**Automated Disease Prediction Using Machine Learning Ensembles**

**Phase 2: Data Preprocessing and Model Design**

**2.1 Overview of Data Preprocessing [Radha]**

In the first part of the study, the data collection was conducted in Phase 1, the initial data exploration was performed. The second phase begins in Phase 2 to get ready the dataset to train the ensemble machine learning models. This phase also makes sure that the dataset is optimized for the models such as Random Forest, Gradient Boosting and Logistic Regression. The following are the general goals that have been addressed in this phase: these are to deal with missing values, deal with outliers and perform feature transformations such as scaling, encoding why and data dimension preparation reduction. This is very important in order to get correct and reliable results in the disease prediction process.

**2. 2 Data Cleaning [Sumeet U Pattan]**

The following are the issues that have been handled in this section: Missing Values, Outliers and Inconsistencies. It is vital to clean the dataset in order to make sure that the data in the dataset is correct and in a condition that can be used for model training. This phase addresses the following issues:

* **Missing Values:**

The missing values can affect the quality of the predictions or even cause some errors in the model. The following strategies were employed:

* **Numerical Features:** Features such as BMI and glucose had missing values which were replaced by the median since the data was skewed.
* **Categorical Features:** The missing values in gender or other categorical columns were filled with the mode (most frequent value) so as to make the data consistent.
* **Outliers:**

Outliers in features such as glucose or BMI may affect predictions especially where models are likely to be affected by outlier values. The following strategies were used in dealing with outliers:

* **Capping**: The model was capped at a certain level to ensure that none of the values was allowed to be beyond the specified limit.
* **Removal:** Data with extremely high or low values which are considered as outliers were deleted after inspection.
* **Inconsistencies:**

This section also dealt with other problems such as duplicate records or records that may be inconsistent such as records with different ages and BMI. Missing or duplicate rows were removed and if there were logical inconsistencies in the data it was edited.

**2. 3 Feature Scaling and Normalization [Shreya Kulkarni]**

Features are scaled and normalized to make sure that all the numerical data contains values of the same range or scale and this is something that is crucial for models such as Gradient Boosting and Logistic Regression.

* **Standardization:**

Some of the continuous variables like glucose and BMI were normalized using Z-score which means that the mean value was set to zero and the standard deviation to one. This is very important for models that are based on distance for example or gradient based optimization.

* **Normalization:**

For the features that had very skewed distribution for example the glucose levels in the blood, Min-Max scaling was done to bring the data into a fixed range of [0, 1]. This is to ensure that the training is smooth and takes less time to converge to a solution.

* **Categorical Features:**

The categorical variables such as gender were one-hot encoded which means that for each category a new column was created. For instance, “Male” and “Female” were encoded as two different binary columns, that way there was no bias of one column being greater than the other.

**2. 4 Feature Transformation and Reduction [Shreya Kulkarni and Sumeet U Pattan]**

Feature selection and feature reduction are required in order to enhance the performance of the ensemble models and to remove noise and irrelevant features.

* **Feature Transformation:**
* Some features were transformed using a log because the data was highly skewed such as the feature glucose.
* Second, the interaction terms were included to analyse the relationship between major variables like age and BMI.
* **Dimensionality Reduction:**
* **Correlation Analysis:** Features which were highly correlated were found out and one of the features was removed and replaced by the other to avoid the dependence.
* **Variance Thresholding:** Features which had low variance (i.e. Principal Component Analysis little (PCA): or First no of contribution all, towards PCA of predicting was features the used while output) in keeping were order the removed to variability reduce of the number dataset. For instance, the 15 initial features were decreased to 10 PCAs, which make the computation faster without negatively impacting the predictive capacity of the model.

**2. 5 Ensemble Model Design [Vasavi C Kulkarni]**

After preprocessing the data, the ensemble models were constructed in order to combine several machine learningalgorithms:

* **Random Forest Classifier:**

It models decision trees and it is less likely to over-fit than other classifiers. It is stable even with noisy data and can deal with non-linear relationships.

* **Gradient Boosting Classifier:**

A forward progressive technique of ensembling where each model attempts to reduce the errors made by the previous model. It produces good results in the presence of multiple features as it is able to identify complex relationships.

* **Logistic Regression:**

A linear model was also included in the ensemble to act as a reference point and to offer a simpler model for comparison. It does not contrast with the other models in the ensemble in any way.

These models were combined using a majority voting mechanism in which the final prediction is based on the combined outcome of all the three models.

**2. 6 Model Training and Validation [Radha]**

The models were trained and the validation was done with the use of the pre-processed dataset. Key steps included:

* **Data Splitting:**

The dataset was split into 80 % training and 20 % testing data sets to determine the performance of the model.

The hyperparameters of each training model were adjusted. Each model was to trained in order to increase the accuracy of training data:

* Random Forest: Number of trees and the maximum number of splits.
* Gradient Boosting: The learning rate and the number of estimators.
* Logistic Regression: The amount of regularization used.
* **Validation Metrics:**

The performances of the models were assessed using the overall correct classification measure, the precision, recall and the F1 measure. For example:

* Random Forest gave an accuracy of 91%.
* Gradient Boosting gave an accuracy of 89%.
* Logistic Regression gave an accuracy of 86%.
* **Ensemble Voting:**

The final predictions were reached with the help of the majority voting in order to make a more stable decision even if one model provided a worse result.

**2.7 Conclusion of Phase 2 [Vasavi C Kulkarni]**

Phase 2 involved in the development of the dataset as well as the design of the ensemble machine learning models for disease prediction. This phase also dealt with the missing values, outliers and inconsistencies that had to be dealt with in order to make the data suitable for modelling in Random Forest, Gradient Boosting and Logistic Regression models. Some of the techniques that were used to reduce the dimensionality included Random Forest, gradient boosting, principal component analysis and Logistic Regression which also increased models computational efficiency.