STEP-BY-STEP GUIDE FOR

DIABETES PREDICTION

AND CLASSIFICATION

By surafel asfawosen

PART 1: PREDICTING GLYHB PERCENTAGES

DATA PRE-PROCESSING

we have imported some important libraries like:

- ✓pandas
- ✓numpy
- ✓train_test_split
- ✓linear regression
- ✓pipeline
- ✓ordinal encoder e.t.c

CONT...

- We have read the data from the given url and save it on the local disk
- load the saved data into memory using phython such as:pandas

cont...

- **Next to this we checked the presence of**
  - ✓ duplication
  - ✓check data type
  - ✓summary stastics
  - ✓ missing value

cont…

- we handled the missing value
- By using heatmap
  - ✓first we droped bp.2s and bp.2d because the had 262 missing values.
  - ✓and then we imput the missing values of the numerical columns like weight,height etc with their mean values for each columns.

cont…

- By using feature correlation matrix we identified the features that expline our target variable(glyhb).
  - ✓**stab.glu'**
  - ✓**'ratio'**
  - ✓**'age**

Feature Scaling: Standardization

- Standardization is a technique to scale the features of a dataset so that they have a mean of 0 and a standard deviation of 1. which helps improve model performance, especially when features have different scales.
- we have used standard scaling for the feature standardization

cont...

- we split the data into train and test data sets by using train_test_split library with 80% of the data given to the training data set and the rest 20% data given for the test data set

Training the Linear Regression Model

- Train a Linear Regression model to predict the target variable using the scaled training data.
- we initialize the model by importing the linearreggression library
- The model learns the relationship between the features and the target, allowing it to make predictions.

cont...

- Use the trained Linear Regression model to make predictions on the scaled test data.
- The model outputs predicted values for the target variable based on the test data, which can be compared to the actual values for evaluation.

Model Evaluation

- Evaluate the performance of the trained Linear Regression model using key metrics,like :
    - Mean Squared Error (MSE)
    - R-squared (R^2)
- These metrics provide insights into the model's accuracy and predictive power.

Displaying Model Coefficients

- Examine the coefficients and intercept of the Linear Regression model to understand the relationship between the features and the target variable.
- The coefficients help interpret the model's behavior and show how each feature contributes to the prediction.

cont...

□**Interpretation of the Histogram**:

- **Histogram**:
  - ✓The histogram shows the frequency distribution of the "stab.glu" values.
  - ✓It tells you how the data is distributed (whether it is skewed, bimodal, etc.) and gives insight into its spread.

cont…

❑VIF Interpretation:

- ✓ It is used to detect multicollinearity in regression models, which occurs when independent variables are highly correlated with each other.
- ✓ High multicollinearity can make the estimation of coefficients unreliable and inflate the standard errors.

➢**Rule of thumb:**

- ✓VIF = 1: No multicollinearity.
- ✓1 < VIF < 5: Moderate correlation with other variables.
- ✓VIF ≥ 5 or 10: High multicollinearity. You may want to consider removing or combining features with high VIF values.

cont…

❑ Boxplot Interpretation:

- **A boxplot** is useful for visualizing the distribution of a dataset.

  ✓ **Distribution**: You can see how each feature is distributed (whether it's skewed, spread out, etc.).

  ✓ **Outliers**: Any outliers in the data will be visually represented.

  ✓ **Spread and Central Tendency**: You can understand the spread (range) and the central tendency (median) for each of the features.

PART 2: CLASSIFICATION MODELS FOR DIABETES

❖here we have done different steps to classify the model for diabets those are:-

- Step 1: Data Preprocessing
  - ✓Convert the diabetes outcome to binary classification (0 or 1).
  - ✓Handle missing values and scale numeric features.
  - ✓Encode categorical variables if necessary.

cont...

- Step 2: Train-Test Split
  - ✓Split data into training (80%) and testing (20%) sets.

cont...

- Step 3: Train and Evaluate Classification Models

➢Train the following classification models:

✓Logistic Regression

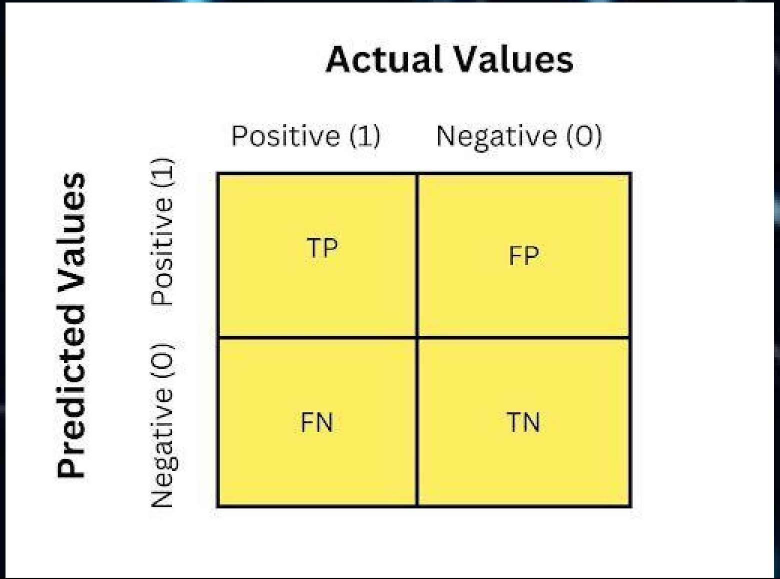
cont...

- **Evaluate models using**:
  - ❑Accuracy
    - ✓accuracy_score(y_test, y_pred): This calculates how many predictions match the actual labels in the test set.
    - ✓ It provides an overall performance measure of the model. For example, if the accuracy is 0.90, this means the model correctly predicted 90% of the test data.

cont...

❑confusion_matrix(y_test, y_pred):

- ✓ This shows a 2x2 matrix (for binary classification) or larger matrix for multiclass classification.
- ✓ It compares the predicted values (y_pred) with the true values (y_test) in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).
- ✓This is helpful for understanding where the model makes mistakes and which types of errors are more frequent.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

cont…

- **classification_report(y_test, y_pred):**
  - This gives more granular metrics like:
    - ✓ **Precision**: The fraction of relevant instances among the retrieved instances (i.e., how many of the positive predictions were actually correct).
    - ✓ **Recall**: The fraction of relevant instances that were retrieved (i.e., how many of the actual positive cases were correctly identified).
    - ✓ **F1-score**: The harmonic mean of precision and recall, balancing both metrics.
    - ✓ **Support**: The number of occurrences of each class in the test set.

cont…

❑Interpretation of Partial Residual Plots:

➢**Linearity Check:** The partial residual plots help assess whether the relationship between each feature and the target variable (log-odds) is linear. Logistic regression assumes a linear relationship between the predictors and the log-odds. If the plot shows a linear pattern (or a smoothed curve), this assumption holds. If the plot shows a non-linear pattern, it might indicate that the feature transformation is needed.

➢**Outliers**: Observations with partial residuals far from zero (either positive or negative) may be outliers and might need further investigation.

CHEAK FOR OUTOCORELATION

❑The Durbin-Watson statistic

- ✓ it is a test used to detect the presence of autocorrelation (correlation of the residuals with themselves) in the residuals of a regression model, particularly in time series data.
- ▪ The Durbin-Watson statistic ranges from 0 to 4:
  - ✓ DW = 2: No autocorrelation (the residuals are uncorrelated).
  - ✓ DW < 2: Positive autocorrelation (residuals are positively correlated).
  - ✓ DW > 2: Negative autocorrelation (residuals are negatively correlated).
  - ✓ DW close to 0: Strong positive autocorrelation.
  - ✓ DW close to 4: Strong negative autocorrelation.

cont...

- Step 4: Model Interpretation
  - ✓Identify the key risk factors associated with diabetes.
  - ✓Explain how the best model makes predictions.

Part 1: Predicting Glycated Hemoglobin (glyhb)

❑ Data Preparation:

✓The dataset was cleaned by removing duplicates and handling missing values. Missing values were filled with the median for numerical columns, and the frame column was filled with the mode ('medium').

✓Features like bp.2s and bp.2d were dropped due to a high number of missing values.

✓Categorical features (gender, location, frame) were encoded using ordinal and one-hot encoding.

cont…

❑**Feature Selection and Transformation:**

- ✓Numerical features (stab.glu, ratio, age) were scaled using StandardScaler.
- ✓A correlation matrix was generated to understand the relationships between features and the target variable (glyhb).

❑Model Training and Evaluation:

- ✓A Linear Regression model was trained on the scaled features.
- ✓The model achieved a Mean Squared Error (MSE) of 1.31 and an R^2 Score of 0.737, indicating a reasonably good fit.
- ✓The coefficients of the model suggest that stab.glu (stable glucose) has the strongest positive influence on glyhb, followed by ratio and age.

cont…

❑Interpretation:

- ✓The range of the target variable (glyhb) is 13.43, indicating a wide variation in glycated hemoglobin levels.
- ✓The model explains 73.7% of the variance in glyhb, with stab.glu being the most significant predictor.

Part 2: Classifying Diabetes

❑Data Preparation:

- ✓A binary target variable (diabetes) was created based on a threshold of 6.5 for glyhb. Values above this threshold were classified as diabetic (1), and below as non-diabetic (0).
- ✓Features were scaled using StandardScaler to prepare for logistic regression.

cont…

❑Model Training and Evaluation:

- ✓A Logistic Regression model was trained on the scaled features.
- ✓The model achieved an accuracy of 96.3%, with a precision of 1.0 and a recall of 0.75.
- ✓The F1 Score was 0.857, indicating a good balance between precision and recall.
- ✓The confusion matrix showed that the model correctly classified 69 non-diabetic cases and 9 diabetic cases, with 3 false negatives.

cont…

❑Interpretation:

- ✓The Logit Regression summary revealed that stab.glu (stable glucose) is the most significant predictor of diabetes, with a positive coefficient (0.0374) and a low p-value ($p < 0.001$).
- ✓Other features like chol, hdl, ratio, and age were not statistically significant in predicting diabetes.
- ✓The Durbin-Watson Statistic of 2.056 suggests no significant autocorrelation in the residuals, indicating a well-specified model.

cont…

❑Visual Insights:

- ✓Partial residual plots for continuous variables (age, stab.glu, chol, bp.1s, waist, hip) were generated to assess the linearity of the log-odds relationship.
- ✓Boxplots of key features (chol, stab.glu, bp.1s, waist, hip) were created to visualize their distributions and identify potential outliers.

Key Takeaways:

❑**For Predicting Glycated Hemoglobin (glyhb):**

- ✓The linear regression model performs well, with stab.glu being the most influential feature. The model explains a significant portion of the variance in glyhb, making it useful for predicting glycated hemoglobin levels.

cont…

❑For Classifying Diabetes:

- ✓The logistic regression model achieves high accuracy and precision, with stab.glu being the most critical predictor. The model is effective in identifying non-diabetic cases but has some limitations in correctly classifying diabetic cases (as indicated by the recall score).

cont…

❑**General Insights:**

- ✓Stable Glucose (stab.glu) is the most important feature in both predicting glyhb and classifying diabetes.

- ✓The dataset has some multicollinearity issues (e.g., high VIF for ratio, weight, and hip), which could be addressed in future iterations to improve model performance.

- ✓The models are robust and well-specified, as indicated by the Durbin-Watson statistic and residual analysis.

cont...

✓This analysis provides a strong foundation for predicting glycated hemoglobin levels and classifying diabetes, with potential for further refinement by addressing multicollinearity and exploring additional features or models.