Pak. J. Biochem. Mol. Biol., 2023, 56(3):95-109

Research Paper

Ancestry Analysis in Population-Scale Genomic Data

Haseeb Zahid

Group M, Karachi, Pakistan.

ABSTRACT

Significant distinctions exist among ethnic groups, encompassing variations in traits such as height, eye color, skin tone, susceptibility to certain illnesses, and responses to specific medications. However, there has been insufficient exploration into the genetic foundations of these differences. The Human Genome Diversity Project has amassed extensive genotypic data from Asian populations. Although Principal Component Analysis (PCA) can aid in discerning disparities among populations, it overlooks variations in individual Single Nucleotide Polymorphisms (SNPs) between populations. Thus, alternative statistical methodologies, such as the "mutual information algorithm," prove valuable in identifying SNPs associated with specific ethnicities and quantifying the discrepancies in SNPs within the Pakistani population. This study endeavors to uncover SNP variations among various ethnic groups in Pakistan. Employing the mutual information algorithm, we statistically compare each SNP across diverse ethnicities within our sample. Subsequently, we construct a classifier capable of determining an individual's ethnicity based on their genetic data, likely through techniques like feature engineering or dimensionality reduction. To assess the classifier's accuracy, we utilize a separate test dataset. The results indicate a 40% success rate in accurately predicting an individual's ethnicity within the test dataset.

KEYWORDS: Principal Component Analysis, Polymorphisms, SNPs.

INTRODUCTION

Single nucleotide polymorphism (SNPs) is the most prevalent type of genetic variation. An SNP, for instance, can change a particular region of the DNA sequence thymine (T) nucleotide to cytosine (C) [1]. Human DNA frequently contains SNPs. The human genome has about 4-5 million SNPs. A mutation must affect 1% or more of the population to be considered an SNP. In diverse communities throughout the world, scientists have found over 600 million SNPs. SNPs differ from copy number variations (CNVs), which are characterized by the duplication or deletion of significant DNA regions or whole genes.

SNPs can be used as biomarkers to find genes linked to illness since they are frequently found in coding and non-coding areas of genes. SNPs can directly cause disease by changing gene function when they are found in a gene or nearby non-coding areas. Even though the great majority of SNPs have no impact on growth or health, some have been proven to be crucial in research on human health. SNPs can be used to forecast a person's response to certain medications, sensitivity to environmental contaminants, and disease susceptibility. SNPs can be used to monitor the generational transmission of disease-related genetic variations [2]. SNPs linked to complex diseases including cancer, diabetes, and heart disease are currently the subject of research.

The dbSNP database, which contains over 60 million SNP sites acquired over millions of years because of mutations and natural selection, has been made available by the NCBI [21]. Most of these SNPs are in the non-coding section of the genome. In the coding region, SNPs can be categorized as synonymous or non-synonymous depending on how they affect the sequence

of a protein. As GWAS and DNA sequencing technologies advance, more SNPs are anticipated to be found in the next years. This might help with the creation of genetic markers and personalized therapy. SNPs, which can act as evolutionary makers, have also aided in our knowledge of how evolution occurs. The Human Genome Project, a multinational endeavor to discover the mysteries of the human genome, has the potential to change the medicine practice in the twenty-first century. It can lead to more accurate techniques of disease prediction, identification, and therapy by providing tools to identify the genetic component of nearly all diseases [2]. However, for these innovations to be effective, they must be accessible to everybody including scientists, clinicians, research participants, and those involved in ethical and social debates. The Human Genome Project has also demonstrated how astonishingly similar people are at the genetic level, with a 99.9% DNA similarity. As a result, using research to justify precise racial bounds is difficult. Examining the remaining 0.1% of genetic diversity, particularly the distribution of single nucleotide polymorphisms (SNPs) between afflicted and unaffected persons, can teach medical researchers a lot about the genetic contributions to complex illnesses like cancer, diabetes, and mental illness. It is critical to carefully balance the risks and benefits of such research to ensure that everyone benefits from its medical achievements. The human genome is 99.9% similar among all individuals, with the remaining 0.1% of genetic variation accounting for unique characteristics. The majority of this variation is due to single nucleotide polymorphisms (SNPs), of which 85% are shared by all human populations and the remaining 15% are population-specific. These SNPs contribute to differences in physical features, susceptibility to illness, and drug responsiveness in populations. For example, variations in SNP alleles can affect the incidence of G6PD deficiency and β-thalassemia in different populations. Efforts have been made to identify key SNPs that influence human pigmentation, but they do not fully explain differences between populations. When doing research, it is critical to consider the genetic variety of communities since findings conducted in one group may not apply to another.

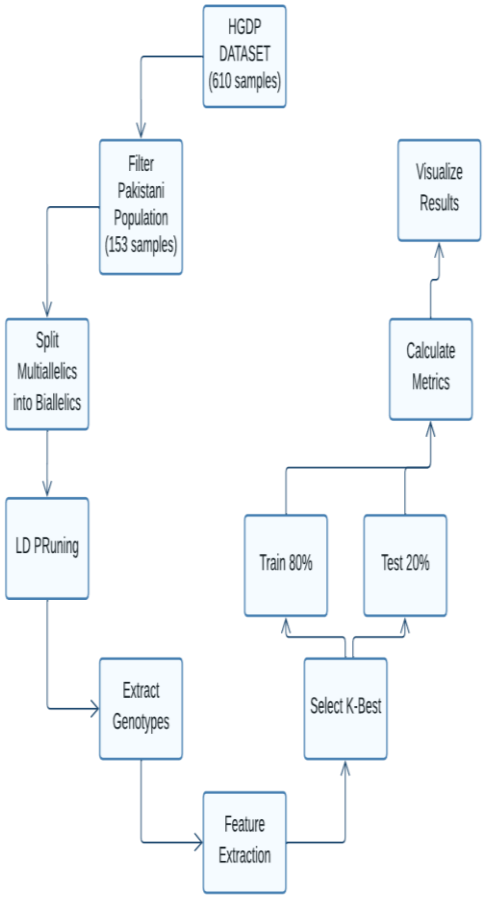
We conducted a comprehensive analysis of genetic variations across ethnic groups using the Human Genome Diversity Project genotype data, which has a larger number of ethnic groups and detailed data. We used advanced feature selection techniques to identify relevant SNPs and found some SNPs that could serve as potential biomarkers for different ethnic groups [5,8]. The nearby genes of these SNPs exhibited noteworthy functions [6,7].

MATERIALS AND METHODS

This study was conducted under the supervision of Dr. M. Kamran Azim, Mohammad Ali Jinnah University, Karachi, Pakistan.

For human ancestry analysis in population-scale genomic data, the whole genome variant data of chromosomes 18, 19, 20, 21, 22, 23, X and Y were downloaded from HGDP project website (release #20190516).

Figure 1: Pipeline processes of the ancestry analysis carried out during this study.



The HGDP dataset is a publicly available dataset containing genetic data for over 1000 individuals from different populations around the world. To start the analysis, the genomic datasets were downloaded from the SFTP site. The dataset contains information about genotypes, which are the genetic variants that an individual carries.

After downloading the data, the datasets belonging to the Pakistani population were extracted. This can be done using the metadata provided with the dataset, which contains information about the population of origin for each individual. In this instance, we choose 153 samples of Pakistani origin.

The genetic variations in the HGDP dataset contain multiallelic variants which were converted into biallelic variants to streamline the analysis. This implies that only "major" and "minor" alleles were considered.

Using the LD pruning method, associated genomic variants are taken out of the dataset. Correlated variations might skew the study and make it hard to understand the findings, thus knowing this is crucial. Linkage disequilibrium (LD) is measured between each pair of variations, and if the LD between them is greater than a predetermined threshold, one of the variants is removed. Up until no pairings of variants have an LD over the threshold, this process is repeated.

The genotypes for each sample were extracted after conducting LD pruning and filtering the dataset. As a result, a matrix in which each row denotes a sample, and each column denotes a genetic variation was obtained.

The practice of producing new features (i.e., variables) from already existing ones is known as feature engineering [9,10]. Calculating summary statistics for each sample, such as the minor allele frequency (MAF), the heterozygosity percentage, or the number of uncommon variations, can be done in the context of genetics. Machine learning algorithms can take these factors into account as input. The pipeline used during this study is represented in Figure 1.

The k-best characteristics that are most relevant for the model used for machine learning must be chosen after completing feature engineering. Techniques like mutual information or feature selection based on correlation can be used for this. Cross-validation or other performance indicators may be used to determine the number of features to pick (k). The dataset has to be divided into training and testing sets to analyze the machine learning

model's performance. The model is trained using a training set, and its performance is assessed on a testing set. No connections of any kind should be made between samples from the training and testing sets.

Following the training of the machine learning model, a table displaying the number of true positives, false positives, true negatives, and false negatives for the classification job may be created. The confusion matrix is what's called this. Accuracy, precision, recall, and F1-score are a few examples of performance metrics that may be computed using this matrix. The analysis's findings may then be shown visually [11-20,22].

RESULTS

In this study, we analyzed variations in single nucleotide polymorphisms (SNPs) among eight ethnic groups in Pakistan i.e. Sindhi, Pathan, Makrani, Kalash, Hazara, Burusho, Brahui and Balochi [21]. We compared each SNP with the various ethnicities in the sample using a statistical method known as the Mutual Information Algorithm (Table 1; Figures 2-8).

A classifier was created that could determine a person's ethnicity based on their genetic data after identifying the disparities between SNPs and ethnicities. Before being fed into the classifier, the genetic data was presumably encoded using a dimensionality reduction or feature engineering approach [3].

Using a test dataset, the classifier's accuracy was assessed. According to the results, the classifier had a 40% accuracy rate when predicting an individual's ethnicity in the test dataset (Table 1; Figures 3-8).

The SNP in chromosome Y provided the best feature significance score of all the chromosomes 18, 19, 20, 21, 22, X, and Y that were examined (Tables 3, 4 and 5).

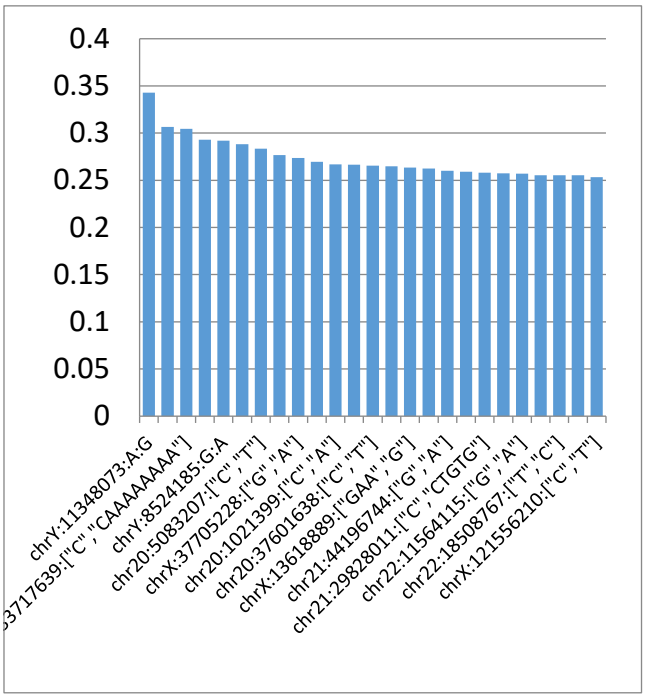


Figure 2: Analyzed SNPs along with their feature scores.

DISCUSSION

Recent years have seen a rise in interest in the field of research that uses genetic information to determine ethnicity. This occurs because of recognized variances in genetic makeup among various ethnic groups, which may be used to determine an individual's ethnicity. However, the use of genetic information for ethnic categorization presents ethical questions, especially in light of potential discrimination based on ethnicity.

To find differences in single nucleotide polymorphisms (SNPs) across various ethnic groups in the South Asian population, we employed a statistical method known as the Mutual Information Algorithm in our study. Then, using genetic data, we created a classifier that could determine a person's ethnicity. The classifier was reported to have a 40% accuracy rate on a test dataset [2,4].

Our study may be hampered by the classifier's relatively low accuracy. This might be due to several factors, such as the size and representativeness of the dataset, the complexity of the underlying genetic variations that affect ethnicity, or the choice of the statistical approach employed for categorization. Without knowing more about the study methods, it is hard to estimate the extent to which these factors may have influenced the findings.

Another potential issue is the ethical concerns raised by ethnic classification based on genetic information. Even while there may be valid justifications for using genetic data to identify an individual's ethnicity, such as in clinical or medical settings, there is also a potential of stigmatization and discrimination based on ethnicity. Since machine learning algorithms have the capacity to reinforce and even magnify biases that already exist in the underlying data, this worry is particularly important when it comes to our study may have a lot of ramifications for more research despite these limitations. Use of statistical methods like the Mutual Information Algorithm, for example, may be useful for finding genetic variations connected to different racial groups. Our ability to produce more specialized treatments and understand more about the underlying genetic causes of sickness may be based on our ability to comprehend these variances. It is essential to consider the possible effects of categorizing people based on their ethnicity using genetic information and to ensure that the required safeguards are in place to prevent misuse or damage. In addition, the development of classifiers that can determine ethnicity from genetic data may be helpful in clinical and medical contexts, particularly for diseases that are more common in specific ethnic groups.

CONCLUSION

It is challenging and morally problematic to use genetic information to categorize people based on their race. It is important to approach this field of research cautiously and to make sure that the required safety precautions are followed to prevent misuse or harm, even if research like ours may be valuable in identifying the genetic basis of disease. The ultimate goal of this research should be to further our understanding of genetics and sickness while simultaneously promoting equity, fairness, and social justice.

Table 1: Examined SNPs Together With their Feature Ratings and Normalized Score

Columns	Scores	Scores Normalized
chrY:11348073:A:G	0.3428350599	1
chrY:8524185:G:A	0.2919141772	0.8514711923
chrY:2954699:A:AAAAT	0.2767626781	0.8072764734
chrY:11338714:A:G	0.2625187222	0.7657289258
chrY:13817435:G:C	0.2590239102	0.7555350677
chrY:16098390:AT:A	0.2553464833	0.7448085483
chrY:20498974:A:G	0.2515315076	0.7336808193
chrY:11648056:G:T	0.2505381019	0.7307831991
chrY:15971266:C:CTG	0.2466528394	0.7194504537
chrY:6358000:TATC:T	0.2419374856	0.7056964525
chrY:25375721:A:AAAAT	0.2336012489	0.6813808627
chrY:12714066:T:C	0.232632013	0.6785537425
chrY:15187353:A:G	0.232361805	0.677765585
chrY:17196338:G:A	0.2318678081	0.676324668
chrY:13665242:G:A	0.2312908947	0.6746418956
chrY:19726907:T:C	0.2305029738	0.6723436449
chrY:18769967:ATG:A	0.2303583717	0.6719218617
chrY:13851402:G:C	0.2264933584	0.66064818

Table 2: Displaying Chromosomal X Findings Along With its Feature Scores

Columns	Score
chrX:37705228:["G","A"]	0.2736115189
chrX:13618889:["GAA","G"]	0.2635025626
chrX:85826394:["G","T"]	0.2573105337
chrX:67135941:["C","T"]	0.2554830838
chrX:121556210:["C","T"]	0.2533517202
chrX:34079098:["CT","C"]	0.2519772036
chrX:67283461:["A","C"]	0.2451618757
chrX:114663098:["C","T"]	0.2439502368
chrX:40946588:["G","A"]	0.2427283723
chrX:49966609:["T","A"]	0.2425120291
chrX:58522976:["C","T"]	0.2398367675
chrX:8462381:["G","T"]	0.2395106586
chrX:90790255:["A","G"]	0.2388189665
chrX:58437519:["A","G"]	0.23591146
chrX:2303877:["G","A"]	0.2350156685
chrX:106449862:["G","A"]	0.2336241439

Table 3: Chromosome 22 Results Along with Its Feature Scores

Columns	Score
chr22:11563939:["A","T"]	0.3066770882
chr22:11564026:["G","A"]	0.2931454646
chr22:12195030:["G","T"]	0.2883005526
chr22:11564115:["G","A"]	0.2571415278
chr22:18508767:["T","C"]	0.2554705315
chr22:26314398:["A","G"]	0.2542385349
chr22:32063317:["G","A"]	0.2511820172
chr22:11340331:["A","T"]	0.2478538097
chr22:50144643:["A","G"]	0.2447913854
chr22:48806988:["C","T"]	0.2435300958
chr22:11564079:["T","C"]	0.2406448146
chr22:15221405:["TC","T"]	0.2359907256
chr22:45052056:["G","C"]	0.2335805207
chr22:11308181:["GGAGCC GGCCGA","G"]	0.2325272008
chr22:18855853:["G","A"]	0.2309802704
chr22:20960993:["A","G"]	0.2271014122

Table 4: Chromosome 21 Results Along With Its Feature Scores

Columns	**Scores**
chr21:16957118:["G","A"]	0.2695256721
chr21:10010625:["A","G"]	0.2664364965
chr21:22267189:["A","T"]	0.2649774515
chr21:44196744:["G","A"]	0.2602320124
chr21:29828011:["C","CTGTG"]	0.2580252458
chr21:22238774:["A","C"]	0.2510015113
chr21:16183560:["G","A"]	0.2474249299
chr21:32371448:["G","A"]	0.2461606888
chr21:20525954:["CT","C"]	0.2372597689
chr21:37776188:["C","T"]	0.232873112
chr21:19986811:["G","T"]	0.2328462296

chr21:14620840:["G","T"]	0.2297010583
chr21:41652046:["G","C"]	0.2288796955
chr21:42983084:["A","G"]	0.226454467
chr21:22610102:["C","T"]	0.2250275302

Table 5: Chromosome 20 with their Feature Scores

Columns	Scores
chr20:33717639:["C","CAAAAAAAA"]	0.3044016004
chr20:5083207:["C","T"]	0.2833216333
chr20:1021399:["C","A"]	0.2668508256
chr20:37601638:["C","T"]	0.265361041
chr20:48724257:["T","C"]	0.2552140577
chr20:867784:["G","A"]	0.2533802114
chr20:7311531:["T","G"]	0.2483500638
chr20:49377102:["A","AAACAAAC"]	0.241491402
chr20:36687946:["CACAA","C"]	0.2401576933
chr20:39640891:["C","A"]	0.2399237843
chr20:36188581:["A","AAC"]	0.2393685614
chr20:60947292:["T","A"]	0.2385916388
chr20:53483842:["C","T"]	0.2383656188
chr20:38763592:["C","T"]	0.2382887183

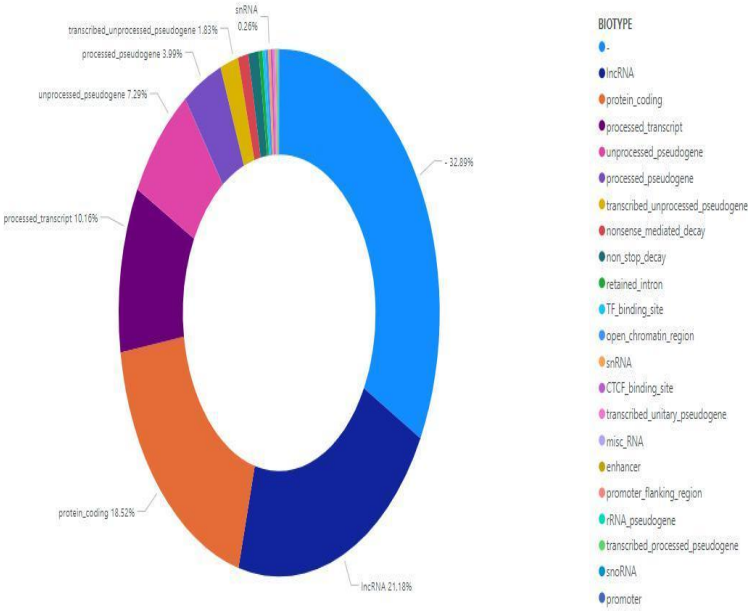


Figure 3: Biotype of Analyzed SNPs

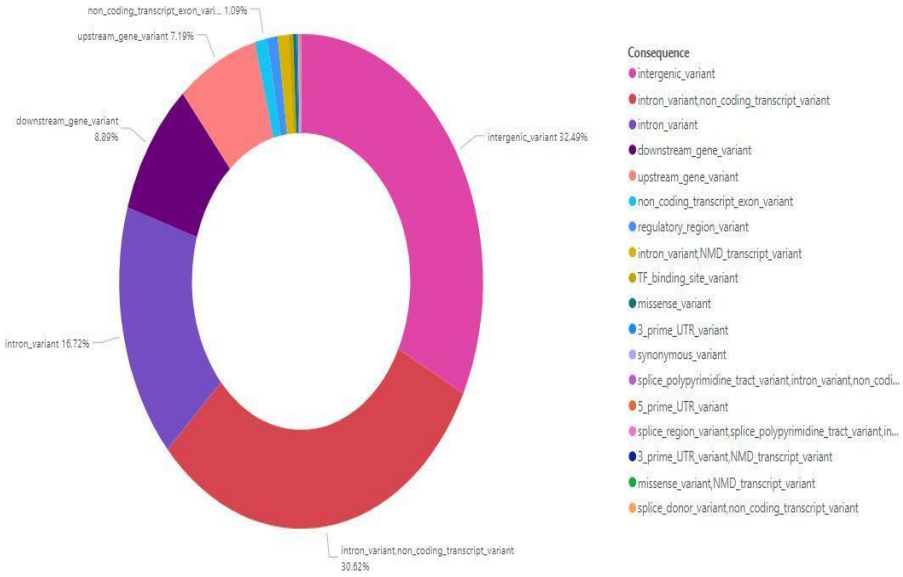


Figure 4: Effects of the Studied SNPS

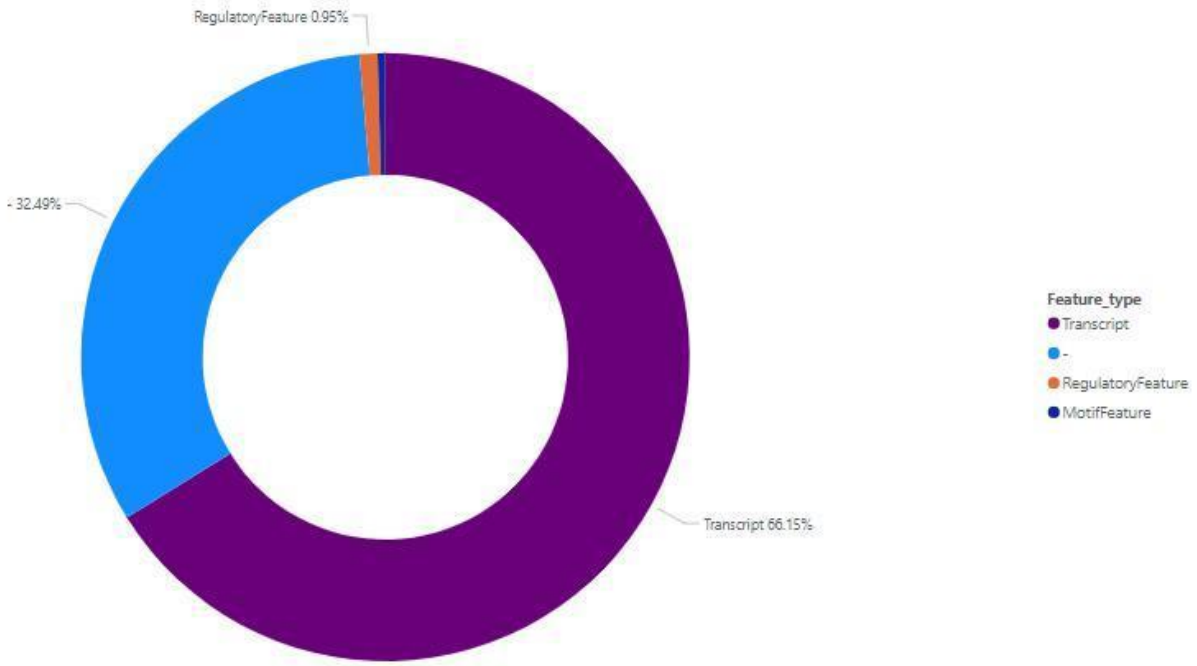


Figure 5: SNPs Analyzed Feature Type.

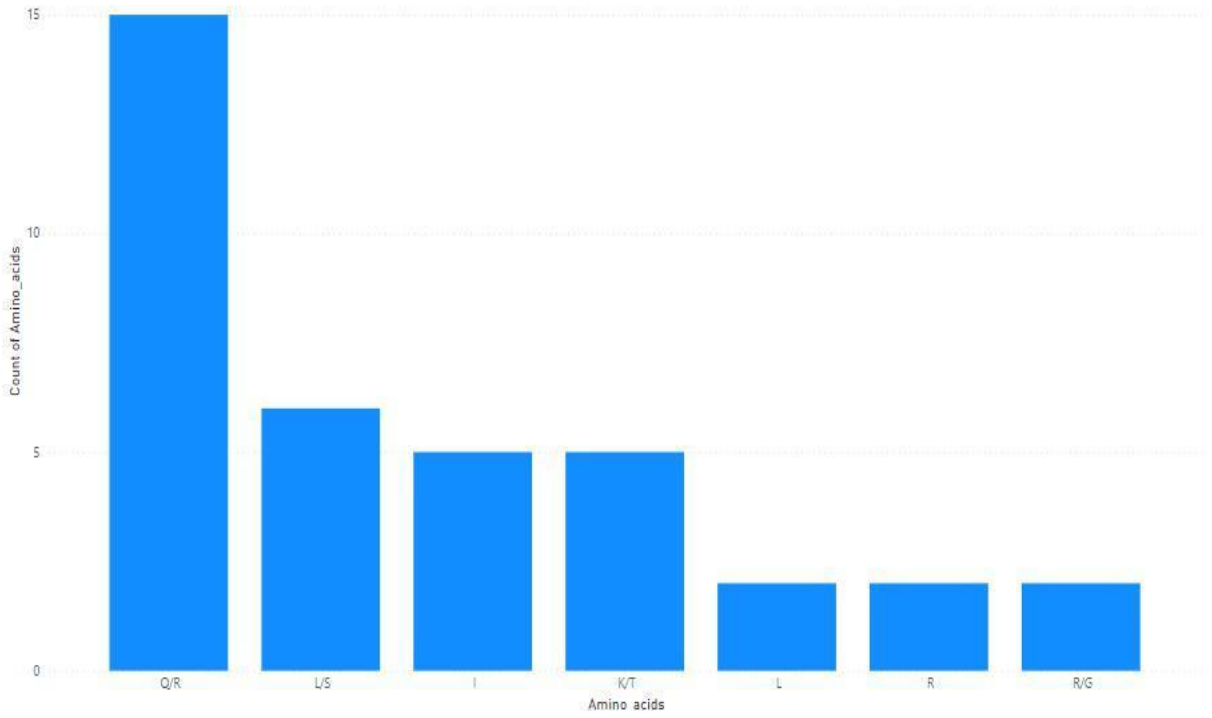


Figure 6: Amino acids of the Examined SNPs

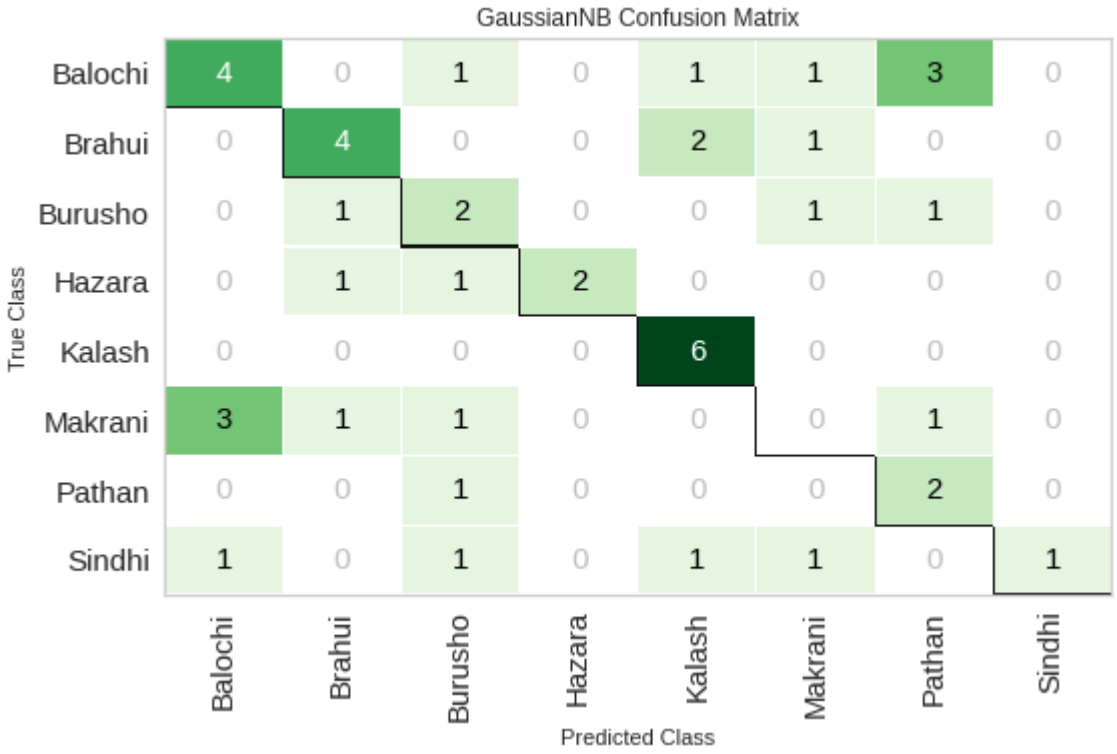


Figure 7: Confusion Matrix

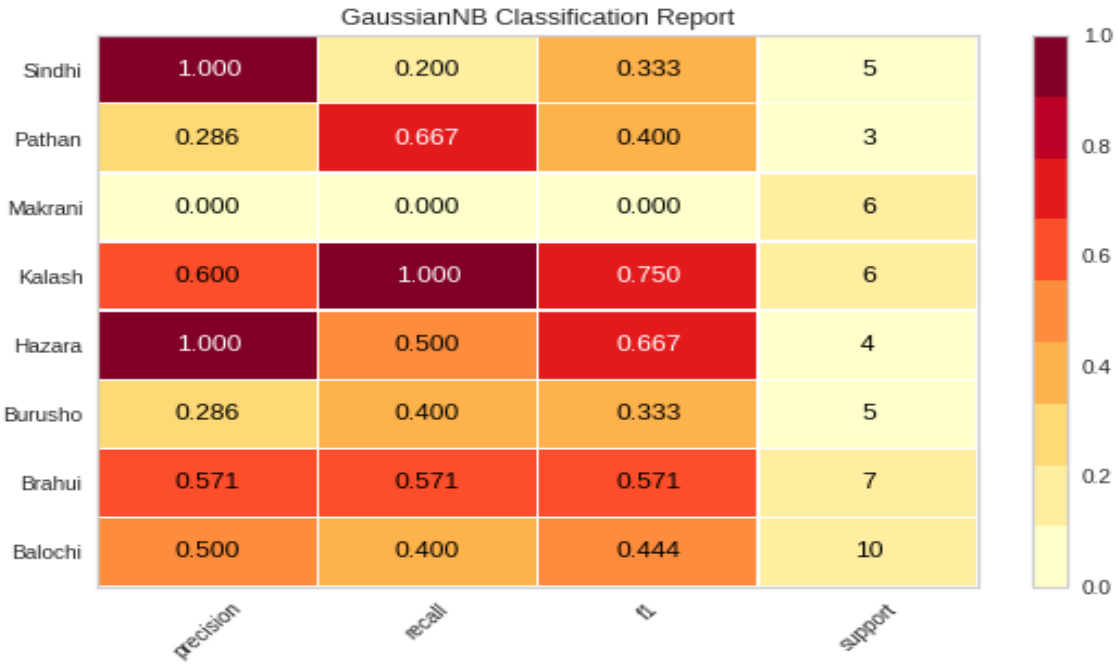


Figure 8: Accuracy, Recall, F1, and Support for the Model's Test Outcomes

REFERENCES

[1] National Library of Medicine. (2023). SNP (Single Nucleotide Polymorphism). MedlinePlus Genetics. April 11, 2023, from https://medlineplus.gov/genetics/understanding/genomicresearch/snp/

[2] Huang, T., Shu, Y., & Cai, Y.-D. (2015). Genetic differences among ethnic groups. BMC Genomics, 16. https://doi.org/10.1186/s12864-015-2328-0

[3] Kang, J. T. L., & Rosenberg, N. A. (2019). Mathematical properties of linkage disequilibrium statistics defined by normalization of the coefficient D = pAB - pApB. Human Heredity, 84(3), 127–143. https://doi.org/10.1159/000504171

[4] Calus, M. P. L., & Vandenplas, J. (2018). SNPrune: An efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. Genetics Selection Evolution, 50(1), 34. https://doi.org/10.1186/s12711-018-0404-z

[5] Vergara, J., & Estevez, P. (2014). A review of feature selection methods based on mutual information. Neural Computing and Applications, 24(1), 175-186. https://doi.org/10.1007/s00521-013-1368-0

[6] Li, L., & Khuri, S. (2000). A Comparison of DNA Fragment Assembly Algorithms. https://www.researchgate.net/publication/220278323_A_Comparison_of_DNA_Fragment_Assembly_Algorithms

[7] Ros, F., & Guillaume, S. (2019). Sampling Techniques for Supervised or Unsupervised Tasks. Springer.

[8] Maitra, S. (2020, December 25). Feature Reduction Strategy to Make Better Generalization Models. https://medium.com/swlh/feature-selection-techniques-to-make-better-generalization-models-6a19dd6dc9b1

[9] Goswami, S. (2020). Using the Chi-Squared test for feature selection with implementation: The lesser the features, the easier to interpret the model. Towards Data Science. https://towardsdatascience.com/using-the-chi-squared-test-for-feature-selection-with-implementation-99e6582aa948

[10] Brown, L., & Tsamardinos, I. (2023). Markov Blanket-Based Variable Selection in Feature Space. arXiv preprint arXiv:2302.02661.

[11] Python Software Foundation. (2020). Python Language Reference, version 3.9. Available at https://docs.python.org/3/

[12] McKinney, W., & others. (2020). Pandas: powerful Python data analysis toolkit. Zenodo. https://doi.org/10.5281/zenodo.3509134

[13] Hail Team. (2020). Hail: Scalable genetic data analysis. https://hail.is/

[14] Rocklin, M., McKinney, W., & others. (2020). Dask: Library for parallel computing in Python. Zenodo. https://doi.org/10.5281/zenodo.4046063

[15] Stenberg, D. (1997). Curl: A client-side URL transfer library. https://curl.haxx.se/

[16] Google. (2021). Google Colaboratory. https://research.google.com/colaboratory/

[17] Bokeh Development Team. (2021). Bokeh: Python library for interactive visualization. Retrieved from https://docs.bokeh.org/en/latest/index.html

[18] Python Software Foundation. (2020). Pprint: Data pretty printer. Retrieved from https://docs.python.org/3/library/pprint.html

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[20] PyCaret Team. (2021). PyCaret: An open source, low-code machine learning library in Python. Retrieved from https://pycaret.org/

[21] Smigielski, E. M., Sirotkin, K., Ward, M., & Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic acids research, 28(1), 352-355. https://pubmed.ncbi.nlm.nih.gov/10592272/

[22] Torres, L., & Hartley, R. (2019). Repositories for academic products/outputs: Latin American and Chilean visions. F1000Research, 8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6820816/