```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         /*****************************************
70          * Data Import and Initial Overview
71          *****************************************/
72
73         /* Import the diabetes dataset */
74         PROC IMPORT DATAFILE="/home/u64112808/sasuser.v94/Diabetes Prediction Project/diabetes.csv"
75             OUT=diabetes_data
76             DBMS=CSV
77             REPLACE;
78             GETNAMES=YES;
79         RUN;

NOTE: Unable to open parameter catalog: SASUSER.PARMS.PARMS.SLIST in update mode. Temporary parameter values will be saved to
WORK.PARMS.PARMS.SLIST.
80          /*************************************************************************
81          *   PRODUCT:   SAS
82          *   VERSION:   9.4
83          *   CREATOR:   External File Interface
84          *   DATE:      23DEC24
85          *   DESC:      Generated SAS Datastep Code
86          *   TEMPLATE SOURCE:  (None Specified.)
87          *************************************************************************/
88             data WORK.DIABETES_DATA    ;
89             %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
90             infile '/home/u64112808/sasuser.v94/Diabetes Prediction Project/diabetes.csv' delimiter = ',' MISSOVER DSD
90       ! lrecl=32767 firstobs=2 ;
91                informat Pregnancies best32. ;
92                informat Glucose best32. ;
93                informat BloodPressure best32. ;
94                informat SkinThickness best32. ;
95                informat Insulin best32. ;
96                informat BMI best32. ;
97                informat DiabetesPedigreeFunction best32. ;
98                informat Age best32. ;
99                informat Outcome best32. ;
100               format Pregnancies best12. ;
101               format Glucose best12. ;
102               format BloodPressure best12. ;
103               format SkinThickness best12. ;
104               format Insulin best12. ;
105               format BMI best12. ;
106               format DiabetesPedigreeFunction best12. ;
107               format Age best12. ;
108               format Outcome best12. ;
109            input
110                        Pregnancies
111                        Glucose
112                        BloodPressure
113                        SkinThickness
114                        Insulin
115                        BMI
116                        DiabetesPedigreeFunction
117                        Age
118                        Outcome
119            ;
120            if _ERROR_ then call symputx('_EFIERR_',1);  /* set ERROR detection macro variable */
121            run;

NOTE: The infile '/home/u64112808/sasuser.v94/Diabetes Prediction Project/diabetes.csv' is:
      Filename=/home/u64112808/sasuser.v94/Diabetes Prediction Project/diabetes.csv,
      Owner Name=u64112808,Group Name=oda,
      Access Permission=-rw-r--r--,
      Last Modified=23Dec2024:12:58:28,
      File Size (bytes)=23873

NOTE: 768 records were read from the infile '/home/u64112808/sasuser.v94/Diabetes Prediction Project/diabetes.csv'.
      The minimum record length was 23.
      The maximum record length was 32.
NOTE: The data set WORK.DIABETES_DATA has 768 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              9388.90k
      OS Memory           37152.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        721  Switch Count  2
      Page Faults                       0
      Page Reclaims                     74
      Page Swaps                        0
      Voluntary Context Switches        16
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264
```

```
768 rows created in WORK.DIABETES_DATA from /home/u64112808/sasuser.v94/Diabetes Prediction Project/diabetes.csv.


NOTE: WORK.DIABETES_DATA data set was successfully created.
NOTE: The data set WORK.DIABETES_DATA has 768 observations and 9 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           0.03 seconds
      user cpu time       0.02 seconds
      system cpu time     0.01 seconds
      memory              9388.90k
      OS Memory           37412.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        721  Switch Count  10
      Page Faults                       0
      Page Reclaims                     789
      Page Swaps                        0
      Voluntary Context Switches        88
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           328


122
123        /* Display dataset structure and metadata */
124        TITLE "Imported Diabetes Dataset Overview";
125        PROC CONTENTS DATA=diabetes_data;
126        RUN;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              1986.65k
      OS Memory           32944.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        722  Switch Count  0
      Page Faults                       0
      Page Reclaims                     91
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           16


127        TITLE;
128
129        /* Display metadata */
130        TITLE "Metadata of Diabetes Dataset";
131        PROC CONTENTS DATA=diabetes_data;
132        RUN;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              975.00k
      OS Memory           32944.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        723  Switch Count  0
      Page Faults                       0
      Page Reclaims                     91
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           24


133        TITLE;
134
135        /* Display the first 10 rows of the dataset */
136        TITLE "Sample of the First 10 Rows in the Dataset";
137        PROC PRINT DATA=diabetes_data(OBS=10);
138        RUN;

NOTE: There were 10 observations read from the data set WORK.DIABETES_DATA.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              679.46k
      OS Memory           32684.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        724  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        1
```

```
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           16


139       TITLE;
140
141       /****************************************
142        * Checking and Handling Missing Values
143        ****************************************/
144
145       /* Summary of missing values and basic statistics */
146       TITLE "Summary of Missing Values and Basic Statistics";
147       PROC MEANS DATA=diabetes_data N NMISS;
148       RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_DATA.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              6194.37k
      OS Memory           37824.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        725  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1344
      Page Swaps                        0
      Voluntary Context Switches        23
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0


149       TITLE;
150
151       /* Frequency distribution of the target variable */
152       TITLE "Frequency Distribution of Outcome Variable";
153       PROC FREQ DATA=diabetes_data;
154           TABLES Outcome / MISSING;
155       RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_DATA.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              822.71k
      OS Memory           32944.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        726  Switch Count  2
      Page Faults                       0
      Page Reclaims                     119
      Page Swaps                        0
      Voluntary Context Switches        16
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


156       TITLE;
157
158       /* Check for invalid zeros in key variables */
159       TITLE "Checking Missing and Invalid Values in Key Variables";
160       PROC MEANS DATA=diabetes_data N NMISS MIN MAX;
161           VAR Glucose BloodPressure SkinThickness Insulin BMI;
162       RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_DATA.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              6254.78k
      OS Memory           37824.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        727  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1344
      Page Swaps                        0
      Voluntary Context Switches        22
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           0


163       TITLE;
164
165       /* Replace biologically invalid zeros with missing values */
166       DATA diabetes_clean;
```

```
167           SET diabetes_data;
168           IF Glucose = 0 THEN Glucose = .;
169           IF BloodPressure = 0 THEN BloodPressure = .;
170           IF SkinThickness = 0 THEN SkinThickness = .;
171           IF Insulin = 0 THEN Insulin = .;
172           IF BMI = 0 THEN BMI = .;
173       RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_DATA.
NOTE: The data set WORK.DIABETES_CLEAN has 768 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              953.71k
      OS Memory           32944.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        728  Switch Count  2
      Page Faults                       0
      Page Reclaims                     106
      Page Swaps                        0
      Voluntary Context Switches        14
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


174
175       /* Validate the dataset after replacing invalid zeros */
176       TITLE "Summary After Replacing Invalid Zeros with Missing Values";
177       PROC MEANS DATA=diabetes_clean N NMISS MIN MAX;
178           VAR Glucose BloodPressure SkinThickness Insulin BMI;
179       RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_CLEAN.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.01 seconds
      system cpu time     0.01 seconds
      memory              6255.65k
      OS Memory           37824.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        729  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1344
      Page Swaps                        0
      Voluntary Context Switches        24
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           24


180       TITLE;
181
182       /* Distribution of missing values after cleaning */
183       TITLE "Checking Distribution of Missing Values After Cleaning";
184       PROC MEANS DATA=diabetes_clean N NMISS;
185       RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_CLEAN.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.01 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              6098.40k
      OS Memory           37824.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        730  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1344
      Page Swaps                        0
      Voluntary Context Switches        22
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           0


186       TITLE;
187
188       /*****************************************
189        * Imputation of Missing Values
190        *****************************************/
191
192       /* Calculate means for missing value imputation */
193       PROC MEANS DATA=diabetes_clean NOPRINT;
194           VAR Glucose BloodPressure SkinThickness Insulin BMI;
195           OUTPUT OUT=mean_values
196               MEAN=Mean_Glucose Mean_BP Mean_ST Mean_Insulin Mean_BMI;
197       RUN;
```

```
NOTE: There were 768 observations read from the data set WORK.DIABETES_CLEAN.
NOTE: The data set WORK.MEAN_VALUES has 1 observations and 7 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              6440.37k
      OS Memory           38084.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        731  Switch Count  3
      Page Faults                       0
      Page Reclaims                     1406
      Page Swaps                        0
      Voluntary Context Switches        31
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


198
199        /* Impute missing values using calculated means */
200        DATA diabetes_imputed;
201            SET diabetes_clean;
202            IF _N_ = 1 THEN SET mean_values;
203            IF MISSING(Glucose) THEN Glucose = Mean_Glucose;
204            IF MISSING(BloodPressure) THEN BloodPressure = Mean_BP;
205            IF MISSING(SkinThickness) THEN SkinThickness = Mean_ST;
206            IF MISSING(Insulin) THEN Insulin = Mean_Insulin;
207            IF MISSING(BMI) THEN BMI = Mean_BMI;
208        RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_CLEAN.
NOTE: There were 1 observations read from the data set WORK.MEAN_VALUES.
NOTE: The data set WORK.DIABETES_IMPUTED has 768 observations and 16 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1301.93k
      OS Memory           33204.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        732  Switch Count  2
      Page Faults                       0
      Page Reclaims                     129
      Page Swaps                        0
      Voluntary Context Switches        14
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


209
210        /* Validate the dataset after imputation */
211        TITLE "Summary After Correcting Imputation Logic";
212        PROC MEANS DATA=diabetes_imputed N NMISS MIN MAX;
213        RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.01 seconds
      memory              6211.68k
      OS Memory           37824.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        733  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1344
      Page Swaps                        0
      Voluntary Context Switches        22
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           0


214        TITLE;
215
216        /* Display the first 10 rows of the final dataset */
217        TITLE "Final Cleaned and Preprocessed Dataset After Correct Imputation";
218        PROC PRINT DATA=diabetes_imputed(OBS=10);
219        RUN;

NOTE: There were 10 observations read from the data set WORK.DIABETES_IMPUTED.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              725.00k
      OS Memory           32684.00k
      Timestamp           12/23/2024 08:28:00 PM
```

```
      Step Count                        734  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           16


220       TITLE;
221
222       /*****************************************
223        * Descriptive Statistics
224        *****************************************/
225
226       /* Generate summary statistics for all numeric variables */
227       TITLE "Descriptive Statistics for Key Variables";
228       PROC MEANS DATA=diabetes_imputed N MEAN STD MIN MAX;
229           VAR Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age;
230       RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              6170.81k
      OS Memory           37824.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        735  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1344
      Page Swaps                        0
      Voluntary Context Switches        22
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0


231       TITLE;
232
233       /*****************************************
234        * Distribution Analysis
235        *****************************************/
236
237       /* Analyze the distribution of glucose levels */
238       TITLE "Distribution of Glucose Levels";
239       PROC SGPLOT DATA=diabetes_imputed;
240           HISTOGRAM Glucose / BINWIDTH=10;
241           DENSITY Glucose;
242           XAXIS LABEL="Glucose Level";
243           YAXIS LABEL="Frequency";
244       RUN;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.11 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              8430.84k
      OS Memory           37172.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        736  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1203
      Page Swaps                        0
      Voluntary Context Switches        172
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           792

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.

245
246       /* Analyze the distribution of BMI */
247       TITLE "Distribution of BMI";
248       PROC SGPLOT DATA=diabetes_imputed;
249           HISTOGRAM BMI / BINWIDTH=2;
250           DENSITY BMI;
251           XAXIS LABEL="Body Mass Index (BMI)";
252           YAXIS LABEL="Frequency";
253       RUN;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.08 seconds
      user cpu time       0.04 seconds
      system cpu time     0.01 seconds
      memory              2147.28k
      OS Memory           37172.00k
      Timestamp           12/23/2024 08:28:00 PM
```

```
      Step Count                        737  Switch Count  1
      Page Faults                       0
      Page Reclaims                     370
      Page Swaps                        0
      Voluntary Context Switches        166
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           472

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.

254
255        /* Analyze the distribution of age */
256        TITLE "Distribution of Age";
257        PROC SGPLOT DATA=diabetes_imputed;
258            HISTOGRAM Age / BINWIDTH=5;
259            DENSITY Age;
260            XAXIS LABEL="Age (Years)";
261            YAXIS LABEL="Frequency";
262        RUN;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.06 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              2339.03k
      OS Memory           37172.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        738  Switch Count  1
      Page Faults                       0
      Page Reclaims                     300
      Page Swaps                        0
      Voluntary Context Switches        167
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           464

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.

263        TITLE;
264
265        /*****************************************
266         * Correlation Analysis
267         *****************************************/
268
269        /* Compute correlations between key variables */
270        TITLE "Correlation Analysis of Key Variables";
271        PROC CORR DATA=diabetes_imputed PLOTS=MATRIX;
272            VAR Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age;
273        RUN;

WARNING: The scatter plot matrix with more than 5000 points has been suppressed. Use the PLOTS(MAXPOINTS= ) option in the PROC CORR
         statement to change or override the cutoff.
NOTE: PROCEDURE CORR used (Total process time):
      real time           0.06 seconds
      user cpu time       0.06 seconds
      system cpu time     0.00 seconds
      memory              1208.18k
      OS Memory           36268.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        739  Switch Count  0
      Page Faults                       0
      Page Reclaims                     52
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           24


274        TITLE;
275
276        /*****************************************
277         * Target Variable Analysis
278         *****************************************/
279
280        /* Analyze the distribution of the target variable */
281        TITLE "Distribution of Outcome (Diabetes vs. Non-Diabetes)";
282        PROC FREQ DATA=diabetes_imputed;
283            TABLES Outcome / PLOTS=FREQPLOT;
284        RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.06 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              2654.03k
      OS Memory           37432.00k
      Timestamp           12/23/2024 08:28:00 PM
```

```
      Step Count                        740  Switch Count  2
      Page Faults                       0
      Page Reclaims                     362
      Page Swaps                        0
      Voluntary Context Switches        161
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           616


285
286       /* Boxplot analysis for glucose by outcome */
287       TITLE "Boxplot of Glucose by Outcome";
288       PROC SGPLOT DATA=diabetes_imputed;
289           VBOX Glucose / CATEGORY=Outcome;
290           XAXIS LABEL="Outcome (0: No Diabetes, 1: Diabetes)";
291           YAXIS LABEL="Glucose Level";
292       RUN;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.06 seconds
      user cpu time       0.03 seconds
      system cpu time     0.01 seconds
      memory              2206.25k
      OS Memory           37172.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        741  Switch Count  1
      Page Faults                       0
      Page Reclaims                     294
      Page Swaps                        0
      Voluntary Context Switches        241
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           424

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.

293
294       /* Boxplot analysis for BMI by outcome */
295       TITLE "Boxplot of BMI by Outcome";
296       PROC SGPLOT DATA=diabetes_imputed;
297           VBOX BMI / CATEGORY=Outcome;
298           XAXIS LABEL="Outcome (0: No Diabetes, 1: Diabetes)";
299           YAXIS LABEL="BMI";
300       RUN;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.06 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              2314.53k
      OS Memory           37172.00k
      Timestamp           12/23/2024 08:28:00 PM
      Step Count                        742  Switch Count  1
      Page Faults                       0
      Page Reclaims                     293
      Page Swaps                        0
      Voluntary Context Switches        232
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           424

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.

301
302       /* Boxplot analysis for age by outcome */
303       TITLE "Boxplot of Age by Outcome";
304       PROC SGPLOT DATA=diabetes_imputed;
305           VBOX Age / CATEGORY=Outcome;
306           XAXIS LABEL="Outcome (0: No Diabetes, 1: Diabetes)";
307           YAXIS LABEL="Age (Years)";
308       RUN;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.06 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2426.65k
      OS Memory           37172.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        743  Switch Count  1
      Page Faults                       0
      Page Reclaims                     293
      Page Swaps                        0
      Voluntary Context Switches        283
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           440

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.
```

```
309        TITLE;
310
311        /*****************************************
312         * Feature Scaling
313         *****************************************/
314
315        /* Calculate mean and standard deviation for Glucose, BMI, and Age */
316        PROC MEANS DATA=diabetes_imputed NOPRINT;
317            VAR Glucose BMI Age;
318            OUTPUT OUT=stats MEAN=Mean_Glucose Mean_BMI Mean_Age
319                             STD=Std_Glucose Std_BMI Std_Age;
320        RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.
NOTE: The data set WORK.STATS has 1 observations and 8 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              6640.37k
      OS Memory           42180.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        744  Switch Count  3
      Page Faults                       0
      Page Reclaims                     1482
      Page Swaps                        0
      Voluntary Context Switches        34
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


321
322        /* Standardize Glucose, BMI, and Age using calculated means and standard deviations */
323        DATA diabetes_scaled;
324            SET diabetes_imputed;
325            IF _N_ = 1 THEN SET stats;
326            Z_Glucose = (Glucose - Mean_Glucose) / Std_Glucose;
327            Z_BMI = (BMI - Mean_BMI) / Std_BMI;
328            Z_Age = (Age - Mean_Age) / Std_Age;
329        RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_IMPUTED.
NOTE: There were 1 observations read from the data set WORK.STATS.
NOTE: The data set WORK.DIABETES_SCALED has 768 observations and 23 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              1412.15k
      OS Memory           37300.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        745  Switch Count  2
      Page Faults                       0
      Page Reclaims                     123
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           520


330
331        /* Verify the scaled variables */
332        TITLE "Summary of Scaled Features (Z_Glucose, Z_BMI, Z_Age)";
333        PROC MEANS DATA=diabetes_scaled N MEAN STD MIN MAX;
334            VAR Z_Glucose Z_BMI Z_Age;
335        RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_SCALED.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              6663.53k
      OS Memory           41920.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        746  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1382
      Page Swaps                        0
      Voluntary Context Switches        22
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


336        TITLE;
337
```

```
338         /*****************************************
339          * Feature Engineering
340          *****************************************/
341
342         /* Add interaction terms and categorize BMI */
343         DATA diabetes_engineered;
344             SET diabetes_scaled;
345
346             /* Interaction term: Glucose and BMI */
347             Interaction_Glucose_BMI = Z_Glucose * Z_BMI;
348
349             /* BMI categories */
350             IF BMI < 18.5 THEN BMI_Category = "Underweight";
351             ELSE IF BMI >= 18.5 AND BMI < 25 THEN BMI_Category = "Normal";
352             ELSE IF BMI >= 25 AND BMI < 30 THEN BMI_Category = "Overweight";
353             ELSE BMI_Category = "Obese";
354         RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_SCALED.
NOTE: The data set WORK.DIABETES_ENGINEERED has 768 observations and 25 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1223.68k
      OS Memory           36784.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        747  Switch Count  2
      Page Faults                       0
      Page Reclaims                     89
      Page Swaps                        0
      Voluntary Context Switches        13
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           520


355
356         /* Verify engineered features */
357         TITLE "Summary of Engineered Features (Interaction_Glucose_BMI and BMI_Category)";
358         PROC MEANS DATA=diabetes_engineered N MEAN STD MIN MAX;
359             VAR Interaction_Glucose_BMI;
360         RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_ENGINEERED.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              7013.96k
      OS Memory           42704.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        748  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1604
      Page Swaps                        0
      Voluntary Context Switches        23
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0


361
362         PROC FREQ DATA=diabetes_engineered;
363             TABLES BMI_Category;
364         RUN;

NOTE: There were 768 observations read from the data set WORK.DIABETES_ENGINEERED.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              970.53k
      OS Memory           36784.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        749  Switch Count  2
      Page Faults                       0
      Page Reclaims                     119
      Page Swaps                        0
      Voluntary Context Switches        21
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


365         TITLE;
366
367         /*****************************************
368          * Logistic Regression Model
369          *****************************************/
```

```
370
371         /* Logistic regression to predict Outcome (Diabetes) */
372         TITLE "Logistic Regression Model: Predicting Diabetes Outcome";
373         PROC LOGISTIC DATA=diabetes_engineered DESCENDING;
374             CLASS BMI_Category (REF="Normal"); /* Specify BMI_Category as a CLASS variable */
375             MODEL Outcome = Z_Glucose Z_BMI Z_Age Interaction_Glucose_BMI BMI_Category / SELECTION=STEPWISE;
376             OUTPUT OUT=logistic_results PREDICTED=Predicted_Prob;
377         RUN;

NOTE: PROC LOGISTIC is modeling the probability that Outcome='1'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 0.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 1.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 2.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 3.
WARNING: There is possibly a quasicomplete separation of data points in step 4. The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood
         iteration. Validity of the model fit is questionable.
NOTE: There were 768 observations read from the data set WORK.DIABETES_ENGINEERED.
NOTE: The data set WORK.LOGISTIC_RESULTS has 768 observations and 27 variables.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.13 seconds
      user cpu time       0.13 seconds
      system cpu time     0.01 seconds
      memory              3315.84k
      OS Memory           37824.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        750  Switch Count  2
      Page Faults                       0
      Page Reclaims                     227
      Page Swaps                        0
      Voluntary Context Switches        17
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           616


378         TITLE;
379
380
381         /*****************************************
382          * Confusion Matrix and Performance Metrics
383          *****************************************/
384         /* Create a binary prediction variable based on a 0.5 threshold */
385         DATA logistic_results;
386             SET logistic_results;
387             Predicted_Class = (Predicted_Prob >= 0.5); /* 1 = Diabetes, 0 = No Diabetes */
388         RUN;

NOTE: There were 768 observations read from the data set WORK.LOGISTIC_RESULTS.
NOTE: The data set WORK.LOGISTIC_RESULTS has 768 observations and 28 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              1252.34k
      OS Memory           36784.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        751  Switch Count  2
      Page Faults                       0
      Page Reclaims                     89
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           520


389
390         /* Generate confusion matrix */
391         TITLE "Confusion Matrix for Logistic Regression Predictions";
392         PROC FREQ DATA=logistic_results;
393             TABLES Outcome * Predicted_Class / CHISQ NOROW NOCOL NOPERCENT;
394         RUN;

NOTE: There were 768 observations read from the data set WORK.LOGISTIC_RESULTS.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              1468.12k
      OS Memory           37300.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        752  Switch Count  4
      Page Faults                       0
      Page Reclaims                     220
      Page Swaps                        0
      Voluntary Context Switches        27
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           528
```

```
395        TITLE;
396
397        /*****************************************
398         * Compute Performance Metrics
399         *****************************************/
400
401        /* Summarize confusion matrix values */
402        PROC SQL;
403            SELECT
404                SUM(CASE WHEN Outcome = 1 AND Predicted_Class = 1 THEN 1 ELSE 0 END) AS TP, /* True Positives */
405                SUM(CASE WHEN Outcome = 0 AND Predicted_Class = 0 THEN 1 ELSE 0 END) AS TN, /* True Negatives */
406                SUM(CASE WHEN Outcome = 0 AND Predicted_Class = 1 THEN 1 ELSE 0 END) AS FP, /* False Positives */
407                SUM(CASE WHEN Outcome = 1 AND Predicted_Class = 0 THEN 1 ELSE 0 END) AS FN /* False Negatives */
408            INTO :TP, :TN, :FP, :FN
409            FROM logistic_results;
410        QUIT;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              5796.46k
      OS Memory           41904.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        753  Switch Count  0
      Page Faults                       0
      Page Reclaims                     68
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           16


411
412        /* Calculate and display performance metrics */
413        DATA performance_metrics;
414            TP = &TP;
415            TN = &TN;
416            FP = &FP;
417            FN = &FN;
418
419            Accuracy = (TP + TN) / (TP + TN + FP + FN);
420            Precision = TP / (TP + FP);
421            Recall = TP / (TP + FN);
422            Specificity = TN / (TN + FP);
423            F1_Score = 2 * (Precision * Recall) / (Precision + Recall);
424
425            OUTPUT;
426        RUN;

NOTE: The data set WORK.PERFORMANCE_METRICS has 1 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              781.75k
      OS Memory           37036.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        754  Switch Count  2
      Page Faults                       0
      Page Reclaims                     86
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


427
428        /* Display performance metrics */
429        TITLE "Performance Metrics for Logistic Regression Model";
430        PROC PRINT DATA=performance_metrics;
431            VAR TP TN FP FN Accuracy Precision Recall Specificity F1_Score;
432            FORMAT Accuracy Precision Recall Specificity F1_Score 8.3; /* Format metrics for readability */
433        RUN;

NOTE: There were 1 observations read from the data set WORK.PERFORMANCE_METRICS.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              743.12k
      OS Memory           37036.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        755  Switch Count  0
      Page Faults                       0
      Page Reclaims                     67
      Page Swaps                        0
```

```
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


434         TITLE;
435
436         /*****************************************
437          * ROC Curve and AUC
438          *****************************************/
439
440         /* Generate ROC curve and calculate AUC */
441         TITLE "ROC Curve and AUC for Logistic Regression";
442         PROC LOGISTIC DATA=diabetes_engineered PLOTS(ONLY)=ROC;
443             CLASS BMI_Category (REF="Normal");
444             MODEL Outcome = Z_Glucose Z_BMI Z_Age Interaction_Glucose_BMI BMI_Category;
445             OUTPUT OUT=roc_results PREDICTED=Predicted_Prob;
446         RUN;

NOTE: PROC LOGISTIC is modeling the probability that Outcome='0'. One way to change this to model the probability that Outcome='1'
      is to specify the response variable option EVENT='1'.
WARNING: There is possibly a quasi-complete separation of data points. The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood
         iteration. Validity of the model fit is questionable.
NOTE: There were 768 observations read from the data set WORK.DIABETES_ENGINEERED.
NOTE: The data set WORK.ROC_RESULTS has 768 observations and 27 variables.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.16 seconds
      user cpu time       0.09 seconds
      system cpu time     0.02 seconds
      memory              4977.09k
      OS Memory           39524.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        756  Switch Count  2
      Page Faults                       0
      Page Reclaims                     540
      Page Swaps                        0
      Voluntary Context Switches        2509
      Involuntary Context Switches      4
      Block Input Operations            0
      Block Output Operations           1240


447         TITLE;
448
449
450         /*****************************************
451          * Refined Logistic Regression Model
452          *****************************************/
453
454         TITLE "Refined Logistic Regression Model: Excluding Non-Significant Interaction Term";
455         PROC LOGISTIC DATA=diabetes_engineered DESCENDING;
456             CLASS BMI_Category (REF="Normal");
457             MODEL Outcome = Z_Glucose Z_BMI Z_Age BMI_Category / SELECTION=STEPWISE;
458             OUTPUT OUT=refined_logistic_results PREDICTED=Predicted_Prob;
459         RUN;

NOTE: PROC LOGISTIC is modeling the probability that Outcome='1'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 0.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 1.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 2.
NOTE: Convergence criterion (GCONV=1E-8) satisfied in Step 3.
WARNING: There is possibly a quasicomplete separation of data points in step 4. The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood
         iteration. Validity of the model fit is questionable.
NOTE: There were 768 observations read from the data set WORK.DIABETES_ENGINEERED.
NOTE: The data set WORK.REFINED_LOGISTIC_RESULTS has 768 observations and 27 variables.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.13 seconds
      user cpu time       0.14 seconds
      system cpu time     0.00 seconds
      memory              3070.71k
      OS Memory           38592.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        757  Switch Count  2
      Page Faults                       0
      Page Reclaims                     241
      Page Swaps                        0
      Voluntary Context Switches        15
      Involuntary Context Switches      4
      Block Input Operations            0
      Block Output Operations           608


460         TITLE;
461
462         /*****************************************
463          * Performance Metrics for Refined Logistic Regression
464          *****************************************/
```

```
465
466         /* Create a binary prediction variable based on a 0.5 threshold */
467         DATA refined_logistic_results;
468             SET refined_logistic_results;
469             Predicted_Class = (Predicted_Prob >= 0.5); /* 1 = Diabetes, 0 = No Diabetes */
470         RUN;

NOTE: There were 768 observations read from the data set WORK.REFINED_LOGISTIC_RESULTS.
NOTE: The data set WORK.REFINED_LOGISTIC_RESULTS has 768 observations and 28 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1249.65k
      OS Memory           37552.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        758  Switch Count  2
      Page Faults                       0
      Page Reclaims                     89
      Page Swaps                        0
      Voluntary Context Switches        17
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           520


471
472         /* Generate confusion matrix */
473         TITLE "Confusion Matrix for Refined Logistic Regression Predictions";
474         PROC FREQ DATA=refined_logistic_results;
475             TABLES Outcome * Predicted_Class / CHISQ NOROW NOCOL NOPERCENT;
476         RUN;

NOTE: There were 768 observations read from the data set WORK.REFINED_LOGISTIC_RESULTS.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              1295.18k
      OS Memory           37812.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        759  Switch Count  4
      Page Faults                       0
      Page Reclaims                     185
      Page Swaps                        0
      Voluntary Context Switches        26
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           544


477         TITLE;
478
479         /* Summarize confusion matrix values */
480         PROC SQL;
481             SELECT
482                 SUM(CASE WHEN Outcome = 1 AND Predicted_Class = 1 THEN 1 ELSE 0 END) AS TP, /* True Positives */
483                 SUM(CASE WHEN Outcome = 0 AND Predicted_Class = 0 THEN 1 ELSE 0 END) AS TN, /* True Negatives */
484                 SUM(CASE WHEN Outcome = 0 AND Predicted_Class = 1 THEN 1 ELSE 0 END) AS FP, /* False Positives */
485                 SUM(CASE WHEN Outcome = 1 AND Predicted_Class = 0 THEN 1 ELSE 0 END) AS FN /* False Negatives */
486             INTO :TP, :TN, :FP, :FN
487             FROM refined_logistic_results;
488         QUIT;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              5908.09k
      OS Memory           42416.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        760  Switch Count  0
      Page Faults                       0
      Page Reclaims                     57
      Page Swaps                        0
      Voluntary Context Switches        4
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           8


489
490         /* Calculate and display performance metrics */
491         DATA refined_performance_metrics;
492             TP = &TP;
493             TN = &TN;
494             FP = &FP;
495             FN = &FN;
496
497             Accuracy = (TP + TN) / (TP + TN + FP + FN);
498             Precision = TP / (TP + FP);
```

```
499           Recall = TP / (TP + FN);
500           Specificity = TN / (TN + FP);
501           F1_Score = 2 * (Precision * Recall) / (Precision + Recall);
502
503           OUTPUT;
504       RUN;

NOTE: The data set WORK.REFINED_PERFORMANCE_METRICS has 1 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              781.62k
      OS Memory           37292.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        761  Switch Count  2
      Page Faults                       0
      Page Reclaims                     86
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


505
506       /*****************************************
507        * Import the Test Dataset
508        *****************************************/
509       PROC IMPORT DATAFILE="/home/u64112808/sasuser.v94/Diabetes Prediction Project/test_data.csv"
510           OUT=new_data
511           DBMS=CSV
512           REPLACE;
513           GETNAMES=YES;
514       RUN;

NOTE: Unable to open parameter catalog: SASUSER.PARMS.PARMS.SLIST in update mode. Temporary parameter values will be saved to
WORK.PARMS.PARMS.SLIST.
515        /*************************************************************************
516        *   PRODUCT:   SAS
517        *   VERSION:   9.4
518        *   CREATOR:   External File Interface
519        *   DATE:      23DEC24
520        *   DESC:      Generated SAS Datastep Code
521        *   TEMPLATE SOURCE:  (None Specified.)
522        *************************************************************************/
523           data WORK.NEW_DATA    ;
524           %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
525           infile '/home/u64112808/sasuser.v94/Diabetes Prediction Project/test_data.csv' delimiter = ',' MISSOVER DSD
525     ! lrecl=32767 firstobs=2 ;
526              informat Pregnancies best32. ;
527              informat Glucose best32. ;
528              informat BloodPressure best32. ;
529              informat SkinThickness best32. ;
530              informat Insulin best32. ;
531              informat BMI best32. ;
532              informat DiabetesPedigreeFunction best32. ;
533              informat Age best32. ;
534              informat Outcome best32. ;
535              format Pregnancies best12. ;
536              format Glucose best12. ;
537              format BloodPressure best12. ;
538              format SkinThickness best12. ;
539              format Insulin best12. ;
540              format BMI best12. ;
541              format DiabetesPedigreeFunction best12. ;
542              format Age best12. ;
543              format Outcome best12. ;
544           input
545                       Pregnancies
546                       Glucose
547                       BloodPressure
548                       SkinThickness
549                       Insulin
550                       BMI
551                       DiabetesPedigreeFunction
552                       Age
553                       Outcome
554           ;
555           if _ERROR_ then call symputx('_EFIERR_',1);  /* set ERROR detection macro variable */
556           run;

NOTE: The infile '/home/u64112808/sasuser.v94/Diabetes Prediction Project/test_data.csv' is:
      Filename=/home/u64112808/sasuser.v94/Diabetes Prediction Project/test_data.csv,
      Owner Name=u64112808,Group Name=oda,
      Access Permission=-rw-r--r--,
      Last Modified=23Dec2024:13:39:27,
      File Size (bytes)=1264

NOTE: 20 records were read from the infile '/home/u64112808/sasuser.v94/Diabetes Prediction Project/test_data.csv'.
```

```
      The minimum record length was 56.
      The maximum record length was 59.
NOTE: The data set WORK.NEW_DATA has 20 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              9279.21k
      OS Memory           42784.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        762  Switch Count  2
      Page Faults                       0
      Page Reclaims                     143
      Page Swaps                        0
      Voluntary Context Switches        18
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


20 rows created in WORK.NEW_DATA from /home/u64112808/sasuser.v94/Diabetes Prediction Project/test_data.csv.



NOTE: WORK.NEW_DATA data set was successfully created.
NOTE: The data set WORK.NEW_DATA has 20 observations and 9 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           0.03 seconds
      user cpu time       0.02 seconds
      system cpu time     0.01 seconds
      memory              9279.21k
      OS Memory           43044.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        762  Switch Count  10
      Page Faults                       0
      Page Reclaims                     1144
      Page Swaps                        0
      Voluntary Context Switches        112
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           288


557
558        /* Display the structure of the imported data */
559        TITLE "Structure of the Imported Test Dataset";
560        PROC CONTENTS DATA=new_data;
561        RUN;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              916.87k
      OS Memory           38064.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        763  Switch Count  0
      Page Faults                       0
      Page Reclaims                     92
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           24


562        TITLE;
563
564
565        /*****************************************
566         * Prepare the Test Dataset
567         *****************************************/
568        DATA new_data_processed;
569            SET new_data;
570            /* Standardize variables using actual training dataset means and standard deviations */
571        Z_Glucose = (Glucose - 121.69) / 30.44;
572        Z_BMI = (BMI - 32.46) / 6.88;
573        Z_Age = (Age - 33.24) / 11.76;
574            /* Assign BMI categories */
575            IF BMI < 18.5 THEN BMI_Category = "Underweight";
576            ELSE IF BMI >= 18.5 AND BMI < 25 THEN BMI_Category = "Normal";
577            ELSE IF BMI >= 25 AND BMI < 30 THEN BMI_Category = "Overweight";
578            ELSE BMI_Category = "Obese";
579        RUN;

NOTE: There were 20 observations read from the data set WORK.NEW_DATA.
NOTE: The data set WORK.NEW_DATA_PROCESSED has 20 observations and 13 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
```

```
      system cpu time     0.00 seconds
      memory              958.75k
      OS Memory           38064.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        764  Switch Count  2
      Page Faults                       0
      Page Reclaims                     119
      Page Swaps                        0
      Voluntary Context Switches        17
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


580
581        /* Verify the processed data */
582        TITLE "Processed Test Dataset";
583        PROC PRINT DATA=new_data_processed(OBS=10);
584        RUN;

NOTE: There were 10 observations read from the data set WORK.NEW_DATA_PROCESSED.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              707.56k
      OS Memory           37804.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        765  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


585        TITLE;
586
587        /*****************************************
588         * Score Processed Test Dataset
589         *****************************************/
590        PROC LOGISTIC INMODEL=refined_model;
591            SCORE DATA=new_data_processed OUT=new_data_predictions;
592        RUN;

NOTE: The data set WORK.NEW_DATA_PREDICTIONS has 20 observations and 17 variables.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1137.34k
      OS Memory           38324.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        766  Switch Count  2
      Page Faults                       0
      Page Reclaims                     148
      Page Swaps                        0
      Voluntary Context Switches        15
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


593
594
595        /* Check if scoring was successful */
596        TITLE "Contents of Scored Data";
597        PROC CONTENTS DATA=new_data_predictions;
598        RUN;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              936.18k
      OS Memory           38064.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        767  Switch Count  0
      Page Faults                       0
      Page Reclaims                     93
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           40


599        TITLE;
```

```
600
601         /*****************************************
602          * Add Predicted Classes
603          *****************************************/
604         DATA new_data_predictions;
605             SET new_data_predictions;
606             Predicted_Prob = P_1; /* Map predicted probability for Outcome=1 */
607             Predicted_Class = (Predicted_Prob >= 0.5); /* Binary classification: 1 = Diabetes, 0 = No Diabetes */
608         RUN;

NOTE: There were 20 observations read from the data set WORK.NEW_DATA_PREDICTIONS.
NOTE: The data set WORK.NEW_DATA_PREDICTIONS has 20 observations and 19 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              964.06k
      OS Memory           38064.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        768  Switch Count  2
      Page Faults                       0
      Page Reclaims                     118
      Page Swaps                        0
      Voluntary Context Switches        13
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


609
610
611         /*****************************************
612          * View Predictions
613          *****************************************/
614         TITLE "Predictions for Processed Test Dataset";
615         PROC PRINT DATA=new_data_predictions(OBS=10);
616             VAR Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age Outcome Predicted_Prob
616       ! Predicted_Class;
617         RUN;

NOTE: There were 10 observations read from the data set WORK.NEW_DATA_PREDICTIONS.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.02 seconds
      system cpu time     0.01 seconds
      memory              759.18k
      OS Memory           37804.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        769  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        1
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


618         TITLE;
619
620         /*****************************************
621          * Export Predictions to CSV
622          *****************************************/
623         PROC EXPORT DATA=new_data_predictions
624             OUTFILE="/home/u64112808/sasuser.v94/Diabetes Prediction Project/predicted_data.csv"
625             DBMS=CSV
626             REPLACE;
627         RUN;

NOTE: Unable to open parameter catalog: SASUSER.PARMS.PARMS.SLIST in update mode. Temporary parameter values will be saved to
WORK.PARMS.PARMS.SLIST.
628          /*************************************************************************
629          *   PRODUCT:   SAS
630          *   VERSION:   9.4
631          *   CREATOR:   External File Interface
632          *   DATE:      23DEC24
633          *   DESC:      Generated SAS Datastep Code
634          *   TEMPLATE SOURCE:  (None Specified.)
635          *************************************************************************/
636             data _null_;
637             %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
638             %let _EFIREC_ = 0;     /* clear export record count macro variable */
639             file '/home/u64112808/sasuser.v94/Diabetes Prediction Project/predicted_data.csv' delimiter=',' DSD DROPOVER
639       ! lrecl=32767;
640             if _n_ = 1 then        /* write column names or labels */
641              do;
642                put
643                   "Pregnancies"
644                ','
645                   "Glucose"
```

```
              ','
                "BloodPressure"
              ','
                "SkinThickness"
              ','
                "Insulin"
              ','
                "BMI"
              ','
                "DiabetesPedigreeFunction"
              ','
                "Age"
              ','
                "Outcome"
              ','
                "Z_Glucose"
              ','
                "Z_BMI"
              ','
                "Z_Age"
              ','
                "BMI_Category"
              ','
                "F_Outcome"
              ','
                "I_Outcome"
              ','
                "P_1"
              ','
                "P_0"
              ','
                "Predicted_Prob"
              ','
                "Predicted_Class"
              ;
            end;
          set  NEW_DATA_PREDICTIONS   end=EFIEOD;
              format Pregnancies best12. ;
              format Glucose best12. ;
              format BloodPressure best12. ;
              format SkinThickness best12. ;
              format Insulin best12. ;
              format BMI best12. ;
              format DiabetesPedigreeFunction best12. ;
              format Age best12. ;
              format Outcome best12. ;
              format Z_Glucose best12. ;
              format Z_BMI best12. ;
              format Z_Age best12. ;
              format BMI_Category $11. ;
              format F_Outcome $12. ;
              format I_Outcome $12. ;
              format P_1 best12. ;
              format P_0 best12. ;
              format Predicted_Prob best12. ;
              format Predicted_Class best12. ;
            do;
              EFIOUT + 1;
              put Pregnancies @;
              put Glucose @;
              put BloodPressure @;
              put SkinThickness @;
              put Insulin @;
              put BMI @;
              put DiabetesPedigreeFunction @;
              put Age @;
              put Outcome @;
              put Z_Glucose @;
              put Z_BMI @;
              put Z_Age @;
              put BMI_Category $ @;
              put F_Outcome $ @;
              put I_Outcome $ @;
              put P_1 @;
              put P_0 @;
              put Predicted_Prob @;
              put Predicted_Class ;
              ;
            end;
           if _ERROR_ then call symputx('_EFIERR_',1);  /* set ERROR detection macro variable */
           if EFIEOD then call symputx('_EFIREC_',EFIOUT);
           run;

NOTE: The file '/home/u64112808/sasuser.v94/Diabetes Prediction Project/predicted_data.csv' is:
      Filename=/home/u64112808/sasuser.v94/Diabetes Prediction Project/predicted_data.csv,
      Owner Name=u64112808,Group Name=oda,
      Access Permission=-rw-r--r--,
      Last Modified=23Dec2024:14:28:01

NOTE: 21 records were written to the file '/home/u64112808/sasuser.v94/Diabetes Prediction Project/predicted_data.csv'.
```

```
      The minimum record length was 127.
      The maximum record length was 190.
NOTE: There were 20 observations read from the data set WORK.NEW_DATA_PREDICTIONS.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              9309.56k
      OS Memory           42784.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        770  Switch Count  0
      Page Faults                       0
      Page Reclaims                     105
      Page Swaps                        0
      Voluntary Context Switches        9
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           16


20 records created in /home/u64112808/sasuser.v94/Diabetes Prediction Project/predicted_data.csv from NEW_DATA_PREDICTIONS.


NOTE: "/home/u64112808/sasuser.v94/Diabetes Prediction Project/predicted_data.csv" file was successfully created.
NOTE: PROCEDURE EXPORT used (Total process time):
      real time           0.03 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              9309.56k
      OS Memory           43044.00k
      Timestamp           12/23/2024 08:28:01 PM
      Step Count                        770  Switch Count  7
      Page Faults                       0
      Page Reclaims                     1055
      Page Swaps                        0
      Voluntary Context Switches        79
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           40


728
729        /* Confirmation */
730        TITLE "Predictions Exported to CSV";
731        RUN;
732
733
734        OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
744
```