2. Dataset Overview

Three datasets were analyzed to extract meaningful genetic markers for MS prediction:

2.1 epi20210205.xlsx (Epigenetic & Gene Expression Data)

- Contains **gene expression and regulatory impact data**.
- Includes data on chromatin accessibility (ATAC-seq), transcription factor binding, and differentially expressed genes (DEGs).
- Helps determine if genetic variants influence gene expression in MS patients.
- Provides insights into **epigenetic modifications** that could affect gene regulation in MS cases.

2.2 gwas-association-downloaded_2025-01-14-MONDO_0005301-withChildTraits.csv (GWAS Data)

- Contains **Genome-Wide Association Study (GWAS) data for MS and related traits**.
- Lists **SNPs (Single Nucleotide Polymorphisms) associated with MS**, their **risk alleles, odds ratios (OR), and p-values** indicating the strength of association.
- Helps identify statistically significant genetic markers strongly linked to MS development.
- Includes **trait names and SNP effect sizes**, enabling the ranking of significant variations based on risk levels.

2.3 pone.0127632.s001.csv (SNP-Level Genetic Risk Factors)

- Contains **SNPs with detailed association statistics**.
- Includes **chromosomal location, confidence intervals, risk alleles, odds ratios, and p-values**.
- Helps assess the predictive power of specific SNPs in MS risk modeling.
- Provides **additional SNP annotation information**, assisting in refining feature selection for machine learning models.

3.1 Filtering Significant SNPs

- From **GWAS Data (gwas-association.csv):**
    - Retained SNPs where **P-value < 0.05** (statistically significant).
    - Extracted relevant columns: **SNPs, Risk Alleles, OR, P-value, and Mapped Gene**.
- From **pone.0127632.s001.csv:**
    - Retained SNPs where **P-value < 0.05** and **Odds Ratio > 1.5** (strong risk association).
    - Extracted: **SNP, Risk Allele, OR, P-value**.

3.2 Merging SNPs with Functional Impact Data

- Combined filtered SNPs from **GWAS** and **pone.0127632.s001.csv**.
- Cross-checked mapped genes with **DEGs in epi20210205.xlsx**.
- Identified SNPs influencing gene expression and regulatory regions.