MN50750 - OPTIMISATION AND SPREADSHEET MODELLING

Coursework: Building Support Vector Machine using Microsoft Excel

Abhratanu Majumder

Word Count - 2045

CONTENTS:

1. INTRODUCTION:

In the wake of the recent occurrences at the local general practitioner's office,
It is getting extremely difficult to deny the rise in patients with heart-related issues. This will significantly lengthen the waiting period for patients who may suffer from missing the best opportunity to receive treatment, in addition to adding to the burden of physicians. Consequently, it is essential to develop a model that would enable physicians to accurately diagnose patients at an early stage. Based on the statistical learning principle, SVM is a dependable classification method. Initially, this method was suggested for jobs involving regression and classification (V.Vapnik, 1995).

The field of medical diagnosis and research has seen a revolutionary change with the application of cutting edge machine learning algorithms. Support Vector Machines (SVM) are one of these approaches that stands out as a potent instrument with significant implications for improving medical decision-making speed and accuracy.The statistical learning theory framework has led to the development of Support Vector Machines (SVM)(Vapnik V., 1998) [2],(Cortes C. and Vapnik V., 1995) [3]. SVM has proven effective in a variety of applications, including time series prediction (Fernandez R., 1999) [4], face recognition (Tefas A., Kotropoulos C., and Pitas I., 1999)[5], and biological data processing for medical diagnosis(Veropoulos K., Cristianini N., and Campbell C., 1999) [6]. Further investigation into their properties and applications are encouraged by their theoretical underpinnings and successful experimental results.Since Support Vector Machine (SVM) performs well in outcome prediction, it may be a commonly used technique in the medical profession for diagnosis and decision-making (Asri, Mousannif, Moatassime and Noel, 2016). Hence, we will be build three different SVM Models in this report, out of which, the best one will be chosen in order to become the final estimation scheme after several comparisons.

Cardiology professionals use four main factors—age, maximal heart rate, cholesterol, and resting blood pressure—while making initial diagnosis for their patients. As a result, the models in this report will be trained and will use the data that cardiologists have collected over the past month to predict the likelihood that a patient would develop heart disease.

The remaining sections are organised as follows. In the next section, model formulation is presented. In section 3, model assessment and model comparison has been shown. In section 4, sensitivity analysis has been demonstrated in details and section 5 explains the advantages and limitations of the models. The report is finally concluded in section 6.

2. MODEL GENERATION:

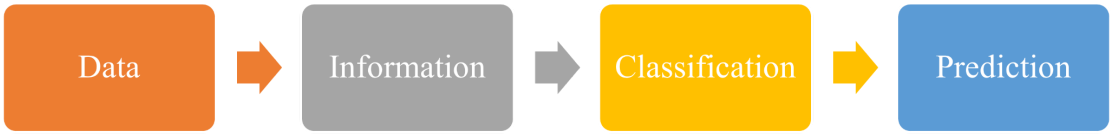

Figure 1: General Flow chart of Diagnosis

In order to translate data into information, it is necessary to apply basic mathematical formulas. These formulas are employed to convert the values of four factors linked to the risk of heart disease into numerical formats recognizable by the computer.

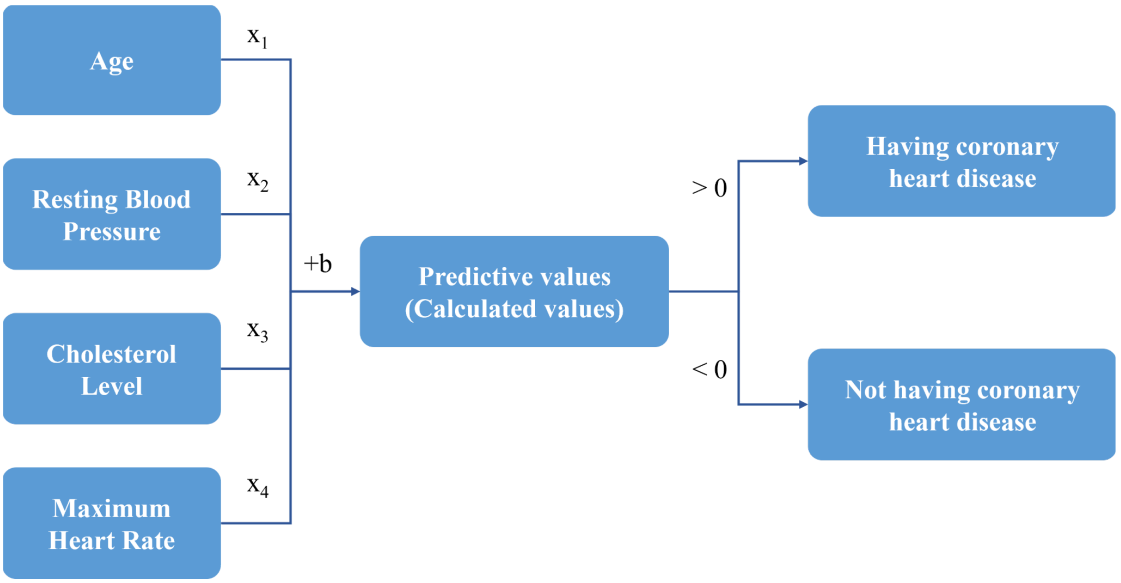

Figure 2: Demonstration of translating patients' data into information

Predictive values for each patient may be derived, as seen in Figure 2, by simply multiplying the values of four components by x1, x2, x3, and x4 and adding the intercept 'b' (SVM1). It feels as though a wall has been constructed to split a big space into two smaller ones. There are patients with heart disease (predictive values >0) in the room on the right and healthy individuals (predictive values < 0) are in the room on the left. Our models seek to improve their predictiveness and dependability by training on the aforementioned characteristics.

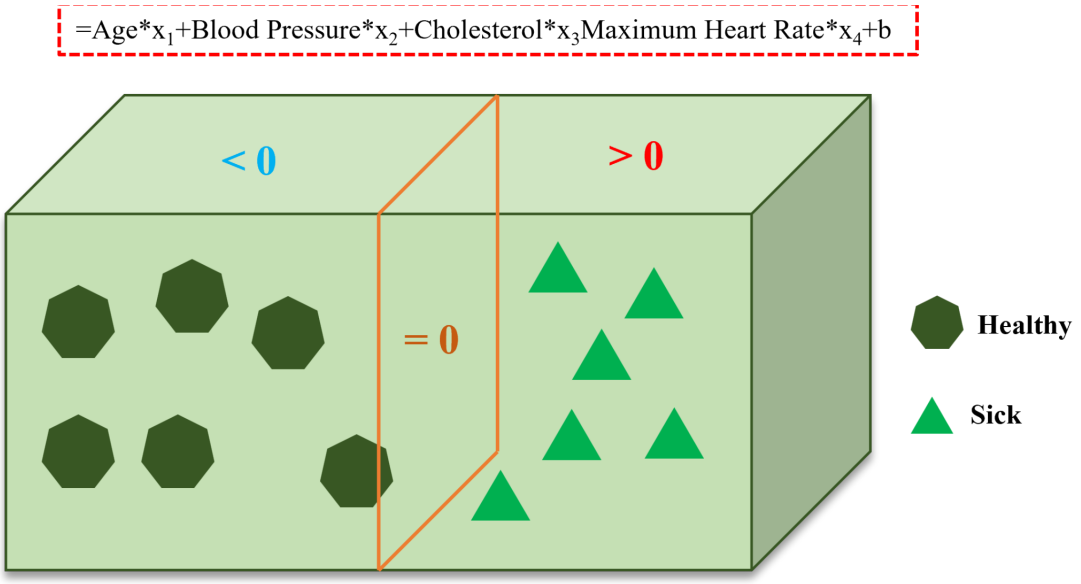

Figure 3: What do SVM models generate

2.1 Improvements in the model:

Nevertheless, even a person in good health might exhibit identical values for the four factors as someone with a medical condition. Hence, **'λ'** is incorporated into the second model (**SVM2**) to signify the acceptable level of misclassification. The optimal **'λ'** will be established based on its performance during the training and testing phases. Moreover, there are indications that the diagnosis for the last five patients in the dataset may lack reliability. Consequently, the third model (**SVM3**) is equipped to autonomously diagnose these cases as part of the training process.

3. ASSESSMENT OF MODELS:

In this section, the performance of each model is discussed through the accuracy metrics in a training set and a test set. The Solver in Microsoft Excel 365 is used for the implementation of the three models and the initial value of each decision variable is kept 1.

3.1. Metrics Description:

Accuracy: It signifies the percentage of samples accurately predicted by the models across the entire dataset. Nevertheless, when the dataset contains a small number of individuals with heart disease, this metric may not accurately reflect the models' performance.

Metrics	SVM 1	SVM 2 (λ = 30)	SVM 3 (λ = 30)
Accuracy	Infeasible	84%	84%

Table 1: Each value of models' metrics in the training set

The Solver in MS Excel is showing the result of SVM 1 as infeasible because the data is not linearly separable. Put otherwise, there is no "wall" that might keep ill and healthy individuals apart. Hence, SVM 1 is not capable of treating and solving these types of issues and can be neglected and therefore, the optimal model is supposed to be selected between SVM 2 and SVM 3.

Metrics	SVM 1	SVM 2
Accuracy	Infeasible	84%

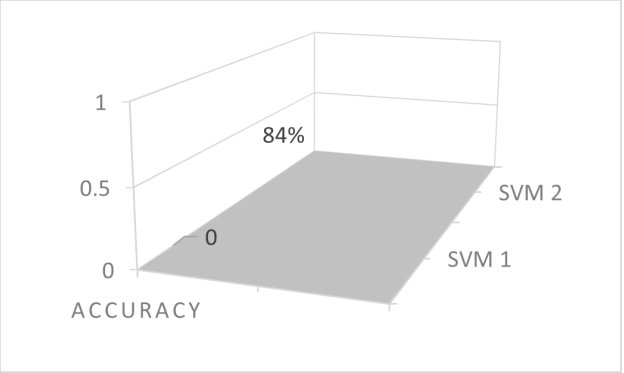

Table 1.1: Accuracy comparison between SVM1 and SVM 2

Table 1.1A: Area Graph for SVM1 and SVM 2

Table 1.1 A shows the area graph where model 1 is infeasible and model 2 is feasible with an accuracy rate of 84%. In SVM 2 and 3, when we keep the values of **'λ'** as 30, the percentage of accuracy is 84% for both the models and the models give a feasible result. Rate of classification for both SVM 2 and SVM 3 is 0.84. Since the accuracy rate is same for both the models, the plan is to check the test using both the models. The accuracy for both the training and test data at **'λ'** = 30, 70 and 10,000, has also been shown below for better understanding.

3.2. Model Comparison:

In order to generate the optimal solution to the problem, comparisons of accuracy metrics above in different models are needed. Now comparison between SVM 2 and 3 has been shown in the following table for both training and test data sets.

SVM 2 (λ)	Training Accuracy	Test Accuracy
30	84%	64%
70	84%	65%
10000	82%	67%

SVM 3 (λ)	Training Accuracy	Test Accuracy
30	84%	64%
70	84%	66%
10000	82%	67%

Table 2: Performance of SVM 2 and SVM 3 on Training and Test

When the training data is run on SVM 2, the accuracy rate is extremely high (84%) when **'λ'= 30,70**. However, when **'λ'=10,000**, the accuracy rate declines by 2%(82%). But, when **'λ'= 30,70,10,000** in the test data, the accuracy rates are 64%, 65% and 67% respectively.

When the training data is run on SVM 3, the accuracy rates are exactly the same as SVM 2 . But, when **'λ'= 30,70,10,000** in the test data, the rates are 64%, 66% and 67% respectively.

Therefore, when test data is run on SVM 2 and **'λ'** is kept 70, the accuracy rate is 65% and the classification rate is 0.65.
Whereas, when the test data is run on SVM 3 and **'λ'** is kept 70, the accuracy rate is 66% and the classification rate is 0.66.

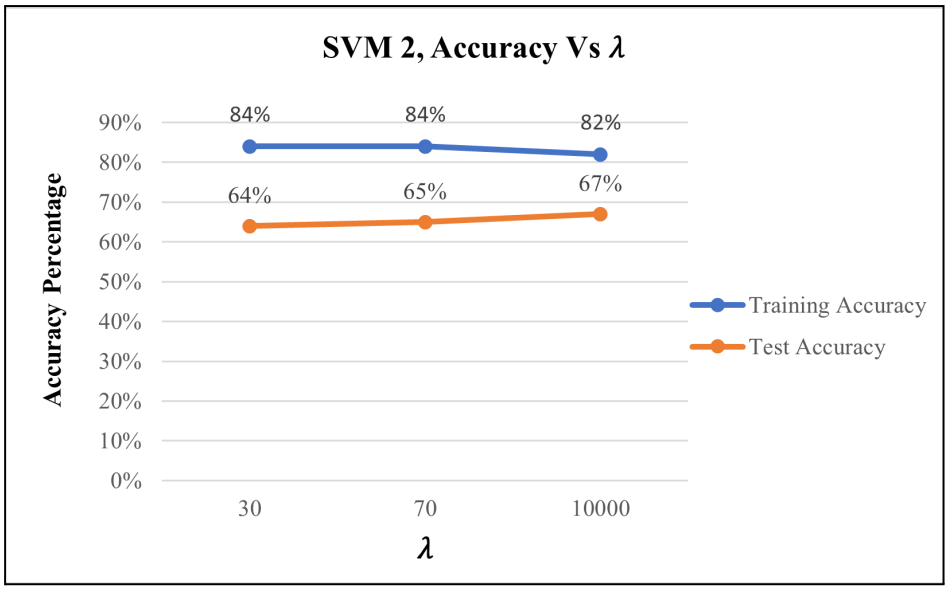

Figure 4: SVM 2, Accuracy Vs λ

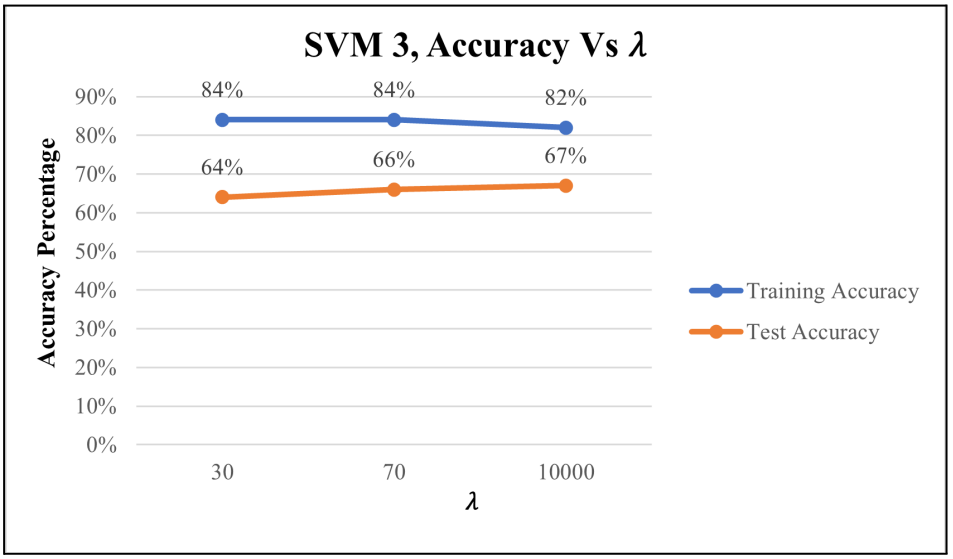

Figure 5: SVM 3, Accuracy Vs λ

The graph clearly demonstrates that when the test data is run on both the models, at **λ = 30,70, 10,000**, where accuracy rate is (64%, 65% and 67%) and (64%, 66% and 67%). So, there is hardly any change in the accuracy percentage. The comparison graph is shown below, which helped in deciding which model we should use in this case. **When SVM 3 has been used for training data, our predicted diagnosis matches the last five patients with that of the doctor's prediction. Therefore, model 3 works best on this data set and can be used to run the test data set.**

3.3. THE OPTIMAL CHOICE:

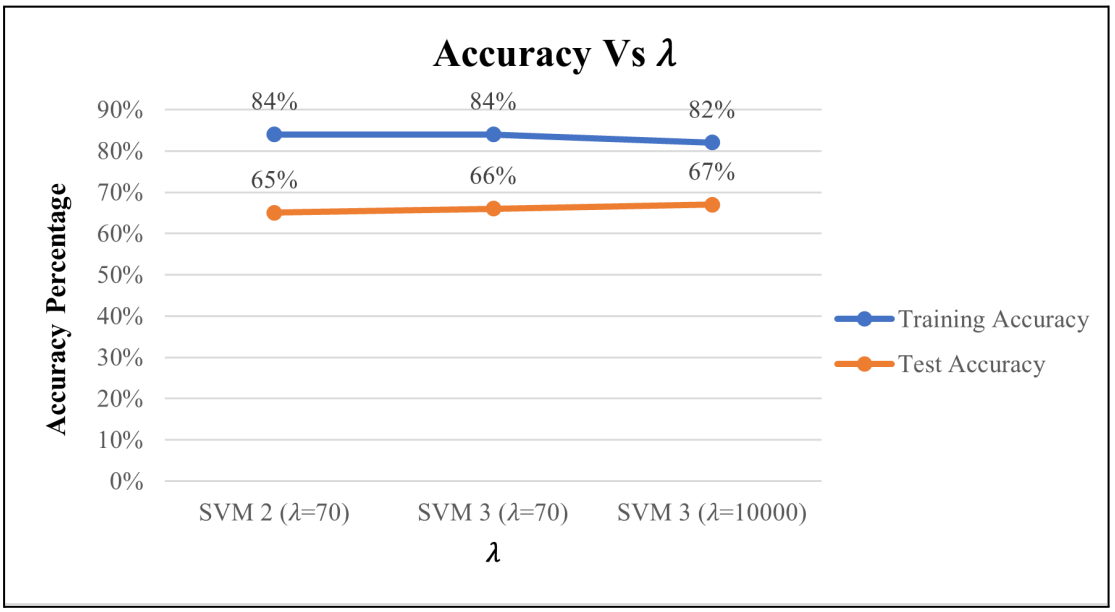

Figure 6: Comparison between SVM 2 and SVM 3

While SVM2 and SVM3's results are very close, SVM3 has a significant edge. This benefit becomes more important in situations when data may be missing or insufficient. As a semi-supervised model, SVM3 exhibits resilience through the efficient use of both labelled and unlabeled data. SVM3's ability to manage scenarios involving partial data makes it a more adaptable and sensible option, guaranteeing dependable performance even in circumstances where full data might not be available.When accuracy rates of SVM 2 and 3 are compared, both the accuracy rates are the same**.** However, it is favourable to use SVM 3 and the respective values of the decision variables are used when **'λ'** is kept at 70.
Therefore, it can also be stated that **λ** can be taken at any values, be it 30, 70 and 10,000. But with increase in **λ**, the accuracy rate decreases for both SVM 2 and SVM 3 model according to the training data set.

4. SENSITIVITY ANALYSIS ON PARAMETERS:

The impact of **'λ'** and the decision variable coefficients are discussed in this section in order to acquire medical insights and further proof. Observations from Table 2 reveal that as **'λ'** increases, there is a decline in the models' accuracy on the training set, indicating a weakening of the models' classification. Nevertheless, there is a simultaneous improvement in accuracy on the test set, as illustrated above (Figure 6).

As anticipated, the chosen solution successfully reached a state of equilibrium in classifying and predicting patients, preventing suboptimal performance in both current and future diagnosis. In addition, the decision variables of the model indicate that age and (in some instances, the cholesterol level) of the patients are more likely to be identified, which is consistent with the actual doctors' diagnosis. For example, we have taken the decision variables at (**λ=70)**.

DECISION VARIABLE	λ=30	λ=70	λ=10,000
Age	0.002856104	0.002714592	0.001309809
Blood Pressure	-0.030141015	-0.029619857	-0.009335712
Cholesterol	0.001301004	0.001296071	0.000540323
Max Heart Rate	-0.038682614	-0.038317371	-0.019809172
Intercept	9.21899013	9.104398722	3.875361293

Table 3: Decision Variables when λ = 30, 70, 10,000

5. ADVANTAGES AND LIMITATIONS:

Advantages when λ = 70

- **Making Decisions more efficiently:** Better informed decision-making processes are facilitated by SVMs' excellent prediction accuracy. The model's classifications may be trusted by medical experts, allowing them to optimise patient treatment programs and speed diagnostic processes.
- **Increasing Classification Accuracy:** SVMs effectively classify cases with high accuracy by utilising the provided dataset. This forecast accuracy reduces workload for nearby General Practitioners (GPs) while also optimising decision-making procedures. Medical personnel can concentrate on essential cases and provide more specialised and effective care if classifications are correct.
- **Optimising Resource Allocation:** Cost savings and resource optimization are closely correlated with the decrease in manual tagging efforts. Reducing the need for manual labour for data labelling allows resources to be allocated to other areas that are vital to medical practice, such as investing in cutting-edge technology or enhancing healthcare services.

Limitations when λ = 70

- **Optimising Model Performance, Addressing λ and Dataset Constraints:** The selected λ value might not be ideal for reaching peak performance because of the modest size of the training and testing datasets and the restricted number of experiments. The model can definitely be improved upon in terms of refinement.

- **False Positive Concerns:** SVM 3 is somewhat more likely than other models to mistakenly diagnose healthy people with the illness, which doesn't make a big difference in how much easier it is for medical personnel to execute their jobs.
- **Insights from Last Five Patients:** With the exception of the last patient, SVM 3 largely agrees with the diagnosis made by physicians after reviewing the last five patients in the training set. The distribution of the data may have an impact on these predictions, and the model finds it difficult to correct its own mistakes. As a result, more examination and testing are required to verify these instances' illness conditions.

6. CONCLUSION:

Medical estimation techniques are becoming more and more necessary due to the increasing incidence of coronary heart disease. In this document, we processed the provided training and test datasets using three different SVM models. To find the best answer, we compared different graphical representations and performance indicators and assessed models with various λ values. Examining the effects of the λ value and the parameters of the selected model further improved the persuasiveness of this report. Lastly, we considered the advantages and limitations of the study to help guide future research on this topic.

To sum up, SVM 3 has demonstrated remarkable performance with a low error rate and high precision in the diagnosis of heart disease, especially when λ is set to 70.

7. APPENDIX:

1. Mathematical Model of SVM 1, SVM 2, SVM 3.

SYMBOL	DESCRIPTION
x_1	Age (variables)
x_2	Resting blood pressure (variables)
x_3	Cholesterol level (variables)
x_4	Maximum heart rate (variables)
x_5	Intercept
t_k	A label to mark whether people are sick or not (-1-healthy,1-sick)
y_k	Classification Error (variables)
z_k	Whether people are diagnosed with disease or not(0-healthy,1-sick
K	visiting patients (50)
J	The last fifth patient (45)

(SVM1) $min \quad x_1^2 + x_2^2 + x_3^2 + x_4^2$

$$s.t \quad x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b} \geq 1 \quad \forall k \in \{1, \dots, K\}: t_k = 1(patient)$$
$$x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b} \leq -1 \quad \forall k \in \{1, \dots, K\}: t_k = -1(healthy)$$
$$x_1, x_2, x_3, x_4, b \ unrestricted$$

(SVM2) $min \quad \lambda(x_1^2 + x_2^2 + x_3^2 + x_4^2) + \sum_{k=1}^{K} y_k^2$

$$s.t \quad y_k \geq 1 - (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = 1(patient)$$
$$y_k \geq 1 + (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = -1(healthy)$$
$$x_1, x_2, x_3, x_4, b \ unrestricted$$
$$y_k \geq 0 \quad \forall k \in \{1, \dots, K\}$$

(SVM3) $min \quad \lambda(x_1^2 + x_2^2 + x_3^2 + x_4^2) + \sum_{k=1}^{K} y_k^2$

$$s.t \quad y_k \geq 1 - (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = 1(patient)$$
$$y_k \geq 1 + (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = -1(healthy)$$
$$y_k \geq (1 - 2z_k)(x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) + 1 \quad \forall k \in \{1, \dots, J\}: t_k = 0(undiagnosed)$$
$$x_1, x_2, x_3, x_4, b \ unrestricted$$
$$y_k \geq 0 \quad \forall k \in \{1, \dots, K\}$$
$$z_k \in \{0,1\} \quad \forall k \in \{J+1, \dots, K\}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

8. REFERENCES:

[1] Vapnik V., The Nature of Statistical Learning Theory, Springer-Verlag, NewYork, 1995.

[2] Vapnik V., ”Statistical Learning Theory”, Wiley, New York, 1998

[3] Cortes C. and Vapnik V., “Support vector networks”, Machine Learning, 20:1-25, 1995

[4] Fernandez R., “Predicting time series with a local support vector regression machine".In (Proceedings of the Workshop on Support Vector Machines Theory and Applications, Advanced Course on Artificial Intelligence (ACAI ’99), Chania, Greece, 1999 (http://www.iit.demokritos.gr/skel/eetn/acai99/Workshops.htm)).

[5] Tefas A., Kotropoulos C., and Pitas I., "Enhancing the performance of elastic graph matching for face authentications by using Support Vector Machines", In (Proceedings of the Workshop on Support Vector Machines Theory and Applications, Advanced Course on Artificial Intelligence (ACAI ’99), Chania, Greece, 1999 (http://www.iit.demokritos.gr/skel/eetn/acai99/Workshops.htm))

[6] Veropoulos K., Cristianini N., and Campbell C., "The Application of Support Vector Machines to Medical Decision Support: A Case Study" In (Proceedings of the Workshop on Support Vector Machines Theory and Applications, Advanced Course on Artificial Intelligence (ACAI ’99), Chania, Greece, 1999 (http://www.iit.demokritos.gr/skel/eetn/acai99/Workshops.htm))