CM52058 - Statistical Data Science

Diabetes Dataset Analysis

1. Load the Diabetes data

Loaded the diabetes dataset from the CSV file into a pandas DataFrame (using method 1), setting the first column as a row label. The dataset has 442 rows (442 patients) and 11 columns (10 predictor columns and 1 target column).

2. Basic dataset information

Column Name	Data Type	Description	Number of missing values
age	int64	Age in years	0
sex	int64	Sex (Female = 0, Male = 1)	
bmi	float64	Body mass index	
map	float64	Blood pressure	
tc	int64	Blood serum measurements	
ldl	float64		
hdl	float64		
tch	float64		
ltg	float64		
glu	int64		
prog	int64	Target: Disease progression one year after baseline	

Figure 1. (Tibshirani et al., 2004) Table describing basic dataset information. Note that 'prog' (disease progression) is the target variable and the other columns are predictor variables.

There are no missing values, so no data imputation or deletion of rows/columns is necessary.

	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	prog
1	59	1	32.1	101.0	157	93.2	38.0	4.0	2.110590	87	151
2	48	0	21.6	87.0	183	103.2	70.0	3.0	1.690196	69	75
3	72	1	30.5	93.0	156	93.6	41.0	4.0	2.029384	85	141
4	24	0	25.3	84.0	198	131.4	40.0	5.0	2.123852	89	206
5	50	0	23.0	101.0	192	125.4	52.0	4.0	1.863323	80	135

Figure 2. First 5 rows of dataset.

3. Descriptive statistics

	mean	std	min	max	range	median	IQR	skew	kurtosis
age	48.518100	13.109028	19.000000	79.000000	60.000000	50.000000	20.750000	-0.231382	-0.671224
sex	0.468326	0.499561	0.000000	1.000000	1.000000	0.000000	1.000000	0.127385	-1.992811
bmi	26.375792	4.418122	18.000000	42.200000	24.200000	25.700000	6.075000	0.598148	0.095094
map	94.647059	13.831204	62.000000	133.000000	71.000000	93.000000	21.000000	0.290664	-0.532780
tc	189.140271	34.608052	97.000000	301.000000	204.000000	186.000000	45.500000	0.378108	0.232948
ldl	115.439140	30.413081	41.600000	242.400000	200.800000	113.000000	38.450000	0.436592	0.601381
hdl	49.788462	12.934202	22.000000	99.000000	77.000000	48.000000	17.500000	0.799255	0.981507
tch	4.070249	1.290450	2.000000	9.090000	7.090000	4.000000	2.000000	0.735374	0.444402
ltg	2.015740	0.226872	1.414973	2.652246	1.237273	2.006461	0.312929	0.291774	-0.134366
glu	91.260181	11.496335	58.000000	124.000000	66.000000	91.000000	14.750000	0.207917	0.236917
prog	152.133484	77.093005	25.000000	346.000000	321.000000	140.500000	124.500000	0.440563	-0.883057

Figure 3. Table showing summary statistics. Computed using .describe() and other pandas methods. Note the higher positive skew of 'hdl' and 'tch' indicating a few patients with high values pulling the mean average up. The median should be used here which is robust to outliers.

Note the high standard deviation of 'hdl', 'ldl' and 'tch' relative to its mean, indicating large spread. The median is useful here since it is a measure of the average which is less affected by outliers (unlike the mean). Also, many of the columns (age etc.) contain only integers and so the median outputs integers (unlike the mean) and so gives an interpretable average value. Skew measures the asymmetry of a distribution: data is positively skewed if its mean > median and negatively skewed if its mean < median. In skewed data (such as 'hdl' and 'tch'), the median

gives a better sense of the average value since in skewed data, a few extreme values greatly alter the mean value. IQR is a measure of spread which only uses the middle 50% of the data and so is less affected by outliers, making it more stable than range and standard deviation: the range of 'glu' is 66 yet the IQR is only 14.75; the range is heavily affected by large outliers here. Kurtosis is a measure of the tailedness of a distribution and so can indicate presence of outliers. 'hdl' and 'prog' both have high (excess) kurtosis values of 0.98 and -0.88 respectively indicating presence of outliers.

4. Statistical test

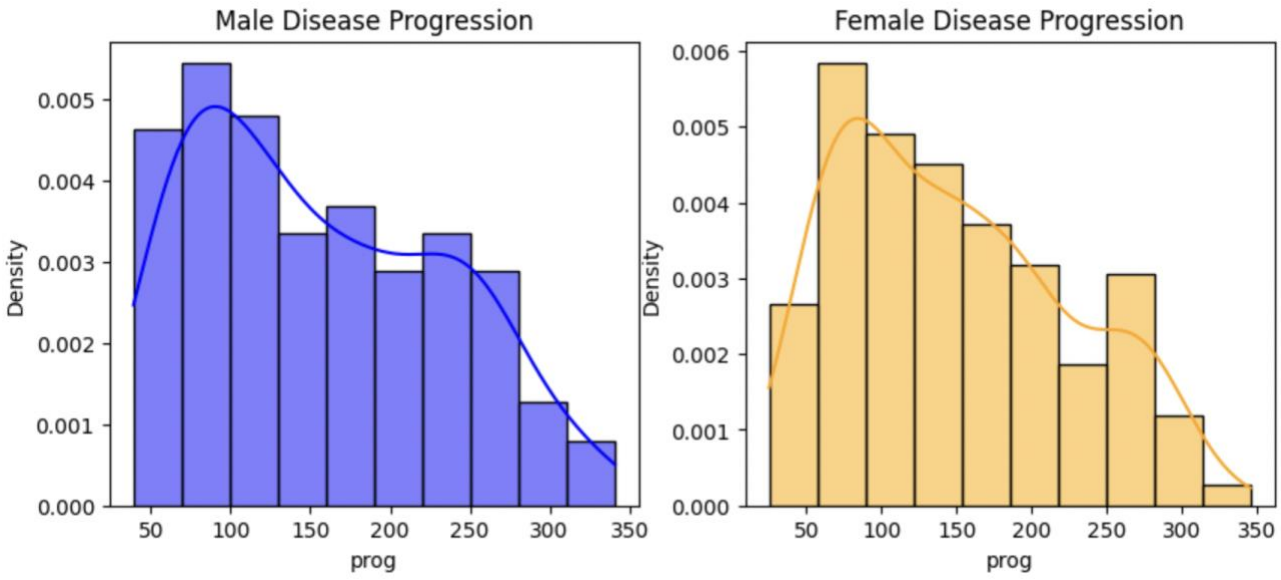

Figure 4. Histograms showing disease progression in males (n=207) and females (n=235). Both curves are positively skewed. The kernel density estimates look very similar in both cases, indicating similar distributions.

The number of bins was determined using Knuth's rule (using astropy.stats.knuth_bin_width) (Markov, 2022) which is a Bayesian optimal bin estimator. I used this here since the distributions do not appear normally distributed (which most other bin estimator algorithms assume) and are very positively skewed and heavy-tailed on the left-hand side. The maximum of each bin number was used for both histograms (same number of bins) in order to allow consistency when comparing with one other. The histograms were plotted using seaborn.histplot.

Using the Shapiro-Wilk test (Brownlee, 2018) for normality (null hypothesis: sample is normally distributed, alternative hypothesis: sample is not normal) for each sample with significance level 1% gives a test statistic

of 0.947 and 0.948 for males and females respectively (using scipy.stats.shapiro). This yields p-values of 6.98x10^-7, 2.01x10^-7 << 0.01 and so we reject the null hypothesis that each sample is normally distributed. Thus, the two-sample t-test which assumes normally distributed samples is not suitable here.

So we use the Mann-Whitney U test, which does not assume normality, with null hypothesis that the distribution of disease progression of males and females are equal and alternative hypothesis that they are not equal and significance level 1%. This gives the U test statistic of 25375 (using scipy.stats.mannwhitneyu). Under the null hypothesis, there is a 0.432 chance of this test statistic or more extreme being observed (a p-value of 0.432). Thus, since p=0.432>0.01 we do not have sufficient evidence to reject the null hypothesis. So, we conclude that disease progression in males and females is equal and is not significantly different in either case.

5. Pearson correlation heatmap

	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
age	1.000000	0.173737	0.185085	0.335427	0.260061	0.219243	-0.075181	0.203841	0.270777	0.301731
sex	0.173737	1.000000	0.088161	0.241013	0.035277	0.142637	-0.379090	0.332115	0.149918	0.208133
bmi	0.185085	0.088161	1.000000	0.395415	0.249777	0.261170	-0.366811	0.413807	0.446159	0.388680
map	0.335427	0.241013	0.395415	1.000000	0.242470	0.185558	-0.178761	0.257653	0.393478	0.390429
tc	0.260061	0.035277	0.249777	0.242470	1.000000	0.896663	0.051519	0.542207	0.515501	0.325717
ldl	0.219243	0.142637	0.261170	0.185558	0.896663	1.000000	-0.196455	0.659817	0.318353	0.290600
hdl	-0.075181	-0.379090	-0.366811	-0.178761	0.051519	-0.196455	1.000000	-0.738493	-0.398577	-0.273697
tch	0.203841	0.332115	0.413807	0.257653	0.542207	0.659817	-0.738493	1.000000	0.617857	0.417212
ltg	0.270777	0.149918	0.446159	0.393478	0.515501	0.318353	-0.398577	0.617857	1.000000	0.464670
glu	0.301731	0.208133	0.388680	0.390429	0.325717	0.290600	-0.273697	0.417212	0.464670	1.000000

Figure 5. Correlation matrix among predictors. Computed using .corr()

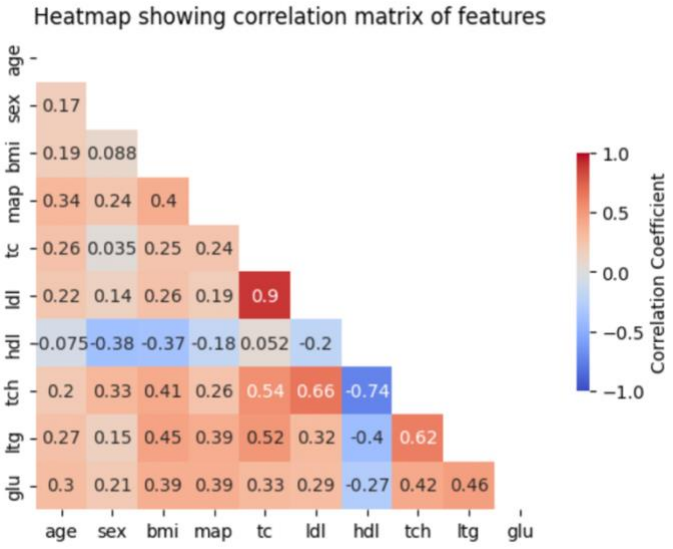

Figure 6. Heatmap showing correlation between different predictors. In particular, we see that 'ldl' and 'tc' have a very strong positive correlation of 0.9 and 'tch' and 'hdl' have a strong negative correlation of -0.74. 'hdl' and 'tc' have correlation very close to 0 (close to uncorrelated).

When plotting the heatmap using seaborn.heatmap, I set vmax and vmin to +-1 so the colourmap would just span the range of possible correlation coefficient values. I removed the upper triangle of correlations to remove unnecessary information and make it easier to read. I also chose to put the correlation coefficient values on the graph for clarity. I used the colourmap 'coolwarm' where red means strong positive correlation and blue means strong negative correlation because of its pre-attentive effect.

6. Correlation with target (bar chart)

The top 2 predictors are 'bmi' and 'ltg' with correlation coefficients of 0.59 and 0.57, respectively. There is a moderate to strong positive correlation (i.e. linear relationship) between disease progression and these factors: as 'bmi'/'ltg' increases, so does disease progression.

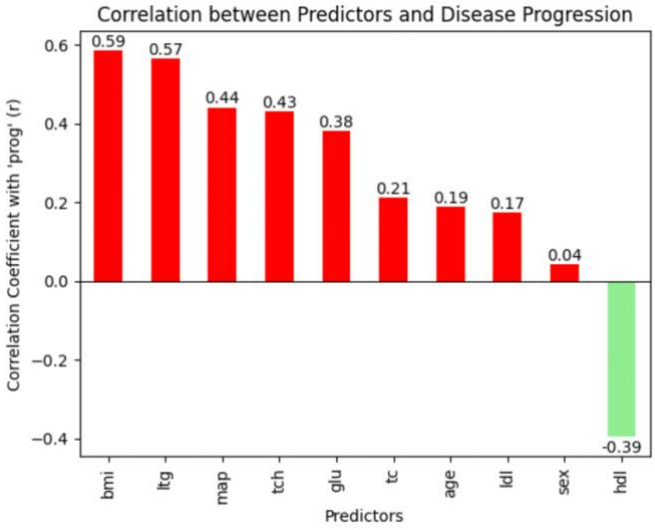

Figure 7. Bar chart showing correlation between predictors and the target disease progression. 'bmi' and 'ltg' have the strongest positive correlation with 'prog'. 'hdl' and 'prog' show a moderate negative correlation.

7. Classification exercise

A new DataFrame was constructed with ‘sex’ as the new target variable and ‘prog’ as a predictor variable.

	age	prog	bmi	map	tc	ldl	hdl	tch	ltg	glu	sex
1	59	151	32.1	101.0	157	93.2	38.0	4.0	2.110590	87	1
2	48	75	21.6	87.0	183	103.2	70.0	3.0	1.690196	69	0
3	72	141	30.5	93.0	156	93.6	41.0	4.0	2.029384	85	1
4	24	206	25.3	84.0	198	131.4	40.0	5.0	2.123852	89	0
5	50	135	23.0	101.0	192	125.4	52.0	4.0	1.863323	80	0

Figure 8. First 5 rows of new DataFrame.

#	Column	Dtype
0	age	int64
1	prog	int64
2	bmi	float64
3	map	float64
4	tc	int64
5	ldl	float64
6	hdl	float64
7	tch	float64
8	ltg	float64
9	glu	int64
10	sex	int64

Figure 9. Table describing new DataFrame.

Fisher LDA (sklearn.discriminant_analysis.LinearDiscriminantAnalysis) was used to classify by ‘sex’ using each new predictor. The first model was trained and tested on the whole dataset and the second model was trained on 80% of the dataset and tested on the remaining 20% using sklearn.model_selection.train_test_split. LDA coefficients and results are shown below.

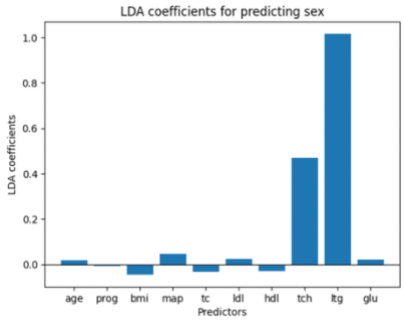

Figure 10. Bar chart showing LDA coefficients learned during training on whole dataset. Shows learned weighting of each feature when classifying sex.

**Figure 11. Confusion matrix obtained testing on training data.
Accuracy of 0.713 achieved.
Model correctly identified 177 females and 138 males but misclassified 127 people.**

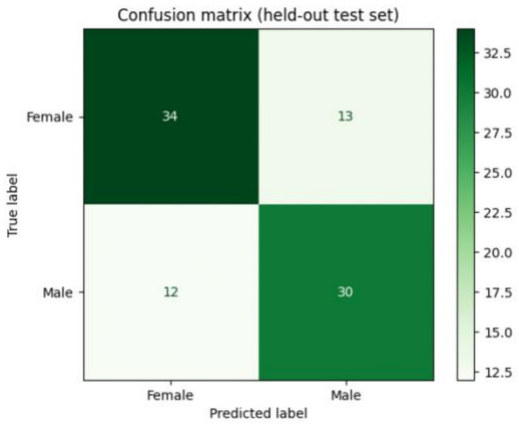

**Figure 12. Confusion matrix obtained using 80%/20% train/test split to test generalisation. Accuracy of 0.719 achieved on unseen test data (marginally better than just testing on training data).
Model correctly identified 34 females and 30 males but misclassified 25 people.**

Accuracy and confusion matrices were calculated and plotted using sklearn.metrics.accuracy_score and sklearn.metrics.ConfusionMatrixDisplay.

Performance will likely generalise well to new data if the new data comes from the same underlying distribution i.e. demographics etc. as the training/test data. The model will also generalise as well on new data as on test if there is enough test data to resemble distribution of real data i.e. through sufficient train-test split and total data. Also, given that the class proportions (of male and female) in the new data are similar (≈ 50/50).

8. Correlation analysis

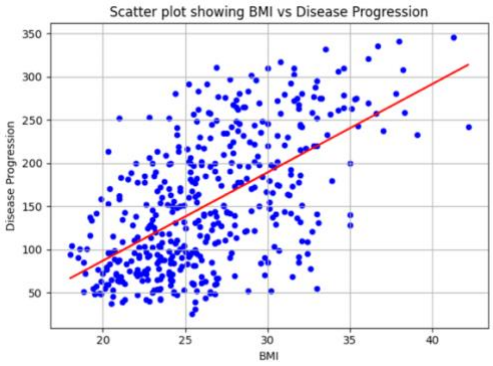

Figure 13. Scatter plot showing BMI vs disease progression with regression line indicating moderate to strong positive correlation/linear relation. Plotted using plt.scatter and np.polyfit

Under the null hypothesis that BMI and disease progression are uncorrelated and alternative hypothesis that they are correlated we perform the Pearson correlation coefficient test (using scipy.stats.pearsonr) with significance level 1%. We obtain a Pearson correlation coefficient of r = 0.586 which gives a p-value of 3.47x10^-42<<0.01. Thus, we reject the null hypothesis and conclude that this relationship is statistically significant. Given how many other factors are also at play and their correlation coefficient of 0.586, they have a moderate to strong positive correlation (linear relation).

This test assumes normally distributed BMI and disease progression which is unlikely to hold here: can test this using the Shapiro-Wilk test as before. It also assumes lack of outliers, a linear relationship between the two variables and both variables to be continuous and also be in pairs. The first two are easily verified using the scatter plot above (see Figure 13). Continuity and pairs holds true for BMI and disease progression in our dataset.

9. Conclusion

We conclude that disease progression is predictable from these indicators, but not all factors are equally useful: ‘bmi’ and ‘ltg’ are the strongest predictors of ‘prog’ with moderate to strong positive correlations ($r \approx 0.6$) and ‘hdl’ has the highest negative correlation with ‘prog’ of -0.39, so as ‘hdl’ increases, ‘prog’ tends to decrease (get better). The scatter plot between BMI and disease progression confirmed a moderate to strong positive linear relation and the p-value of $3.47 \times 10^{-42} << 0.01$ suggests with high confidence that this relation holds.

However, factors like ‘ldl’ and ‘age’ have weaker correlations with ‘prog’ ($r \approx 0.2$) and so are less predictive of disease progression.

Additionally, disease progression cannot be predicted from sex ($r = 0.04$). Disease progression does not differ significantly in males and females, evidenced by the Mann-Whitney U test ($p = 0.432 > 0.01$) and the visually similar-looking histograms. So, the two groups should be treated as one when looking at factors affecting disease progression.

Sidenote: It is important to note that although there is correlation between these factors and disease progression, further studies (like randomised control trials) are necessary to prove a causal relationship. If a causal relationship is proven, lifestyle of patients may be changed to decrease factors like ‘bmi’ and ‘ltg’ and increase ‘hdl’ in order to improve (decrease) disease progression.

10. References

Tibshirani, R., Johnstone, I., Hastie, T. and Efron, B. (2004). Least angle regression. *The Annals of Statistics*, [online] 32(2), pp.407–499. doi: https://doi.org/10.1214/009053604000000067.

Markov, M. (2022). *Optimal number of bins for a histogram*. [online] Medium. Available at: https://medium.com/@maxmarkovvision/optimal-number-of-bins-for-histograms-3d7c48086fde.

Brownlee, J. (2018). *17 Statistical Hypothesis Tests in Python (Cheat Sheet)*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/.

11. GenAI Statement

AI was used to debug some code related to Figure 7 and Figure 13. No AI was used outside of this.