Using a random forest classifier to generate hair and eye colour predictions from SNP data

Goal

This project evaluates the ability of a random-forest classifier to correctly assign hair colour and eye colour labels to samples using genotype data collected from 24 phenotype-informative SNP loci. While the performance of phenotype classifiers is quite high in homogeneous populations, the dataset used to test the RF classifier here includes admixed individuals only. The exploration of admixed datasets is essential to ensuring phenotype prediction can be used as a fair and accurate tool in forensic contexts.

Hair and Eye Colour Classes

There are three possible hair colours that can be assigned to a sample: 1. Brown 2. Black 3. Blonde

There are three possible eye colours that can be assigned to a sample: 1. Brown 2. Blue 3. Intermediate - This includes any non-brown and non-blue eye colour, such as grey, green, and hazel.

The true hair colour and eye colour for each sample will be used to determine if the classifier correctly identifies the group a sample belongs to.

Random Forest Classifier Performance

The performance of hair and eye colour predictions were assessed separately.

Hair Colour

Due to the small number of blonde samples, the metrics in Table 2 for the blonde class are heavily skewed.

Table 1: Hair Colour Confusion Matrix

True	Predicted		
	Black	Blonde	Brown
Black	1	0	2
Blonde	0	0	2
Brown	3	2	6

Table 2: Hair Colour Classifier Metrics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
Class: Black	0.25	0.833	0.333	0.769	0.333	0.25	0.286

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
Class: Blonde	0.00	0.857	0.000	0.857	0.000	0.00	NaN
Class: Brown	0.60	0.167	0.545	0.200	0.545	0.60	0.571

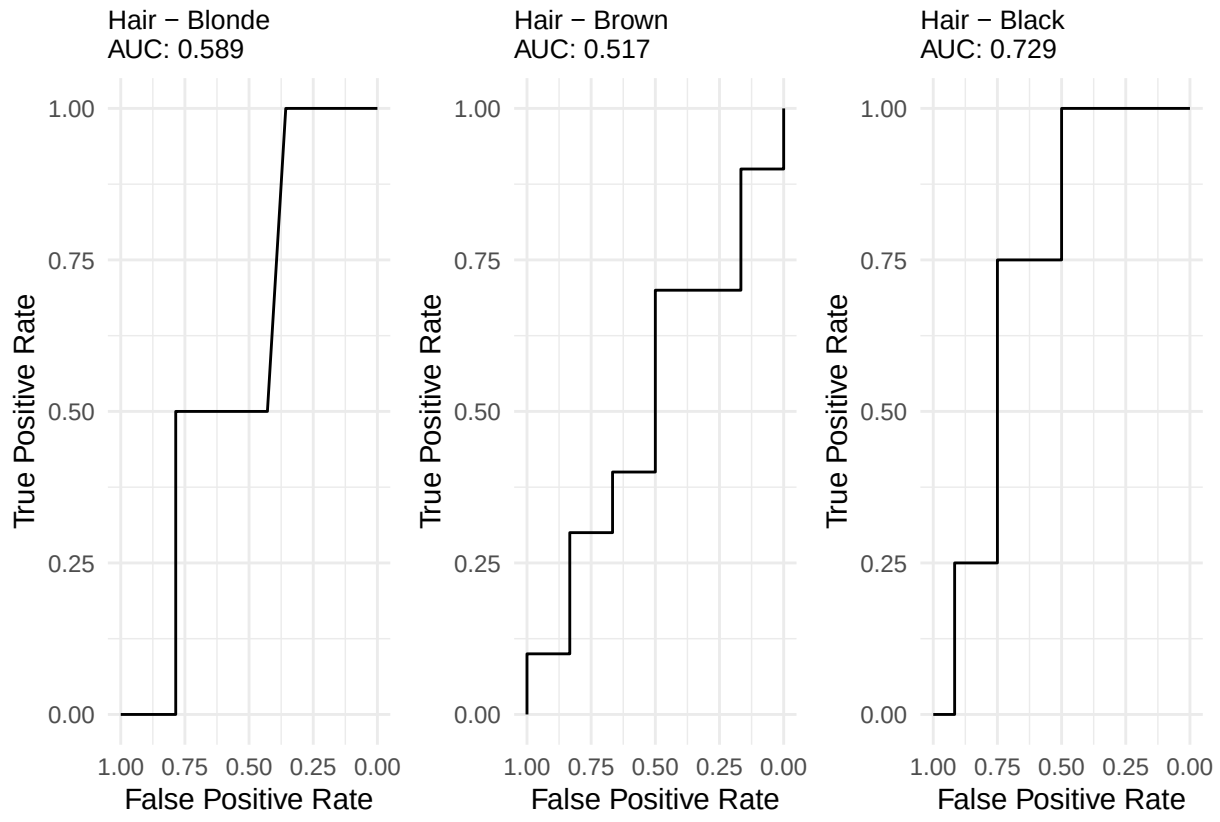

Table 3: Hair Colour - AUC per Class

	AUC
Black	0.7291667
Blonde	0.5892857
Brown	0.5166667

Eye Colour

Due to a low number of samples in the intermediate class, the performance metrics for this class shown in Table 5 are heavily skewed.

Table 4: Eye Colour Confusion Matrix

True	Predicted		
	Brown	Intermediate	Blue

Brown	11	2	1
Intermediate	0	1	0
Blue	0	0	0

Table 5: Eye Colour Classifier Metrics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
Class: Brown	1.000	0.25	0.786	1.000	0.786	1.000	0.88
Class: Intermediate	0.333	1.00	1.000	0.857	1.000	0.333	0.50
Class: Blue	0.000	1.00	NaN	0.933	NA	0.000	NA

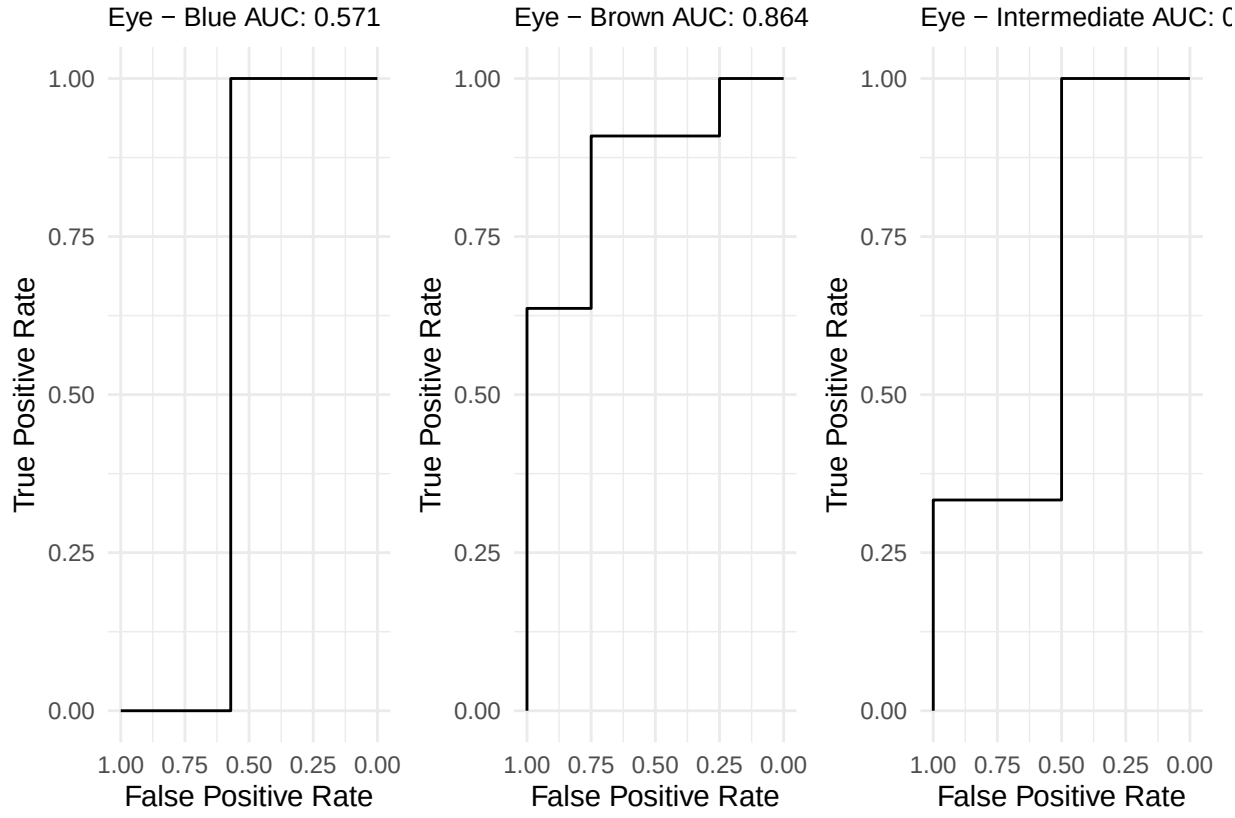

Table 6: Eye Colour - AUC per Class

	AUC
Brown	0.8636364
Intermediate	0.6666667
Blue	0.5714286

Conclusions

The interpretation of these results is limited due to the small sample size overall. The current metrics show the RF model is most effective at classifying brown eyes and black hair. If this trend holds true in a larger sample, this would indicate that the 24 SNPs utilized in this study are most informative for brown eyes and black hair, but less informative for other hair and eye colour classes. The addition of more SNPs to phenotyping panels may assist in more effective classification of blonde and brown hair, and intermediate and blue eyes.