GROUP:19

TITLE: MACHINE LEARNING FOR GENETIC DATA ANALYSIS

Overview:
In our project we are going to develop a machine learning framework which can process a large amount of genetic data to identify patterns and mutations which are associated with specific diseases.

Brief Introduction:
Beginning with genetic marker data, this code pipeline organizes information into features and a target variable then uses a Random Forest Regressor to learn patterns. It predicts disease likelihood from genetic markers in a validation set and evaluates accuracy using Mean Squared Error, showcasing machine learning's role in disease prediction from genetic data.

GENETIC DATA ANALYSIS

- **MACHINE LEARNING IN GENETIC ANALYSIS:** IT BLENDS COMPUTER SCIENCE WITH BIOLOGY, TEACHING COMPUTERS TO SPOT PATTERNS IN VAST GENETIC DATA TO COMPREHEND HOW GENES IMPACT HEALTH.

- **DETECTIVE WORK FOR GENES:** SCIENTISTS EMPLOY THIS TECHNOLOGY TO CONNECT SPECIFIC GENES TO TRAITS OR DISEASES, ESSENTIALLY CREATING A DETECTIVE TASK TO UNDERSTAND GENETIC INFLUENCES.

- **PREDICTING HEALTH RISKS:** BY ANALYZING GENES, THIS APPROACH HELPS PREDICT THE LIKELIHOOD OF SOMEONE GETTING SICK, OFFERING INSIGHTS INTO POTENTIAL HEALTH RISKS.

- **TECHNOLOGY UNVEILING GENETIC SECRETS:** UTILIZING TECHNOLOGY TO UNCOVER THE HIDDEN INFORMATION WITHIN OUR GENES, IT EQUIPS DOCTORS TO OFFER MORE PERSONALIZED CARE BASED ON INDIVIDUALS' GENETIC PROFILES.

- **REVOLUTIONIZING PERSONALIZED MEDICINE:** ULTIMATELY, THIS TECHNOLOGY HAS THE POTENTIAL TO REVOLUTIONIZE HEALTHCARE BY ENABLING HIGHLY PERSONALIZED MEDICAL CARE TAILORED TO EACH PERSON'S UNIQUE GENETIC CHARACTERISTICS.

UNDERSTANDING OUR DATA SET

- **DATASET COMPONENTS:** IT INCLUDES 'DISEASE LIKELIHOOD' (THE CHANCE OF A DISEASE) AND GENETIC MARKERS (MARKER_1 TO MARKER_10) WITH NUMERICAL VALUES, REPRESENTING TRAITS.

- **DISEASE LIKELIHOOD DEFINITION:** THE 'DISEASE LIKELIHOOD' COLUMN SIGNIFIES THE PROBABILITY OR SUSCEPTIBILITY OF A DISEASE BASED ON GENETIC MARKERS.

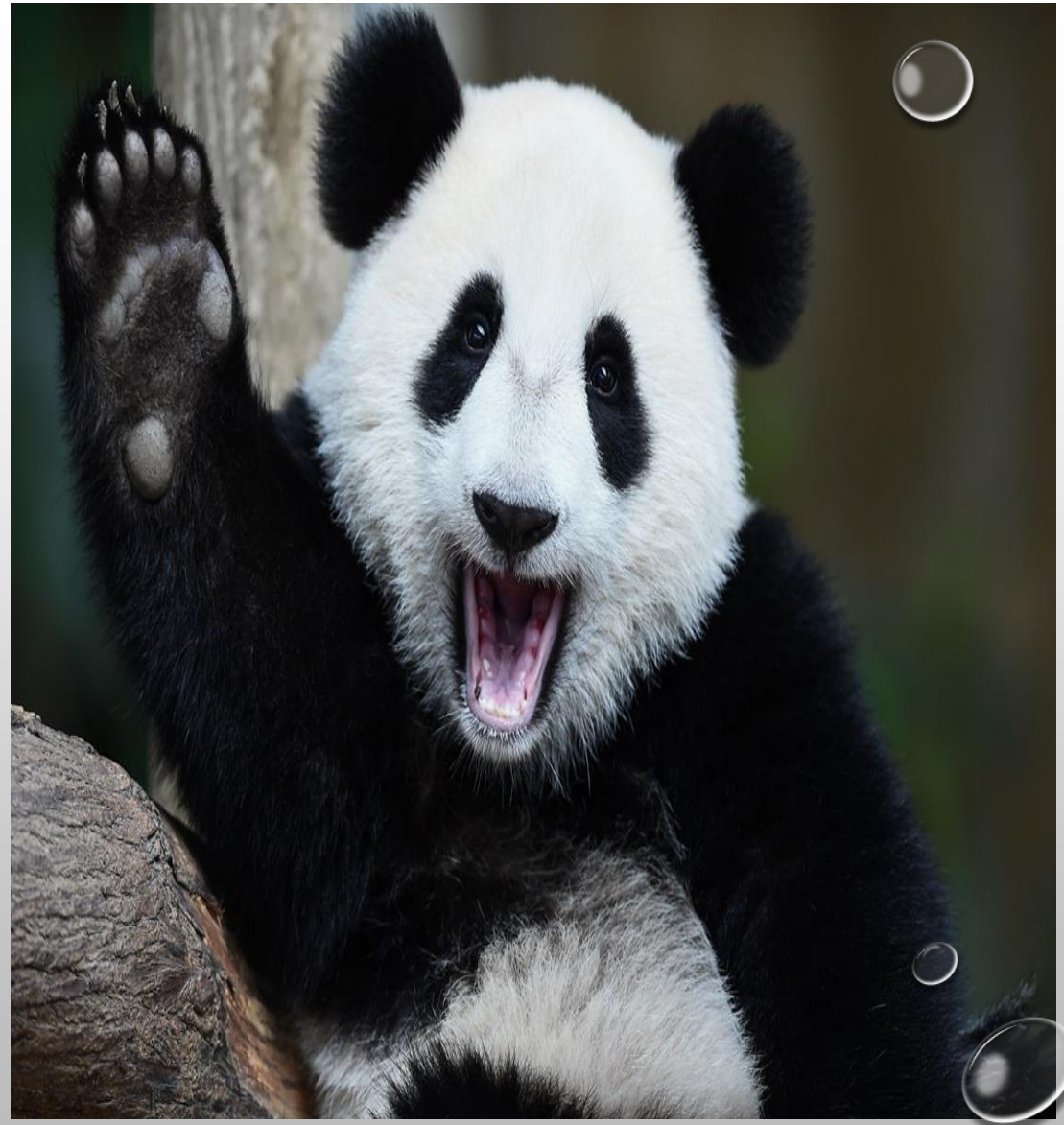
- **DATA STRUCTURE:** ABOUT 300 ROWS REPRESENT INDIVIDUALS, WITH COLUMNS (MARKERS) HAVING VALUES RANGING FROM 0 TO 1, INDICATING ACTIVITY LEVELS OF THESE GENETIC MARKERS.

- **MARKERS' ROLE:** THESE GENETIC MARKERS INDICATE SPECIFIC TRAITS, AND WE'RE INVESTIGATING HOW THEIR VALUES LINK TO THE LIKELIHOOD OF THE DISEASE.

- **DETERMINING MARKER IMPORTANCE:** EXPLORING THE CONNECTION BETWEEN THESE MARKERS AND THE DISEASE LIKELIHOOD HELPS IDENTIFY WHICH MARKERS ARE CRUCIAL IN PREDICTING THE DISEASE.

STEP BY STEP PROCESS

- **IMPORTING NECESSARY LIBRARIES:**
- **PANDAS:** IT'S AN OPEN-SOURCE LIBRARY BUILT ON TOP OF THE PYTHON PROGRAMMING LANGUAGE. IT PROVIDES DATA STRUCTURES AND FUNCTIONS SPECIFICALLY DESIGNED TO WORK WITH STRUCTURED AND TABULAR DATA, MAKING IT EASIER TO HANDLE AND ANALYZE.
- **UNDERSTANDING FILE PATHS**: EXPLAINING THE DIFFERENCE BETWEEN ABSOLUTE AND RELATIVE FILE PATHS, CLARIFYING HOW THEY SPECIFY LOCATIONS IN A FILE SYSTEM.
- **FILE LOADING PROCESS**: DEMONSTRATING HOW TO LOAD A CSV FILE USING PANDAS (PD.READ_CSV()), CREATING A DATAFRAME CALLED 'DATASET' THAT HOLDS THE CSV DATA FOR EASY MANIPULATION AND ANALYSIS.
- **DATASET CHECK:** PRINTING THE ENTIRE DATASET (PRINT(DATASET)) AND DISPLAYING ITS COLUMN NAMES (PRINT(DATASET.COLUMNS)) TO PROVIDE AN OVERVIEW OF THE DATA AND CONFIRM THE PRESENCE OF SPECIFIC COLUMNS, LIKE 'TARGET_COLUMN'.



MODEL TRAINING AND VALIDATION WORKFLOW

- **DATA SPLITTING:** SEGREGATING DATA INTO 'FEATURES' (X) AND 'TARGET VARIABLE' (Y), WHERE X REPRESENTS COLUMNS USED FOR PREDICTION (EXCLUDING 'DISEASE_LIKELIHOOD') AND Y HOLDS THE VALUES TO PREDICT ('DISEASE_LIKELIHOOD').
- **TRAIN-TEST SPLIT:** USING TRAIN_TEST_SPLIT FUNCTION TO DIVIDE DATA INTO TRAINING (70%) AND VALIDATION (30%) SETS, ENSURING THE MODEL LEARNS FROM ONE SET AND VALIDATES ON ANOTHER, UNSEEN SET.
- **MODEL INITIALIZATION:** INITIALIZING A RANDOM FOREST REGRESSOR MODEL USING RANDOMFORESTREGRESSOR() FROM THE SCIKIT-LEARN LIBRARY FOR PREDICTIVE ANALYSIS.
- **MODEL TRAINING:** TRAINING THE MODEL USING MODEL.FIT(X_TRAIN, Y_TRAIN), WHERE THE MODEL LEARNS PATTERNS AND RELATIONSHIPS BETWEEN FEATURES (X_TRAIN) AND THE TARGET VARIABLE (Y_TRAIN).
- **MODEL EVALUATION:** EMPLOYING MEAN SQUARED ERROR (MSE) VIA MEAN_SQUARED_ERROR() TO EVALUATE MODEL PERFORMANCE, CALCULATING THE AVERAGE SQUARED DIFFERENCE BETWEEN PREDICTED AND ACTUAL VALUES ON THE VALIDATION SET.

MEAN_SQUARED_ERROR

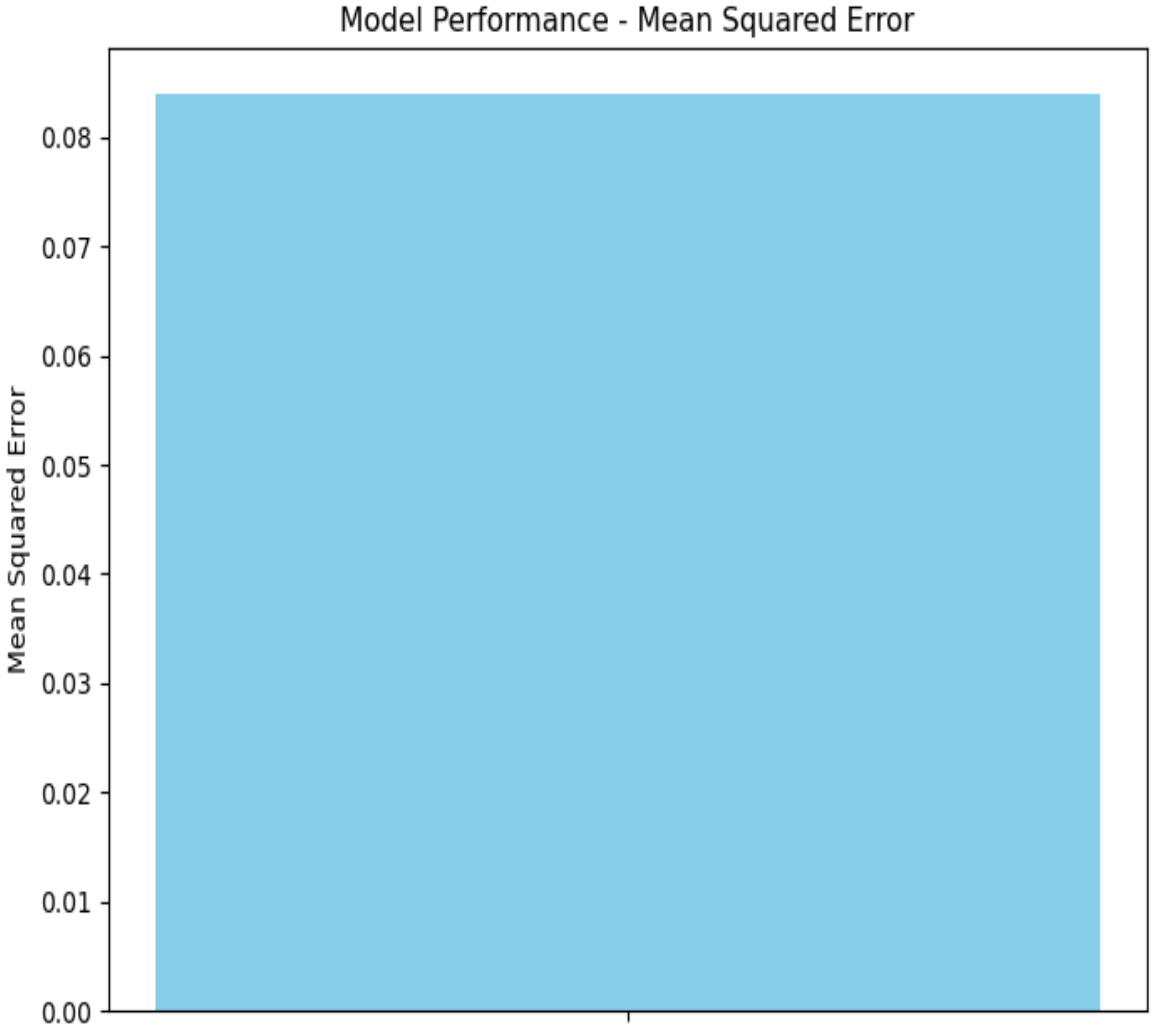
MSE(mean_squared_error):

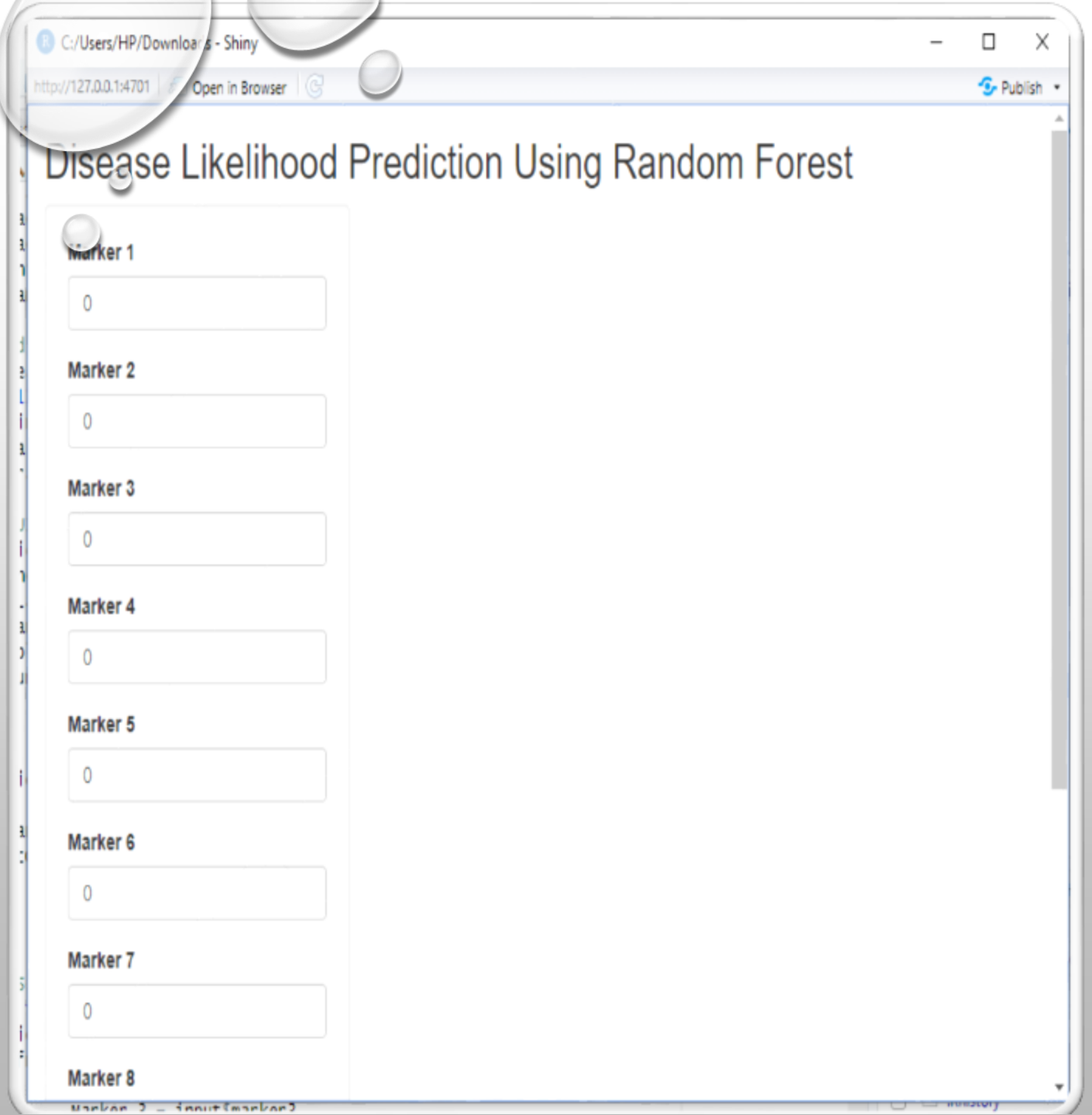
For me the simple understanding of mse is the squares of the difference between the predicted valuea and the actual values. It is a common metric used in regression problems to assess how well the model prediction align with the true value

Mean Squared Error on Validation Set:

0.08719870332954016

A mean squared error (MSE) of 0.0872 suggests that, on average, the squared difference between the predicted values generated by the model and the actual target values in the validation set is approximately 0.0872. In regression tasks, a lower MSE indicates that the model's predictions are closer to the actual values, which generally signifies better performance





DEPLOYMENT:

WE CREATED SIMPLE SHINY WEB APPLICATION THAT PREDICTS DISEASE LIKELIHOOD USING A RANDOM FOREST MODEL BASED ON GENETIC MARKERS.

- **MODEL CONSTRUCTION:** UTILIZES 'SHINY' & 'RANDOMFOREST' FOR WEB-BASED DISEASE PREDICTION VIA GENETIC MARKERS.
- **USER INTERFACE DESIGN:** CREATES AN INTUITIVE INTERFACE WITH 10 MARKER INPUT FIELDS AND A PREDICTION BUTTON.
- **SERVER LOGIC SETUP:** IMPLEMENTS REACTIVE FUNCTIONS FOR USER INPUTS & 'OBSERVEEVENT' TO TRIGGER PREDICTIONS.
- **APPLICATION EXECUTION:** INTEGRATES UI AND LOGIC VIA 'SHINYAPP' FOR REAL-TIME PREDICTIONS.
- **PREDICTIVE FUNCTIONALITY:** GENERATES DISEASE LIKELIHOOD PREDICTIONS USING USER-PROVIDED GENETIC DATA.

THANK YOU

BEST
GROUP 19 :
KASHAPPA OMKAR JADHAV
LIKITH SATYA SAI MANIKANTA