MN50750 - OPTIMISATION & SPREADSHEET MODELLING

Word count:2.041

ABSTRACT

This paper presents an Excel-based automated system for predicting patient health outcomes, which is crucial for small medical practices. It leverages key health indicators—age, blood pressure, cholesterol, and maximum heart rate—to enhance early intervention and patient care. The system minimises human error through consistent, algorithm-driven assessments and can be customised to individual practice needs. Additionally, the document introduces a second model, SVM2, incorporating a regularisation parameter, lambda (λ), to allow a degree of misclassification, thus optimising the predictive accuracy: a third model, SVM3, counters potential diagnostic errors within the dataset. The approach based on Excel is a significant technological progress towards managing healthcare proactively.

Leon Mihaescu

Coursework I - Building a Support Vector Machine Using Excel

Table of contents

Table of Figures

Introduction

As per Wayne L. Winston, a Professor Emeritus of Decision Sciences at the Indiana University School of Business, whether you work for a Fortune 500 corporation, a small company, a government agency, or a not-for-profit organisation, knowing how to use Microsoft Excel in your daily work can save you hours. At the same time, Excel can provide you with new and different approaches to analysing significant business problems, making your work environment more efficient. (Winston,2021).

The importance of an Excel-based system

An automated system that uses Excel to predict patient health outcomes is crucial for small medical practices. It helps to streamline operations and improve patient care. The system incorporates critical health indicators such as age, blood pressure, cholesterol, and maximum heart rate.

By analysing patient data quickly and efficiently, doctors can intervene early and predict potential health issues based on critical metrics. This improves patient outcomes and can even save lives. An automated system saves time by reducing manual data entry and analysis, which enables healthcare providers to focus more on patient care and less on administrative tasks.

Excel's robust computational capabilities can help minimise human error in predictions, ensuring that diagnoses are based on consistent, algorithm-driven assessments. Such systems can be tailored to individual practice needs, allowing for personalised healthcare strategies that consider their patient base's unique demographic and health profiles.

By identifying patients who are likely to be sick (Diagnosis "1") and those not needing immediate medical attention (Diagnosis "-1"), practices can prioritise resources and attention to those in need.

With empirical data, medical practitioners can make well-informed decisions about their patients' treatment plans and preventive measures. When patients understand the metrics behind their health assessments, they can become more involved in managing their health. Over time, the accumulated data can reveal health trends among the patient population, which can assist in guiding preventive health strategies. In summary, implementing an automated health prediction system using Excel is about adopting technology and embracing a proactive approach to healthcare management that benefits patients and healthcare providers in numerous ways.

With research so far, it has been discovered that Microsoft Excel is one of those applications that ensure the smooth running of organisations and will boost the organisation's work efficiency within and outside the institution. (Jordan,2021).

Development of Analytical Model

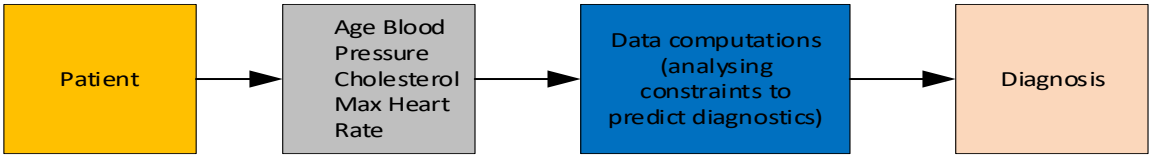


Figure 1 Development of the Analytical Mode

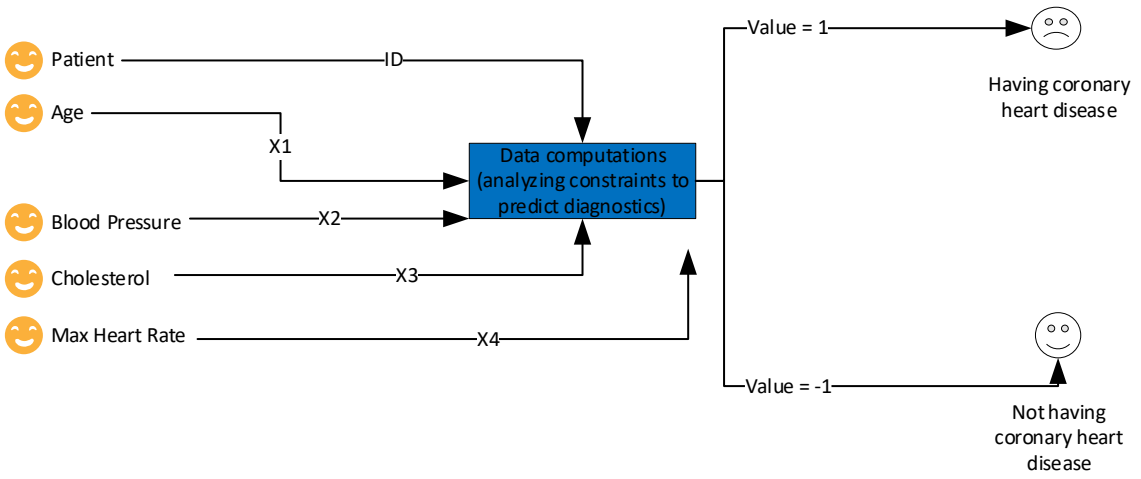


Figure 2 The Excel system described.

We introduce the data provided, data obtained from each patient, like age, blood pressure, cholesterol, and maximum heart rate, into an Excel format. As depicted in Figure 2, the predictive values for each patient are identified with x1, x2,x3, and x4. All this data is divided into two categories; the first category is recognised by the value of "1" and referred to as "Having coronary heart disease", and the second group is identified with the value of "-1" and referred to as "Not having coronary heart disease."

We aimed to divide these two into two perfect groups on our first try, using a "hard-margins" approach. Figure three describes what we venture on to accomplish. The critical point about our "hard-margin" approach is that the lines in the middle must separate the "round emojis" ideally, with no red emoji on the green side and no green emoji on the red side. If we manage to do that, we can say the data is "linearly separable". The "line in the middle of the margins" refers to the best possible place to put the data divider so that it's as far away from the nearest red and green emoji as possible. This line is like a safe zone or buffer area that tries to keep the separation as clear as possible. This is called the "maximum margin", and the line in the middle is the decision boundary.

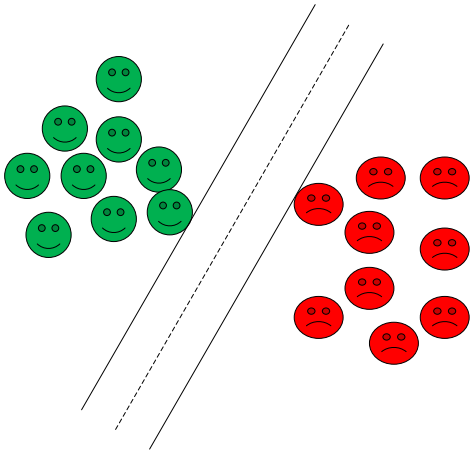


Figure 3 Linear separable data

In the field of medical diagnostics, it is widely recognised that healthy individuals may display similar clinical readings as patients suffering from various ailments in terms of our four key health indicators. When data is divided, it will look more like the data displayed in Figure 4. So, our first model, as seen in Appendix 1, is infeasible.

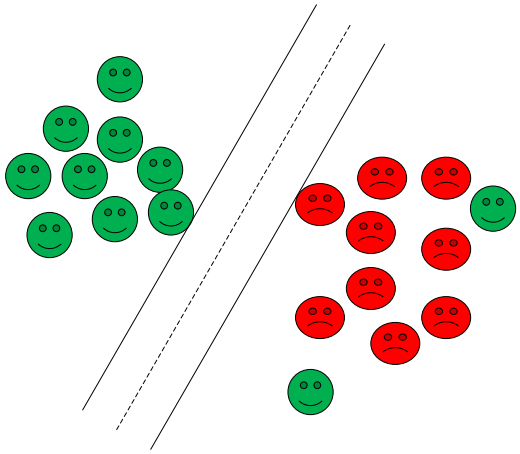


Figure 4 Non separable data

To improve the predictive accuracy and account for this overlap, we have developed a second model called SVM2. This model includes a regularisation parameter, denoted as lambda (λ), which measures the extent of permissible misclassification, allowing for a degree of flexibility in distinguishing between healthy and unhealthy states.

To obtain the optimal value of λ, we follow a thorough calibration process, using both training and testing datasets to ensure the model's effectiveness.

Moreover, at the suggestion of Dr Swede Hart, the diagnostic outcomes for the final five patients in the dataset may carry a potential margin of error. Our third model, SVM3, is designed to address this issue with an automated diagnostic feature. This feature is specifically tailored to refine the training process, ensuring a reliable analysis, and reducing the impact of any questionable data points.

Evaluation of Predictive Analytic Models

To evaluate the models, we had to decide upon metrics and some evaluation criteria, and we agreed upon accuracy, precision, recall, and specificity.

Accuracy represents the overall correctness of the model, calculated as the total number of correct predictions divided by the total number of predictions. And even if our sample could be more impressive in many data inputs, according to Mr. Christopher M. Bishop, in principle, high accuracy may be achievable with a relatively small number of samples. (Bishop,2006). Precision is the Positive Predictive Value (PPV). It measures the ratio of true positive predictions to the total positive predictions made. Recall, often called Sensitivity or True Positive Rate (TPR), measures the ratio of true positive predictions to the actual positive cases. Specificity, known as the True Negative Rate (TNR), measures the ratio of true negative predictions to the actual negative cases. (Wiesbauer, 2019)
In short:

- **Accuracy**: How often the model is correct.
- **Precision**: Of the times the model predicted '1', how often was it right?
- **Recall**: Of all the actual 'disease' cases, how many did the model catch?

- **Specificity**: Of all the actual "no disease" cases, how many did the model correctly identify as 'no'?

Evaluative criteria	SVM1 - Model 1	SVM2 - Model2 /λ=0	SVM3 - Model3 /λ=0
Accuracy	Infeasible	84%	88%
Precision	Infeasible	95.45%	95.65%
Recall	Infeasible	87.50%	91.67%
Specificity	Infeasible	80%	75%

Figure 5 Lambda = 0 and SVM1

SVM1 - Model 1: The model did not produce a viable solution (see Appendix 1).
SVM2 - Model 2 (λ=0): This model is functional. With λ set to 0, indicates there was no regularisation applied.
SVM3 - Model 3 (λ=0): This model is similar to SVM2 in that it also has λ set to 0, but it shows improved performance in most criteria except for Specificity.

- **Accuracy**: 88% - A higher accuracy indicates that this model is correct 88% of the time across all predictions, which is better than SVM2.
- **Precision**: 95.65% - This is slightly higher than SVM2, suggesting a marginal improvement in the model's ability to predict positive outcomes correctly.
- **Recall**: 91.67% - This is a notable improvement over SVM2, indicating this model is better at identifying true positive outcomes.
- **Specificity**: 75% - This is lower than SVM2, which means this model is not as good at identifying true negative outcomes.

Statistical Analysis and Trends for λ =0 and SVM1 :

- **Trend in Performance**: Moving from SVM1 to SVM3, there is a clear trend of improvement in the model performance.
- **Importance of Regularization (λ)**: Both SVM2 and SVM3 have λ set to 0, which means they might be more prone to overfitting. Regularisation can help prevent overfitting by penalising larger coefficients in the model.
- **Balancing Precision and Recall**: SVM3 has a higher recall than SVM2 but at the cost of lower specificity. This trade-off is common in classification tasks, and adjustments to the model should consider the importance of each metric in the context of its application.
- **Comparative Performance**: The improvement in accuracy and recall suggests that SVM3 is likely a more refined model.

Statistical Analysis and Trends for SVM2 & SVM3

SVM2 - Model2	Training				Test			
λ	Accuracy	Precision	Recall	Specificity	Accuracy	Precision	Recall	Specificity
7	84%	95.45%	87.50%	80%	64%	78.05%	78.50%	67%
10000	82%	95.35%	85.42%	82%	67%	82.72%	77.91%	70%
25000	76%	95%	79.17%	86%	66%	85.71%	74.16%	76%

SVM3 - Model3	Training				Test			
λ	Accuracy	Precision	Recall	Specificity	Accuracy	Precision	Recall	Specificity
7	88%	95.65%	91.67%	75%	68%	80.00%	81.93%	65%
10000	78%	97.50%	79.59%	92%	67%	84.81%	76.14%	73%
25000	74%	97.37%	75.51%	93%	64%	87.67%	70.33%	80%

Figure 6 SVM 2&3 Training and Test

SVM2 - Model2:Training Performance:

- As λ increases, there is a general decrease in Accuracy and Recall, suggesting that the model is becoming less sensitive to the training data, potentially underfitting as it focuses more on regularisation (simplification of the model).
- Precision remains relatively stable, which means that the proportion of true positive predictions to total positive predictions is less affected by the changes in λ.
- Specificity increases with higher λ values, indicating that the model is improving in identifying true negatives.

Test Performance:

- Accuracy improves slightly as λ increases, but not in a strictly linear fashion. This suggests that while the model may be underfitting on the training data with higher λ values, it generalises somewhat better on unseen data (test data).
- The highest Specificity on test data is observed at λ=25000, which is consistent with the model being more conservative and less likely to predict positive outcomes.

SVM3 - Model3:Training Performance:

- There is a clear downward trend in Accuracy and Recall with increasing λ, indicating a move towards underfitting the training data as the model becomes simpler.
- Precision increases with λ, indicating that the model is becoming more certain about its positive predictions but at the cost of missing out on some true positive outcomes (as seen by the decrease in Recall).
- Specificity shows a general upward trend with higher λ values, aligning with the model's increasing focus on identifying true negatives correctly.

Test Performance:

- There is a drop in Accuracy and Recall as λ increases, suggesting that the model's ability to generalise to unseen data is decreasing with stronger regularisation.
- Precision generally improves with higher λ, which again might be reflective of the model making fewer positive predictions but being more accurate when it does.
- The highest Specificity on test data is observed at λ=25000, similar to SVM2, indicating better identification of true negatives.

An increase in λ tends to decrease the model's ability to fit the training data closely (underfitting) but can improve its ability to generalise to unseen data to a certain extent.

There's a trade-off observed between Precision and Recall as λ changes. As λ increases, the models tend to predict positive outcomes less frequently but more accurately, which is why Precision increases, but Recall decreases. The increase in Specificity with higher λ suggests that the models become better at predicting negative outcomes correctly. This is typical with increased regularisation, where models may predict positive outcomes more conservatively.

SVM3 - Model3 appears to be superior in terms of Accuracy and Recall for training data when λ is low, but as λ increases, both models seem to trend similarly in performance.

Model Performance on Test Data:

SVM2 - Model2:

- The best accuracy on the test data is observed at λ=25000 (66%), which also has the highest specificity (76%).
- Precision peaks at λ=25000 (85.71%), but the recall is the lowest at this point (74.16%).

SVM3 - Model3:

- The best accuracy is at λ=7 (68%), which is higher than the best accuracy of SVM2.
- Precision is highest at λ=25000 (87.67%), but this model has a better balance between Precision and recall at λ=7 (80% precision and 81.93% recall).
- Specificity is best at λ=25000 (80%), but it is not much lower at λ=7 (65%).

Best Choice:

SVM3 - Model3 at λ=7 seems to be the best choice. It has the highest accuracy on the test data compared to SVM2 at any λ value.

It also has a very good balance between Precision and recall, which is crucial in a medical context where both false positives and false negatives have significant consequences.

While its specificity at λ=7 is lower compared to SVM2 at λ=25000, the overall balance of metrics makes it more reliable for general use.

In summary, while SVM3 - Model3 at λ=7 may not be the best in every single metric, it offers the best overall performance on the test data, making it the preferable model based on our data.

Conclusion

The first model demonstrates the foundational concept of linear separability in patient data classification while recognising its limitations when faced with the complex nature of medical diagnostics. The advanced SVM2 model addresses these limitations by introducing lambda (λ). This innovative approach embraces the nuances of patient data, allowing for a realistic and adaptable prediction system. The culmination of this work is seen in SVM3. This model predicts and learns from potential inaccuracies, ensuring continuous improvement and reliability. This shift from a rigid classification system to an intelligent, self-optimising model exemplifies Excel's adaptive capabilities in solving real-world issues. As healthcare continues to evolve, incorporating such advanced data analytics models ensures that medical professionals can offer more personalised, efficient, and informed patient care, solidifying the role of technology as an indispensable ally in health management and disease prevention.

References

C. M. Bishop. (2006). *Pattern Recognition and Machine Learning*. New York: Springer. p.524.

Franz Wiesbauer. (2019). *Sensitivity & Specificity: Comprehensive Insights*. [Online Video]. 12 March 2019. Available from: https://www.youtube.com/watch?v=8J_i2C4elnk. [Accessed: 1 November 2023].

James Jordan. (2021). *Excel 2021 for beginners*. Not known: Independently published. pp.18-21.

W. Winston. (2021). *Microsoft Excel Data Analysis and Business Modelling*. 7th ed. London: Pearson Education. pp.23-27.

Bibliography

Accuracy, Precision, Recall, TPR, FPR, Specificity, Sensitivity, F1 Score in Machine Learning. www.youtube.com, https://www.youtube.com/watch?v=601Sy5rWm5M. Accessed 06 Nov. 2023.

Excel Solver - Example and Step-By-Step Explanation. www.youtube.com, https://www.youtube.com/watch?v=dRm5MEoA3OI. Accessed 06 Nov. 2023.

How To Use Regularization To Set Lambda? Bias and Variance in Machine Learning. www.youtube.com, https://www.youtube.com/watch?v=GuFgfKtQnqg. Accessed 06 Nov. 2023.

Introduction to Support Vector Machine (SVM) and Kernel Trick (How Does SVM and Kernel Work?). www.youtube.com, https://www.youtube.com/watch?v=ikt7Qze0czE. Accessed 06 Nov. 2023.

Kim, Elven. 'How Bias-Variance Relate to Accuracy-Precision'. Medium, 27 Nov. 2022, https://medium.com/@elvenkim1/bias-and-variance-f160a23226d7.Accessed 06 Nov. 2023.

Never Forget Again! // Precision vs Recall with a Clear Example of Precision and Recall. www.youtube.com, https://www.youtube.com/watch?v=qWfzIYCvBqo. Accessed 06 Nov. 2023.

Precision and Recall in Classification Models | Built In. https://builtin.com/data-science/precision-and-recall. Accessed 06 Nov. 2023.
Precision-Recall versus Accuracy and the Role of Large Data Sets. www.youtube.com, https://www.youtube.com/watch?v=s3zydjHy7WM. Accessed 06 Nov. 2023.
Regularization - Explained! www.youtube.com, https://www.youtube.com/watch?v=bOIPwdWso_0. Accessed 06 Nov. 2023.

Appendix 1

	x1	x2	x3	x4		
Patient #	Age	Blood Pressure	Cholesterol	Max Heart Rate	Diagnosis	F(x,k)
1	63	145	233	150	-1	3.984009
2	37	130	250	187	-1	-8.79665
3	41	130	204	172	-1	-3.0744
4	56	120	236	178	-1	-1
5	57	120	354	163	-1	-1
6	57	140	192	148	-1	5.027021
7	56	140	294	153	-1	0.508465
8	44	120	263	173	-1	-3.27177
9	52	172	199	162	-1	-3.35475
10	57	150	168	174	-1	-1.4014
11	54	140	239	160	-1	0.176499
12	48	130	275	139	-1	3.796646
13	49	130	266	171	-1	-3.04327
14	58	150	283	162	-1	-1.99607
15	50	120	219	158	-1	2.818881
16	58	120	340	172	-1	-2.41406
17	43	150	247	171	-1	-6.17155
18	69	140	239	151	-1	5.466394
19	59	135	234	161	-1	1.784248
20	44	130	233	179	-1	-4.93091
21	42	140	226	178	-1	-6.1281
22	61	150	243	137	-1	5.617068
23	40	140	199	178	-1	-5.7226
24	71	160	302	162	-1	-1
25	59	150	212	157	-1	1.562851
26	65	140	417	157	-1	-2.27946
27	53	130	197	152	-1	4.307678
28	41	105	198	168	-1	1.059944
29	44	130	219	188	-1	-6.56012
30	67	160	286	108	1	11.01026
31	67	120	229	129	1	12.81307
32	63	130	254	147	1	5.840603
33	56	130	256	142	1	5.418505
34	48	110	229	168	1	0.999999
35	60	130	206	132	1	10.11869
36	40	110	167	114	1	13.57363
37	43	120	177	120	1	11.3217
38	57	150	276	112	1	9.46434
39	55	132	353	132	1	4.249838
40	65	150	225	114	1	12.30597
41	58	112	230	165	1	3.565556
42	50	150	243	128	1	5.312978
43	44	112	290	153	1	1.446177
44	60	130	253	144	1	5.913632
45	54	124	266	109	1	12.96915
46	50	140	233	163	1	-1.18591
47	65	120	177	140	-1	11.47153
48	62	140	268	160	1	1
49	58	120	284	160	1	2.069609
50	60	117	230	160	1	4.53537

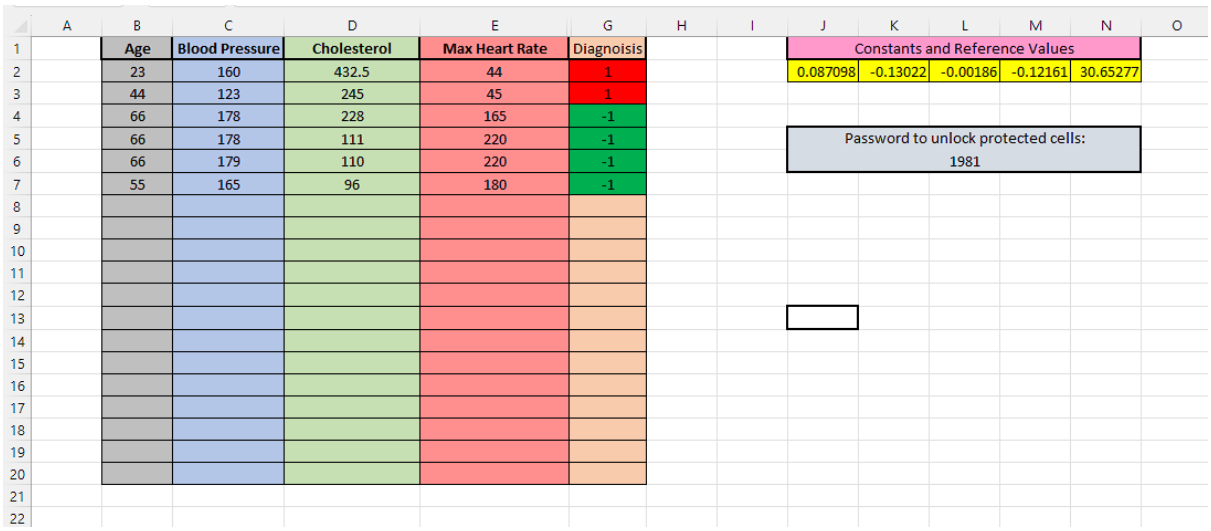
x1	x2	x3	x4	Intercept
0.215153	-0.12128	-0.03096	-0.22918	49.60373

0.1145 Objective function
(SVM1) min : x 2 1 + x 2 2 + x 2

*there is no feasible solution that can be obtained with Solver.

decision variables
costrains
data
data that needs verification
diagnosis

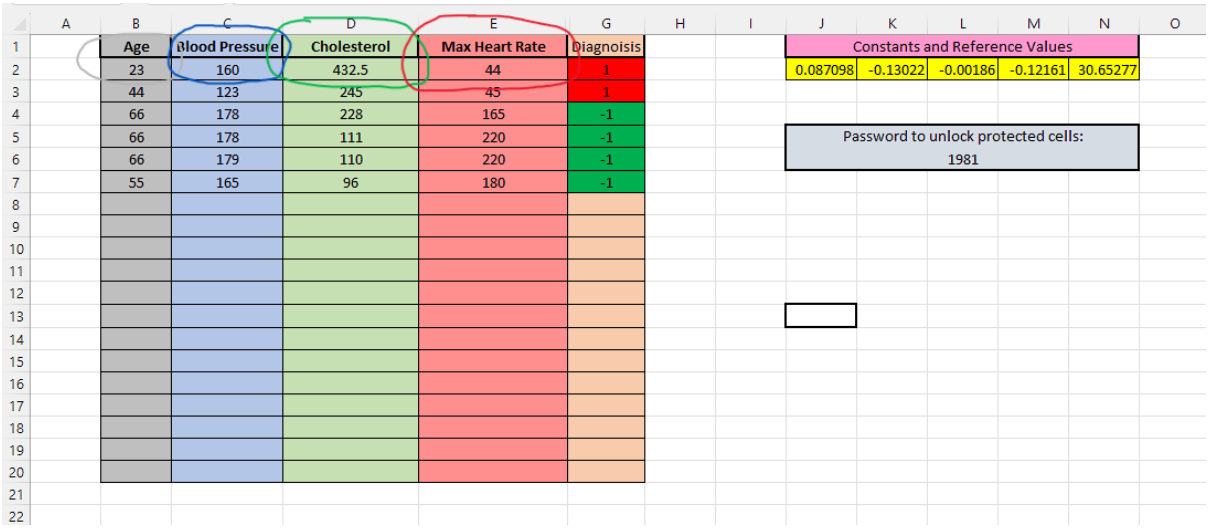
Annexe 1 – User Manual *Patient Health Analysis: Predictive Diagnosis Tool using Age, Blood Pressure, Cholesterol, and Max Heart Rate*



Age	Blood Pressure	Cholesterol	Max Heart Rate	Diagnoisis
23	160	432.5	44	1
44	123	245	45	1
66	178	228	165	-1
66	178	111	220	-1
66	179	110	220	-1
55	165	96	180	-1

Constants and Reference Values				
0.087098	-0.13022	-0.00186	-0.12161	30.65277

Password to unlock protected cells:
1981



Age	Blood Pressure	Cholesterol	Max Heart Rate	Diagnoisis
23	160	432.5	44	1
44	123	245	45	1
66	178	228	165	-1
66	178	111	220	-1
66	179	110	220	-1
55	165	96	180	-1

Constants and Reference Values				
0.087098	-0.13022	-0.00186	-0.12161	30.65277

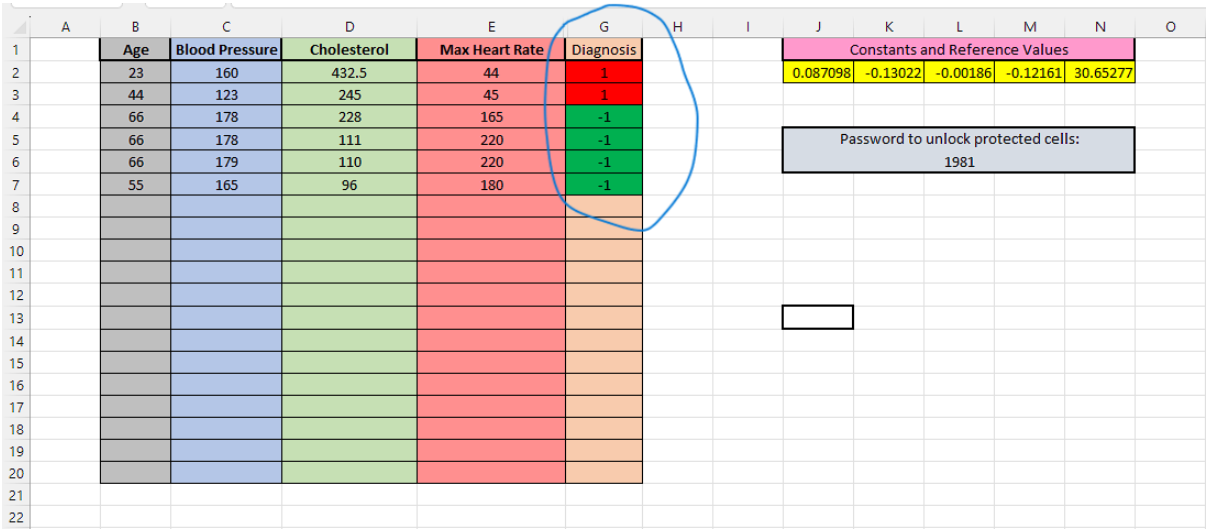
Password to unlock protected cells:
1981

Age: will take whole numbers from 0 to 125

Blood pressure: will take numbers (decimals allowed) from a range between 60 and 200, representing mm Hg (millimetres of mercury) and a range that is recognised as the min and max for blood pressure on record.

Cholesterol: will take numbers (decimals allowed) from a range between 25 - 550, milligrams per decilitre, a range that is recognised as the min and max for Cholesterol on record.

Max Heart Rate: will take numbers (decimals allowed) from 35 - 220 beats per minute, a range recognised as the min and max for Max Heart Rate on record.

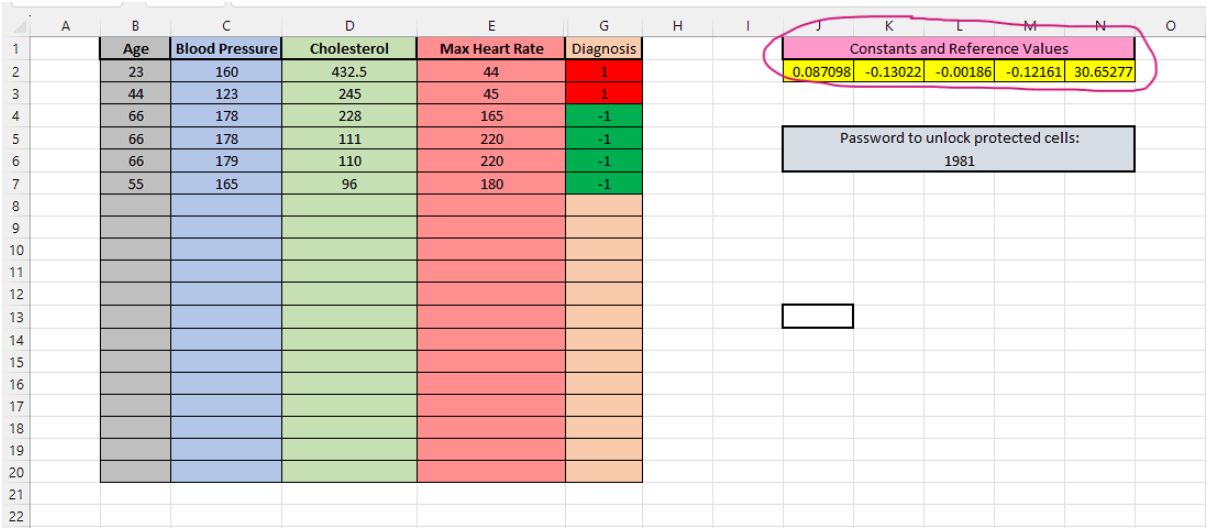


Age	Blood Pressure	Cholesterol	Max Heart Rate	Diagnosis
23	160	432.5	44	1
44	123	245	45	1
66	178	228	165	-1
66	178	111	220	-1
66	179	110	220	-1
55	165	96	180	-1

Constants and Reference Values				
0.087098	-0.13022	-0.00186	-0.12161	30.65277

Password to unlock protected cells:
1981

Diagnosis : -the value of "1" will be accompanied by the colour red and refers to a patient with disease.
-the value of "1" will be accompanied by the colour green, referring to a patient with no disease.

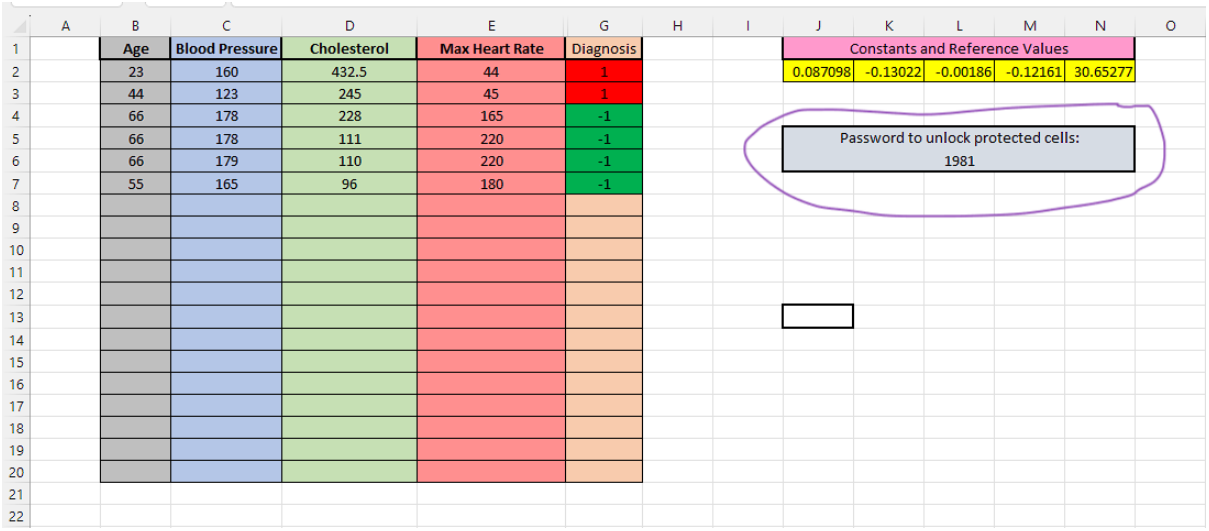


Age	Blood Pressure	Cholesterol	Max Heart Rate	Diagnosis
23	160	432.5	44	1
44	123	245	45	1
66	178	228	165	-1
66	178	111	220	-1
66	179	110	220	-1
55	165	96	180	-1

Constants and Reference Values				
0.087098	-0.13022	-0.00186	-0.12161	30.65277

Password to unlock protected cells:
1981

Constraints obtained from the chosen. SVM 3 model with lambda having a value of 7

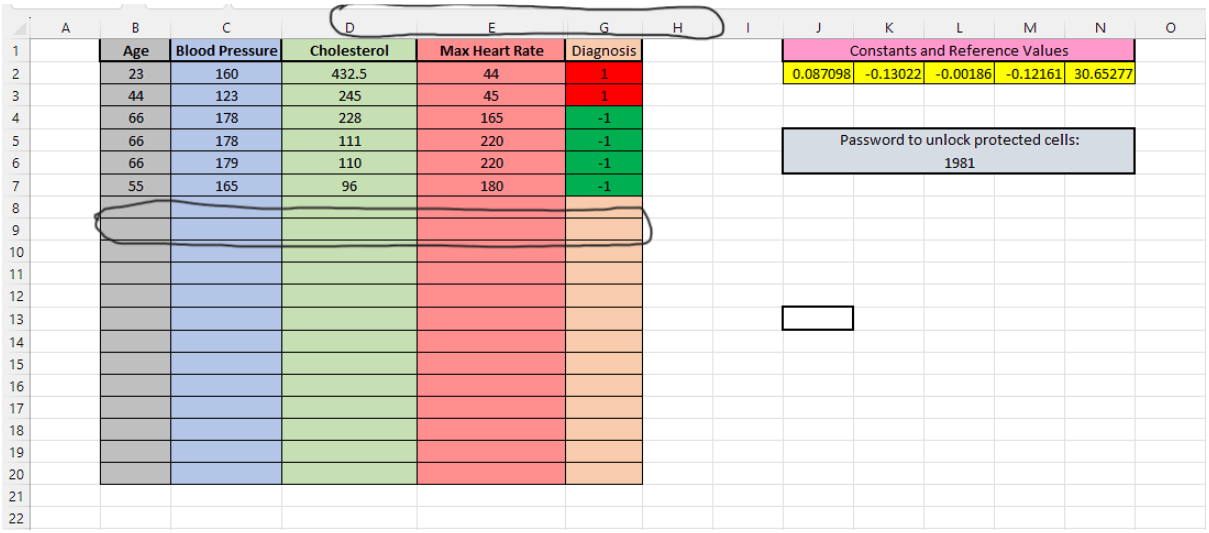


Age	Blood Pressure	Cholesterol	Max Heart Rate	Diagnosis
23	160	432.5	44	1
44	123	245	45	1
66	178	228	165	-1
66	178	111	220	-1
66	179	110	220	-1
55	165	96	180	-1

Constants and Reference Values				
0.087098	-0.13022	-0.00186	-0.12161	30.65277

Password to unlock protected cells:
1981

The **cells are protected**, so the formulas and the cell constraints should not be accidentally modified. The password is: 1981



Age	Blood Pressure	Cholesterol	Max Heart Rate	Diagnosis
23	160	432.5	44	1
44	123	245	45	1
66	178	228	165	-1
66	178	111	220	-1
66	179	110	220	-1
55	165	96	180	-1

Constants and Reference Values				
0.087098	-0.13022	-0.00186	-0.12161	30.65277

Password to unlock protected cells:
1981

Column F is hidden, so the interface is more user-friendly.

The diagnosis will be displayed only when the correct data is entered in all rows, corresponding to Age, Blood pressure, Cholesterol and Max Heart Rate.