How Non-genetic Genetic Affect Semen Quality - An Analysis Using Simple Machine Learning Models

Minh Phan - s3335814

ABSTRACT

This report investigates how some non-genetic factors such as medical history and lifestyle affect semen quality. The data was taken from UCI machine learning model repository. The analysis used K-nearest neighbour, decision tree and random forest machine learning models. Overall, the results indicated that combinations of numbers of hours sitting in a day, childish diseases, surgical intervention, alcohol consumptions and accident are useful in predicting which individuals have normal semen. The report concludes that, these non-genetic factors were only useful in assisting the semen evaluation process; the models cannot replace semen evaluation at this stage. It is recommended that further investigation with larger data sets and more sophisticated models will improve the predicting power of these non-genetic factors in semen quality assessment.

INTRODUCTION

Carlsen, E., Giwercman, A., Keiding, N., & Skakkebaek, N. E. (1992) suggested that semen quality has been reduced for the last ten years, even though other metrics of well-being has been increased. In addition, there are studies suggesting that medical history, lifestyles and other environmental factors contributed to the decline of semen quality (Yang, H., Chen, Q., Zhou, N., Sun, L., Bao, H., Tan, L.,Cao, J. , 2015). Traditionally, physicians collect information about semen quality based on semen analysis (Joffe, M. ,2010). The following study attempted to use simple models of machine learning to investigate the relationship between some of the non-genetic factors and evaluated semen results of one hundred participants. The purpose of the study is assessing how well machine learning can assist the process of predicting semen quality based on participants medical history and lifestyle.

METHODOLOGY

Dataset

The data was collected and shared by Lucentia Research Group, University of Alicante on UCI Machine Learning Repository (Gil. D, Girela, J. L, Juan, J. D, Gomez-Torres, M. J and Johnsson. M, 2012). The participants are male aged from eighteen to thirty-six. The data set consisted of nine variables: season in which the analysis was performed, age, childish diseases, accident, surgical intervention, high fevers in the last, frequency of alcohol consumption, smoking habit and number of hours sitting in a day. The target variable is semen diagnosis.

All the categorial variables were numerated by assigning specific values for all the levels. Figure 1.0

Variables	Level 1	Level 2	Level 3	Level 4	Level 5
Season	Winter : -1	Spring: -0.33	Summer: 0.33	Fall: 1	
Childish diseases	Yes: 0	No :1			
Accident	Yes: 0	No :1			
Surgical intervention	Yes: 0	No :1			
High fevers in the last year	Less than 3 months: -1	More than three months: 0	None: 1		
Frequency of alcohol consumption	several times a day: 0.2	every day: 0.4	several times a week: 0.6	once a week: 0.8	hardly ever or never: 1.0
Smoking habit	Never: -1	Occasionally: 0	Daily: 1		

The number of hours sitting is proportion of sixteen hours (0-1)

Machine learning models

The data was split by a factor of 0.5 into training set and test set. The training set was fitted in to a simple machine learning model. The trained model is subsequently used to predict the test set. The result was recorded and compared to the target variables of the test set using confusion matrix and classification report. In addition, a simple hill climbing was applied to each trained model; the process investigated which features were used by the according models to make the predictions (Mourad. A. ,2018). The analysis repeated the hill climbing process in order to find the common features used by the models. Lastly, the analysis also invested the reliability of the models by using a k-fold cross-validation.

1. K-Nearest neighbour (Ren, Y. , 2018)
2. Decision trees (Ren, Y. ,2018)
3. Random forest (Pedregosa.F., Varoquaux. G., Gramfort. A., Michel.V., Thirion. B., Grisel. O., Blondel. M., Prettenhofer. P., Weiss. R., Dubourg. V., Vanderplas. J., Passos. A., Cournapeau. D., Brucher.M., Perrot. M., Duchesnay.E. ,2011)

RESULTS

K-Nearest neighbour

Confusion metric

	0	1
0	42	1
1	5	2

Classification report

	precision	recall	f1-score	support
0	0.89	0.98	0.93	43
1	0.67	0.29	0.40	7
avg / total	0.86	0.88	0.86	50

Decision tree

Confusion metric

	0	1
0	36	7
1	3	4

Classification report

	precision	recall	f1-score	support
0	0.92	0.84	0.88	43
1	0.36	0.57	0.44	7
avg / total	0.84	0.80	0.82	50

Random Forest

Confusion metric

	0	1
0	43	0
1	6	1

Classification report

	precision	recall	f1-score	support
0	0.88	1.00	0.93	43
1	1.00	0.14	0.25	7
avg / total	0.89	0.88	0.84	50

Hill climbing

K-Nearest neighbour [8, 4, 3, 2, 1, 6]
Decision tree [8, 7, 4, 3, 2, 6]
Random forest [8, 7, 4, 3, 2]

*These are example results; different random split of the dataset would provide different values.

K-fold cross-validation

[fold 0] score: 0.8235
[fold 1] score: 0.6471
[fold 2] score: 0.9412
[fold 3] score: 0.7647
[fold 4] score: 0.9375
[fold 5] score: 0.8125

The three models were generally efficient at predicting the semen with normal quality. However, the prediction for altered semen quality was not sufficient. The decision tree performed best when it came to predicting which individuals would have altered semen quality. The other models performed worse than chance, two out of seven for both models.

The hill climbing suggested that all the models use the following common features to make predictions 8- number of hours sitting in a day, 3-Accident, 4- Surgical intervention, 2-Childish diseases and possible 6- Alcohol consumption.

In addition, all the models used the feature surgical intervention, however by looking the visualisation of the relationship between accident and the semen quality outcome, there was a distinct difference between the two groups, and this could explain why the models were inefficient predicting altered semen quality.

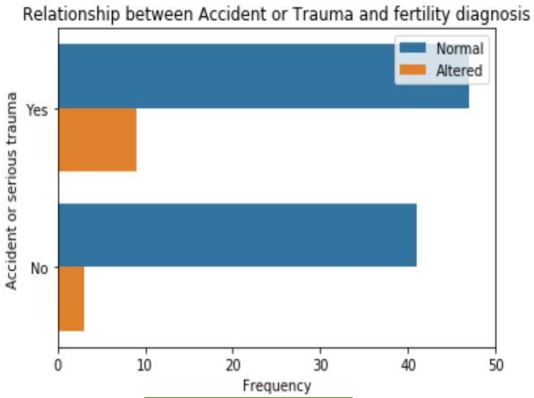

Figure 1.0

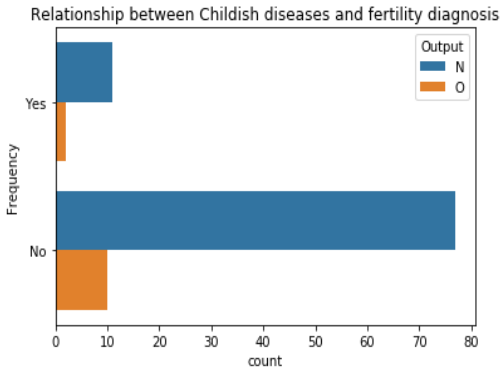

Figure 1.1

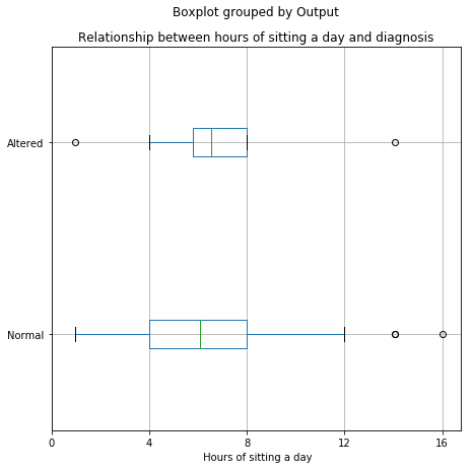

Figure 1.2

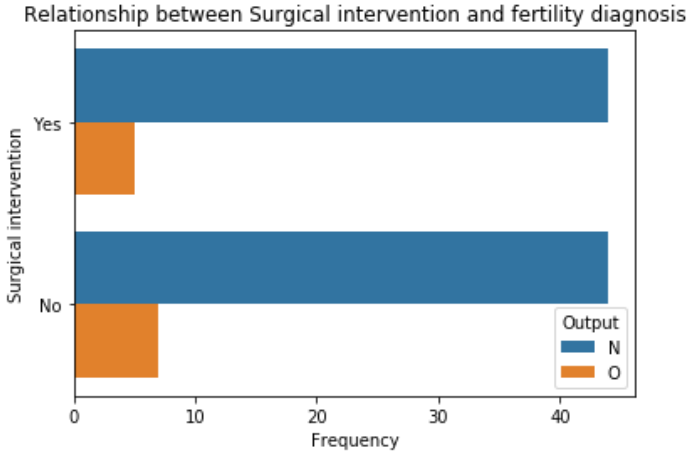

Figure 1.3

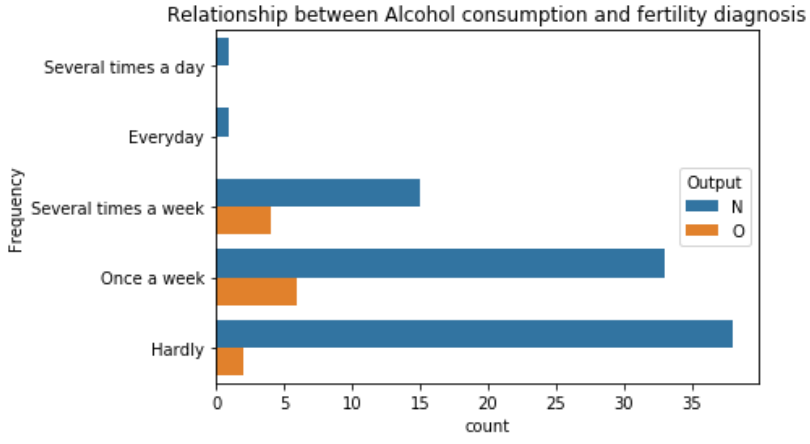

Figure 1.4

DISCUSSION

The results of the three models suggested that it is possible to predict which individuals were more likely to have healthy sperm based on their medical history and lifestyle factors. However, the models were insufficient at predicting which individuals had a higher chance of having altered sperm.

The analysis addressed that individuals who had altered semen quality were most likely to sit from six to eight hours a day. The analysis also suggested that individuals who did not have childish diseases had a higher chance of having the semen quality altered; the same case applied to individuals who had accidents or trauma to the relating region. The study also suggested that surgical intervention has very little effect on semen quality. Lastly, individuals who consumed the most alcohol were at the lowest risk of having altered semen quality.

CONCLUSION

The analysis validated the predictability of semen quality based on some non-genetic factors, including individuals' medical history and lifestyle choices. However, the predicting powers of the models suggested that prediction using machine learning is only good an assisting tool. The models did not perform well during the predicting process of individuals which altered sperm quality.

The study suggested that there is potential to develop machine learning techniques, which accurately predict individuals with normal sperm quality based on medical history and lifestyle choices. The process can be used to improve the screening process for sperm donors. The study also indicates that machine learning can assist the semen evaluation process for fertility treatment. For example, an application that track individuals' daily routine, this will help individual monitoring the sitting time daily or amount of alcohol consumption in order to minimise the negative effects to sperm quality.

The study used only one-hundred participants; a study with larger numbers of participants can further solidify these findings. In addition, the majority of variables used in the study are self-reported, possibly creating bias.

In conclusion the analysis shows that a study of how sitting hours effects semen quality is successful using machine learning. This will help reveal how the modern lifestyle has been contributing to the decline in semen quality for the past decade.

References

1. Carlsen, E., Giwercman, A., Keiding, N., & Skakkebaek, N. E. (1992). Evidence for decreasing quality of semen during past 50 years. *BMJ : British Medical Journal*, *305*(6854), 609–613.
2. Gil. D, Girela, J. L, Juan, J. D, Gomez-Torres, M. J and Johnsson. M (2012). Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, 39(16):12564 â€“ 12573.
3. Joffe, M. (2010). Semen quality analysis and the idea of normal fertility. *Asian Journal of Andrology*, *12*(1), 79–82. http://doi.org/10.1038/aja.2009.3
4. Mourad. A. (2018). *COSC2670: Practical Data Science (051637), week 5 tutorial* [Power Point slides]. Computer Science and Information Technology, the School of Computer Science, RMIT University, Melbourne, Australia.
5. Mourad. A. (2018). *COSC2670: Practical Data Science (051637), week 6 tutorial* [Power Point slides]. Computer Science and Information Technology, the School of Computer Science, RMIT University, Melbourne, Australia.
6. Mourad. A. (2018). *COSC2670: Practical Data Science (051637), week 7 tutorial* [Power Point slides]. Computer Science and Information Technology, the School of Computer Science, RMIT University, Melbourne, Australia.
7. Pedregosa.F., Varoquaux. G., Gramfort. A., Michel.V., Thirion. B., Grisel. O., Blondel. M., Prettenhofer. P., Weiss. R., Dubourg. V., Vanderplas. J., Passos. A., Cournapeau. D., Brucher.M., Perrot. M., Duchesnay.E. (2011) *Scikit-learn: Machine Learning in Python*, JMLR 12, pp. 2825-2830, 2011.
8. Ren, Y. (2018). *COSC2670: Practical Data Science (051637), week 6 notes-Classification* [Power Point slides]. Computer Science and Information Technology, the School of Computer Science, RMIT University, Melbourne, Australia.
9. Yang, H., Chen, Q., Zhou, N., Sun, L., Bao, H., Tan, L.,Cao, J. (2015). Lifestyles Associated With Human Semen Quality: Results From MARHCS Cohort Study in Chongqing, China. *Medicine*, *94*(28), e1166. http://doi.org/10.1097/MD.0000000000001166