e a curva ROC do classificador. As Figuras 16 a 19 trazem esses parâmetros de avaliação para os modelos de maior (A) e menor acurácia (B).

Figura 16. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor olhos em escuro (0) e claro (1) para os marcadores CORR.



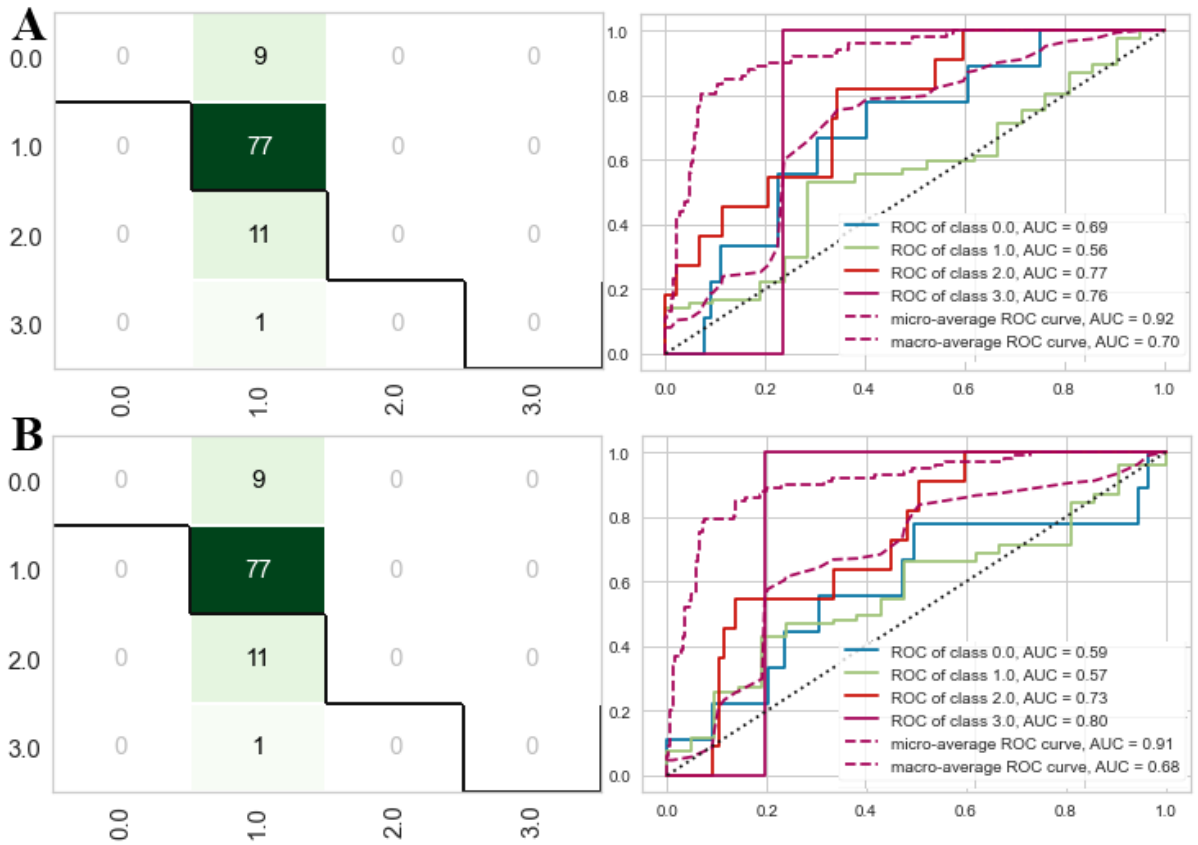
Fonte: Autor (2022).

Figura 17. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor olhos em castanho (0), intermediário (1) e azul (2) para os marcadores CORR.



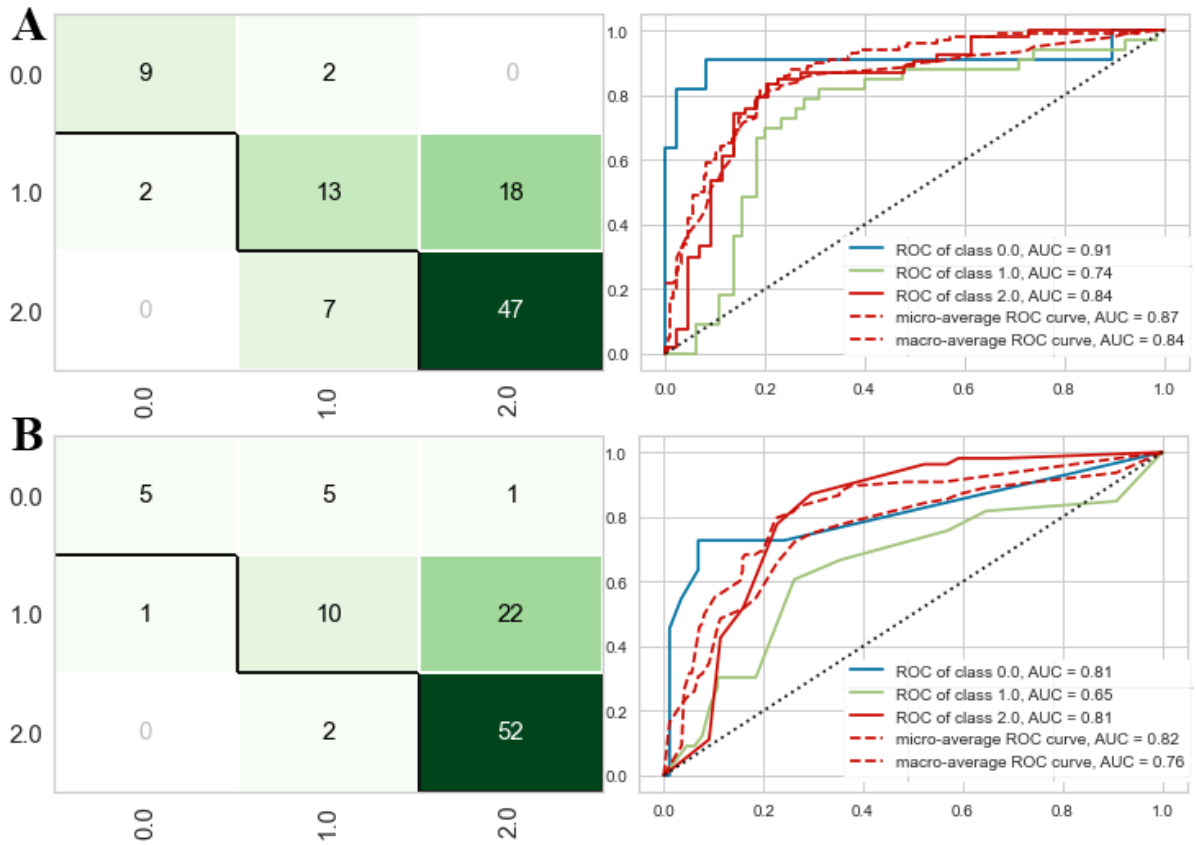
Fonte: Autor (2022).

Figura 18. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor cabelos em preto (0), castanho (1), loiro (2) e ruivo (3) para os marcadores CORR.



Fonte: Autor (2022).

Figura 19. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de tom de pele em negro (0), intermediário (1) e branco (2) para os marcadores CORR.



Fonte: Autor (2022).

5.2.1 Modelos MI

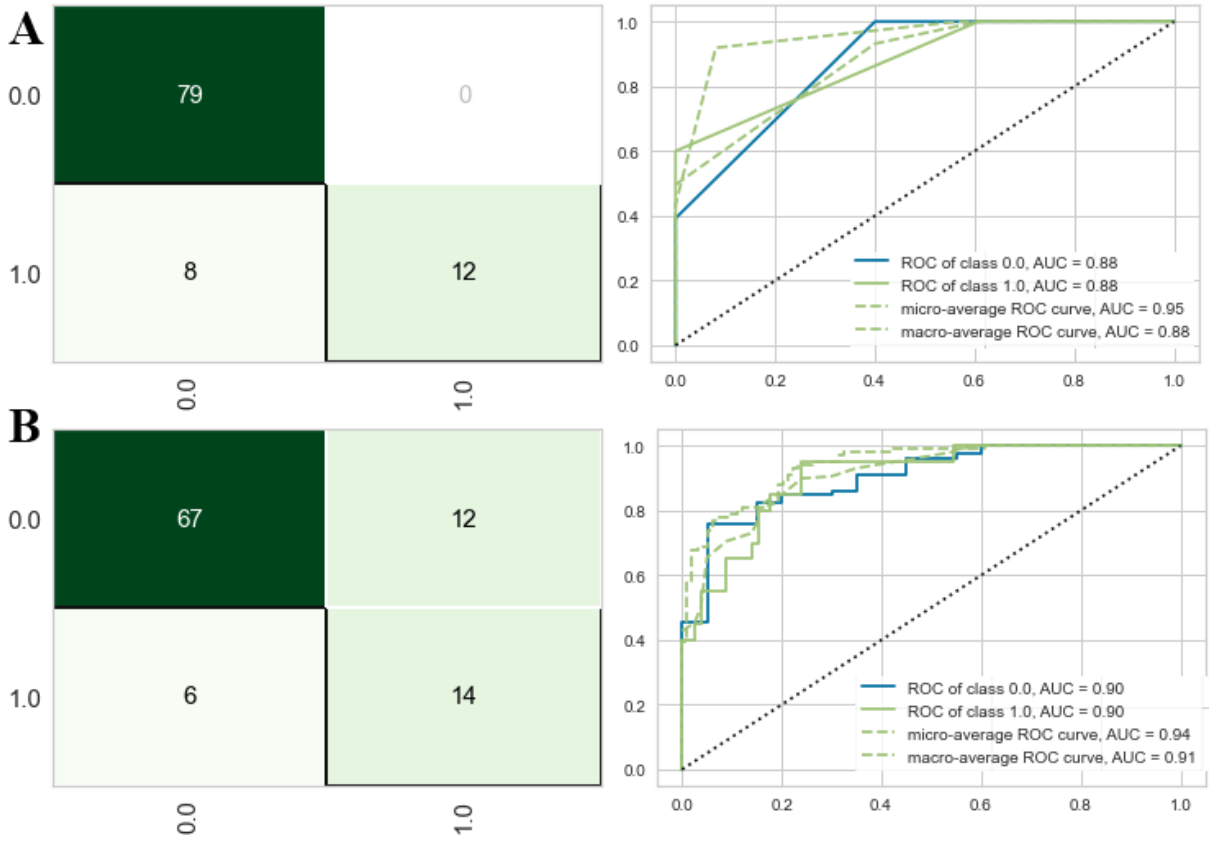
Os algoritmos de maior e menor acurácia aplicados para as predições de cor de olhos (em 2 e 3 categorias), cabelo e tom de pele a partir do grupo de marcadores MI estão presentes na Tabela 3. Os resultados das matrizes de confusão e curvas ROC estão nas Figuras 20 a 23.

Tabela 3. Maior e menor acurácia dos modelos MI para os quatro fenótipos considerando os algoritmos de Árvore de Decisão (DT), *Random Forest* (RF), *Support Vector Machines* (SVM), Regressão Logística (LR) e *Naive Bayes* (NB).

	Maior Acurácia	**Menor Acurácia**
Olhos (2 categorias)	SVM/RF/DT (86,43%)	LR (82,18%)
Olhos (3 categorias)	RF (76,11%)	LR (71,26%)
Cabelo	SVM/RF/NB (79,75%)	LR (77,32%)
Pele	RF (74,49%)	DT (70,23%)

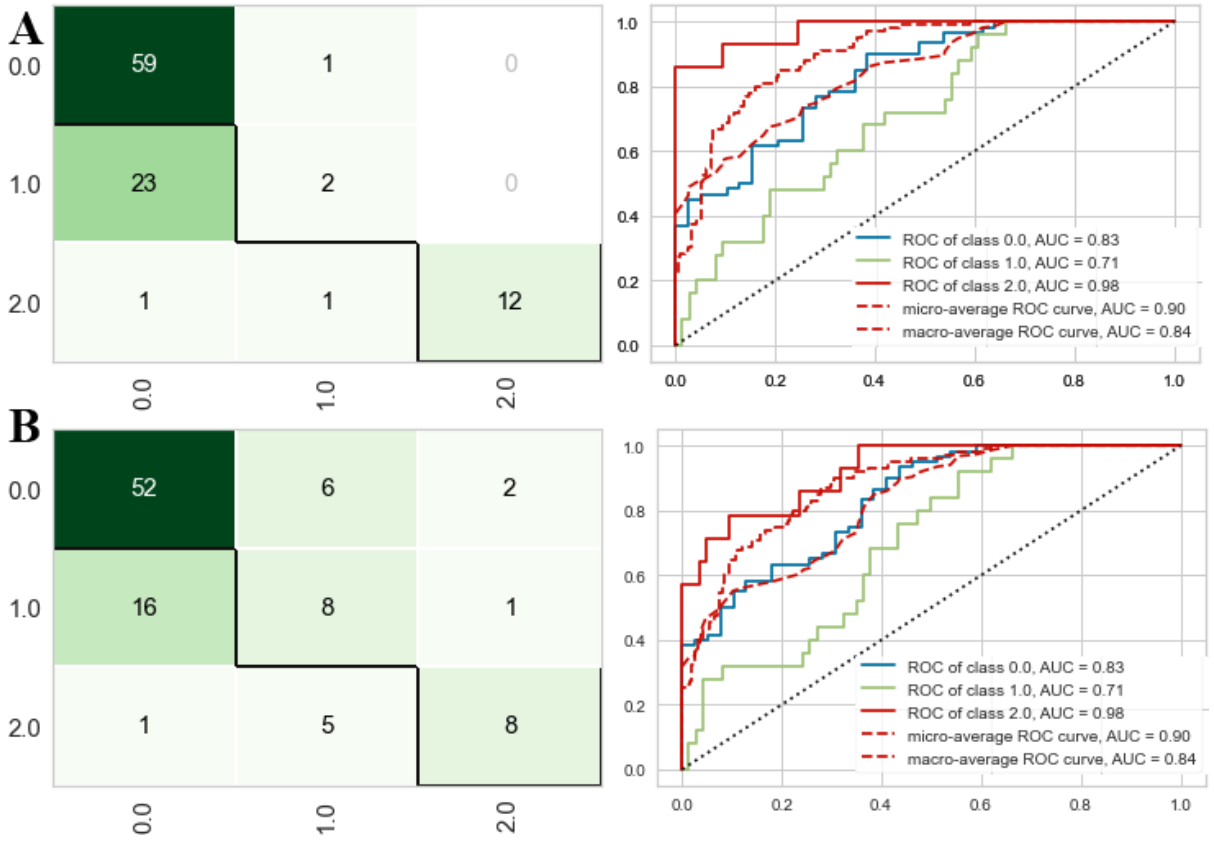
Fonte: Autor (2022).

Figura 20. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor olhos em escuro (0) e claro (1) para os marcadores MI.



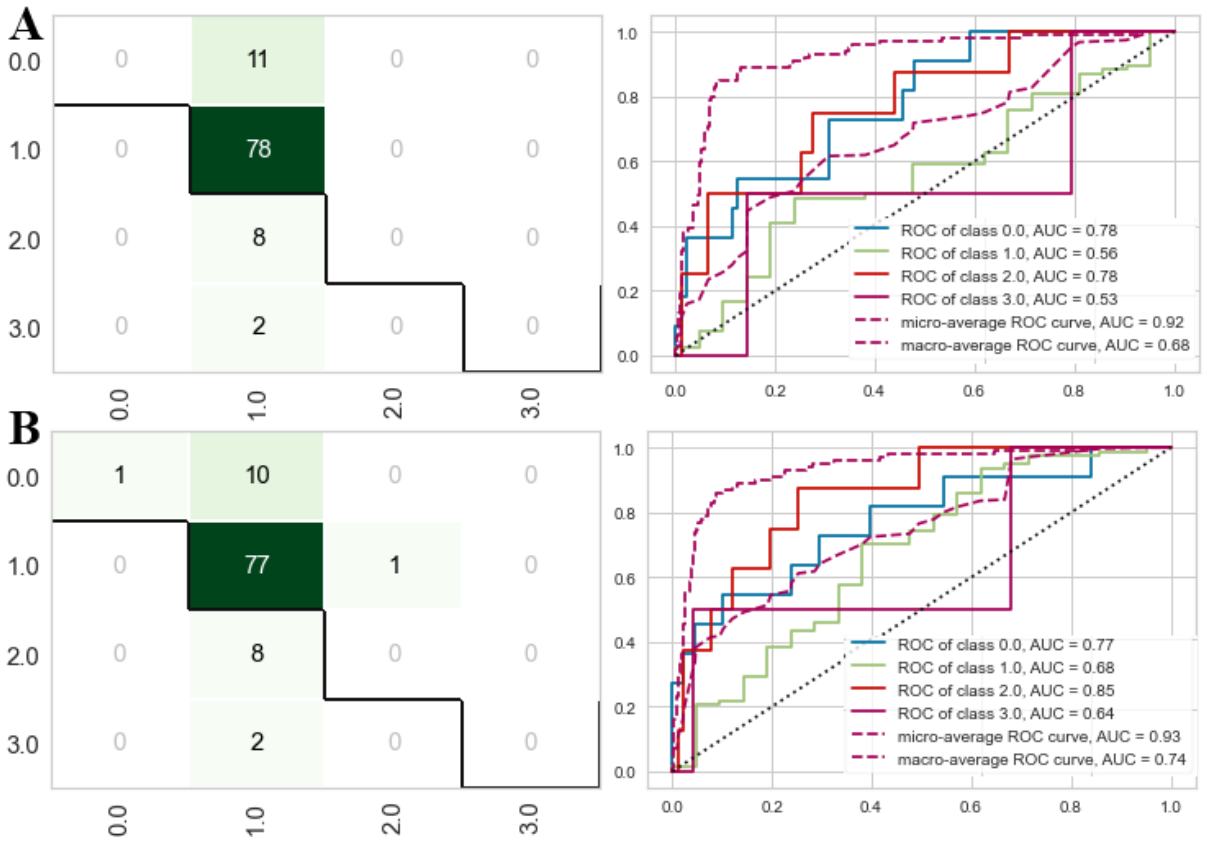
Fonte: Autor (2022).

Figura 21. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor olhos em castanho (0), intermediário (1) e azul (2) para os marcadores MI.



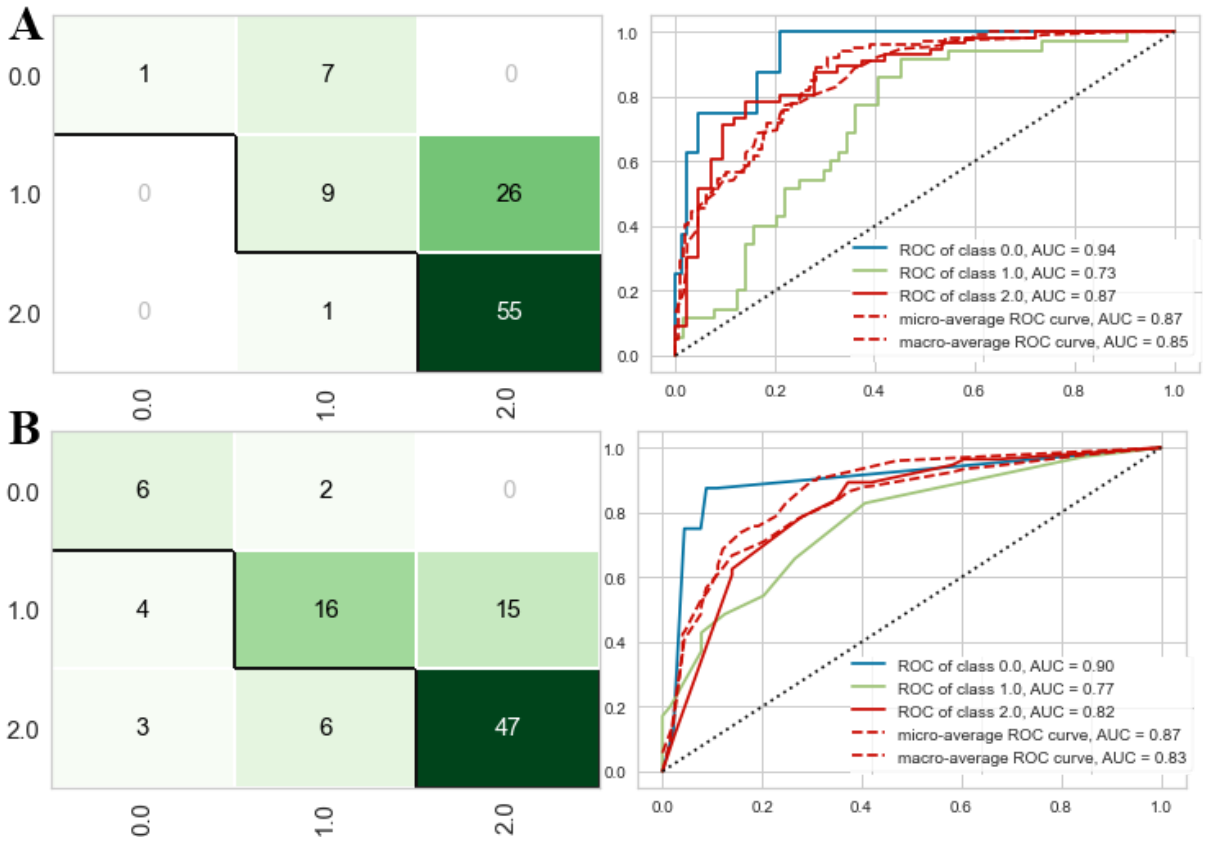
Fonte: Autor (2022).

Figura 22. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor cabelos em preto (0), castanho (1), loiro (2) e ruivo (3) para os marcadores MI.



Fonte: Autor (2022).

Figura 23. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de tom de pele em negro (0), intermediário (1) e branco (2) para os marcadores MI.



Fonte: Autor (2022).

5.2.1 Modelos MULTICOL

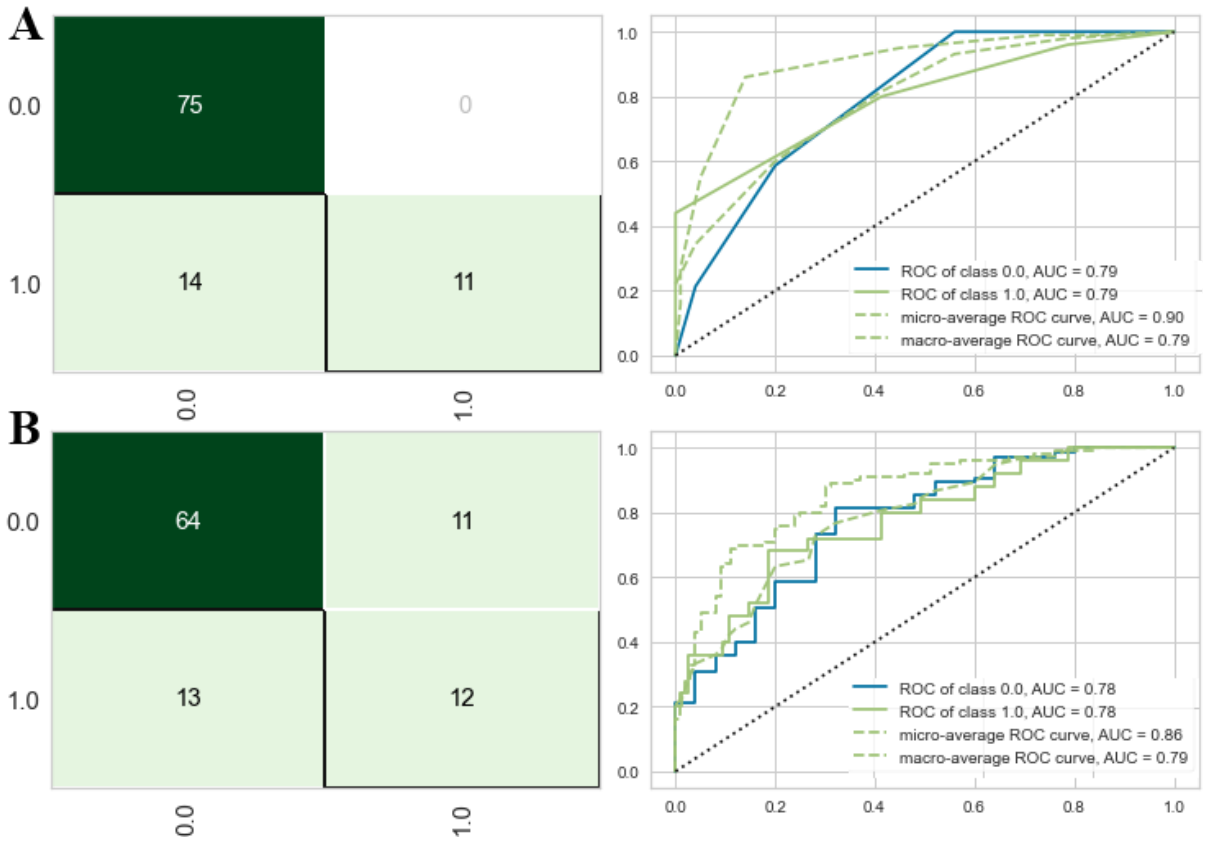
Os algoritmos de maior e menor acurácia aplicados para as predições de cor de olhos (em 2 e 3 categorias), cabelo e tom de pele a partir do grupo de marcadores MULTICOL estão presentes na Tabela 4. Os resultados das matrizes de confusão e curvas ROC estão nas Figuras 24 a 27.

Tabela 4. Maior e menor acurácia dos modelos MULTICOL para os quatro fenótipos considerando os algoritmos de Árvore de Decisão (DT), *Random Forest* (RF), *Support Vector Machines* (SVM), Regressão Logística (LR) e *Naive Bayes* (NB).

	Maior Acurácia	**Menor Acurácia**
Olhos (2 categorias)	DT (86,49%)	LR (80,03%)
Olhos (3 categorias)	RF (75,40%)	NB (70,56%)
Cabelo	SVM/RF/NB/DT (79,83%)	LR (76,20%)
Pele	RF (73,99%)	DT (69,76%)

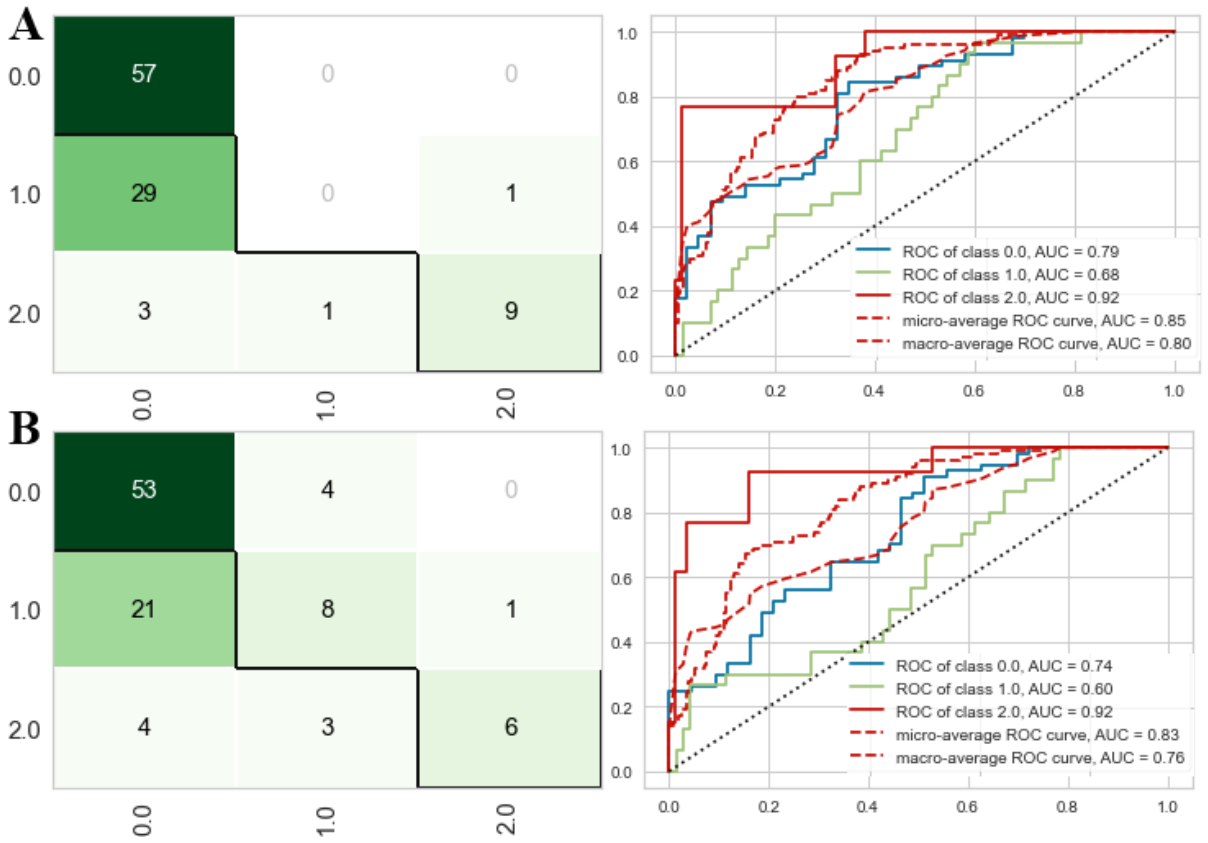
Fonte: Autor (2022).

Figura 24. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor olhos em escuro (0) e claro (1) para os marcadores MULTICOL.



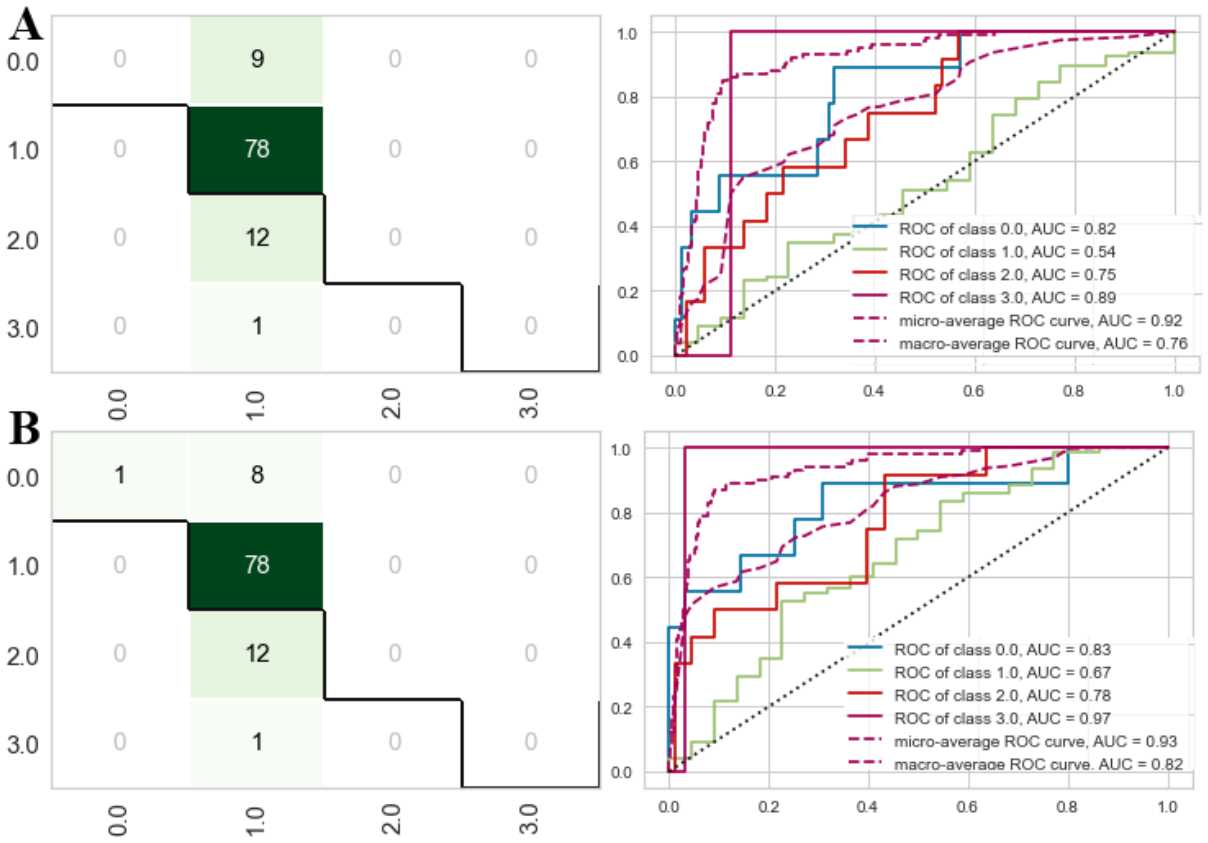
Fonte: Autor (2022).

Figura 25. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor olhos em castanho (0), intermediário (1) e azul (2) para os marcadores MULTICOL.



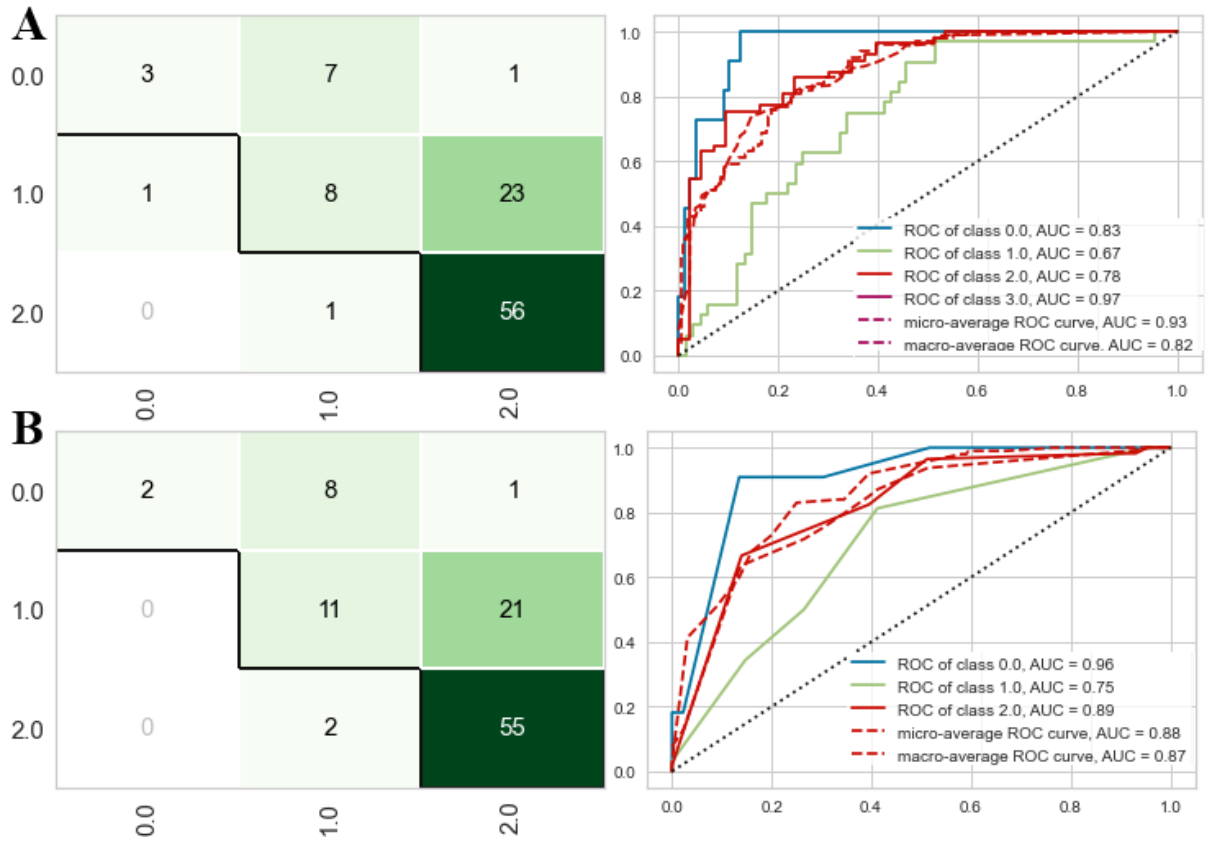
Fonte: Autor (2022).

Figura 26. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de cor cabelos em preto (0), castanho (1), loiro (2) e ruivo (3) para os marcadores MULTICOL.



Fonte: Autor (2022).

Figura 27. Matriz de confusão e curvas ROC dos modelos de maior (A) e menor (B) acurácia para classificação de tom de pele em negro (0), intermediário (1) e branco (2) para os marcadores MULTICOL.



Fonte: Autor (2022).

A segunda etapa deste trabalho focou na aplicação de modelos de predição dos fenótipos de interesse a partir dos subconjuntos de marcadores, utilizando cinco algoritmos: Árvore de Decisão (DT), *Random Forest* (RF), *Support Vector Machines* (SVM), Regressão Logística (LR) e *Naive Bayes* (NB). O desempenho dos modelos foi medido a partir de sua acurácia e sensibilidade. Em geral, houve pouca variação em ambos os parâmetros de avaliação de desempenho dos modelos, independente de qual grupo de variantes estava sendo utilizado e do algoritmo aplicado.filtros de triagem. Pode-se afirmar também que os algoritmos estavam bem adaptados aos dados da amostra.

6. DISCUSSÃO

6.1 MARCADORES

Ao todo, 49 marcadores foram analisados, sendo que seis desses apresentaram uma taxa de variação inferior a 1% dentro do recorte populacional. Quatro dos SNPs tidos como monomórficos estão integrados ao sistema HIrisPlex-S (MC1R rs312262906, MC1R rs11547464, MC1R rs1805006 e MC1R rs201326893), enquanto que as outras duas variantes (OCA2 rs1800416 e SLC24A5 rs16960620) foram adicionadas à análise com base nos resultados de trabalhos previamente publicados na área.

A variante rs1800416 é citada na literatura como componente de um haplótipo de OCA2 que abarca também rs1800407 e rs1800404 (ANDRADE et al., 2017; TOZZO et al., 2021). O alelo C de rs1800416 foi vinculado à manifestação de traços físicos mais escuros (ANDRADE et al., 2017; TOZZO et al., 2021). Ao avaliarem diferentes marcadores dos genes OCA2 e HERC2 para fins de predição fenotípica na população brasileira, Andrade e colaboradores (2017), encontraram uma frequência baixa do alelo C na amostra analisada. Os autores explicaram tal fato como decorrente da influência de populações europeias e ameríndias na composição genética de brasileiros, tendo em vista especialmente que esse alelo não é encontrado em europeus. O monomorfismo desse marcador dentro do recorte amostral analisado no presente trabalho pode também ser esclarecido por tal afirmativa.

O SNP rs16960620, além de estar incluso em um haplótipo com outras variantes do gene SLC24A5, é reconhecido como um dos indicadores biogeográficos utilizados para estimar a ancestralidade de um indivíduo (GIARDINA et al., 2008; DURSO et al., 2014). Supõe-se que haja uma relação entre origem biogeográfica e EVCs associados à pigmentação de estruturas (GIARDINA et al., 2008; DURSO et al., 2014), porém, devido a baixa variação desse SNP na amostra, incluí-lo nos modelos preditivos não acarretaria em novas informações quanto a este ponto.

Os quatro marcadores monomórficos do gene MC1R observados neste trabalho estão fortemente associadas ao fenótipo de cabelos ruivos (BOX et al., 2001; ZORINA-LICHTENWALTER et al., 2019). De fato, dentro do HIrisPlex-S, o efeito cumulativo de tais SNPs, somados a outros do mesmo gene, são os principais indicadores de indivíduos com tal característica (WALSH et al., 2013). Essas variantes são mais prevalentes no Norte da Europa (ZORINA-LICHTENWALTER et al., 2019), região com a maior proporção mundial de ruivos (CUNNINGHAM et al., 2010) e que historicamente teve menor influência nos processos de colonização e, consequentemente, na composição genética de

coortes da América do Sul (ONGARO et al., 2019). A amostra analisada apresenta somente 1,63% indivíduos ruivos (n = 10), de modo que a ausência de variação desses marcadores faz-se justificável.

Após a verificação dos SNPs polimórficos, uma nova triagem foi realizada a partir dos testes de correlação linear de Pearson e de Informação Mútua, gerando assim dois subconjuntos de marcadores. Tanto no subgrupo CORR quanto no MI, 36 marcadores foram selecionados a partir desta etapa. Houve a sobreposição de duas variantes removidas em ambos os casos: DEF8 rs8051733 e MC1R rs1805008. Contudo, outras quatro variantes do gene MC1R foram excluídas ao todo; uma em CORR (rs1805005) e três em MI (rs885479, rs1805009 e 1805007). Novamente, questões relacionadas ao fenótipo de cabelos ruivos podem estar por trás da performance deficiente desses marcadores nos filtros estatísticos aplicados.

Outros quatro marcadores foram excluídos do grupo CORR: RALY rs6059655, OCA2 rs1545397, OCA2 rs1800407, KITLG rs12821256. As duas primeiras variantes neste caso foram adicionadas na última versão do sistema HIrisPlex-S (CHAITANYA et al., 2018). Já o SNP rs12821256 apareceu inicialmente no modelo elaborado para a predição de cor de cabelo, possuindo discreta associação com tons loiros (WALSH et al., 2013). Em contraste com os achados deste trabalho, o marcador rs1800407, constituinte do haplótipo do gene OCA2 anteriormente citado, foi apontado por Walsh e colaboradores (2011) como um dos seis SNPs utilizados para inferências sobre cor de olhos. Tal variante é descrita na literatura como regulador da penetrância de HERC2 rs12913832, que será discutido adiante (STURM et al., 2008; WALSH et al., 2011; ANDERSEN et al., 2016). Entretanto, estudos abrangendo diferentes grupos populacionais mundiais chegaram em resultados similares aos aqui obtidos (ALGHAMDI et al., 2019; SHAPTURENKO et al., 2019; LONA-DURAZO et al., 2022), de modo que mais estudos sobre a eficiência desse SNP na fenotipagem forense se fazem necessários.

Excetuando as variantes de MC1R, apenas mais duas foram removidas do subgrupo MI: TYRP1 rs683, SLC24A4 rs17128291. A primeira delas foi incluída inicialmente no modelo HIrisPlex (WALSH et al., 2013) por ser um marcador independentemente informativo em relação à predição de cor de cabelos (BRANICKI et al., 2011). Os próprios autores admitiram uma baixa significância dentro do contexto de modelo preditivo (BRANICKI et al., 2011), o que pode ser corroborado pelos achados do presente trabalho. Algo semelhante acontece com o SNP rs17128291, primeiramente elencado no sistema HIrisPlex-S

(CHAITANYA et al., 2018) após ter sido associado com processos de pigmentação por Liu e colaboradores (2015), e que não passou no filtro de seleção de Informação Mútua.

Finalmente, no subconjunto de marcadores MULTICOL, uma variante de cada par com alta correlação entre si foi removida: SLC45A2 rs28777 (par com rs16891982), TYR rs1126809 (par com rs1393350) e HERC2 rs1129038 (par com rs12913832). Em todos os casos, foram retirados os marcadores adicionados em atualizações do sistema HIrisPlex-S, os respectivos pares de cada um estavam presentes desde o primeiro modelo (WALSH et al., 2011). Duas hipóteses foram elaboradas para explicar os valores elevados de correlação linear de Pearson entre esses SNPs: ligação gênica e efeito pleiotrópico. Devido ao fato de ambas as variantes dos pares estarem localizadas no mesmo gene, a hipótese de pleiotropia foi descartada. Essa mesma característica também reforça a ideia de ligação gênica. Não foram encontradas evidências na literatura sobre a ligação entre os marcadores de SLC45A2. Já no caso de TYR, Grønskov e colaboradores (2019) caracterizam um haplótipo que envolve rs1393350 e outra variante (rs1042602) diferente dos resultados obtidos neste estudo. Por fim, o haplótipo entre os SNPs do gene HERC2 é bem documentado (EIBERG et al., 2008; RUIZ et al., 2013; ANDERSEN et al., 2016; ANDRADE et al., 2017; TOZZO et al., 2021) e reforçado também pelos achados aqui descritos.

Ambas as variantes que compõem o haplótipo do gene HERC2 foram previamente vinculadas aos polos do espectro de variação de cor de olhos (EIBERG et al., 2008; STURM et al., 2008; RUIZ et al., 2013; ANDERSEN et al., 2016; ANDRADE et al., 2017; TOZZO et al., 2021). Em especial o SNP rs12913832, uma constante nos sistemas de predições fenotípicas desde a implementação do IrisPlex (WALSH et al., 2011) e cujo papel regulatório do sítio promotor do gene vizinho OCA2 lhe confere o título de marcador fundamental para a determinação de olhos castanhos e azuis em caucasianos; o alelo ancestral A associa-se ao primeiro, já o alelo polimórfico G, ao segundo (STURM et al., 2008; WALSH et al., 2011; AMOS et al., 2011; ANDRADE et al., 2017). Estudos com as populações brasileira e venezuelana confirmaram a influência do haplótipo como um todo na diferenciação de indivíduos latinoamericanos com olhos castanhos e não-castanhos (FREIRE-ARADAS et al., 2014; ANDRADE et al., 2017). De fato, rs1129038 e rs12913832 obtiveram os maiores valores nos filtros do teste de correlação de Pearson e de Informação Mútua aqui aplicados em relação ao fenótipo de pigmentação da íris.

Ainda no âmbito da predição de cor de olhos, além dos marcadores citados anteriormente, nossos dados apontam a influência considerável de todas as variantes de HERC2 e pelo menos mais três de OCA2. Fato que, mais uma vez, corrobora os achados de

estudos prévios sobre a importância desses dois genes em tal fenótipo (EIBERG et al., 2008; ANDERSEN et al., 2016; ANDRADE et al., 2017; TOZZO et al., 2021). Contudo, pode-se observar também que variantes de outros genes se destacaram nos filtros estatísticos aplicados. O SNP rs16891982 se faz presente desde o IrisPlex e foi associado a olhos azuis (WALSH et al., 2011). Já rs28777 e rs1426654, apesar de terem sido adicionados nos sistemas posteriores do HIrisPlex (WALSH et al., 2013) e HIrisPlex-S (CHAITANYA et al., 2018), respectivamente, obtiveram valores de expressão na correlação linear com a característica em questão. Por fim, salientamos que SLC24A5 rs2555364, HERC2 rs1133496 e HERC2 rs11636232, adicionados nesta análise devido a estudos (DURSO et al., 2014; ANDRADE et al., 2017) que fogem do eixo dos principais modelos preditivos até o momento, despontam como potenciais novos marcadores de interesse para a predição EVCs relacionadas à pigmentação na população brasileira.

Em relação à cor de cabelo dos indivíduos amostrados, observamos que os SNPs que estavam mais associados ao fenótipo compõem uma tríade já citada anteriormente; dois deles encontram-se no gene SLC45A2 (rs16891982 e rs28777) e o remanescente está no gene SLC24A5 (rs2555364). A variante ASIP rs6058017, que não está incluída em nenhum dos sistemas que compõem o HIrisPlex-S, foi identificada como de interesse para a característica em questão. Estudos prévios culminaram em conclusões conflitantes quanto a este marcador. Em um recorte populacional dos Estados Unidos, o alelo ancestral (G) foi encontrado em indivíduos com olhos e cabelos castanhos. Já o alelo polimórfico (A) estava presente tanto em caucasianos com traços mais claros quanto em indivíduos com traços intermediários, como asiáticos, e em nativos aborígenes que apresentavam EVCs mais escuras. Sob o ponto de vista biogeográfico, europeus em geral apresentam o alelo A, enquanto que populações africanas e do leste asiático têm maior tendência de possuir o alelo G (KANETSKY et al., 2002; ZEIGLER-JOHNSON et al., 2004; VOISEY et al., 2006; LIMA; GONÇALVES; FRIDMAN, 2015). ASIP rs6058017, apesar de ter sido previamente analisado na população brasileira, não apresentou associação direta com o fenótipo de cor de cabelos, e foi reportado como componente genético de tons de pele mais claros. Os autores ressaltam ainda a necessidade de estudos adicionais nos processos de pigmentação em populações com alto grau de miscigenação (LIMA; GONÇALVES; FRIDMAN, 2015). De qualquer forma, vale registrar que, por meio de nossas análises, ASIP rs6058017 pode, em potencial, ser também um marcador informativo para brasileiros. Curiosamente, no resultado do teste de correlação linear de Pearson, rs1805007 e rs1110400 destacaram-se entre as outras variantes do gene

MC1R, as quais, até o momento, estão todas descritas como determinantes para cabelos ruivos (BOX et al., 2001; ZORINA-LICHTENWALTER et al., 2019).

As predições de tom de pele parecem estar mais "diluídas" entre os marcadores na amostra. Os filtros estatísticos aplicados apontam a influência de um número maior de SNPs nessa característica, diferentemente do que foi observado para cor de olhos, por exemplo, onde duas variantes pareciam estar muito mais relacionadas ao traço do que as outras. Marcadores dos genes SLC24A5 (rs1426654 e rs2555364) e SLC45A2 (rs16891982 e rs28777) novamente figuraram entre os destaques para a predição de cor de pele, de modo que pode-se extrapolar a importância dos mesmos para todos as EVCs relacionadas à pigmentação dentro do recorte populacional analisado. A influência de outras variantes como rs6119471, rs1800404 e rs1126809, além das variantes de HERC2, também foram observadas no fenótipo de tom de pele. ASIP rs6119471 é um dos marcadores adicionados ao HIrisPlex-S (CHAITANYA et al., 2018) e cujo alelo ancestral G serve como marcador biogeográfico devido a sua prevalência em populações africanas (SPICHENOK et al., 2011; CHAITANYA et al., 2018). Tal alelo tem importância também na definição de EVCs escuras: olhos castanhos e tons de pele que divergem do claro (PNEUMAN et al., 2012; HART et al., 2013; MUSHAILOV et al., 2015). OCA2 rs1800404, um dos SNPs que não se faz presente nos sistemas de predição citados até o momento, foi reportado como componente genético na coloração da pele de populações africanas, latino-americanas, leste-asiáticas, europeias e de afro-descendentes nos Estados Unidos (NORTON et al., 2006; CRAWFORD et al., 2017; ADHIKARI et al., 2019; BATAI et al., 2021; FENG; MCQUILLAN; TISHKOFF, 2021). Tons mais claros estão associados ao alelo variante T, enquanto que o alelo C é observado em indivíduos com pele mais escura e, em especial, aqueles originários de regiões da África e Melanésia (NORTON et al., 2006; BINO; DUVAL; BERNERD, 2018; FENG; MCQUILLAN; TISHKOFF, 2021). Souza e colaboradores (2021) chegaram em conclusões similares ao avaliar o efeito desta variante em uma amostra populacional de Pernambuco no Brasil. Já TYR rs1126809, em sua forma polimórfica A, resulta no funcionamento insuficiente da enzima catalisadora inicial dos processos de melanogênese, inibindo assim a síntese de pigmento e, consequentemente, resultando em fenótipos mais claros (BERSON et al., 2000; SULEM et al., 2007; NAN et al., 2009; JAGIRDAR et al., 2014; CHAITANYA et al., 2018; MEYER et al., 2020; REIS et al., 2020). Vale ressaltar que esse SNP foi identificado como fator de risco para o desenvolvimento de carcinomas basais e de células escamosas (KHORUDDIN et al., 2021). Andersen e colaboradores (2020) realizaram um estudo de associação de marcadores de pigmentação ao tom de pele em indivíduos brasileiros,

gerando um ranking com os nove polimorfismos de maior interesse. Entre tais encontram-se algumas das variantes analisadas no presente trabalho: rs16891982, rs6119471, rs12913832 e rs1426654.

6.2 MODELOS

Cada um dos subgrupos resultantes da triagem de marcadores (CORR: correlação linear de Pearson, MI: teste de informação mútua e MULTICOL: testes de multicolinearidade) serviu como base para os processos de aprendizagem e eventuais predições dos algoritmos aplicados. Em linhas gerais, observou-se pouca diferença na acurácia dos modelos aplicados, independentemente de quais fenótipos e marcadores estavam sendo analisados. Tal constatação sugere que a eficiência das predições advém, em maior parte, de variantes comuns aos três casos; destacando-se então: rs12913832 (olhos em duas ou três classes) e rs16891982, rs28777, rs2555364 e rs1426654 (cabelos e pele), todas discutidas anteriormente.

Ainda assim, alguns padrões foram encontrados na comparação entre a acurácia dos modelos aplicados. Algoritmos de LR obtiveram a menor acurácia em sete dos 12 testes realizados. Logo depois vêm as árvores de decisão, as quais tiveram a performance menos eficiente neste ponto nos três casos de predição de pele. Os modelos NB figuram como os preditores menos acurados em dois dos casos de classificação de olhos em três categorias. Em contrapartida, os algoritmos de RF aparecem como os melhores preditores em cerca de 75% dos casos (nove vezes em 12 testes).

Olson (2017) e colaboradores, ao avaliarem metodologias de MI dentro do contexto da bioinformática, relatam que os modelos de LR e DT tendem a ter uma performance inferior em relação a algoritmos mais complexos, como, por exemplo, as Random Forest. Tal relato está em concordância com os dados gerados no trabalho, exceto quando se trata das DT nos casos de predição de olhos em duas categorias, que funciona essencialmente como uma classificação binária. A própria estrutura do algoritmo de árvore de decisão (ALZBERG, 1994, POLAKA; TOM; BORISOV, 2010; HAN; KAMBER; PEI, 2012; SARKER, 2021) pode ser responsável pelos resultados mais satisfatórios neste caso. Por fim, vale ressaltar que expressões de regressão logística são os preditores nos modelos do sistema HIrisPlex-S (WALSH et al., 2011; WALSH et al., 2013; CHAITANYA et al., 2018) e que, no recorte populacional analisado, não aparentam ser a melhor estratégia a ser utilizada em termos de acurácia.

A despeito das nuances relativas aos algoritmos, os modelos aplicados obtiveram valores consideráveis de acurácia, variando entre 86,49% e 69.76%. Sob enfoque dos quatro

fenótipos, uma quantidade maior de acertos ocorreu na classificação binária de olhos, enquanto que os resultados menos favoráveis estavam associados às predições de tom de pele. Esse fenômeno é, de fato, uma tendência quando se trata da fenotipagem forense de EVCs relacionadas aos traços de pigmentação. Palmal e colaboradores (2021) exploraram diferentes abordagens para predições dentro de uma amostra latino-americana. A primeira delas envolveu estudos de associação e um modelo próprio, culminando em 89% de acurácia para olhos, 85 % para cabelos e 75% para pele. Já a segunda utilizou a mesma técnica estatística porém com os marcadores do HIrisPlex-S, obtendo 89% de acurácia para olhos, 84% para cabelos e 82% para pele. Finalmente, os autores aplicaram os dados na plataforma online do sistema HIrisPlex-S, chegando nos valores conflitantes de 87% para olhos, 56% para cabelos e apenas 26% para tom de pele. Carratto e colaboradores (2019) foram mais fundo ao analisar indivíduos brasileiros, seus achados sendo ainda mais alarmantes; somente 19,16% das predições foram corretas ao aplicar a ferramenta digital baseada nas variantes e metodologias do HIrisPlex-S. Com tudo isso em mente, pode-se afirmar que as boas taxas de acurácia encontradas no presente trabalho mostram que os algoritmos estão bem ajustados à amostra analisada.

Os modelos foram também avaliados em relação a sua sensibilidade (AUC) e quantidade de predições corretas (matriz de confusão) para cada uma das categorias específicas dos fenótipos. Em geral, o conjunto de marcadores MI resultou em valores mais expressivos de AUC. Contudo, a discrepância em comparação a CORR e MULTICOL é pequena na maioria das EVCs analisadas. Não houve também grande variação na quantidade de predições corretas nos modelos como um todo.

Certas especificidades da predição de olhos, tal como a dificuldade na identificação correta de olhos intermediários observada também no presente trabalho, já foram amplamente reportadas na literatura. Estudos prévios de fenotipagem forense apontam a subjetividade da categorização de tons intermediários frente às classes castanho e azul, além das escassas informações sobre os componentes genético relacionados à pigmentação da íris em tons esverdeados e mel como possíveis explicações para tais falhas (LIU et al., 2009; WALSH et al., 2011; PNEUMAN et al., 2012; RUIZ et al., 2013; DEMBINSKI; PICARD, 2014; KAYSER, 2015; SALVORO et al., 2019). Uma outra questão que se faz importante em nossa amostra é o desequilíbrio do número de indivíduos alocados nos grupos fenotípicos possíveis. Tanto na classificação em duas classes quanto em três, olhos escuros e castanhos estão consideravelmente mais representados do que olhos claros e intermediários/azuis,

respectivamente. Ressaltamos aqui principalmente o traço de olhos azuis, o qual está presente em menos de 100 indivíduos dos 611 que compõem o recorte populacional analisado.

A problemática dos tons intermediários se estende também ao fenótipo de cor de cabelos, especialmente no que diz respeito à diferença entre loiro escuro e castanho claro. As mudanças naturais que acontecem ao longo da vida de certos indivíduos cujos cabelos originalmente claros vão escurecendo com o passar do tempo surgem como principal causa de erros nas predições neste caso. Até que se descubra os mecanismos biológicos por trás de tal processo, a distinção entre esses fenótipos permanece complexa. Outros agravantes pontuais envolvem cabelos grisalhos e a facilidade para se tingir e assim alterar a natural cor de cabelos (WALSH et al., 2013; KAYSER, 2015). Novamente, a inconsistência na representatividade de indivíduos nas categorias de predição desta EVC foram o maior desafio para os modelos aplicados. A amostra contava com cerca de 450 indivíduos de cabelo castanho, um número muito superior à soma de todas as outras categorias, fato que gerou um padrão bastante peculiar para as predições realizadas. Por não haver informações suficientes sobre três dos quatro possíveis fenótipos, os algoritmos tiveram dificuldade em identificar padrões e, consequentemente, estimaram a esmagadora maioria dos indivíduos, e em alguns casos todos, como pertencentes ao grupo de cabelos castanhos, justamente por este ser o mais representado dentro dos dados de treino e de teste. Até mesmo cabelos ruivos, os quais possuem um padrão de expressão essencialmente ligado ao gene MC1R (BOX et al., 2001; WALSH et al., 2013; ZORINA-LICHTENWALTER et al., 2019) e, portanto, diverge do espectro de tons entre loiro e preto, tiveram suas predições prejudicadas devido à quantidade ínfima de indivíduos com tal característica.

Kayser (2015) descreve a predição de tons de pele como a mais intricada entre as EVCs relacionadas à pigmentação de estruturas. Segundo o autor, a grande barreira que estudos nessa área devem ultrapassar é que o surgimento de variações no tom de pele não foi restrita a apenas uma localidade ou população, como é o caso dos olhos azuis na Europa, por exemplo. O mapeamento dos componentes genéticos responsáveis pela determinação desse traço devem considerar a heterogeneidade mundial, o que o torna ainda mais complexo sob o ponto de vista de identificar padrões em uma população específica e o extrapolar para as demais (KAYSER, 2015). O primeiro trabalho de destaque neste tópico foi publicado por Marona e colaboradores (2014) e, a partir de 29 SNPs, resultou em uma boa diferenciação entre indivíduos de pele branca e de pele intermediária/negra. Posteriormente, os idealizadores do HIrisPlex-S basearam seu sistema de predição na escala de Fitzpatrick de tons de pele, de forma a considerar cinco possíveis categorias para classificação em vez de

apenas três. O aumento da especificidade neste caso gerou uma queda na acurácia das classes que compõem o que se considera como branco (muito pálida, pálida e intermediária), enquanto que os tons mais escuros de pele (morena escura e negra) obtiveram uma performance melhor (WALSH et al., 2017; CHAITANYA et al., 2018). A divisão dos fenótipos no presente estudo levou em consideração as diretrizes expressas por Marona e colaboradores (2014), o que pode possivelmente justificar o fato de que os resultados obtidos aqui se aproximam mais desse estudo do que os dados reportados na elaboração do HIrisPlex-S. Entretanto, temos neste caso também uma amostra com desbalanço entre as classes representadas; a presença de indivíduos com o tom de pele negro é a menor entre os três fenótipos considerados.

Como um último adendo, vale salientar outros detalhes que podem eventualmente ter contribuído para a performance dos modelos aplicados, além do que já foi discutido anteriormente. Dentro dos padrões de estudos envolvendo ML, a quantidade de amostras individuais analisadas estava um pouco aquém do que é recomendado para problemas de classificação tão complexos quanto estes. Isso se reflete bem no caso da predição do fenótipo de cabelos, principalmente por causa do desbalanço nas categorias em questão. Dos 611 genótipos iniciais, mais de um sexto foi perdido dependendo do conjunto de marcadores analisados devido à quantidade de informações vazias presentes para determinados SNPs. Por fim, é importante frisar que, apesar das ressalvas citadas, os modelos aplicados tiveram desempenho satisfatório e têm potencial para serem aperfeiçoados ainda mais.

7. CONCLUSÃO

O presente trabalho utilizou os preceitos e modelos de ML visando encontrar soluções para os problemas de classificação envolvendo as três EVCs relacionadas à pigmentação de estruturas (cor de olhos, cabelos e tom de pele) em um recorte da população brasileira. Para tal, marcadores clássicos da fenotipagem forense foram somados a SNPs explorados em estudos locais e mais específicos com o intuito de testar sua eficácia dentro da amostra, e sua consequente associação aos quatro fenótipos estipulados. De um total de 49 marcadores, seis não apresentaram variação na população e foram removidos da análise final. Posteriormente três novas triagens atuaram em paralelo para restringir a quantidade de marcadores de acordo com abordagens estatísticas específicas. Cinco algoritmos (DF, RT, LR, NB E SVM) foram adaptados de acordo com as variantes selecionadas, testados na amostra e seu desempenho foi aferido por meio do cálculo da acurácia, da curva ROC/AUC e das matrizes de confusão geradas.

Via de regra, os marcadores rs1426654 e rs2555364, rs16891982 e rs28777 demonstram influência em todos os fenótipos. Variantes do gene HERC2, especialmente rs1129038 e rs12913832, foram novamente corroborados como de grande importância para a predição de olhos, tanto na categoria clássica de três categorias quanto na separação claro/escuro aqui abordada. Os dois SNPs pertencentes ao gene ASIP também mostraram seu valor na definição dos fenótipos de cor de cabelo e tom de pele. Seis dos oito marcadores incluídos na análise e que não compõem o sistema HIrisPlex-S (ASIP rs6058017, HERC2 rs1133496, HERC2 rs11636232, OCA2 rs7495174, OCA2 rs1800404, SLC24A5 rs2555364) apresentaram, em geral, relação com as características de interesse, de modo que faz-se a sugestão de que sejam elaborados mais estudos com os mesmos a fim de determinar sua eventual inclusão em novos sistemas de predição.

Houve pouca variação no desempenho dos modelos aplicados, independentemente das abordagens estatísticas utilizadas e do grupo de marcadores selecionados. Isso reforça a ideia de que os principais marcadores específicos de cada fenótipo passaram em todos os filtros de triagem. Pode-se afirmar também que os algoritmos estavam bem adaptados aos dados da amostra.

Observou-se que, assim como reportado anteriormente na literatura, a predição da categoria de olhos intermediários foi problemática. Além disso, a amostra em questão não abarcava uniformemente as categorias pertencentes aos fenótipos analisados, especificamente olhos claros/azuis, pele negra e outros tons de cabelo que divergem do castanho. Sugere-se que estudos futuros foquem não somente na expansão quantitativa do recorte amostral como

um todo, mas também na inclusão de indivíduos que se enquadrem nessas classificações que não foram tão contempladas até o momento.

Em suma, as análises desenvolvidas demonstraram que os processos de calibragem de modelos e triagem dos marcadores estão bem ajustados e podem continuar sendo utilizados no aperfeiçoamento dos modelos preditivos. Os algoritmos aqui aplicados obtiveram uma boa performance, apesar de certas particularidades que devem ser trabalhadas em futuros estudos. A elaboração de ferramentas de fenotipagem forense que contemplem a população brasileira é promissora e tem grande potencial para complementar investigações criminais nos mais diversos contextos.

8. REFERÊNCIAS

ADHIKARI, Kaustubh *et al*. A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. **Nature Communications**, [S.L.], v. 10, n. 1, p. 1-16, 21 jan. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41467-018-08147-0.

ALBERTS, Bruce et al. DNA, cromossomos e genoma. In: ALBERTS, Bruce et al. **Biologia Molecular da Célula**. 8. ed. Porto Alegre: Artmed, 2017. Cap. 4. p. 173-236.

ALGHAMDI, Jahad et al. Eye color prediction using single nucleotide polymorphisms in Saudi population. **Saudi Journal Of Biological Sciences**, [S.L.], v. 26, n. 7, p. 1607-1612, nov. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.sjbs.2018.09.011.

ALLWOOD, Julia S.; HARBISON, Sallyann. SNP model development for the prediction of eye colour in New Zealand. **Forensic Science International: Genetics**, [S.L.], v. 7, n. 4, p. 444-452, jul. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2013.03.005.

ALALUF, Simon *et al*. The Impact of Epidermal Melanin on Objective Measurements of Human Skin Colour. **Pigment Cell Research**, [S.L.], v. 15, n. 2, p. 119-126, abr. 2002. Wiley. http://dx.doi.org/10.1034/j.1600-0749.2002.1o072.x.

AMANKWAA, Aaron Opoku; MCCARTNEY, Carole. The effectiveness of the UK national DNA database. **Forensic Science International**: Synergy, [S.L.], v. 1, p. 45-55, 2019. Elsevier BV. http://dx.doi.org/10.1016/j.fsisyn.2019.03.004.

AMOS, Christopher I. *et al*. Genome-wide association study identifies novel loci predisposing to cutaneous melanoma†. **Human Molecular Genetics**, [S.L.], v. 20, n. 24, p. 5012-5023, 17 set. 2011. Oxford University Press (OUP). http://dx.doi.org/10.1093/hmg/ddr415.

ANDERSEN, Jeppe D. et al. Importance of nonsynonymous OCA 2 variants in human eye color prediction. **Molecular Genetics & Genomic Medicine**, [S.L.], v. 4, n. 4, p. 420-430, 11 mar. 2016. Wiley. http://dx.doi.org/10.1002/mgg3.213.

ANDERSEN, Jeppe D. *et al*. Skin pigmentation and genetic variants in an admixed Brazilian population of primarily European ancestry. **International Journal Of Legal Medicine**, [S.L.], v. 134, n. 5, p. 1569-1579, 9 maio 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00414-020-02307-y.

ANDRADE, Edilene S. et al. Associations of OCA2 - HERC2 SNPs and haplotypes with human pigmentation characteristics in the Brazilian population. **Legal Medicine**, [S.L.], v. 24, p. 78-83, jan. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.legalmed.2016.12.003.

BARSH, Gregory s *et al*. What Controls Variation in Human Skin Color? Plos Biology, [S.L.], v. 1, n. 1, p. 27-31, 13 out. 2003. **Public Library of Science** (PLoS). http://dx.doi.org/10.1371/journal.pbio.0000027.

BATAI, Ken *et al*. Genetic loci associated with skin pigmentation in African Americans and their effects on vitamin D deficiency. **Plos Genetics**, [S.L.], v. 17, n. 2, p. 1-18, 18 fev. 2021. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pgen.1009319.

BERSON, Joanne F. *et al*. A Common Temperature-sensitive Allelic Form of Human Tyrosinase Is Retained in the Endoplasmic Reticulum at the Nonpermissive Temperature. **Journal Of Biological Chemistry**, [S.L.], v. 275, n. 16, p. 12281-12289, abr. 2000. Elsevier BV. http://dx.doi.org/10.1074/jbc.275.16.12281.

BINO, Sandra del; DUVAL, Christine; BERNERD, Françoise. Clinical and Biological Characterization of Skin Pigmentation Diversity and Its Consequences on UV Impact. **International Journal Of Molecular Sciences**, [S.L.], v. 19, n. 9, p. 2668, 8 set. 2018. MDPI AG. http://dx.doi.org/10.3390/ijms19092668.

BITTRICH, Sebastian *et al*. Application of an interpretable classification model on Early Folding Residues during protein folding. **Biodata Mining**, [S.L.], v. 12, n. 1, p. 1-17, 5 jan. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s13040-018-0188-2.

BRANICKI, Wojciech et al. Model-based prediction of human hair color using DNA variants. **Human Genetics**, [S.L.], v. 129, n. 4, p. 443-454, 4 jan. 2011. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00439-010-0939-8.

BREIMAN, Leo. Random Forests. **Machine Learning**, [S.L.], v. 45, n. 1, p. 5-32, 2001. Springer Science and Business Media LLC. http://dx.doi.org/10.1023/a:1010933404324.

BOX, Neil F. et al. Melanocortin-1 Receptor Genotype is a Risk Factor for Basal and Squamous Cell Carcinoma. **Journal Of Investigative Dermatology**, [S.L.], v. 116, n. 2, p. 224-229, fev. 2001. Elsevier BV. http://dx.doi.org/10.1046/j.1523-1747.2001.01224.x

BUTLER, John M. The future of forensic DNA analysis. **Philosophical Transactions Of The Royal Society B: Biological Sciences**, [S.L.], v. 370, n. 1674, p. 20140252-20140262, 5 ago. 2015. The Royal Society. http://dx.doi.org/10.1098/rstb.2014.0252.

BUTLER, John M.; WILLIS, Sheila. Interpol review of forensic biology and forensic DNA typing 2016-2019. **Forensic Science International: Synergy**, [S.L.], v. 2, p. 352-367, 2020. Elsevier BV. http://dx.doi.org/10.1016/j.fsisyn.2019.12.002.

CARRATTO, T.M.T. *et al*. Evaluation of the HIrisPlex-S system in a Brazilian population sample. **Forensic Science International: Genetics Supplement Series**, [S.L.], v. 7, n. 1, p. 794-796, dez. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.fsigss.2019.10.180.

CESSIE, S. Le; VAN HOUWELINGEN, J. C.. Ridge Estimators in Logistic Regression. **Applied Statistics**, [S.L.], v. 41, n. 1, p. 191, 1992. JSTOR. http://dx.doi.org/10.2307/2347628.

CHAITANYA, Lakshmi *et al*. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: introduction and forensic developmental validation. **Forensic Science International: Genetics**, [S.L.], v. 35, p. 123-135, jul. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2018.04.004.

CHAUHAN, Nagesh Singh. **Naïve Bayes Algorithm: Everything You Need to Know**. Disponível em: https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html. Acesso em: 23 jun. 2022.

CRAWFORD, Nicholas G. *et al*. Loci associated with skin pigmentation identified in African populations. **Science**, [S.L.], v. 358, n. 6365, p. 1-26, 17 nov. 2017. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/science.aan8433.

CUNNINGHAM, A. L. et al. Red for danger: the effects of red hair in surgical practice. **Bmj**, [S.L.], v. 341, n. 092, p. 6931-6931, 9 dez. 2010. BMJ. http://dx.doi.org/10.1136/bmj.c6931.

DARIO, Paulo *et al*. Assessment of IrisPlex-based multiplex for eye and skin color prediction with application to a Portuguese population. **International Journal Of Legal Medicine**, [S.L.], v. 129, n. 6, p. 1191-1200, 20 ago. 2015. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00414-015-1248-5.

DEMBINSKI, Gina M.; PICARD, Christine J.. Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population. **Forensic Science International: Genetics**, [S.L.], v. 9, p. 111-117, mar. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2013.12.003.

D'ISCHIA, Marco *et al*. Melanins and melanogenesis: from pigment cells to human health and technological applications. **Pigment Cell & Melanoma Research**, [S.L.], v. 28, n. 5, p. 520-544, 16 ago. 2015. Wiley. http://dx.doi.org/10.1111/pcmr.12393.

DURSO, Danielle Fernandes et al. Association of Genetic Variants with Self-Assessed Color Categories in Brazilians. **Plos One**, [S.L.], v. 9, n. 1, p. 83926-83940, 8 jan. 2014. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0083926.

EIBERG, Hans *et al*. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. **Human Genetics**, [S.L.], v. 123, n. 2, p. 177-187, 3 jan. 2008. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00439-007-0460-x.

FENG, Yuanqing; MCQUILLAN, Michael; TISHKOFF, Sarah. Evolutionary genetics of skin pigmentation in African populations. **Human Molecular Genetics**, [S.L.], v. 30, n. 1, p. 88-97, 12 jan. 2021. Oxford University Press (OUP). http://dx.doi.org/10.1093/hmg/ddab007.

FREIRE-ARADAS, A. *et al*. Exploring iris colour prediction and ancestry inference in admixed populations of South America. **Forensic Science International**: Genetics, [S.L.], v. 13, p. 3-9, nov. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2014.06.007.

GHANEM, Ghanem; FABRICE, Journé. Tyrosinase related protein 1 (TYRP1/gp75) in human cutaneous melanoma. **Molecular Oncology**, [S.L.], v. 5, n. 2, p. 150-155, 3 fev. 2011. Wiley. http://dx.doi.org/10.1016/j.molonc.2011.01.006.

GIARDINA, Emiliano et al. Haplotypes in SLC24A5 Gene as Ancestry Informative Markers in Different Populations. **Current Genomics**, [S.L.], v. 9, n. 2, p. 110-114, 1 abr. 2008. Bentham Science Publishers Ltd.. http://dx.doi.org/10.2174/138920208784139528.

GILL, Peter; JEFFREYS, Alec J.; WERRETT, David J.. Forensic application of DNA ‘fingerprints’. **Nature**, [S.L.], v. 318, n. 6046, p. 577-579, dez. 1985. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/318577a0.

GRØNSKOV, Karen *et al*. A pathogenic haplotype, common in Europeans, causes autosomal recessive albinism and uncovers missing heritability in OCA1. **Scientific Reports**, [S.L.], v. 9, n. 1, p. 1-7, 24 jan. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-018-37272-5.

GWIRTZ, K *et al*. Can one use Earth’s magnetic axial dipole field intensity to predict reversals? **Geophysical Journal International**, [S.L.], v. 225, n. 1, p. 277-297, 13 nov. 2020. Oxford University Press (OUP). http://dx.doi.org/10.1093/gji/ggaa542.

HART, Katie L. et al. Improved eye- and skin-color prediction based on 8 SNPs. **Croatian Medical Journal**, [S.L.], v. 54, n. 3, p. 248-256, jun. 2013. Croatian Medical Journals. http://dx.doi.org/10.3325/cmj.2013.54.248.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining**: concepts and techniques. 3. ed. Waltham: Morgan Kaufmann Publishers, 2012.

HIDA, Tokimasa *et al*. Elucidation of Melanogenesis Cascade for Identifying Pathophysiology and Therapeutic Approach of Pigmentary Disorders and Melanoma. **International Journal Of Molecular Sciences**, [S.L.], v. 21, n. 17, p. 6129-6153, 25 ago. 2020. MDPI AG. http://dx.doi.org/10.3390/ijms21176129.

HOO, Zhe Hui; CANDLISH, Jane; TEARE, Dawn. What is an ROC curve? **Emergency Medicine Journal**, [S.L.], v. 34, n. 6, p. 357-359, 16 mar. 2017. BMJ. http://dx.doi.org/10.1136/emermed-2017-206735.

HONG, Sae Rom *et al*. DNA methylation-based age prediction from saliva: high age predictability by combination of 7 cpg markers. **Forensic Science International: Genetics**, [S.L.], v. 29, p. 118-125, jul. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2017.04.006.

HORRELL, Erin M. Wolf; BOULANGER, Mary C.; D’ORAZIO, John A.. Melanocortin 1 Receptor: structure, function, and regulation. **Frontiers In Genetics**, [S.L.], v. 7, p. 1-16, 31 maio 2016. Frontiers Media SA. http://dx.doi.org/10.3389/fgene.2016.00095.

HUNT, Gillian *et al*. Eumelanin and Phaeomelanin Contents of Human Epidermis and Cultured Melanocytes. **Pigment Cell Research**, [S.L.], v. 8, n. 4, p. 202-208, ago. 1995. Wiley. http://dx.doi.org/10.1111/j.1600-0749.1995.tb00664.x.

IMESCH, Pascal D.; WALLOW, Ingolf H.L.; ALBERT, Daniel M.. The color of the human eye: a review of morphologic correlates and of some conditions that affect iridial pigmentation. **Survey Of Ophthalmology**, [S.L.], v. 41, p. 117-123, fev. 1997. Elsevier BV. http://dx.doi.org/10.1016/s0039-6257(97)80018-5.

INTERPOL. **How we work**: Forensics, DNA Forensics, DNA. Disponível em: https://www.interpol.int/en/How-we-work/Forensics/DNA. Acesso em: 21 jun. 2022.

ITO, Shosuke. A Chemist's View of Melanogenesis. **Pigment Cell Research**, [S.L.], v. 16, n. 3, p. 230-236, jun. 2003. Wiley. http://dx.doi.org/10.1034/j.1600-0749.2003.00037.x.

JAGIRDAR, Kasturee *et al*. Molecular analysis of common polymorphisms within the humanTyrosinaselocus and genetic association with pigmentation traits. **Pigment Cell & Melanoma Research**, [S.L.], v. 27, n. 4, p. 552-564, 12 maio 2014. Wiley. http://dx.doi.org/10.1111/pcmr.12253.

JORDAN, M. I.; MITCHELL, T. M.. Machine learning: trends, perspectives, and prospects. **Science**, [S.L.], v. 349, n. 6245, p. 255-260, 16 jul. 2015. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/science.aaa8415.

KAELBLING, L. P.; LITTMAN, M. L.; MOORE, A. W.. Reinforcement Learning: a survey. **Journal Of Artificial Intelligence Research**, [S.L.], v. 4, p. 237-285, 1 maio 1996. AI Access Foundation. http://dx.doi.org/10.1613/jair.301

KALIYADAN, Feroze; KULKARNI, Vinay. Types of variables, descriptive statistics, and sample size. **Indian Dermatology Online Journal**, [S.L.], v. 10, n. 1, p. 82, 2019. Medknow. http://dx.doi.org/10.4103/idoj.idoj_468_18.

KANETSKY, Peter A. *et al*. A Polymorphism in the Agouti Signaling Protein Gene Is Associated with Human Pigmentation. **The American Journal Of Human Genetics**, [S.L.], v. 70, n. 3, p. 770-775, mar. 2002. Elsevier BV. http://dx.doi.org/10.1086/339076.

KATOCH, Sourabh; CHAUHAN, Sumit Singh; KUMAR, Vijay. A review on genetic algorithm: past, present, and future. **Multimedia Tools And Applications**, [S.L.], v. 80, n. 5, p. 8091-8126, 31 out. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11042-020-10139-6.

KAYSER, Manfred. Forensic DNA Phenotyping: predicting human appearance from crime scene material for investigative purposes. **Forensic Science International: Genetics**, [S.L.], v. 18, p. 33-48, set. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2015.02.003.

KHORUDDIN, Nurul Ain *et al*. Pathogenic nsSNPs that increase the risks of cancers among the Orang Asli and Malays. **Scientific Reports**, [S.L.], v. 11, n. 1, p. 1-22, 9 ago. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-021-95618-y.

KIMPTON, Colin et al. Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. **International Journal Of Legal Medicine**, [S.L.], v. 106, n. 6, p. 302-311, nov. 1994. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/bf01224776.

LONA-DURAZO, Frida *et al*. Investigating the genetic architecture of eye colour in a Canadian cohort. **Iscience**, [S.L.], v. 25, n. 6, p. 104485, jun. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.isci.2022.104485.

LE, James. **Support Vector Machines in R**. Disponível em: https://www.datacamp.com/tutorial/support-vector-machines-r. Acesso em: 23 jun. 2022.

LEERUNYAKUL, Kanchana; SUCHONWANIT, Poonkiat. Asian Hair: a review of structures, properties, and distinctive disorders. Clinical, **Cosmetic And Investigational Dermatology**, [S.L.], v. 13, p. 309-318, abr. 2020. Informa UK Limited. http://dx.doi.org/10.2147/ccid.s247390.

LIMA, Felícia de Araújo; GONÇALVES, Fernanda de Toledo; FRIDMAN, Cintia. SLC24A5 and ASIP as phenotypic predictors in Brazilian population for forensic purposes. **Legal Medicine**, [S.L.], v. 17, n. 4, p. 261-266, jul. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.legalmed.2015.03.001.

LINACRE, Adrian; TEMPLETON, Jennifer E. L.. Forensic DNA profiling: state of the art. **Research And Reports In Forensic Medical Science**, [S.L.], p. 25, ago. 2014. Informa UK Limited. http://dx.doi.org/10.2147/rrfms.s60955.

LIU, Fan *et al*. Eye color and the prediction of complex phenotypes from genotypes. **Current Biology**, [S.L.], v. 19, n. 5, p. 192-193, mar. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.cub.2009.01.027.

LIU, Fan *et al*. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. **Human Genetics**, [S.L.], v. 134, n. 8, p. 823-835, 12 maio 2015. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00439-015-1559-0

LUQUE, Amalia *et al*. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. **Pattern Recognition**, [S.L.], v. 91, p. 216-231, jul. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.patcog.2019.02.023.

MARANO, Leonardo Arduino; FRIDMAN, Cintia. DNA phenotyping: current application in forensic science. **Research And Reports In Forensic Medical Science**, [S.L.], v. 9, p. 1-8, fev. 2019. Informa UK Limited. http://dx.doi.org/10.2147/rrfms.s164090.

MAROÑAS, O. *et al*. The Genetics of Skin, Hair, and Eye Color Variation and Its Relevance to Forensic Pigmentation Predictive Tests. **Forensic Science Review**, [S.L], v. 27, n. 1, p. 13-40, jan. 2015.

MENON, Adarsh. **Logistic Regression in Machine Learning using Python.** Disponível em: https://towardsdatascience.com/logistic-regression-explained-and-implemented-in-python-880955306060. Acesso em: 23 jun. 2022.

MEYER, Olivia S. *et al*. Association between brown eye colour in rs12913832: gg individuals and snps in tyr, tyrp1, and slc24a4. **Plos One**, [S.L.], v. 15, n. 9, p. 1-15, 11 set. 2020. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0239131.

MUSHAILOV, Vladimir *et al*. Assay Development and Validation of an 8-SNP Multiplex Test to Predict Eye and Skin Coloration. **Journal Of Forensic Sciences**, [S.L.], v. 60, n. 4, p. 990-1000, 17 mar. 2015. Wiley. http://dx.doi.org/10.1111/1556-4029.12758.

NAN, Hongmei *et al*. Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. **International Journal Of Cancer**, [S.L.], v. 125, n. 4, p. 909-917, 15 ago. 2009. Wiley. http://dx.doi.org/10.1002/ijc.24327.

NASIR, Inzamam Mashood *et al*. Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training. **Sensors**, [S.L.], v. 20, n. 23, p. 6793-6811, 27 nov. 2020. MDPI AG. http://dx.doi.org/10.3390/s20236793.

NIMS, Raymond W. *et al*. Short tandem repeat profiling: part of an overall strategy for reducing the frequency of cell misidentification. **In Vitro Cellular & Developmental Biology - Animal**, [S.L.], v. 46, n. 10, p. 811-819, 7 out. 2010. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11626-010-9352-9.

NORTON, H. L. *et al*. Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. **Molecular Biology And Evolution**, [S.L.], v. 24, n. 3, p. 710-722, 5 dez. 2006. Oxford University Press (OUP). http://dx.doi.org/10.1093/molbev/msl203.

NOWROOZI, Ehsan *et al*. A survey of machine learning techniques in adversarial image forensics. **Computers & Security**, [S.L.], v. 100, p. 1-50, jan. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.cose.2020.102092.

NUSSBAUM, Robert L.; MCINNES, Roderick R.; WILLARD, Huntington F.. Diversidade Genética Humana: Mutação e Polimorfismo. In: NUSSBAUM, Robert L. MCINNES, Roderick R.; WILLARD, Huntington F.. **Thompson & Thompson Genética Médica**. 8. ed. Rio de Janeiro: Elsevier, 2016. Cap. 4. p. 43-54.

OBITE, C. P. *et al*. Multicollinearity Effect in Regression Analysis: a feed forward artificial neural network approach. **Asian Journal Of Probability And Statistics**, [S.L.], p. 22-33, 9 jan. 2020. Sciencedomain International. http://dx.doi.org/10.9734/ajpas/2020/v6i130151.

OLSON, Randal S. *et al*. Data-driven advice for applying machine learning to bioinformatics problems. **Biocomputing 2018**, [S.L.], p. 1-12, 17 nov. 2017. WORLD SCIENTIFIC. http://dx.doi.org/10.1142/9789813235533_0018.

ONGARO, Linda et al. The Genomic Impact of European Colonization of the Americas. **Current Biology**, [S.L.], v. 29, n. 23, p. 3974-3986, dez. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.cub.2019.09.076.

PALMAL, Sagnik *et al.* PREDICTION OF EYE, HAIR AND SKIN COLOUR IN LATIN AMERICANS. **Forensic Science International: Genetics**, [S.L.], p. 102517, abr. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2021.102517.

PARSON, W. *et al*. DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. **Forensic Science International: Genetics**, [S.L.], v. 13, p. 134-142, nov. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2014.07.010.

PEDREGOSA, Fabian *et al*. Scikit-learn: machine learning in python. **Arxiv**, [S.L.], p. 1-6, 2012. ArXiv. http://dx.doi.org/10.48550/ARXIV.1201.0490.

PHILLIPS, Chris. The Golden State Killer investigation and the nascent field of forensic genealogy. **Forensic Science International: Genetics**, [S.L.], v. 36, p. 186-188, set. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2018.07.010.

PICARDO, Mauro; CARDINALI, Giorgia. The Genetic Determination of Skin Pigmentation: kitlg and the kitlg/c-kit pathway as key players in the onset of human familial pigmentary diseases. **Journal Of Investigative Dermatology**, [S.L.], v. 131, n. 6, p. 1182-1185, jun. 2011. Elsevier BV. http://dx.doi.org/10.1038/jid.2011.67.

PINHEIRO, M. Fátima. Criminalística Biológica. In: CORTE-REAL, Francisco; VIEIRA, Duarte Nuno. **Princípios da Genética Forense**. Coimbra: Imprensa da Universidade de Coimbra, 2015. Cap. 2. p. 43-73.

PNEUMAN, Amanda *et al*. Verification of eye and skin color predictors in various populations. **Legal Medicine**, [S.L.], v. 14, n. 2, p. 78-83, mar. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.legalmed.2011.12.005.

POLAKA, Inese; TOM, Igor; BORISOV, Arkady. Decision Tree Classifiers in Bioinformatics. **Scientific Journal Of Riga Technical University. Computer Sciences**, [S.L.], v. 42, n. 1, p. 118-123, 1 jan. 2010. Walter de Gruyter GmbH. http://dx.doi.org/10.2478/v10143-010-0052-4.

POŚPIECH, Ewelina *et al*. Gene–gene interactions contribute to eye colour variation in humans. **Journal Of Human Genetics**, [S.L.], v. 56, n. 6, p. 447-455, 7 abr. 2011. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/jhg.2011.38.

QADIR, Abdalbasit Mohammed; VAROL, Asaf. The Role of Machine Learning in Digital Forensics. **2020 8Th International Symposium On Digital Forensics And Security (Isdfs)**, [S.L.], p. 1-5, jun. 2020. IEEE. http://dx.doi.org/10.1109/isdfs49300.2020.9116298.

RACHMIN, Inbal *et al*. Topical treatment strategies to manipulate human skin pigmentation. **Advanced Drug Delivery Reviews**, [S.L.], v. 153, p. 65-71, jan. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.addr.2020.02.002.

RANBIR *et al*. Machine Learning-Based Analytical Systems: food forensics. **Acs Omega**, [S.L.], v. 7, n. 51, p. 47518-47535, 16 dez. 2022. American Chemical Society (ACS). http://dx.doi.org/10.1021/acsomega.2c05632.

RASHID, Omar Fitian; OTHMAN, Zulaiha Ali; ZAINUDIN, Suhaila. A Novel DNA Sequence Approach for Network Intrusion Detection System Based on Cryptography Encoding Method. **International Journal On Advanced Science**, Engineering And Information Technology, [S.L.], v. 7, n. 1, p. 183, 25 fev. 2017. Insight Society. http://dx.doi.org/10.18517/ijaseit.7.1.1569.

REIS, Larissa B. *et al*. Skin pigmentation polymorphisms associated with increased risk of melanoma in a case-control sample from southern Brazil. **Bmc Cancer**, [S.L.], v. 20, n. 1, p. 1-11, 9 nov. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s12885-020-07485-x.

RUIZ, Y. *et al*. Further development of forensic eye color predictive tests. **Forensic Science International: Genetics**, [S.L.], v. 7, n. 1, p. 28-40, jan. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2012.05.009.

RUIZ-LINARES, Andrés *et al*. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. **Plos Genetics**, [S.L.], v. 10, n. 9, p. 1-13, 25 set. 2014. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pgen.1004572.

SAFRAN, Marilyn *et al*. The GeneCards Suite. **Practical Guide To Life Science Databases**, [S.L.], p. 27-56, 2021. Springer Singapore. http://dx.doi.org/10.1007/978-981-16-5812-9_2.

SALZBERG, Steven L.. C4.5: programs for machine learning by J. Ross quinlan. morgan kaufmann publishers, inc., 1993. **Machine Learning**, [S.L.], v. 16, n. 3, p. 235-240, set. 1994. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/bf00993309.

SALVORO, Cecilia *et al*. Performance of four models for eye color prediction in an Italian population sample. **Forensic Science International: Genetics**, [S.L.], v. 40, p. 192-200, maio 2019. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2019.03.008.

SARKER, Iqbal H.. Machine Learning: algorithms, real-world applications and research directions. **Sn Computer Science**, [S.L.], v. 2, n. 3, p. 1-21, 22 mar. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s42979-021-00592-x.

SHAPTURENKO, M.N. et al. HERC2 (rs12913832) and OCA2 (rs1800407) genes polymorphisms in relation to iris color variation in Belarusian population. **Forensic Science International: Genetics Supplement Series**, [S.L.], v. 7, n. 1, p. 331-332, dez. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.fsigss.2019.09.127.

SILVA JUNIOR, Ronaldo Carneiro da *et al*. Development of DNA databases in Latin America. **Forensic Science International**, [S.L.], v. 316, p. 110540, nov. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.forsciint.2020.110540.

SCHNEIDER, Peter M.; PRAINSACK, Barbara; KAYSER, Manfred. The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. **Deutsches Aerzteblatt Online**, [S.L.], p. 873-880, 23 dez. 2019. Deutscher Arzte-Verlag GmbH. http://dx.doi.org/10.3238/arztebl.2019.0873.

SCHOBER, Patrick; BOER, Christa; SCHWARTE, Lothar A.. Correlation Coefficients. **Anesthesia & Analgesia**, [S.L.], v. 126, n. 5, p. 1763-1768, maio 2018. Ovid Technologies (Wolters Kluwer Health). http://dx.doi.org/10.1213/ane.0000000000002864.

SLOMINSKI, Andrzej; PAUS, Ralf. Melanogenesis Is Coupled to Murine Anagen: toward new concepts for the role of melanocytes and the regulation of melanogenesis in hair growth.. **Journal Of Investigative Dermatology**, [S.L.], v. 101, n. 1, p. 90-97, jul. 1993. Elsevier BV. http://dx.doi.org/10.1111/1523-1747.ep12362991.

SLOMINSKI, Andrzej *et al*. Melanin Pigmentation in Mammalian Skin and Its Hormonal Regulation. **Physiological Reviews**, [S.L.], v. 84, n. 4, p. 1155-1228, out. 2004. American Physiological Society. http://dx.doi.org/10.1152/physrev.00044.2003.

SLOMINSKI, Andrzej et al. Hair Follicle Pigmentation. **Journal Of Investigative Dermatology**, [S.L.], v. 124, n. 1, p. 13-21, jan. 2005. Elsevier BV. http://dx.doi.org/10.1111/j.0022-202x.2004.23528.x.

SOUZA, Juliana Maria de *et al*. Forensic DNA Phenotyping: starting point to prediction model in pernambuco population, brazil. **Research, Society And Development**, [S.L.], v. 10, n. 13, p. 1-30, 11 out. 2021. Research, Society and Development. http://dx.doi.org/10.33448/rsd-v10i13.20955.

SPARKES, R. et al. The validation of a 7-locus multiplex STIR test for use in forensic casework. **International Journal Of Legal Medicine**, [S.L.], v. 109, n. 4, p. 195-204, dez. 1996. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/bf01225518.

SPICHENOK, Olga *et al*. Prediction of eye and skin color in diverse populations using seven SNPs. **Forensic Science International**: Genetics, [S.L.], v. 5, n. 5, p. 472-478, nov. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2010.10.005.

STURM, R. Eye colour: portals into pigmentation genes and ancestry. **Trends In Genetics**, [S.L.], v. 20, n. 8, p. 327-332, ago. 2004. Elsevier BV. http://dx.doi.org/10.1016/j.tig.2004.06.010.

STURM, Richard A. et al. A Single SNP in an Evolutionary Conserved Region within Intron 86 of the HERC2 Gene Determines Human Blue-Brown Eye Color. **The American Journal Of Human Genetics**, [S.L.], v. 82, n. 2, p. 424-431, fev. 2008. Elsevier BV. http://dx.doi.org/10.1016/j.ajhg.2007.11.005.

STURM, Richard A.; LARSSON, Mats. Genetics of human iris colour and patterns. **Pigment Cell & Melanoma Research**, [S.L.], v. 22, n. 5, p. 544-562, out. 2009. Wiley. http://dx.doi.org/10.1111/j.1755-148x.2009.00606.x.

STURM, Richard; DUFFY, David L. Human pigmentation genes under environmental selection. **Genome Biology**, [S.L.], v. 13, n. 9, p. 248-263, 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/gb-2012-13-9-248.

SULEM, Patrick *et al*. Genetic determinants of hair, eye and skin pigmentation in Europeans. **Nature Genetics**, [S.L.], v. 39, n. 12, p. 1443-1452, 21 out. 2007. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/ng.2007.13.

THORVALDSDOTTIR, H.; ROBINSON, J. T.; MESIROV, J. P.. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. **Briefings In Bioinformatics**, [S.L.], v. 14, n. 2, p. 178-192, 19 abr. 2012. Oxford University Press (OUP). http://dx.doi.org/10.1093/bib/bbs017.

TOZZO, Pamela et al. External visible characteristics prediction through SNPs analysis in the forensic setting: a review. **Frontiers In Bioscience-Landmark**, [S.L.], v. 26, n. 10, p. 828, 2021. IMR Press. http://dx.doi.org/10.52586/4991.

TURCHETTO-ZOLET, Andreia Carina *et al*. Polimorfismo de Nucleotídeo único (SNP): metodologias de identificação, análise e aplicações. In: TURCHETTO-ZOLET, Andreia

Carina *et al*. (org.). **Marcadores Moleculares na Era Genômica: Metodologias e Aplicações**. Ribeirão Preto: Comissão Editorial Sociedade Brasileira de Genética, 2017. Cap. 8. p. 132-179.

UDOGADI, Nwawuba Stanley *et al*. Forensic DNA Profiling: autosomal short tandem repeat as a prominent marker in crime investigation. **Malaysian Journal Of Medical Sciences**, [S.L.], v. 27, n. 4, p. 22-35, 2020. Penerbit Universiti Sains Malaysia. http://dx.doi.org/10.21315/mjms2020.27.4.3.

VIDEIRA, Inês Ferreira dos Santos; MOURA, Daniel Filipe Lima; MAGINA, Sofia. Mechanisms regulating melanogenesis. **Anais Brasileiros de Dermatologia**, [S.L.], v. 88, n. 1, p. 76-83, fev. 2013. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0365-05962013000100009.

VIRMOND, Marina Barreiros *et al*. Fenotipagem forense pelo DNA através de SNPs. **Revista Brasileira de Criminalística**, [S.L.], v. 5, n. 2, p. 37-47, 28 jul. 2016. Associação Brasileira de Criminalística - ABC. http://dx.doi.org/10.15260/rbc.v5i2.128.

VOISEY, J. *et al*. A polymorphism in the agouti signalling protein (ASIP) is associated with decreased levels of mRNA. **Pigment Cell Research**, [S.L.], v. 19, n. 3, p. 226-231, jun. 2006. Wiley. http://dx.doi.org/10.1111/j.1600-0749.2006.00301.x.

WAKAMATSU, Kazumasa; ZIPPIN, Jonathan H.; ITO, Shosuke. Chemical and biochemical control of skin pigmentation with special emphasis on mixed melanogenesis. **Pigment Cell & Melanoma Research**, [S.L.], v. 34, n. 4, p. 730-747, 22 mar. 2021. Wiley. http://dx.doi.org/10.1111/pcmr.12970.

WALSH, Susan *et al*. IrisPlex: a sensitive dna tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. **Forensic Science International: Genetics**, [S.L.], v. 5, n. 3, p. 170-180, jun. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2010.02.004.

WALSH, Susan *et al*. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. **Forensic Science International: Genetics**, [S.L.], v. 7, n. 1, p. 98-115, jan. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.fsigen.2012.07.005.

WALSH, Susan *et al*. Global skin colour prediction from DNA. **Human Genetics**, [S.L.], v. 136, n. 7, p. 847-863, 12 maio 2017. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00439-017-1808-5.

WARD, William H. *et al*. Clinical Presentation and Staging of Melanoma. In: W.H., Ward; J.M., Farma (ed.). **Cutaneous Melanoma: Etiology and Therapy**. Brisbane: Codon Publications, 2017. Cap. 6. p. 79-89.

WEBB, Geoffrey I. Naïve Bayes. **Encyclopedia Of Machine Learning And Data Mining**, [S.L.], p. 1-2, 2016. Springer US. http://dx.doi.org/10.1007/978-1-4899-7502-7_581-1.

WERRETT, David J. The National DNA Database. **Forensic Science International**, [S.L.], v. 88, n. 1, p. 33-42, jul. 1997. Elsevier BV. http://dx.doi.org/10.1016/s0379-0738(97)00081-9.

WICKENHEISER, Ray A.. Forensic genealogy, bioethics and the Golden State Killer case. **Forensic Science International: Synergy**, [S.L.], v. 1, p. 114-125, 2019. Elsevier BV. http://dx.doi.org/10.1016/j.fsisyn.2019.07.003.

YANG, Yaran; XIE, Bingbing; YAN, Jiangwei. Application of Next-generation Sequencing Technology in Forensic Science. **Genomics, Proteomics & Bioinformatics**, [S.L.], v. 12, n. 5, p. 190-197, out. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.gpb.2014.09.001.

ZANNA, Paola T. *et al*. Mechanism of dimerization of the human melanocortin 1 receptor. **Biochemical And Biophysical Research Communications**, [S.L.], v. 368, n. 2, p. 211-216, abr. 2008. Elsevier BV. http://dx.doi.org/10.1016/j.bbrc.2008.01.060.

ZANELLA, Camila Martini *et al*. Microssatélites: Metodologias de identificação e análise. In: TURCHETTO-ZOLET, Andreia Carina *et al*. (org.). **Marcadores Moleculares na Era Genômica: Metodologias e Aplicações**. Ribeirão Preto: Comissão Editorial Sociedade Brasileira de Genética, 2017. Cap. 6. p. 94-117.

ZAVALA, Elena I. *et al*. Impact of DNA degradation on massively parallel sequencing-based autosomal STR, iiSNP, and mitochondrial DNA typing systems. **International Journal Of Legal Medicine**, [S.L.], v. 133, n. 5, p. 1369-1380, 2 jul. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00414-019-02110-4.

ZEIGLER-JOHNSON, Charnita *et al*. Population Differences in the Frequency of the Agouti Signaling Protein g.8818A>G Polymorphism. **Pigment Cell Research**, [S.L.], v. 17, n. 2, p. 185-187, abr. 2004. Wiley. http://dx.doi.org/10.1111/j.1600-0749.2004.00134.x.

ZHOU, Hongfang; WANG, Xiqian; ZHU, Rourou. Feature selection based on mutual information with correlation coefficient. **Applied Intelligence**, [S.L.], v. 52, n. 5, p. 5457-5474, 12 ago. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10489-021-02524-x.

ZORINA-LICHTENWALTER, Katerina et al. A study in scarlet: MC1R as the main predictor of red hair and exemplar of the flip-flop effect. **Human Molecular Genetics**, [S.L.], v. 28, n. 12, p. 2093-2106, 16 jan. 2019. Oxford University Press (OUP). http://dx.doi.org/10.1093/hmg/ddz018.

ZOU, Kelly H.; TUNCALI, Kemal; SILVERMAN, Stuart G.. Correlation and Simple Linear Regression. **Radiology**, [S.L.], v. 227, n. 3, p. 617-628, jun. 2003. Radiological Society of North America (RSNA). http://dx.doi.org/10.1148/radiol.2273011499.

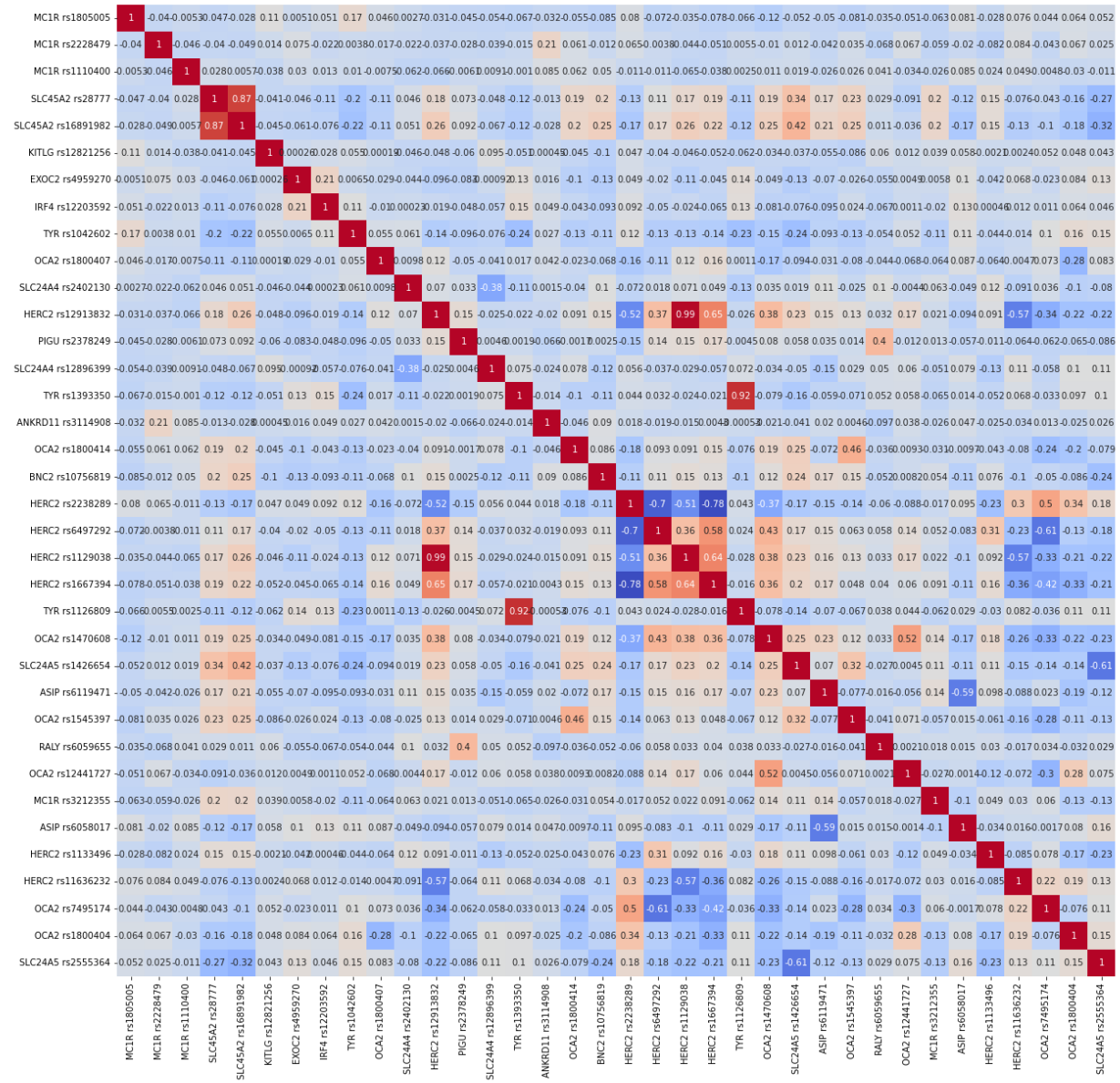
APÊNDICE A - MARCADORES DO HIRISPLEX-S

MC1R rs312262906, MC1R rs11547464, MC1R rs885479, MC1R rs1805008, MC1R rs1805005, MC1R rs1805006, 'MC1R rs1805007, MC1R rs1805009, MC1R rs201326893, MC1R rs2228479, MC1R rs1110400, SLC45A2 rs28777, SLC45A2 rs16891982, KITLG rs12821256, EXOC2 rs4959270, IRF4 rs12203592, TYR rs1042602, OCA2 rs1800407, SLC24A4 rs2402130, HERC2 rs12913832, PIGU rs2378249, SLC24A4 rs12896399, TYR rs1393350, TYRP1 rs683, ANKRD11 rs3114908, OCA2 rs1800414, BNC2 rs10756819, HERC2 rs2238289, SLC24A4 rs17128291, HERC2 rs6497292, HERC2 rs1129038, HERC2 rs1667394, TYR rs1126809, OCA2 rs1470608, SLC24A5 rs1426654, ASIP rs6119471, OCA2 rs1545397, RALY rs6059655', OCA2 rs12441727, MC1R rs3212355, DEF8 rs8051733.

APÊNDICE B - MARCADORES EXTRAS

ASIP rs6058017, HERC2 rs1133496, HERC2 rs11636232, OCA2 rs7495174, OCA2 rs1800404, OCA2 rs1800416, SLC24A5 rs2555364, SLC24A5 rs16960620.

APÊNDICE C - MAPA DE CALOR DO TESTE DE CORRELAÇÃO LINEAR DE PEARSON ENTRE PARES DE MARCADORES



Fonte: Autor (2022).

APÊNDICE D - HIPERPARÂMETRO DOS CLASSIFICADORES

Marcadores: CORR, fenótipo: olhos (duas classes)

	model	best_score	best_params
0	svm	0.858618	{'C': 4, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.862708	{'bootstrap': False, 'max_depth': 2, 'max_leaf...
2	logistic_regression	0.803277	{'C': 3, 'solver': 'lbfgs', 'warm_start': False}
3	decision_tree	0.864766	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.838105	{'alpha': 53, 'fit_prior': False}

Fonte: Autor (2022).

Marcadores: CORR, fenótipo: olhos (três classes)

	model	best_score	best_params
0	svm	0.739781	{'C': 2, 'kernel': 'poly', 'gamma': 'auto', 's...
1	random_forest	0.754122	{'bootstrap': True, 'max_depth': 6, 'max_leaf_...
2	logistic_regression	0.727410	{'C': 36, 'solver': 'liblinear', 'warm_start':...
3	decision_tree	0.750006	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.721326	{'alpha': 10, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: CORR, fenótipo: cabelos

	model	best_score	best_params
0	svm	0.799174	{'C': 4, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.801219	{'bootstrap': False, 'max_depth': 12, 'max_lea...
2	logistic_regression	0.768436	{'C': 4, 'solver': 'liblinear', 'warm_start': ...
3	decision_tree	0.797129	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.797129	{'alpha': 29, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: CORR, fenótipo: pele

	model	best_score	best_params
0	svm	0.735666	{'C': 8, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.731551	{'bootstrap': True, 'max_depth': 6, 'max_leaf_...
2	logistic_regression	0.698831	{'C': 8, 'solver': 'lbfgs', 'warm_start': False}
3	decision_tree	0.698768	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.739769	{'alpha': 0, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: MI, fenótipo: olhos (duas classes).

	model	best_score	best_params
0	svm	0.862894	{'C': 2, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.862882	{'bootstrap': False, 'max_depth': 3, 'max_leaf...
2	logistic_regression	0.800341	{'C': 4, 'solver': 'liblinear', 'warm_start': ...
3	decision_tree	0.864902	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.822587	{'alpha': 9, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: MI, fenótipo: olhos (três classes).

	model	best_score	best_params
0	svm	0.739978	{'C': 2, 'kernel': 'rbf', 'gamma': 'scale', 's...
1	random_forest	0.754083	{'bootstrap': True, 'max_depth': 10, 'max_leaf...
2	logistic_regression	0.727833	{'C': 1, 'solver': 'liblinear', 'warm_start': ...
3	decision_tree	0.752051	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.705696	{'alpha': 15, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: MI, fenótipo: cabelos.

	model	best_score	best_params
0	svm	0.798394	{'C': 1, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.798394	{'bootstrap': True, 'max_depth': 15, 'max_leaf...
2	logistic_regression	0.762091	{'C': 6, 'solver': 'liblinear', 'warm_start': ...
3	decision_tree	0.798394	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.798394	{'alpha': 63, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: MI, fenótipo: pele.

	model	best_score	best_params
0	svm	0.725751	{'C': 2, 'kernel': 'rbf', 'gamma': 'scale', 's...
1	random_forest	0.739917	{'bootstrap': False, 'max_depth': 9, 'max_leaf...
2	logistic_regression	0.709614	{'C': 34, 'solver': 'saga', 'warm_start': False}
3	decision_tree	0.697603	{'criterion': 'gini', 'splitter': 'random', 'm...
4	multinomial_naive_bayes	0.713679	{'alpha': 3, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: MULTICOL, fenótipo: olhos (duas classes).

	model	best_score	best_params
0	svm	0.860323	{'C': 3, 'kernel': 'rbf', 'gamma': 'scale', 's...
1	random_forest	0.864363	{'bootstrap': False, 'max_depth': 8, 'max_leaf...
2	logistic_regression	0.821877	{'C': 2, 'solver': 'liblinear', 'warm_start': ...
3	decision_tree	0.864363	{'criterion': 'entropy', 'splitter': 'best', '...
4	multinomial_naive_bayes	0.840096	{'alpha': 41, 'fit_prior': False}

Fonte: Autor (2022).

Marcadores: MULTICOL, fenótipo: olhos (três classes).

	model	best_score	best_params
0	svm	0.742966	{'C': 8, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.761111	{'bootstrap': False, 'max_depth': 20, 'max_lea...
2	logistic_regression	0.712602	{'C': 7, 'solver': 'lbfgs', 'warm_start': False}
3	decision_tree	0.746957	{'criterion': 'entropy', 'splitter': 'best', '...
4	multinomial_naive_bayes	0.728776	{'alpha': 9, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: MULTICOL, fenótipo: cabelos.

	model	best_score	best_params
0	svm	0.797573	{'C': 3, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.797573	{'bootstrap': True, 'max_depth': 9, 'max_leaf_...
2	logistic_regression	0.773245	{'C': 1, 'solver': 'liblinear', 'warm_start': ...
3	decision_tree	0.799593	{'criterion': 'entropy', 'splitter': 'best', '...
4	multinomial_naive_bayes	0.797573	{'alpha': 76, 'fit_prior': True}

Fonte: Autor (2022).

Marcadores: MULTICOL, fenótipo: pele.

	model	best_score	best_params
0	svm	0.738914	{'C': 6, 'kernel': 'rbf', 'gamma': 'auto', 'sh...
1	random_forest	0.744949	{'bootstrap': False, 'max_depth': 7, 'max_leaf...
2	logistic_regression	0.714597	{'C': 2, 'solver': 'liblinear', 'warm_start': ...
3	decision_tree	0.702377	{'criterion': 'entropy', 'splitter': 'random',...
4	multinomial_naive_bayes	0.718675	{'alpha': 8, 'fit_prior': True}

Fonte: Autor (2022).