Introduction to Machine Learning and Data Mining
Course no. 02450

Regression and classification applied to a medical dataset

Students:
Siddhanta Gupta - s210217
Hiten Chadha - s210208
Martin Blanck - s202281

October 28, 2022

Contents

1 Regression 2

1.1 Study of a simple linear regression model 2

1.2 Comparison of three regression models 2

1.3 Statistical analysis and conclusions 4

2 Classification 5

2.1 Description of the classification models 5

2.2 Statistical analysis and conclusions 7

3 Discussion 8

3.1 Comparison with original Research paper 8

3.2 Conclusion 8

4 Exam problems for the project 9

4.1 Question 3 9

4.2 Question 4 9

4.3 Question 5 9

4.4 Question 6 10

References 10

A Table of contribution 11

Introduction

This document reports on the second assignment for the project ”Supervised learning: Classification and regression” for the course 02450 Introduction to Machine Learning and Data Mining and solves a relevant classification and regression problem for our dataset.

This project report is an extension of the first project report on ”Data: Feature extraction, and visualization”. The main purpose of the report is to perform regression analysis and classification on our dataset using models/methods suggested in the project problem statement and statistically evaluate the result. The dataset chosen is a medical dataset whose aim is to predict if a patient is likely to have diabetes or not. As a reminder, after filtering, the dataset ended up with 369 observations and 8 attributes: number of pregnancies, blood glucose level, blood pressure, skin thickness, insulin level, body mass index, diabetes pedigree function [1], the age and whether the subject is diabetic or not.

The major part of the code used within the range of this work is inspired by the scripts provided with the course 02450 Introduction to Machine Learning and Data Mining and downloaded from DTU Learn.

1 Regression

1.1 Study of a simple linear regression model

This section discusses the linear regression analysis of the PIMA Indian Diabetes dataset. From theory, it is known that a continuous variable needs to be selected to perform linear regression. So, after careful consideration of the dataset in hand, the continuous variable *BMI (Body Mass Index)* has been chosen since it is a great indicator for diabetes. The primary reason behind choosing the variable *BMI* is because it showed the highest correlation rate with the binary variable *Diabetes* in Report 1 (along with skin thickness). Also, it is clinically known that the probability of acquiring diabetes is higher in patients having high BMI. Consequently, this section will try to evaluate how the remaining attributes will affect the prediction of BMI in the test dataset. All the remaining eight variables are considered as independent variables in the models.

Before proceeding with the regression analysis, feature transformation was applied on the dataset to remove the corrupted value and the outliers. The one-of-K-coding technique was not relevant regarding this dataset. So the data was standardized simply by subtracting the mean and diving it with the standard deviation (zero normalization). The standardized data has a mean of 0 and a standard deviation of 1 in each column of the dataset.

The concerned linear regression model equations are as follows -

$$\begin{aligned} y_i = f(x_i, w) = \tilde{x}_i^T w \\ E(w) = ||y - X^T w||^2 \end{aligned} \tag{1.1}$$

Based on this formula, the error is minimized and the optimal weights w is estimated. In MATLAB, it is computationally achieved by using the least squares model.

Once the model seemed like a good fit for the above-mentionned regression objectives, a regularization parameter was introduced λ to the model. The equation of the parameter is as follows -

$$E_\lambda(w, w_0) = ||y - w_0 1 - \hat{X} w||^2 + \lambda ||w||^2, \lambda \geq 0 \tag{1.2}$$

Furthermore, λ was initialized and set in the range from 10^{-5} to 10^{10} and an optimal value has been determined using cross-validation as equal to 0.1. It can be seen in the figure 1 that the generalization error drops for $\lambda =$ 0.1 and increases for higher values of λ. Therefore the optimal value is selected as 0.1. It is a relatively high value of the regularization strength which is probably a sign that the model has low variance but high bias.

1.2 Comparison of three regression models

In this part of the report, the same variable *BMI* was selected to perform the regression as in the previous section. Standardization was performed within each cross-validation fold by subtracting the mean and dividing

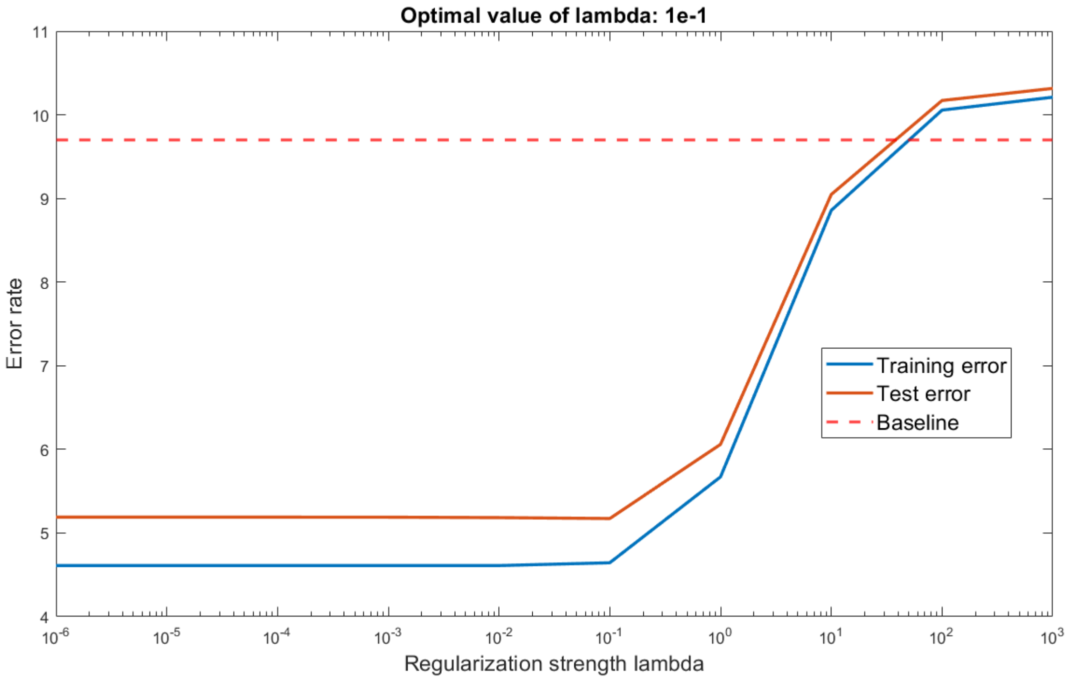


Figure 1: Training and test errors in function of the regularization strength lambda. The optimal value of λ is chosen when the test error is minimal

the current training set by its standard deviation. The next task in hand is to compare the following three models: the regularized linear regression model from the previous section, an artificial neural network (ANN) and a baseline.

In order to compute these models, a two-level cross-validation with K1 = K2 = 5 folds is implemented to obtain a feasible ANN model with maximum iterations and hidden units. Initially in order to compute the optimal number of hidden units (or layers), the number of trains of the outer loop is set to 2 and the inner loop to 5, then check the necessary number of iterations where it converges and lastly fine tune the model (by changing the range of λ, K1 and K2) for better results. During this initial testing, it has been noted that by increasing the number of hidden layers, the error of the model increased. In the final computation, the number of hidden units ranges from 5 to 15 and the optimal number is determined is determined with the inner cross-validation loop.

For λ, a narrower interval than in the previous section is used, since the optimal value was expected to be around 0.1 as discussed in the previous section:

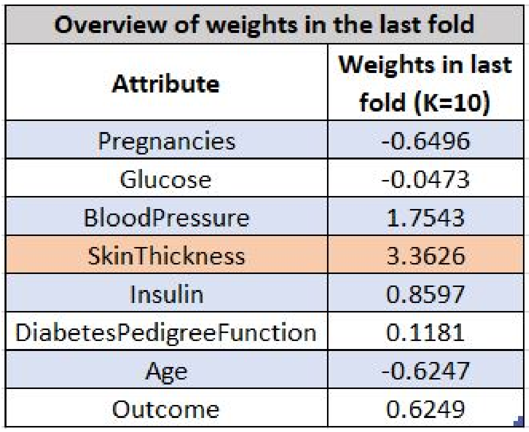
$$\lambda \in [10^{-6} : 10^{3}] \tag{1.3}$$

Also, the error formula used for the regression model is the squared loss per observation error which is as follows:

$$E = \frac{1}{N^{test}} \sum_{i=1}^{N^{test}} (y_i - \tilde{y}_i)^2 \tag{1.4}$$

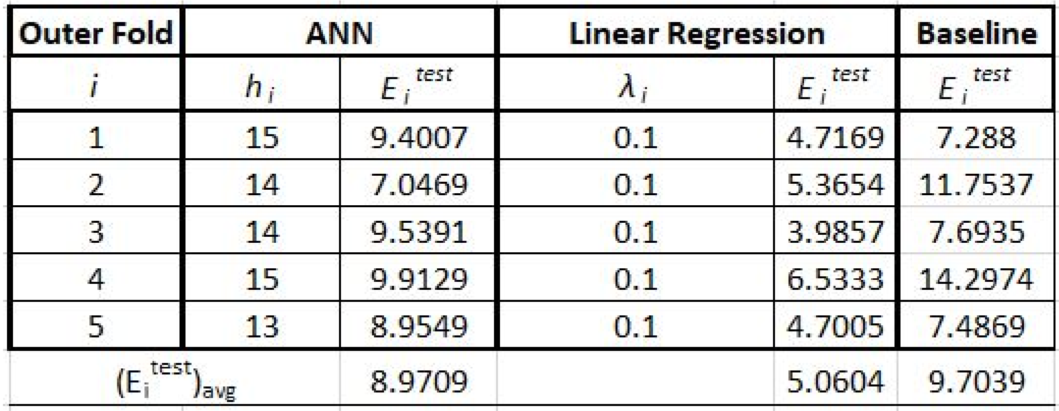
Furthermore, as per the requirement of the project description, the baseline model is computed as a simple linear regression with no features, i.e. the mean of the target attribute on the training data is used to predict the y on the test data.

Figure 2 denotes the weights in the last fold. The highest values of weights are obtained for attributes viz. *Skin Thickness (3.3626), Blood Pressure (1.7543) and Insulin (0.8597).* This supports the clinical assumption also that *BMI* is highly related to *Skin Thickness* and *Blood Pressure* of a patient. Indeed, BMI is proportional to the weight of the person; and it is known that overweight people have higher blood pressure (hypertension). Furthermore, it seems that levels of *Insulin* also contributes directly to BMI of the patient which might not be evident biologically. If we look at it holistically, *Skin Thickness* has the highest effect on the value of BMI. On the other hand, attributes like blood glucose levels or diabetes pedigree function seems less relevant for the regression task.



Overview of weights in the last fold	
Attribute	**Weights in last fold (K=10)**
Pregnancies	-0.6496
Glucose	-0.0473
BloodPressure	1.7543
SkinThickness	3.3626
Insulin	0.8597
DiabetesPedigreeFunction	0.1181
Age	-0.6247
Outcome	0.6249

Figure 2: Overview of the weights *'w'* of each attribute in the last fold



Outer Fold	**ANN**		**Linear Regression**		**Baseline**
i	h_i	E_i^{test}	λ_i	E_i^{test}	E_i^{test}
1	15	9.4007	0.1	4.7169	7.288
2	14	7.0469	0.1	5.3654	11.7537
3	14	9.5391	0.1	3.9857	7.6935
4	15	9.9129	0.1	6.5333	14.2974
5	13	8.9549	0.1	4.7005	7.4869
$(E_i^{test})_{avg}$		8.9709		5.0604	9.7039

Figure 3: Summary of 5-fold Cross Validation for Regression

By looking at the optimal generalization error parameter for all the three models in Figure 3, it can be concluded that the **'best performing model'** is the regression model with an optimal estimated generalization error of 7.0469 for $h_i = 14$ hidden units. The **'second best performing model'** is the ANN with an optimal estimated generalization error of 8.97 for $\lambda = 0.01$ and the **'worst performing model'** is the baseline with an optimal estimated generalization error of 9.70. The poor performance of the baseline was expected as it is created by computing the average of all the data.

1.3 Statistical analysis and conclusions

Since the dataset involved here takes place in a medical context, it is only logical to perform a statistical test according to the **setup II** described in the course's book. Indeed, being able to use these results only for this precise dataset would be irrelevant and redundant in terms of usability. On the other hand it would be of great interest to be able to apply the hereby trained models on other similar datasets from any hospital in the world.

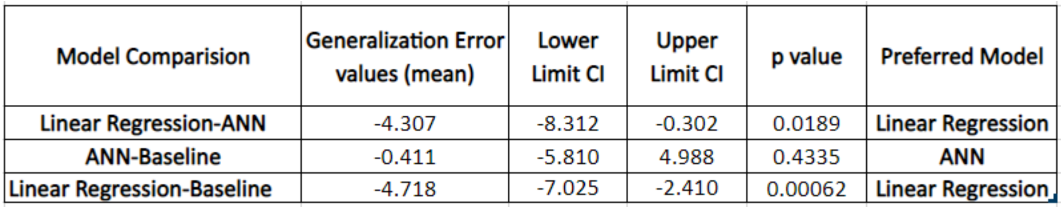
For performance evaluation a K-fold (K1=K2=5) two-layer cross-validation is used to statistically compare the three models with each other in terms of accuracy.Then the difference of generalization error of the two models is computed and then compared using the squared loss measure. For the statistical test, the cross-validation is repeated twice to obtain J = 10 splits (because of greatly important computational times). The generalization error of the three models was then estimated for each split. The 1 - α confidence intervals were calculated with alpha = 0.05 to find the lower and the upper limit of the parameter for different possible model comparisons.

According to figure 4, it can be seen that Linear Regression performs better than the ANN as well as the baseline. This conclusion can be supported by the low p-values obtained i.e p = 0.0189 and p = 0.00062 respectively showing great significance of the obtained results. It can also be concluded that the distribution of

r_j is between the confidence interval given by the lower and the upper limit, which further verifies the results obtained.

Besides, the comparison of the ANN model to the baseline appears to be problematic. They seem to be similar with close performance which is not desirable at all. However this result might not be significant since the p-value is very high, indicating that no conclusion can be done regarding these models.

In conclusion, it can surely inferred that out of the three models, the linear regression model seems to be the more fitted to perform the regression task.



Model Comparision	Generalization Error values (mean)	Lower Limit CI	Upper Limit CI	p value	Preferred Model
Linear Regression-ANN	-4.307	-8.312	-0.302	0.0189	Linear Regression
ANN-Baseline	-0.411	-5.810	4.988	0.4335	ANN
Linear Regression-Baseline	-4.718	-7.025	-2.410	0.00062	Linear Regression

Figure 4: Summary of **Setup II** statistical test for Regression

2 Classification

This chapter of the report will tackle the modelling of a classification problem with this Diabetes dataset. As discussed in the first report of the semester, the aim here seems to be pretty obvious: to classify whether a subject has diabetes or not depending on the attributes listed in the introduction of the report. This problem is therefore a binary classification problem where the outcome will be 0 if the patient is not diabetic or 1 if she is. Furthermore, it appears to be a relevant problem since it would represent a gain of time and money in the diagnosis procedure.

2.1 Description of the classification models

By the end of this section, three types of models will be tackled just as in the first section. Here, the three methods will be trained and their performance statistically evaluated again regarding the methods related to **setup II**.

First, a logistic regression model is computed using the regularization strength λ as a complexity-controlling parameter and taking 50 values between 10^{-5} to 10.

$$\lambda \in [10^{-5} : 10] \tag{2.1}$$

Then, the second method used is k-nearest neighbor classification. Here the number of neighbors used is chosen as the complexity-controlling parameter. It ranges from 1 to 50 neighbors.

$$K_{KNN} \in [1 : 50] \tag{2.2}$$

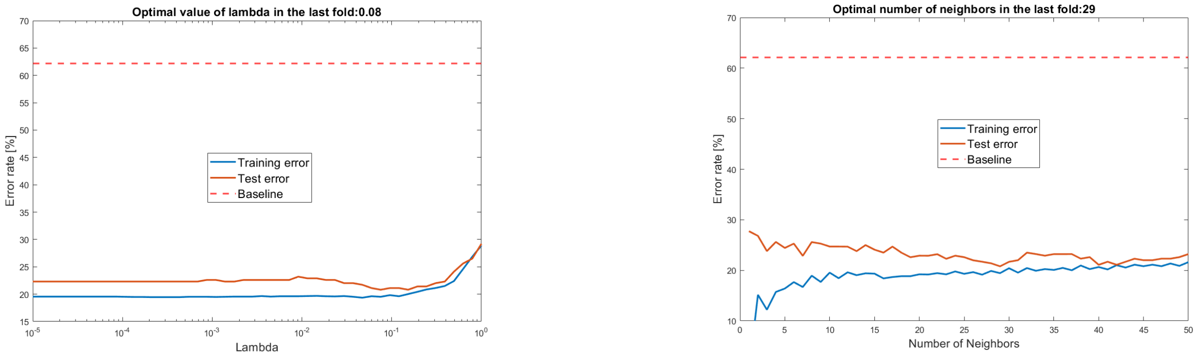
Besides, it is worth mentioning that the different distance measures included in MATLAB were tested. After trials not reported here, the euclidean distance has been chosen since it lead to the lowest error rates.

Finally the last method is just a baseline to compare with. The main objective of the baseline is to see easily if the models are useful at all. A model's performance worse than the baseline would indicate that such models are useless. The baseline is computed as assigning every observation in the test set to the highest class. Here it means that it assumes every patient is diabetic.

The three methods are compared using a two-level cross-validation with $K1 = K2 = 10$ folds. The inner fold is meant to find the optimal values of the complexity-controlling parameters, i.e. respectively the number of neighbours and the regularization strength λ. The error measure used here is the error rate:

$$E = \frac{\text{Number of misclassified observations}}{N^{test}} \tag{2.3}$$

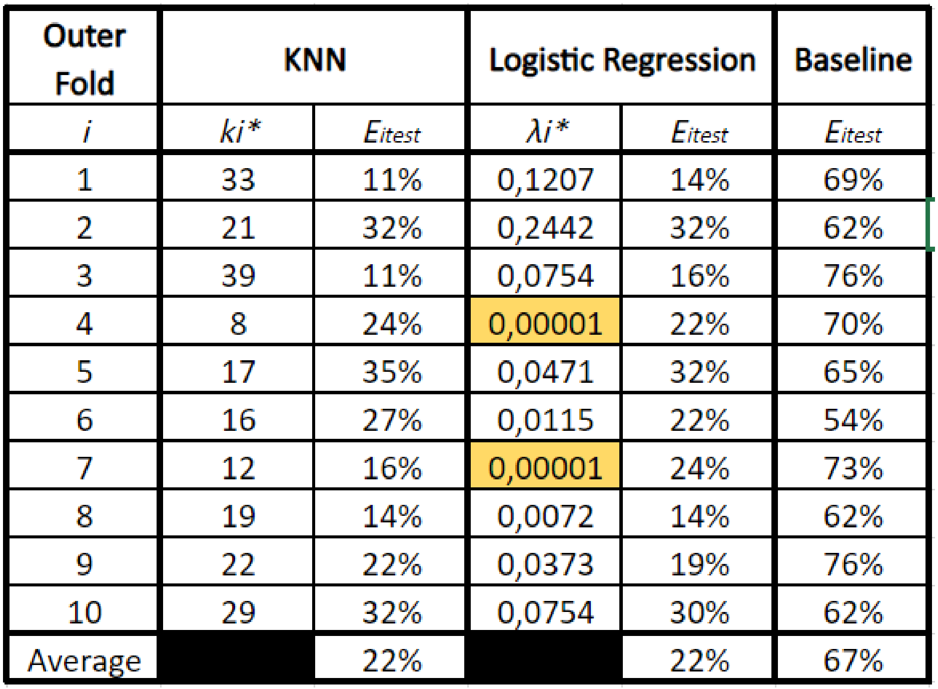
As an example, on figure 5 is illustrated how the optimal value of the complexity-controlling parameter can be chosen. The actual optimal value is displayed on top of each sub-figure.



(a) Training and test errors in function of the regularization strength

(b) Training and test errors in function of the number of neighbors

Figure 5: Plot of the training and test errors computed on the last classification fold. This illustrates how the optimal value of the complexity-controlling parameters is chosen.



Outer Fold	KNN		Logistic Regression		Baseline
i	*ki**	*Eitest*	*λi**	*Eitest*	*Eitest*
1	33	11%	0,1207	14%	69%
2	21	32%	0,2442	32%	62%
3	39	11%	0,0754	16%	76%
4	8	24%	0,00001	22%	70%
5	17	35%	0,0471	32%	65%
6	16	27%	0,0115	22%	54%
7	12	16%	0,00001	24%	73%
8	19	14%	0,0072	14%	62%
9	22	22%	0,0373	19%	76%
10	29	32%	0,0754	30%	62%
Average		22%		22%	67%

Figure 6: Two-level cross-validation table used to compare the three models in the classification problem.

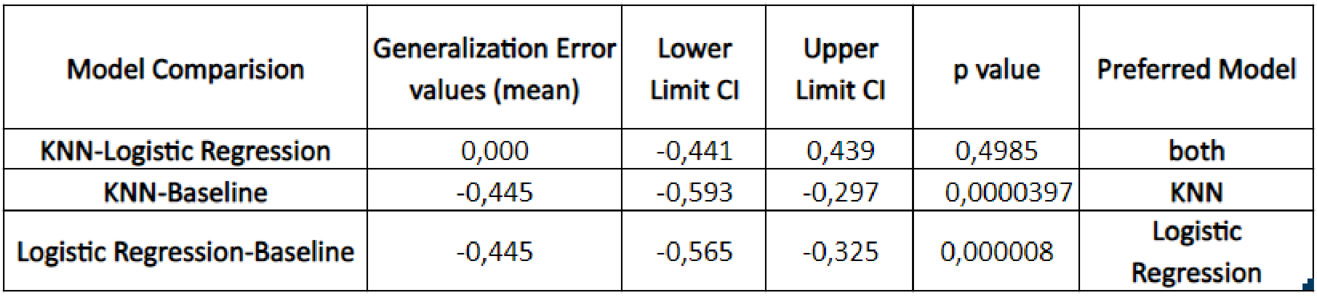
The table on figure 6 shows the generalization error obtained for each fold of the outer loop. Note that the optimal λ values highlighted in yellow are assumed to be corrupted value that could not be corrected within this work.

It is believed they are due to a lack of resolution in the choice of the range of the λ values.

On average, it appears that the optimal regularization strength is smaller than in the regression task, with a magnitude more in the order of 0.01 (vs 0.1 in the regression task).

2.2 Statistical analysis and conclusions

Just like in the previous section, the statistical analysis is performed in vision of the methods described in the course's book related to **Setup II**. A summary of the analysis performed is found in figure 7. Again, the confidence intervals were calculated for $\alpha = 0.05$.



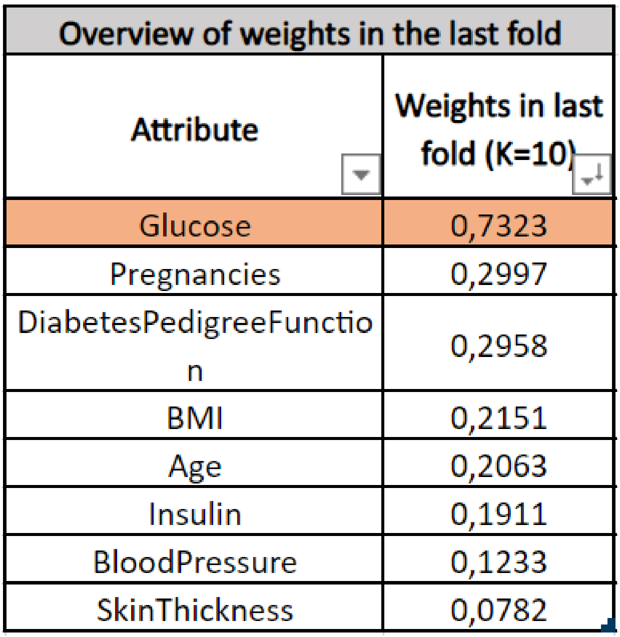
Model Comparision	Generalization Error values (mean)	Lower Limit CI	Upper Limit CI	p value	Preferred Model
KNN-Logistic Regression	0,000	-0,441	0,439	0,4985	both
KNN-Baseline	-0,445	-0,593	-0,297	0,0000397	KNN
Logistic Regression-Baseline	-0,445	-0,565	-0,325	0,000008	Logistic Regression

Figure 7: Summary of **Setup II** statistical test for classification

On one hand, it can be concluded that the KNN and the logistic regression approaches are both better than the baseline model. Indeed, the difference in generalization error is relatively great and the p-value associated is very low in both cases.

On the other hand, it could be supposed that KNN and Logistic Regression models are identical since they have the same generalization error. But the p-value is extremely high so it is greatly unlikely that this comparison result is significant. Thus, we can say that the KNN as well as the Logistic Regression models are equally accurate.

Additionally, on the table from figure 8 can be seen the weights of the optimal logistic regression model generated.



Overview of weights in the last fold

Attribute	Weights in last fold (K=10)
Glucose	0,7323
Pregnancies	0,2997
DiabetesPedigreeFunctio n	0,2958
BMI	0,2151
Age	0,2063
Insulin	0,1911
BloodPressure	0,1233
SkinThickness	0,0782

Figure 8: Overview of the weights *'w'* of each attribute in the last fold

Unlike in the regression section, the weights are here relatively close to each other. Still, it can be noticed that the blood glucose level's weight is slightly higher than the other weights. It is assumed that this statement is significant since hyperglycemia (or too high blood glucose levels) is one of the main symptoms of diabetes. Therefore, high glucose levels could be more relevant an attribute to classify a subject as diabetic or not.

Furthermore, hypothesis on pregnancies and diabetes pedigree function can be made. Indeed, there is a type of diabetes that appears first during pregnancy and called gestational diabetes[2]. Genetics is also believed to be an important risk factors of diverse types of diabetes[3].

Overall, it is not the same features that are deemed relevant as for the regression part of the report. The weights magnitudes and signs are significantly different from the regression one.

3 Discussion

From the regression section of the report, it can be concluded that the regularization parameter λ_{opt} is the most significant parameter in determining the behaviour of the model. The λ_{opt} parameter helped to maintain a low error value and optimized the complexity of the model as desired. Also it's understood that it is very important to perform statistical analysis tests, as it paints a clearer comparative picture regarding which model performs better and not get easily biased by just seeing the generalization error values. But critically speaking, it is definitely required to test the realized model further with new test data and improve it. Additionally we also noted that the computations took a lot of time to execute and hence great attention was required while choosing the final parameters of the model (like K1, K2, lambda range etc.). Besides, even though classification and methods can present similarities in their approach, it has been seen that the attributes of the dataset contributed differently to the predictions in both cases.

3.1 Comparison with original Research paper

The original research paper by Smith, J.W. et. al. (1988) uses a more exhaustive dataset (with 576 training cases) compared to the training set used within this report (369 objects). The authors of the paper used 75% percent of the data to train the model (learning mode) and remaining 25% data for the test phase whereas here the data was filtered by removing outliers and missing values. That point could be a limitation from the work of this report: the missing values could have been replaced with the mean of the respective attributes to obtain more training data which in turn would likely have resulted in better tuning of the models. Nevertheless, even with 369 observations in the training data, it appears that some of the hereby implemented models performed at par (with an accuracy of 78%) with the ADAP model used in the research paper (with an accuracy of 76%). Also, access to more test data from new patients could help to further investigate potential improvements in the accuracy of the models leveraging the new data. Lastly, other statistical analysis methods like AUC-ROC curves could be investigated for a better understanding of the estimated classification outputs (whether it is better or worse than the findings of the paper).

3.2 Conclusion

As a conclusion, this data set is well suited for the context of this report, being the application of regression and classification methods. It permitted to get familiar with their main aspects, complexities and possibilities. However, drawing conclusion on a medical dataset such as this one should be done really carefully. Being able to predict if a patient is diseased or not would be the supreme aim of the application of machine learning methods in a medical context, and is of course not what is intended within this report.

4 Exam problems for the project

4.1 Question 3

The number of parameters of the model is equal to the number of connections between the different layers of the network.

The model presented in the question has:
- 7 attributes in the input layer
- a single hidden layer with 10 units
- 4 units in the output layer

We should also consider the bias introduced in the hidden and output layer. Every unit there is said to add one bias.

That gives:

$$N = 7 \times 10 + 10 \times 4 + (10 + 4) = 124 \tag{4.1}$$

The correct answer is answer A.

4.2 Question 4

We start by looking at the node D. It makes the split between class 1 and 3. Looking on figure 4, we can see that the prediction is class 3 as:
- b1 is greater than-0.76 while b2 is smaller than 0.01
- or b2 is smaller than 0.01 while b1 is greater than -0.76

That allows us to discard answer C.

Then we can do the same for node C which represents the vertical boundary at b1 = -0.16. Answers A and B cannot be true since node C separate classes 1 and 3 from class 4. This boundary has nothing to do with b2 as proposed in proposition A and B.

The correct answer is the proposition D.

4.3 Question 5

For each outer fold, 5 neural network models and 5 logistic regression models are trained. This is made $K1 \times K2 = 20$ times. So we have:

$$(5 \times 25 + 5 \times 9) \times 4 \times 5 = 3400ms \tag{4.2}$$

Additionally, in each outer fold the optimal model needs to be trained in the end. Therefore, we need to add $25 + 9 = 34ms$ for each outer fold.

In the end we have an algorithm running for:

$$3400 + 34 \times 5 = 3570ms \tag{4.3}$$

The correct answer is C

4.4 Question 6

We answered this question by doing trial and error and doing the calculations until we found the corresponding answer.

First we calculated all the $\hat{y}_k$. For example, if we use the vector b from proposition A, we would have:

$$\hat{y}_1 = 1.2 - 2.1 \times (-1.4) + 3.2 \times 2.6 = 12.46 \tag{4.4}$$

Then we can calculate all the per-class probabilities following the equation of the question. We did that until we found a situation where $P(y = 4|\hat{y})$ is greater than the 3 other probabilities.

This happened in situation B with $P(y = 4|\hat{y}) = 0.73$.

As a consequence, **the correct answer is B.**

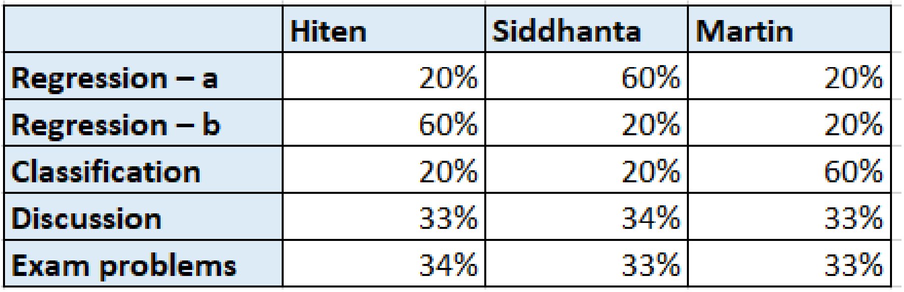
References

[1] Kaggle datasets, *Pima Indians Diabetes Database*, URL: https://www.kaggle.com/uciml/pima-indians-diabetes-database?select=diabetes.csv, Accessed: 08-03-2021

[2] MayoClinic Website, *Gestational Diabetes*, URL: https://www.mayoclinic.org/diseases-conditions/gestational-diabetes/symptoms-causes/syc-20355339, Accessed: 18-04-2021

[3] MayoClinic Website, *Diabetes*, URL: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444, Accessed: 18-04-2021

A Table of contribution



	Hiten	Siddhanta	Martin
Regression – a	20%	60%	20%
Regression – b	60%	20%	20%
Classification	20%	20%	60%
Discussion	33%	34%	33%
Exam problems	34%	33%	33%