Identification of genes related to migraine

Ruthu Gandal Shankare Gowda
Clemson University
SC,29631
rgandal@clemson.edu
Contribution:
(Models: K-Means, Linear
Writing: Report, Fall-2023)

Danqi Xu
Clemson University
SC,29631
danqix@clemson.edu
Contribution:
(Models: DBSCAN, Logistic
Writing: Report, Fall-2023)

1. INTRODUCTION

1.1 Problem Statement

The identification of genes implicated in migraines represents a paramount challenge in elucidating the intricate genetic architecture of this multifaceted neurological condition. This pioneering study endeavors to meticulously uncover the specific genes or genetic markers, characterized by Single Nucleotide Polymorphisms (SNPs), exhibiting the most conspicuous association with either the susceptibility to or prevalence of migraines. This comprehensive investigation scrutinizes a vast dataset, encompassing an expansive cohort comprising 102,084 migraine cases and 771,257 controls derived from five independent studies. By conducting an exhaustive and in-depth analysis of the genetic attributes and intricate statistical correlations ingrained within this extensive dataset, the primary aim of this research is to pinpoint and elucidate the pivotal genetic determinants most profoundly linked to both the onset and predisposition to migraines.

1.2 Motivation

Understanding the genetic roots of migraines holds profound implications for healthcare. Migraines affect millions globally, causing immense suffering and impacting quality of life. My team's motivation for delving into this study is deeply personal as one of our teammates battles migraines, spurring a passionate commitment to unraveling its genetic components. This firsthand experience fuels our dedication to identifying specific genes or genetic markers linked to migraines, offering hope for improved treatments and a deeper understanding of this complex neurological condition. Unraveling the genetic factors underlying this condition offers a promising avenue for personalized treatments, potentially revolutionizing how we approach migraine management. Moreover, as genetics significantly influences susceptibility, this research not only aids in early detection but also provides critical insights into the biological mechanisms at play, potentially paving the way for tailored therapies based on an individual's genetic profile, enhancing treatment efficacy, and improving patient outcomes. The extensive dataset, spanning diverse studies and encompassing a substantial number of cases and controls, presents an unparalleled opportunity to uncover intricate genetic associations that could significantly advance our understanding of migraine genetics, driving innovation in diagnosis and therapeutic interventions.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

1.3 Dataset Overview

The dataset utilized in this study comprises 8,117 entries extracted from a collective pool of 102,084 migraine cases and 771,257 control samples originating from five distinct research studies. Each entry corresponds to a specific Single Nucleotide Polymorphism (SNP) location on a chromosome, uniquely identified by an 'rs_number'. Alongside this identification, the dataset contains five attributes crucial for genetic analysis. Notably, the 'chromosome' attribute designates the chromosome number, while the 'eaf' attribute reveals the frequency of the effect allele within the study population. Additionally, 'reference_allele' and 'other_allele' delineate the positions of the respective alleles.

Moreover, this dataset incorporates 14 attributes that provide statistical insights into the relationship between these genetic markers and the occurrence of migraines. Notably, the dataset was not readily available or accessible through open sources. Dr. Nina Hugib graciously provided this valuable dataset, encompassing both migraine occurrence data and associated genetic information, facilitating the comprehensive investigation into the genetic underpinnings of migraines.

2. Summary of EDA

2.1 Unity of Analysis

The unit of analysis in a dataset refers to the individual observations that are being studied. Single nucleotide polymorphisms (SNPs), which refers to differences in human DNA sequences, each SNP was denoted by a unique 'rs_number' in our dataset.

2.2 Total Number of Observations

In our dataset there are total of 8117 observations.

2.3 Unique Observations

Footnotes Every observation is unique in our dataset as there is no duplicates. In total there are 8060 rows x 20 columns after removing NAN values.

2.4 Time Period

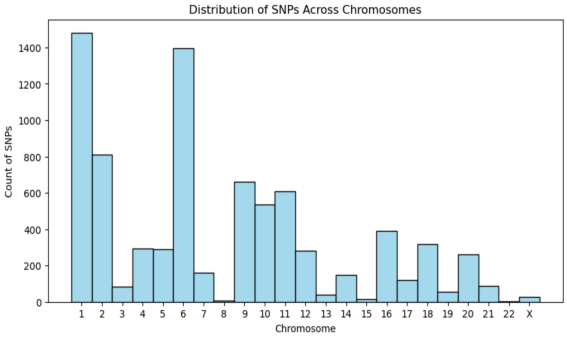
The dataset does not include information on the relevant time-period.

2.5 Data Cleaning

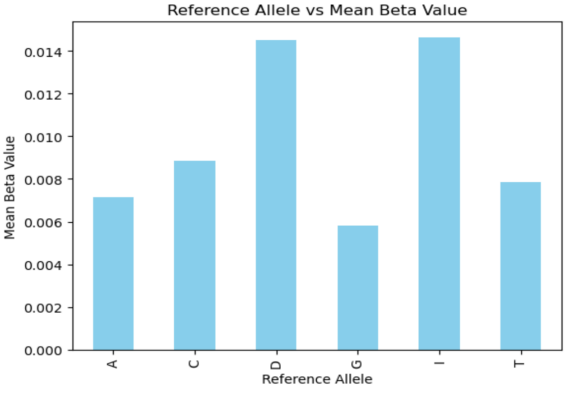
First handling the missing values checked for NaN values in the 'i2' column and dropped the rows containing 57 NaN values in that column in our dataset.

Deleting Duplicates for this first we check for any duplicate data and delete if any duplicates are present. Since there were no duplicates in our dataset, we listed unique counts. Performed range check for ‘n_studies’ and ‘n_samples’for further analysis.

2.6 Visualization of the response



The histogram depicted below showcases the distribution of SNPs (Single Nucleotide Polymorphisms) across the 23 chromosomes in Homo sapiens. Evidently, chromosomes 1 and 6 stand out with the highest counts of SNPs, indicating a denser concentration of genetic variations in these regions. Conversely, chromosomes 22 and 8 display the least number of SNPs in this representation of the dataset. This visualization underscores the uneven distribution of SNPs across the human genome, highlighting distinct variations in SNP frequency among different chromosomes.



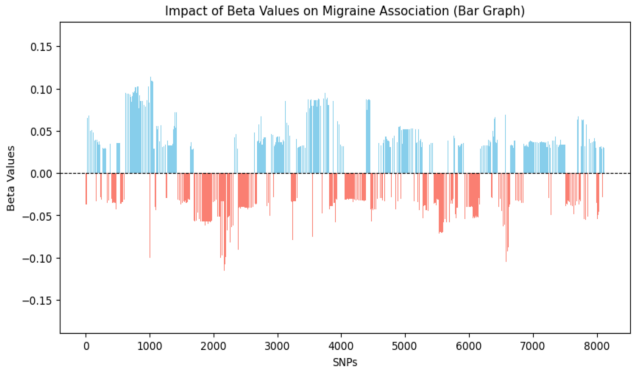
The bar graph provided illustrates the relationship between reference alleles and their corresponding mean beta values. Notably, the D and I reference alleles showcase notably higher mean beta values, hinting at a potential positive correlation with migraine. Conversely, the G reference allele stands out with the lowest mean beta value among the alleles observed, suggesting a comparatively weaker association with migraine in this dataset. This visualization underscores potential differential associations between specific reference alleles and their likelihood of being linked to migraine susceptibility or occurrence.

2.7 Visualization of Key predicators

Two pivotal predictors in our analysis are 'beta' and 'p.value'.

Beta: The 'beta' parameter serves as the effect size metric, quantifying the strength and direction of the relationship between a Single Nucleotide Polymorphism (SNP) and migraine. A positive beta value indicates a positive association, suggesting that individuals carrying this allele are more prone to experiencing migraines. Conversely, a negative beta value signifies a negative association, implying that individuals with this allele are less likely to suffer from migraines.

p.value: The 'p.value' metric signifies the statistical significance of the association between the SNP and migraine. A smaller p-value corresponds to a more statistically significant association. In the context of SNP analysis, this smaller p-value indicates stronger evidence supporting the hypothesis that the presence of the SNP is associated with the occurrence of migraines. The visual representation of p-values is depicted in the plots above, highlighting the statistical significance of these associations between SNPs and migraines.



3. Summary of Machine Learning Models

After conducting thorough Exploratory Data Analysis (EDA), we've opted to employ three distinct machine learning models those are the Linear regression, K-means clustering and DBSCAN Clustering.

3.1 Linear Regression

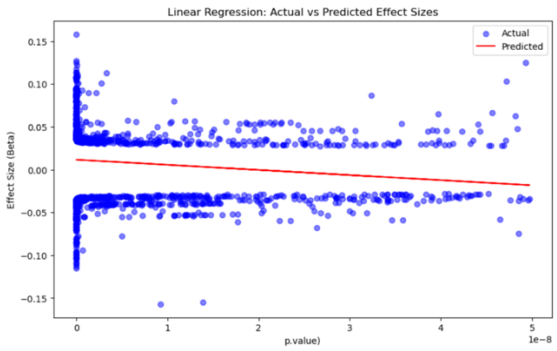
Our investigation focuses on understanding the relationship between genes and migraine susceptibility, with the continuous variable 'beta' representing the likelihood of experiencing migraines. A higher 'beta' value signifies an increased chance of

migraine occurrence, while a negative 'beta' suggests a reduced likelihood. To uncover significant predictors influencing 'beta', we conducted a linear regression analysis using the model: $Y = \beta 0 + \beta 1\ X1 + \beta 2\ X2 + \beta 3\ X3 + \varepsilon$. We specifically selected six predictors—'eaf', 'p.value', 'se', 'z', 'n_samples', and 'i2'—for inclusion in the regression model. Notably, our assumption is that 'eaf' and 'p.value' exhibit positive correlations with 'beta', indicating that higher values of 'eaf' and 'p.value' are associated with an elevated likelihood of experiencing migraines. This regression analysis aims to discern and quantify the influence of these selected predictors on the likelihood of migraine occurrences, offering insights into potential genetic factors contributing to migraine susceptibility.

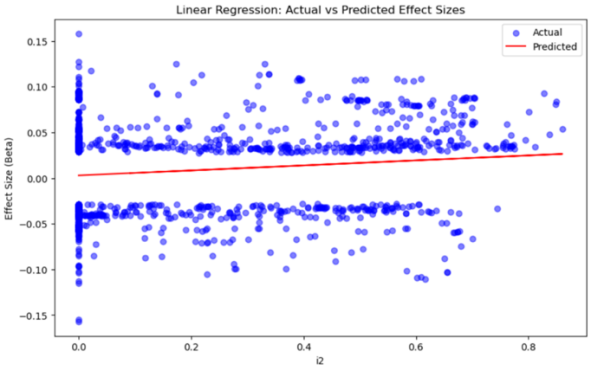
3.1.1 Case Studies

Case 1: Y = Beta, X = eaf

Case 2: Y = Beta, X = p.value



Case 3: Y = Beta, X = i2



We have considered three different case studies for this analysis shown above.

3.1.2 Model's Test Error Rate

Initially we applied the test dataset to the model to generate predictions. Subsequently, we computed key evaluation metrics on the test dataset, revealing the following results: Mean Squared Error (MSE) of 0.000204965, Root Mean Squared Error (RMSE) of 0.014316604, and an R-squared value of 0.918237233. This signifies a strong predictive capability of the model on the test dataset, showcasing its ability to explain approximately 91.82% of the variance in the target variable.

Additionally, we conducted a cross-validation procedure to further assess the model's performance. The MSE result obtained through cross-validation stood at 0.00048512291013213203. This cross-validation analysis provides a supplementary evaluation of the model's predictive accuracy, showcasing its consistency in delivering reliable predictions across different subsets of the data.

3.1.3 Model Performance

The preliminary analysis highlights promising indicators of our model's performance. The Mean Squared Error (MSE) nearing zero suggests minimal bias compared to the actual data, showcasing the model's closeness to accurate predictions. Moreover, the Root Mean Squared Error (RMSE) results signify a high level of precision in our predictive capabilities. Our Ordinary Least Squares (OLS) regression illustrates an R-squared value of 90.7% for the training dataset, closely aligning with the 91.8% R-squared attained for the test dataset. This strong R-squared value indicates that our selected predictors effectively elucidate the variance in the 'beta' variable.

Furthermore, the cross-validation outcomes validate our initial findings, reinforcing the accuracy and robustness of our model in effectively capturing the underlying patterns within the data. These collective results underscore the model's reliability and its adeptness in fitting the data, providing substantial confidence in its predictive capacity for understanding migraine susceptibility.

3.1.4 Model Performance Selected case prediction

Upon selecting three 'rs_numbers' randomly, our model's predictions for their beta values closely match the actual beta values. This alignment suggests a high level of accuracy in our model's predictive capabilities.

Case Index	Actual Beta	Predicted Beta
5682	0.039290	0.041877
6495	-0.040910	-0.056033
3033	0.036404	0.042260

3.2 K-Means Clustering

K-means clustering, a widely utilized unsupervised machine learning algorithm, proves invaluable for partitioning datasets into cohesive, non-overlapping clusters. This technique is particularly well-suited for datasets dominated by numeric attributes, precisely the case with the dataset at hand, which features essential genetic information such as position, Effect Allele Frequency (EAF), beta values, standard errors, z-scores, and p-values. The algorithm excels in grouping similar data points, unveiling inherent patterns within the data that might otherwise remain obscured. If clusters are formed, then identification of genes would be much easier.

In the context of genetic data, the K-means algorithm offers a powerful approach for the identification of genes associated with conditions like migraine. The algorithm's ability to produce clusters, each characterized by a centroid representing the mean of the data points within that cluster, facilitates the interpretation and understanding of the underlying structure of Single Nucleotide Polymorphisms (SNPs) related to migraine.

By forming distinct clusters based on the genetic attributes such as EAF, beta values, and other relevant features, K-means provides a systematic means to categorize and analyze SNPs. The resulting centroids serve as representative points, offering insights into the average genetic characteristics of each identified cluster.

3.2.1 Model's Test Error Rate

The silhouette score, ranging from -1 to 1, serves as a metric to gauge the clarity and definition of clusters in our data. A higher silhouette score implies more well-defined and distinct clusters, while lower scores suggest potential overlap or ambiguity in cluster boundaries. In essence, the silhouette score provides a quick and interpretable assessment of the quality and separability of clusters in the context of our K-means model.

3 Cases:

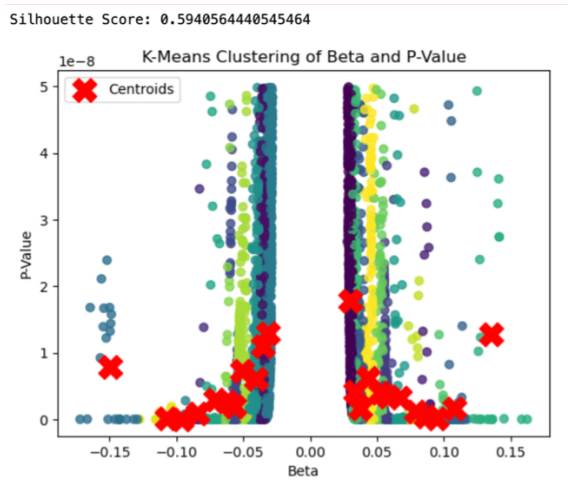
1. Silhouette Score: 0.5940564440545464(k-means clustering beta & p-values with distinct chromosome clusters)
2. Silhouette Score: 0.7168206201734997 (k-means clustering beta & p-values with distinct Reference_allele clusters)
3. Silhouette Score: 0.7168206201734997(k-means clustering beta & p-values with distinct Other_allele clusters)

3.2.2 Model Performance

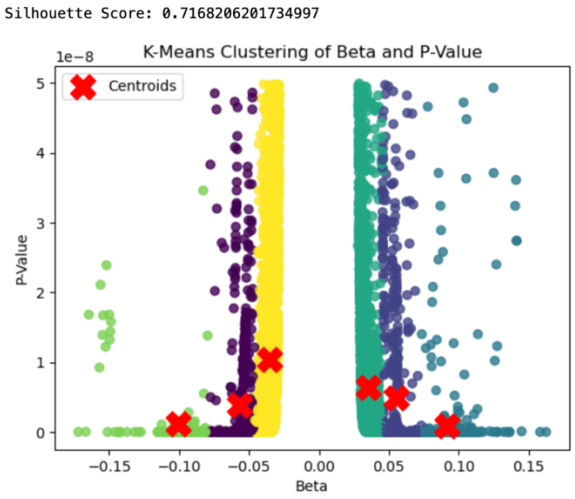
The Silhouette Score: 0.7168206201734997 (k-means clustering beta & p-values with distinct Reference_allele clusters) and Silhouette Score: 0.7168206201734997(k-means clustering beta & p-values with distinct Other_allele clusters) seems close to 1 and are well clustered compared to Silhouette Score: 0.5940564440545464(k-means clustering beta & p-values with distinct chromosome clusters) .

3.2.3 Model Performance Selected case prediction

The K-means clustering Beta & p-values with distinct count of chromosome clusters (k=23)



The K-means clustering Beta & p-values with distinct count of Reference_allele clusters (k=6)



The K-means clustering Beta & p-values with distinct count of Reference_allele clusters (k=6) obtain similar clustering graph like the above one because k=6 in both the cases.

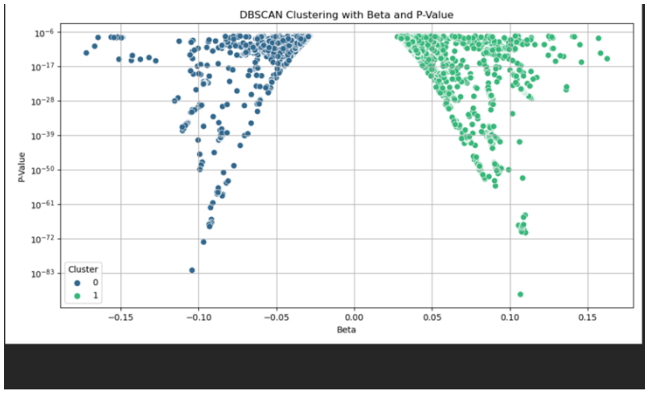
Third Silhouette Scores are relatively high (0.7168206201734997), indicating strong cluster structures with well-defined and cohesive clusters. The first score (0.5940564440545464) is still reasonable but not as strong as the other two. Overall, these scores suggest that our clustering model is performing well, and the data points are appropriately grouped into clusters with meaningful distinctions between the clusters.

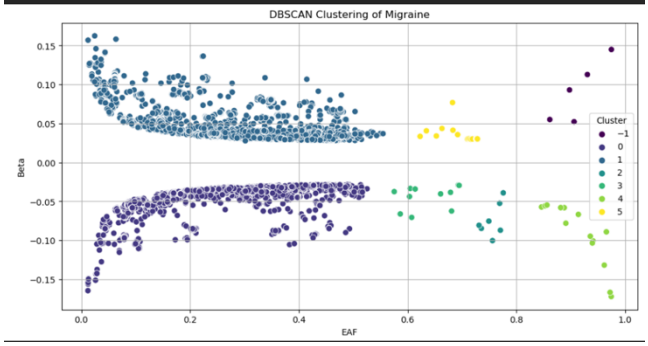
3.3 DBSCAN Clustering AND Logistic Regression

Ideally, our study would encompass individual-specific gene locus information for those with and without migraines. However, our dataset comprises genetic data across approximately 800,000 samples for each gene location. To evaluate accuracy, we've

introduced a new response variable named 'migraine_association'. Here, the 119 identified gene locations associated with migraines are labeled as '1', while the rest are marked as '0'.Overall, with the creation of the binary response variable, utilizing a logistic regression model seems promising in providing insights into gene locations that potentially elevate the risk of migraines. However, we encounter two problems. Firstly, selecting beta as the response variable and simply categorizing a beta greater than 0.05 as relevant to migraine oversimplifies the model. A single variable, beta, could not adequately define the relevance to migraine, as gene analysis is far more complex than this. Secondly, as we have too many attributes in different scale, it affects our model performance. To address this issue, we scaled our dataset to standardize it. To better visualize the dataset, we implemented DBSCAN clustering.

Below graphs represents DBSCAN Clustering with the key predicators 'Beta' and 'p.value' also DBSCAN Clustering with migraine.



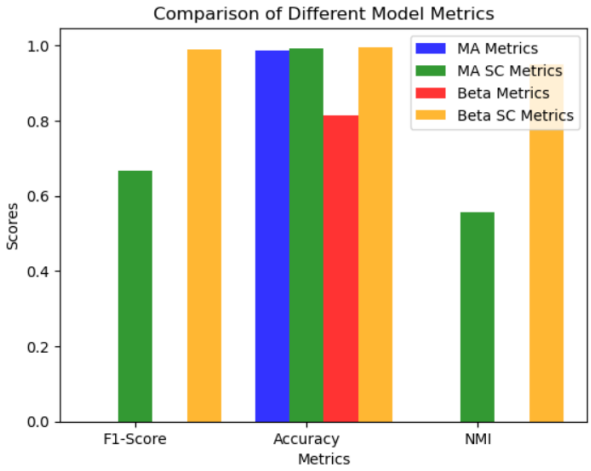


3.3.1 Model's Test Error Rate[logistic regression]

F1 score, accuracy, and NMI (Normalized Mutual Information Score) are measured here to evaluate models for 4 cases. F1 measures how well the model could predict migraine association class. Accuracy shows the portion of our model correctly predicts, though its reliability is influenced by our sample data imbalance. In addition, NMI reflects the correlation between the model's results and the actual data.

Cases	F1-Score	Accuracy	NMI
1) Beta as response variable	0.0	0.81430934	0.0
2)Beta as response variable after scaling	0.989966555	0.99627791	0.95085617
3)Migraine association as response variable	0.0	0.986765922	0
4)Migraine association as response variable after scaling	0.666666666	0.9933829611	0.5559282534

The below graph plot shows the comparisons of different model metrics.



3.3.2 Model Performance

The logistic regression model is more fit to the data after scaling. As previously mentioned, our dataset is not standardized. Consequently, the f1-score and NMI are 0 for case 1 and case 3, indicating that the model struggles to predict the migraine association class accurately. Although the accuracy rate is high, this is mainly due to the majority class of our database being the "no migraine" association class, leading to the biased outcome. However, scaling the dataset improves the performance. F1-score in case 2 demonstrates that model could correctly distinguish true positive and true negative cases and 95% NMI suggests a high correlation between model and actual data. Case 4 also performs better after scaling though not as good as case 2. Case 4 is considered more reflective of real-world scenarios as it uses identified gene locations as response variable while in Case 2,we are assuming that beta > 0.05 is migraine relevant.

3.3.3 Model Performance Selected case prediction

3 gene locations ("rs1542668", "rs2078371", "rs13078967") from identified gene locations are selected to make predictions for all cases. While Case 1 and 3 did not make any correct prediction, case 2 and 4 successfully predicted 2 of 3 are correlated with migraine, which shows that scaled our dataset do improve our model performance significantly.

3.4 Best Model

The In these four models, the scaled logistic regression model is the one that best fits our data. Firstly, we know which genes are related to migraines and which are not. This is a set of binary data, which makes it very suitable for analysis using a logistic model.

Additionally, the logistic regression model can effectively provide the predictive success rate of our model. In our tests, we observed that before scaling, our model's prediction success rate was zero. After scaling, there was a significant improvement in f1 score and NMI, and the model successfully predicted whether a gene is related to migraines or not.

4. Summary and Conclusion

4.1 Convergence of Results and Motivation

Through our project, we embarked on a journey to uncover the genetic foundations of migraines, recognizing the profound impact on global health and quality of life. Motivated by a deeply personal connection to migraine sufferers within our team, our study aimed to delve into the genetic components of this neurological condition. We harnessed an extensive dataset spanning diverse studies and a substantial number of cases and controls, leading to valuable insights.

Our analysis of a dataset encompassing nearly 800,000 individuals, including both migraine sufferers and non-sufferers, unearthed 123 gene loci potentially associated with migraines. Notably, the 'beta' variable stood out as a significant indicator, showcasing a stronger association with migraine risk when higher. Leveraging a linear model, we successfully predicted beta values from other dataset variables, demonstrating the predictive capacity of our analysis. Additionally, our logistic regression model uncovered consistent clustering patterns among these gene loci, suggesting a predictable relationship with migraine susceptibility.

In essence, our project's findings offer a glimpse into the genetic landscape of migraines, identifying specific gene loci and highlighting the 'beta' variable's importance in understanding the risk factors associated with this complex condition. This analysis provides a foundation for further research and potential avenues for targeted therapies and improved management strategies in the realm of migraine treatment.

4.2 Key Findings

Our project's findings provide a substantial contribution to domain experts in migraine genetics. The identification of 123 potential gene loci associated with migraines lays a robust foundation for focused investigations, offering a roadmap for further exploration. By highlighting the 'beta' variable's significance as a strong indicator of migraine risk, our study directs attention towards specific genetic markers deserving deeper analysis. Additionally, showcasing the effectiveness of a linear model in predicting 'beta' values underscores the utility of statistical modeling, presenting a viable approach for identifying and validating genetic markers in migraine studies. Ultimately, our insights could empower experts in conducting more targeted research, potentially leading to advanced diagnostics and personalized treatments for migraines. Notably, our emphasis on the 'beta' value and p-value reaffirms their crucial role in determining the relevance of gene locations to migraines, providing a dual confirmation of their significance in this context.

4.3 Future Works

Given more time, our extended approach would entail the collection of individualized data for each sample within two distinct groups: one comprising individuals diagnosed with migraines and the other consisting of individuals without this condition. This meticulous categorization would serve to augment our model's precision in pinpointing precise genetic attributes influencing migraine risk. By segregating the datasets based on migraine status, we aim to refine our analysis, moving beyond statistical results derived from a mixed group of migraine and non-migraine individuals. Additionally, we propose leveraging a neural network model to estimate and further enhance our understanding of the complex genetic interplay contributing to migraine susceptibility. This methodological shift would enable a more targeted and detailed investigation into the specific genetic factors underlying migraines, fostering a more nuanced comprehension of this intricate neurological condition.