A New Machine Learning Framework for Understanding the Link between Cannabis Use and First-Episode Psychosis

Wajdi ALGHAMDI[a] , Daniel STAMATE[a], Daniel STAHL[b], Alexander ZAMYATIN [c], Robin MURRAY[d] and Marta DI FORTI[e]

[a] Data Science & Soft Computing Lab, and Department of Computing, Goldsmiths, University of London.

[b] Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London.

[c] Faculty of Informatics, Department of Applied Informatics, National Research Tomsk State University.

[d] Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London.

[e] MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London.

Abstract. Lately, several studies started to investigate the existence of links between cannabis use and psychotic disorders. This work proposes a refined Machine Learning framework for understanding the links between cannabis use and 1st episode psychosis. The novel framework concerns extracting predictive patterns from clinical data using optimised and post-processed models based on Gaussian Processes, Support Vector Machines, and Neural Networks algorithms. The cannabis use attributes' predictive power is investigated, and we demonstrate statistically and with ROC analysis that their presence in the dataset enhances the prediction performance of the models with respect to models built on data without these specific attributes.

Keywords. eHealth, Machine Learning, First-Episode Psychosis, Gaussian Processes, Support Vector Machine, Neural Networks.

1. Introduction

According to the World Health Organisation (WHO), eHealth is any secure, cheap, and efficient use of information and communications technologies in order to support health [1]. These days, more health care providers are replacing traditional paper notes with electronic patient records. In addition, the usage of advanced technologies such as computers, personal digital assistants, smart phones, etc. had enabled information to become more available and accurate. This lead to a tremendous increase in the electronic health data making an ideal promising land for applying machine learning algorithms to extract insights from data.

Currently machine learning algorithms are in the process of revolutionizing health. In just the same way as machine learning has made an enormous difference to business and industry, it will just as undoubtedly enhance medical research, and improve the

Corresponding author Email: map01wa@gold.ac.uk

practice of healthcare providers. For instance, machine learning algorithms have been successfully used in understanding the different manifestations of asthma [2], diagnosing psychosis [3][4], classifying leukaemia [5], detecting heart conditions in electrocardiogram (ECG) data [6], etc. In particular, machine learning algorithms have been proven to be capable of dealing with complicated medical data such as ECG signal data, where they show some outstanding results compared to traditional statistical approaches.

These studies suggest that machine learning can provide medical research with powerful techniques beyond the traditional statistical approaches mostly used, which include the conventional statistical tests, linear and logistic regression. Also, in biomedical engineering, several recent papers explored the potential for machine learning algorithms to detect various diseases. This has led to the publication of more guidance for medical researchers on how to infer and question such findings [7]. Finally, there is a tremendous interest in current interdisciplinary research into exploiting the power of machine learning to enable further progress in the new area of Precision Medicine, in which predictive modelling plays a vital role in forecasting treatment outcomes, and thus decisively contributes in optimising and personalising treatments for patients [8].

The medical field is considered as a critical area of research, yet there are still many difficult tasks that need to be carried out precisely and efficiently. The future success of health sector planning, and of healthcare in general, will be the adoption of intelligent systems where robotics and machine learning intersect. In order for health sector planning to catch up in this fast changing environment, machine learning must be put at the core of most strategies. For example, new developments in Psychiatry concern the so called Data-driven Computational Psychiatry, which relies heavily on the use of machine learning.

In this study, we propose a new computational psychiatry and machine learning framework based on developing optimised models for predicting the onset of first-episode psychosis with Gaussian Processes (GP), Support Vector Machines (SVM), and Neural Networks (NNET). In particular, our aim is the predictive modelling approach to help understanding the link between the first-episode psychosis and cannabis use. The dataset on which we based our study was collected by psychiatry practitioners and was used in previously conducted studies, such as [3][9]. It comprises an extensive set of variables, including demographic, drug-related and other variables, with specific information on participants' history of cannabis use - some of which being illustrated in Table 1.

Table 1. Cannabis use attributes in the analysed dataset [3].

Attribute	Description
lifetime_cannabis_user	Ever used cannabis: yes or no
age_first_cannabis	Age upon first use of cannabis: 7 to 50
age_first_cannabis_under15	Age less than 15 when first used cannabis: yes, no or never used
age_first_cannabis_under14	Age less than 14 when first used cannabis: yes, no or never used
current_cannabis_user	Current cannabis user: yes or no
cannabis_fqcy	Pattern of cannabis use: never used, only on weekends or daily
cannabis_measure	Cannabis usage measure: none, hash less than once per week, hash on weekends, hash daily, skunk less than once per week, skunk on weekends, skunk daily
cannabis_type	Cannabis type: never used, hash or skunk
duration	Cannabis use duration: 0 to 41 (months)

The framework we present here integrates: data pre-processing, model tuning, model post-processing with receiver operating characteristic (ROC) optimisation based on the maximum accuracy cut-off threshold, and model evaluation with k-fold cross-testing. This sequence of enumerated phases is repeated 500 times for each GP, SVM with radial and polynomial kernels, and NNET algorithms, to study the potential variation of the performances of the resulting models. We investigated also the cannabis use attributes' predictive power by establishing statistically that their presence in the dataset augments the models' performances.

2. Methods

2.1. Description of study population

The data we used to build our predictive models were a part of a case-control study [9]. The clinical data comprise 1106 records divided into 489 patients, 370 controls and 247 unlabelled records. The patients were individuals who presented with first-episode psychosis to the inpatient units of the South London & Maudsley Mental Health National Health Service (NHS) Foundation Trust. The controls were healthy people recruited from the same area served by the Trust. Each record in the data refers to a participant in the study and has 255 possible attributes divided into four groups. The first group consists of demographic attributes, which represent general features like gender, race and level of education. The second group of drug-related attributes contains information on the use of non-cannabis drugs, such as tobacco, stimulants and alcohol. The third group contains genetic attributes. These were removed from the analysis since they were out of this study scope. The final group contains cannabis-related attributes, such as the duration of use, initial date of use, frequency, cannabis type, etc.

For the purpose of building prediction models, we first removed any attribute with more than 50% missing value. Then, we perform the same high-level simplification of the dataset that was proposed in [3]. The resulting dataset, after the transformations, contained 783 records and 78 attributes. The records are divided into 451 patients and 332 controls. A summary of some of these fields—specifically, those that relate to cannabis use, such as type, age of first use and duration—can be seen in Table 1.

2.2. Data pre-processing

The quality of data may significantly affect the performance of the predictive models [10]. In order to help improve the quality of the data and, consequently, of the predictive models, the clinical data is pre-processed. Data pre-processing usually deals with the preparation and transformation of the initial dataset. In this study, we applied numerous pre-processing techniques such as missing values imputation, class balancing and feature selection to improve the efficiency and ease the modelling process.

Firstly, in term of missing values imputation, we applied random forest imputation from the *randomForest* package [11]. Although this method is computationally expensive, it enhanced the predictive power of the final models.

Secondly, the synthetic minority over-sampling technique (SMOTE) [12] was selected to treat the unbalanced classes existed in the data. SMOTE chooses a data point randomly from the minority class, de-terminus the K nearest neighbours to that point and then uses these neighbours to generate new synthetic data points using slight alterations. Our analysis used five neighbours. The results show that SMOTE leads to an increase in both the area under the ROC curve (AUC) and the accuracy.

Finally, we applied a feature selection technique based on the information gain [13]. To do so, we evaluate the information gain for each attribute with respect to the class. Such techniques are often used with forward selection or backward elimination, which considers only removing the feature subset with least ranking values. In this study, we apply information gain to filter out the attribute that does not have predictive power regarding information gain. Figure 1 illustrates some (due to lack of space) attribute predictive power with respect to the information gain. Figure 1 shows that attributes such as *riskcan* and *typefrq2,* which are cannabis measures are the highest ranked attribute when attributes such as gender are among the least ranked attributes. Initially, this indicates that some of the cannabis use attributes have more productivity power than some of the demographic attributes

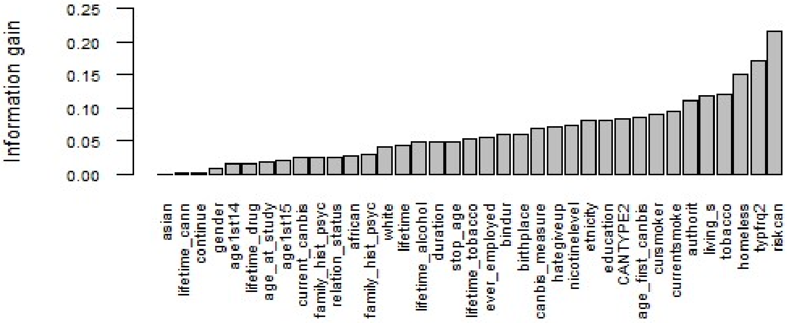

Figure 1. Attributes' predictive power with respect to Information Gain.

2.3. Predictive modelling

To develop optimised predictive models for first-episode psychosis, we controlled the values of the parameters for each of the considered algorithms using chosen grids. Predictive models have been fitted in a five-fold cross-validation procedure, on each training set after pre-processing techniques were applied on the same training set, and have been tested on each test set. Models based on SVM, NNET, and GP were optimised to maximise AUC.

First, SVM models were tuned with different kernels such as SVM with the radial kernel (SVMR) and SVM with the polynomial kernel (SVMP). The optimal SVM models were obtained with SVMR, after tuning the parameters cost and gamma over 10 values. The optimal values for cost and gamma were 16 and 0.004, respectively.

Then, GP models were tuned with different kernels such as GP with the radial kernel (GPR) and GP with the polynomial kernel (GPP). The optimal GP models were obtained with GPP, with the parameters degree and scale, tuned over 10 values. The optimal values for degree and cost were 3 and 0.01, respectively.

Finally, NNET models were tuned over 15 values for the size (i.e. the number of hidden units) and 15 values for the decay (i.e. the weight decay), which is the parameter in the penalisation method for model regularisation. The optimal values were 13 and 0.01, respectively.

2.4. Predictive model post-processing

ROC curves allow visual analyses of the trade-offs between a predictive model's sensitivity and specificity regarding various probability cut-offs. The curve is obtained by measuring the sensitivity and specificity of the predictive model at every cutting point and plotting the sensitivity against 1-specificity. Figure 2 shows the ROC curves obtained for three of our predictive models, which are SVMR, NNET and GPR. The curve shows that SVMR performs better than other models regarding the evaluation dataset.

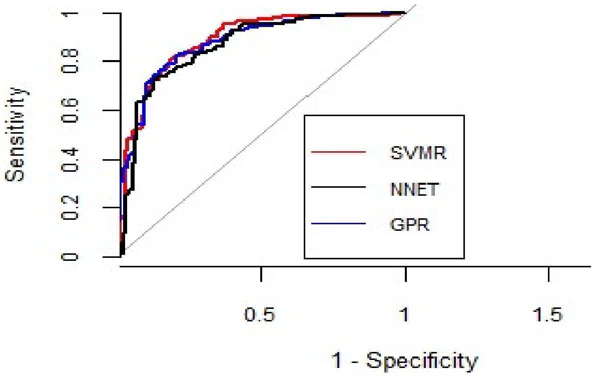

Figure 2. ROC curves for 3 models: SVMR, NNET and GPR.

Several methods exist for finding a new cut-off threshold on the ROC curve. In this study, we find the point on the ROC curve corresponding to the highest accuracy.

2.5. Overall modelling procedure

The overall modelling procedure, which is based on data pre-processing, model optimisation, model post-processing and model evaluation, is inspired by [4] and reformulated and adapted to the context of the present framework. First, the dataset is randomly split, with stratification, in 60% and 40% parts denoted here by P1 and P2, respectively. Then, P1 is used for training and for optimising the model, as explained in Subsection 2.3. Different pre-processing methods that were explained in Subsection 2.2 were appropriately integrated into the cross-validation. Finally, in order to further enhance the model performance, the post-processing and model evaluation methods were applied to the optimised model using k-fold cross testing on the P2 dataset. In the k-fold cross testing procedure, we produce k post-processed model variants of the original optimised model. First, we create k stratified folds of the P2 dataset. Then, k-1 folds are used to find an alternative probability cut-off, that corresponds to the highest accuracy, on the ROC curve. The remaining one-fold is scored with the post-processed model based on the newly found cut-off point. This procedure enhances the predictive models and ensures that the datasets for post-processing and scoring are always distinct.

3. Results

Due to expected potential variations of the predictive models' performance, we conducted extensive repeated experiments simulations to study these variations and the models' stability. The simulations consisted of 500 iterations of the procedure explained in Subsection 2.5. The models' performances concerning accuracy, AUC, sensitivity and specificity were evaluated for each iteration.

The aggregation of all iterations yielded various distributions of the above performance measures. These distributions were then visualised using box plots in Figure 3 to capture the models' performance capability and stability.

Also, estimations of the predictive neural networks' performances regarding means and standard deviation (SD) are shown in Table 2. We report results regarding models which are post-processed with ROC optimisation based on the largest accuracy cut-off method, as explained in Subsection 2.5. The results show that SVMR achieved the best results with a mean accuracy of 0.83 (95% CI [0.79, 0.87]) and a mean sensitivity of 0.87 (95% CI [0.81, 0.93]), similar to the results achieved by GPP. The rest of the predictive models scored a mean accuracy of 81%, which is better than all performances scored at [3].

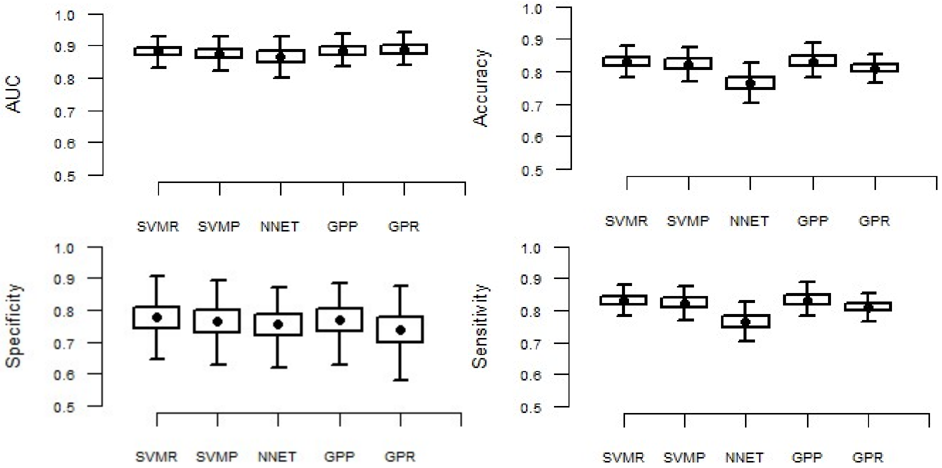

Figure 3. 500 repeated experiments simulations on Support Vector Machines with Radial (SVMR) and Polynomial kernels (SVMP), Gaussian Processes with Radial (GPR) and Polynomial kernels (GPR) and Single Layer Neural Networks (NNET).

Overall, we find that the models, especially SVMR and GPP, have good predictive power and stability, based on an acceptable level of variation in their performance measures evaluated across extensive repeated experiments simulations. Also, the results indicate that the performance differences between the different methods for selecting the ROC cutting points are not significant regarding the 4 performances.

Table 2. Estimations of the predictive model's performances.

Model	Accuracy		AUC		Sensitivity		Specificity	
	Mean	SD	mean	SD	mean	SD	mean	SD
SVMR	0.83	0.02	0.88	0.02	0.87	0.03	0.77	0.04
SVMP	0.81	0.02	0.87	0.02	0.81	0.04	0.80	0.04
NNET	0.81	0.02	0.86	0.03	0.81	0.04	0.80	0.04
GPR	0.81	0.02	0.86	0.04	0.86	0.03	0.73	0.03
GPP	0.83	0.02	0.89	0.02	0.88	0.03	0.77	0.04

After performing the repeated experiments simulations, we further investigated the predictive models in order to better comprehend the predictive power of the cannabis-related attributes over first-episode psychosis via statistical tests. To do so, we re-fit our performing models but removed the cannabis-related attributes, represented in Table 1, from the dataset. Then, we compared the performances of the models built with and without the cannabis-related attributes using t-test. We thereby demonstrated the predictive value of cannabis-related attributes with respect to first-episode psychosis by showing that there is a statistically significant difference between the performances of the predictive models built with and without the cannabis variables.

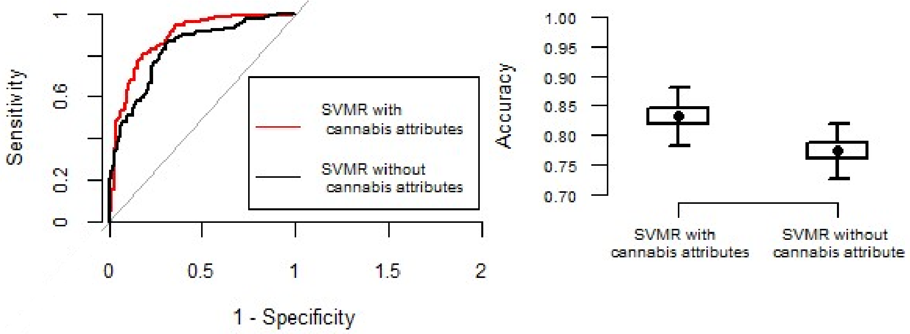

Figure 4. Left: ROC curves for optimised SVMR, with and without the cannabis attributes. Right: boxplots for 500 repeated experiments simulations for optimised SVMR, with and without the cannabis attributes.

Our analysis showed that the accuracy of SVMR decreased by 6% if the cannabis-related attributes were dropped from the process of building the predictive models as shown in the right image in Figure 4. If we compare, for instance, the accuracies of the SVMR models built on the data sets with and without the cannabis use attributes, the p-value obtained for the one-tailed t-test was 0.0002. This means that the predictive models with cannabis attributes have higher predictive accuracy than the models that were built without the cannabis attributes. This leads us to conclude that the additional cannabis variables jointly account for predictive information over first-episode psychosis. These results are consistent with findings from [3]. Also, we demonstrated that there is a significant difference between the ROC curves of the predictive models built with and without the cannabis variables as shown in the lift image in Figure 4. This also confirms the idea that the predictive models with cannabis attributes have higher predictive power than the models that were built without the cannabis attributes.

4. Conclusion

The advent of machine learning has so far proved to be of prime importance and capability in various fields, and recently in medical research and healthcare. This paper proposes a novel computational psychiatry and machine learning framework based on developping predictive models for the onset of first-episode psychosis in presence of clinical data including also cannabis use. We explored three types of machine learning algorithms, namely Gaussian Processes, Support Vector Machines, and feed-forward neural networks. Models are tuned and further optimised via post-processing, and evaluated with a k-fold cross testing methodology. In order to study the variation of the performances of the prediction models, the framework incorporates 500 repetitions of the model building, optimising, testing sequence. Experimental results show that the 3 machine learning algorithms lead to comparable models, with a slight advantage for Support Vector Machines and Gaussian Processes in front of neural networks.

Our best models score an average accuracy of 83%, which is above all the accuracy performances achieved in previous studies such as [3]. This paper extends on previous work as [3] by proposing a new machine learning framework based on a novel methodology in which models are post-processed based on ROC optimisation, and evaluated with the recent method of k-fold cross testing which we adapt after [4]. Moreover, in this new methodology, we developed optimized models with other powerful techniques such as Gaussian Processes and artificial neural networks not addressed in [3]. We also demonstrate statistically that the best models' performance decreases if cannabis attributes are removed from the analysis. This fact is also confirmed and illustrated by ROC analysis.

References

[1] World Health Organization. [online] Available at: http://www.who.int/en/ [Accessed 10 Feb. 2018].

[2] D. Belgrave, R. Cassidy, D. Stamate, et al., Predictive Modelling Strategies to Understand Heterogeneous Manifestations of Asthma in Early Life, 16th IEEE International Conference on Machine Learning and Applications, 2017.

[3] W. Alghamdi, D. Stamate, et al., A Prediction Modelling and Pattern Detection Approach for the First Episode Psychosis Associated to Cannabis Use, 15th IEEE International Conference on Machine Learning and Applications, 2016. pp.825-830.

[4] A. Katrinecz, D. Stamate, et al., Predicting Psychosis Using the Experience Sampling Method with Mobile Apps, 16th IEEE International Conference on Machine Learning and Applications, 2017.

[5] M. Adjouadi, M. Ayala, et al., Classification of Leukaemia Blood Samples Using Neural Networks. Ann Biomed Engineering, 2010. 1473-82.

[6] Y. Yan, X. Qin, et al., A Restricted Boltzmann Machine Based Two-Lead Electrocardiography Classification, in Proc. 12th Int. Conf. Wearable Implantable Body Sens, 2015.

[7] K. Foster, R. Koprowski, et al., Machine learning, medical diagnosis, and biomedical engineering research-commentary, BioMedical Engineering OnLine, 2014. pp. 94-95.

[8] R. Iniesta, D. Stahl, and P. McGuffin, Machine learning, statistical learning and the future of biological research in psychiatry, Psychological Medicine, 2016. pp. 2455–2465.

[9] M. Di Forti, A. Marconi, et al., Proportion of Patients in South London with First-Episode Psychosis Attributable to Use of High Potency Cannabis: a Case-Control Study, The Lancet Psychiatry, 2015.

[10] M. Kuhn, K. Johnson, Applied Predictive Modelling. Springer, 2013.

[11] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2(3), 2002.

[12] N. Qazi, K. Raza, Effect of Feature Selection, SMOTE and Under Sampling on Class Imbalance Classification, 2012 UKSim 14th, 2012. pp. 145-150.

[13] P. Tan, M.Steinbach, V.Kumar, Introduction to Data Mining, 2016.

[14] M. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction, New York: Oxford University Press, 2003.