ABORDAGEM PARA A PREVISÃO DE ABANDONO NUM GINÁSIO
APROXIMACIÓN PARA LA PREVISIÓN DE ABANDONO EN UN GIMNASIO
APPROACH FOR PREDICTING DROPOUT IN A HEALTH CLUB

Pedro Sobreiro *
sobreiro@esdrm.ipsantarem.pt

Abel Santos *
abelsantos@esdrm.ipsantarem.pt

* Escola Superior de Desporto de Rio Maior, Instituto Politécnico de Santarém
Unidade de Investigação do Instituto Politécnico de Santarém (UIIPS) – Portugal

Resumo Resumen Abstract

Este estudo pretende apresentar um modelo para prever o abandono dos clientes num ginásio, baseado em dados existentes no software de gestão Health Center. As variáveis selecionadas, identificadas de acordo com a sua relevância em estudos realizados e disponibilidade de dados, foram: idade, género, tempo de inscrição, média de visitas mensais, faturação realizada ao cliente, número de aulas frequentadas e distância a percorrer para chegar ao clube. O número de clientes utilizados para a previsão de abandono foram de 810, através da utilização de um algoritmo de *Machine Learning Two-class logistic regression* para a classificação. A aproximação realizada permitiu prever com uma exatidão de 83% se o cliente abandonava ou ficava no ginásio. Os resultados obtidos sugerem que pode ser vantajoso a utilização da aproximação realizada para prever o abandono e explorar medidas adicionais para contrariar o abandono de clientes em risco.

PALAVRAS CHAVE: Gestão do desporto; Ginásios; Machine Learning; Previsão de abandono.

...

Este estudio pretende presentar un modelo para prever el abandono de los clientes en un gimnasio, basado en datos existentes en el software de gestión Health Center. Las variables seleccionadas, identificadas de acuerdo con la su relevancia identificada en estudios realizados y disponibilidad de datos, fueron: edad, género, tiempo de inscripción, promedio de visitas mensuales, facturación realizada al cliente, número de clases frecuentadas y distancia a recorrer para llegar al club. El número de clientes utilizados para la previsión de abandono fue de 810, mediante el uso de un algoritmo de *Machine Learning Two-class logistic regression* para la clasificación. La aproximación realizada permitió prever con una exactitud del 83% si el cliente abandonaba o quedaba lo gimnasio. Los resultados obtenidos sugieren que puede ser ventajoso el uso de la aproximación realizada para prever el abandono y explotar medidas adicionales para contrarrestar el abandono de clientes en riesgo.

PALABRAS CLAVE**:** Gestión deportiva; Gimnasios; Machine learning; Previsión abandono.12

...

This study aimed to develop a model to predict customers dropouts in a health club, using existing data in the software Health Center used to manage the health club. The variables selected, identified according to the relevance in studies performed and availability of data, were: age, gender, enrollment time, average monthly visits, customer billing, number of classes attended and distance to reach the club. The number of customers used to develop the dropout prediction were 810, using a Machine Learning algorithm Two-class logistic regression for the classification. The approach adopted allowed to predict with an accuracy of 83% if the client left or stayed in the gym. The results suggest that can be useful tool to predict dropout and to use additional approaches to

counteract clients in risk to dropout.

KEYWORDS: Dropout prediction; Health club; Machine learning; Sport management.

1. Introdução

A competividade existente atualmente exige a exploração de abordagens inovadoras que permitam às organizações ganharem vantagens competitivas, em cidades com muitas organizações similares, onde a competição ainda se torna mais crítica. A rendibilidade no desenvolvimento da atividade é uma preocupação com que os ginásios se deparam, em que a retenção surge como uma área fundamental para o sucesso de qualquer negócio (Hurley, 2004) e para a sua rendibilidade (Ferrand, Robinson, & Valette-Florence, 2010). A retenção é compreendida como uma intenção de voltar a comprar e manter-se membro de um ginásio (Bodet, 2012), sendo um problema com que se deparam os ginásios atualmente (MacIntosh & Law, 2015), onde apenas 50% dos clientes se mantém após o primeiro ano (Emeterio, Iglesias-Soler, Gallardo, Rodriguez-Cañamero, & García-Unanue, 2016; MacIntosh & Law, 2015), aspeto reforçado com os custos menores na manutenção de um cliente em relação à angariação de novos clientes (Ahmad & Buttle, 2002).

A retenção não tem uma solução simples e necessita de várias estratégias (Gonçalves, 2012). Os métodos tradicionais de avaliação da retenção baseiam-se na análise de informação obtida em questionários; em alternativa pode utilizar-se aproximação analítica, com dados obtidos em bases de dados de organizações (Delen, 2010). O *Machine Learning* pode ser utilizado para suportar o desenvolvimento de estratégias de retenção de acordo com os dados existentes (Verbeke, Martens, Mues, & Baesens, 2011), através do estudo e construção de algoritmos que podem aprender com os dados e fazer previsões (Ron Kohavi & Foster Provost, 1998). Esta área apresenta um grande potencial de crescimento, onde 50% das empresas planeia a sua utilização para compreender os clientes (MIT Technology Review, 2017).

Contudo, apesar da sua importância, não temos conhecimento de um estudo que realize a previsão de abandono em ginásio, utilizando uma abordagem baseada no *Machine Learning*; podemos, no entanto, comprovar a sua utilização na previsão de abandono de alunos no ensino (Aulck, Velagapudi, Blumenstock, & West, 2016; Dekker, Pechenizkiy, & Vleeshouwers, 2009; Delen, 2010). Deste modo, a utilização desta abordagem, recorrendo à utilização dos dados gerados pela utilização dos serviços no ginásio por parte dos seus clientes, poderá ser uma aproximação vantajosa. Estes dados podem ser utilizados, periodicamente, para realizar uma previsão de abandono específica por cliente. Assim, a previsão poderá permitir aos responsáveis dos ginásios desenvolverem ações para contrariar o abandono. Considerando que evitar o abandono é muito mais rentável do que angariar novos clientes (Edward & Sahadev 2011).

A eficiência na previsão do abandono está relacionada com a seleção das variáveis preditoras (Hall, 1998), que pode ser suportada em estudos existentes. O género é considerado importante para previsão do abandono (Cooil, Keiningham, Aksoy, & Hsu, 2007; Kamakura & Wedel, 1995) e retenção (Pawlowski, Breuer, Wicker, & Poupaux, 2009), aspeto reforçado com a identificação da diferença entre géneros na realização do exercício físico (Pridgeon & Grogan, 2012). Por outro lado, segundo Pawlowski et al. (2009), a idade surge como um elemento relevante para a compreensão do consumo desportivo. Pawlowski et al. (2009) também consideram o rendimento mensal como uma variável determinante. A rotina na realização da prática desportiva é fundamental para a manutenção do cliente (Ferrand et al., 2010; Pridgeon & Grogan, 2012), aumentando a disponibilidade para realizarem deslocações maiores (MacIntosh & Law, 2015; Pawlowski et al., 2009). Apesar da importância de uma avaliação das variáveis consideradas relevantes, a base de dados existente na organização limita a sua seleção.

Considerando a inexistências de estudos explorando a utilização do *Machine Learning* na previsão do abandono em ginásios, propõe-se uma abordagem inovadora que pode ser realizada para a previsão do abandono dos clientes num ginásio, através da análise dos dados gerados decorrentes da utilização dos serviços do ginásio pelos os clientes.

2. Método

A amostra é composta por 810 clientes de um ginásio que está localizado na região centro do país, no distrito de Santarém. Os dados utilizados para o desenvolvimento do presente estudo foram disponibilizados pelo responsável de um ginásio em três *datasets*: dados de entradas dos clientes, dados dos clientes e dados de pagamentos. Os dados foram obtidos através da exportação da base de dados do *software* de gestão *Health Center* da *PROINF Software*, uma solução informática para gerir ginásios que permite efetuar o controlo de pagamentos e os acessos do ginásio, bem como a gestão de atividades. Foram exportados os registos dos acessos dos clientes, registos dos clientes e registos com a informação dos pagamentos recebidos. Os dados de acessos continham 28680 registos (linhas) e 19 atributos (colunas), representando a informação gravada quando era efetuada uma entrada ou saída de um cliente no ginásio. Os dados dos clientes incluíam a informação referente a 810 clientes (linhas) e 96 atributos referentes a cada cliente (colunas). Os dados dos pagamentos eram compostos por 7783 registos de pagamento e 41 colunas com atributos referentes a cada pagamento efetuado. O tratamento de dados foi realizado com o *Anaconda* e *IPython* (Continuum Analytics, 2016), recorrendo ao *Pandas* (McKinney & Others, 2010) e *NumPy* (Walt, Colbert, & Varoquaux, 2011). A previsão foi realizada recorrendo ao algoritmo de classificação *Two-class Logistic Regression* do Scikit-learn (Pedregosa et al., 2011).

A metodologia usada para o desenvolvimento deste estudo utilizou as seguintes etapas: (1) pré-processamento; (2) extração de variáveis preditoras; (3) construção do modelo e validação; e (4) cálculo da exatidão da previsão.

O pré-processamento consistiu na transformação e remoção de dados redundantes. O passo seguinte consistiu na extração nas variáveis preditoras. As variáveis foram selecionadas de acordo com os dados disponíveis na base de dados e considerando simultaneamente a sua relevância para a previsão do abandono. Os atributos extraídos encontram-se representados na Tabela I. Por último, também foi associado um uma classificação a cada cliente representando se tinha abandonado ou não ginásio até momento que os dados foram extraídos.

Tabela 1. Definição das variáveis preditoras utilizadas no estudo

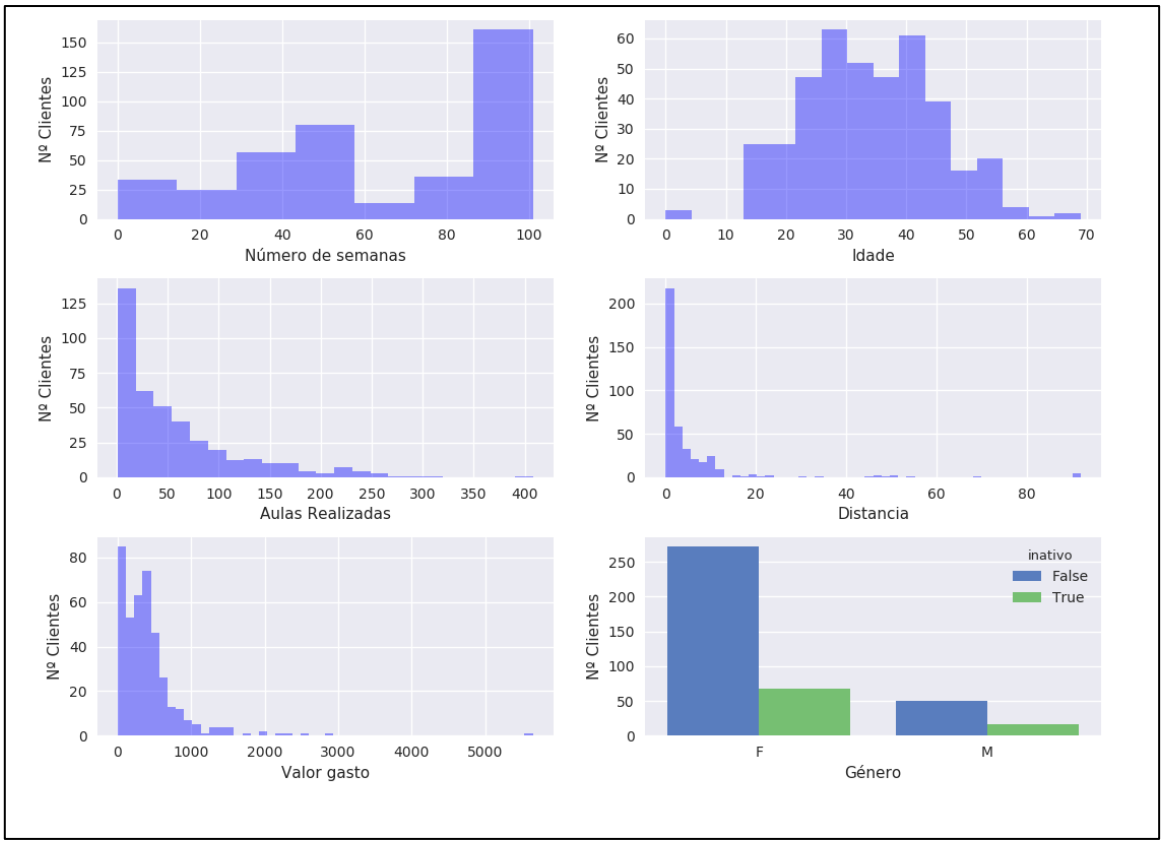
Variável	**Definição**
Idade	Representa a idade do sócio e foi calculada a partir da data de nascimento existente nos dados dos clientes.
Género	Representa o género do sócio e foi obtida a partir dos dados dos clients.
Tempo de inscrição em semanas	Tempo de inscrição do sócio em semanas calculado a partir da data de inscrição disponível nos dados dos clientes e o registo de acessos disponível nos dados de entrada.
Média de visitas mensais	Número de visitas médias por semana, calculadas a partir do total acessos por semana dividido pelo tempo de inscrição em semanas.
Faturação realizada ao cliente	Valor faturado ao cliente desde a sua inscrição, calculado a partir dos dados dos pagamentos.
Número de aulas frequentadas	Número de aulas que o cliente realizou desde que se inscreveu no ginásio, calculado pelo a partir dos dados dos clientes.
Distância a percorrer para chegar ao clube	Distância até ao ginásio. A distância foi calculada a partir do código postal dos dados dos clientes através de georreferenciação num algoritmo desenvolvido para o efeito.
Situação	Abandonou ou não o ginásio.

Depois de ter sido executada a extração das variáveis, foi realizada a construção do modelo preditivo, através da classificação binária do abandono do cliente (abandonou ou não abandonou). Para criar o modelo preditivo, os dados foram divididos em 70% para treinar o modelo e 30% para testar o modelo. Após a aprendizagem do modelo foi realizado o seu teste onde se efetivou o cálculo da exatidão da previsão, confrontando os dados utilizados para treinar o modelo com os dados reservados para testar o modelo.

3. Resultados e Discussão

O tratamento da informação extraída foi complementado com a construção de histogramas (Figura I) e diagramas de dispersão (Figura 2) para visualização dos dados obtidos e análise dos dados.

Figura 1. Histogramas dos dados extraídos



A exatidão do algoritmo *two-class logistic regression* é de 0.826, ou seja, o sucesso na previsão se o cliente abandona ou fica no ginásio é de aproximadamente 83%. Este valor foi calculado a partir dos dados utilizados para testar o modelo confrontando a previsão com a situação real do cliente (se abandonou ou não abandonou o ginásio), conforme podemos verificar na Tabela 2, onde se obtém uma taxa de sucesso de 114 previsões corretas num total de 138. A matriz de confusão (*Confusion Matrix*) representa os *True Positive* (TP - Não abandonou com resultado previsto de não abandono), *True Negative* (TN - Abandonou com resultado previsto de abandonar), *False Positive* (FP - Não abandonou com resultado previsto de abandonar), *False Negative* (FN - Abandonou com resultado previsto de não abandonar).

Figura 2. Diagramas de dispersão das variáveis numéricas, com a representação dos clientes que abandonaram e que foram retidos

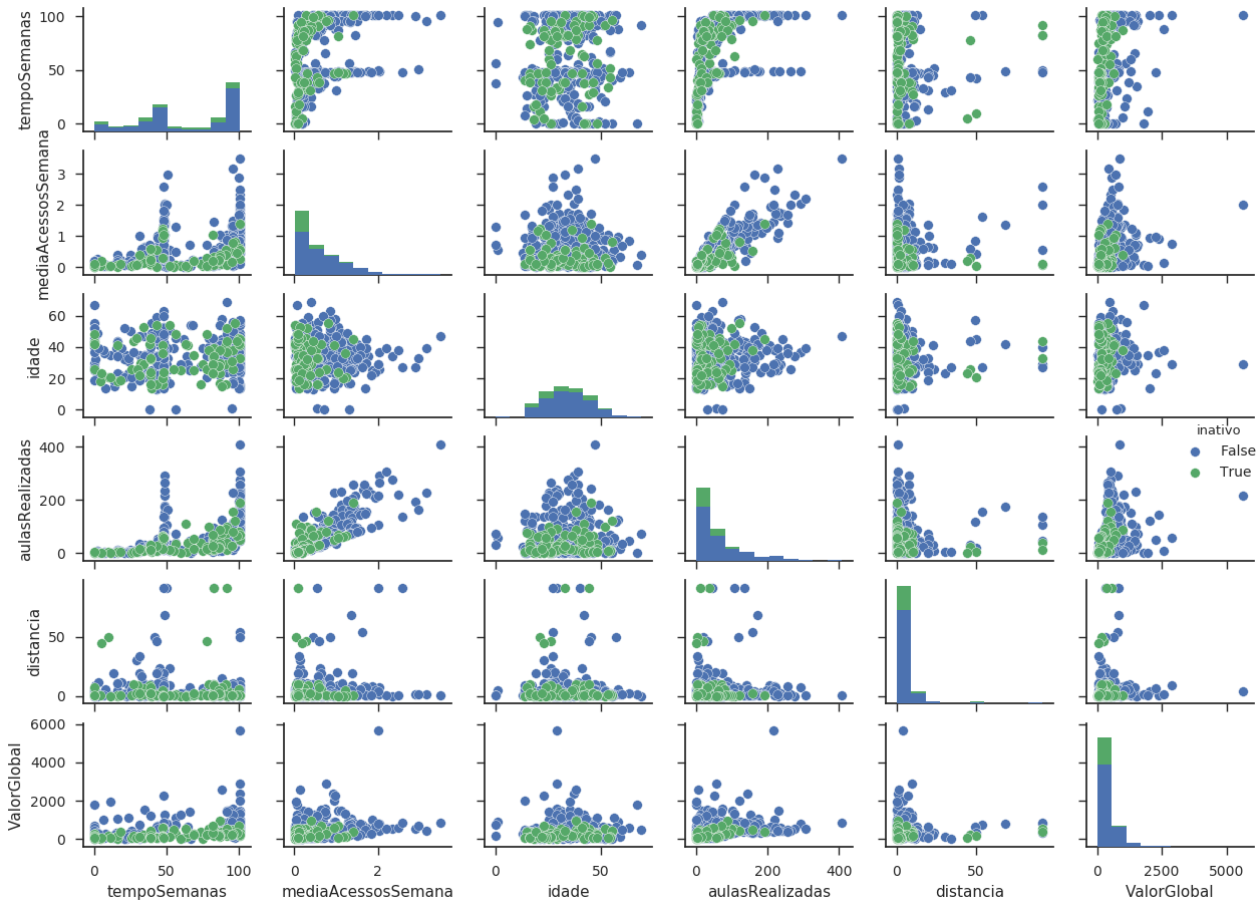


Tabela 2. Matriz confusão com real versus previsto

		Previsto	
		Não Abandonou	Abandonou
Real	Não Abandonou	True Positive 112	False Negative 1
	Abandonou	False Positive 23	True Negative 2

A precisão da previsão calculada com TP/(TP+FP) referente aos clientes que não abandonaram foi de 83% e de 67% para os que abandonaram. A exatidão da previsão do algoritmo de 83% é superior a um estudo desenvolvido para prever o abandono num ginásio por Emeterio et al. (2016), que apresentou uma exatidão de 70%, apesar de não utilizar uma aproximação baseada

em *Machine Learning*, o que demonstra a eficiência deste tipo de aproximação. O modelo foi treinado com 70% dos dados, correspondendo a um total de 320 clientes, que apesar de corresponder a valores aceitáveis (Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012), apenas 73 tinham abandonado. Este valor pode ser insuficiente para treinar o modelo de forma a se obter uma maior exatidão dos resultados. Desta forma, o desenvolvimento da mesma aproximação no contexto de um ginásio com maior número de clientes, pode ser interessante no sentido de testar se a exatidão do modelo ainda pode melhorar. A vantagem na realização da previsão do abandono num ginásio permite que os gestores identifiquem os clientes em risco de abandono e utilizem esta informação para contrariar a deserção dos clientes.

Em face do conhecimento obtido pela pesquisa efetuada, este é o primeiro estudo a aplicar uma aproximação de *Machine Learning* para prever o abandono num ginásio através da utilização da informação existente que regista o comportamento dos clientes, o que permitirá aos responsáveis dos ginásios utilizarem esta aproximação para preverem o abandono de uma forma recorrente sem necessitarem de inquirirem os seus clientes, possibilitando também a determinação da sua probabilidade de abandono. Não existe uma garantia de prever corretamente o futuro testando com dados já passados (abandonou ou não abandonou), aspeto também evidenciado Moseley e Mead (2008), mas os resultados são promissores.

4. Conclusões e Recomendações

O estudo desenvolvido demonstra que as bases de dados podem ser utilizadas para a obtenção de variáveis que permitam prever o abandono sem necessidade de recolher informação adicional. Contudo existe informação que poderia aumentar a qualidade da previsão, como a satisfação do consumidor ou variáveis externas que não estavam disponíveis através da abordagem realizada.

Como futura investigação seria interessante analisar o desempenho de outros algoritmos – tais como: *Two-Class Boosted decision*, *Two-Class Neural Network*, *Two-Class Support Vector* e *Two-Class Decision Forest* – e avaliar se apresentam uma previsão com exatidões superiores na identificação de clientes em risco de abandono. Outra área interessante para investigar, no seguimento deste estudo, seria a segmentação dos clientes recorrendo a algoritmos não supervisionados para a análise de *clusters*. Por último, considera-se que a aplicação do algoritmo na construção de uma solução que permitisse a avaliação da probabilidade de abandono dos clientes ativos do ginásio periodicamente, para um acompanhamento mais próximo dos que apresentam um risco maior de abandono, seria uma ferramenta vantajosa para os responsáveis dos ginásios no que respeita ao aumento da taxa de retenção.

Referências

Ahmad, R., & Buttle, F. (2002). Customer retention management: a reflection of theory and practice. Marketing Intelligence & Planning, 20(3), 149–161. doi:10.1108/02634500210428003

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting Student Dropout in Higher Education. arXiv preprint arXiv:1606.06364. Obtido de https://arxiv.org/abs/1606.06364

Bodet, G. (2012). Loyalty in Sport Participation Services: An Examination of the Mediating Role of Psychological Commitment. Journal of Sport Management, 26(1), 30–42.

Continuum Analytics. (2016). Anaconda Software Distribution. Obtido de https://continuum.io

Cooil, B., Keiningham, T. L., Aksoy, L., & Hsu, M. (2007). A Longitudinal Analysis of Customer Satisfaction and Share of Wallet: Investigating the Moderating Effect of Customer Characteristics. Journal of Marketing, 71(1), 67–83. doi:10.1509/jmkg.71.1.67

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009). Predicting students drop out: A case study. Em Educational Data Mining 2009. Obtido de http://www.educationaldatamining.org/conferences/index.php/EDM/2009/paper/download/1467/1433

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498–506. doi:10.1016/j.dss.2010.06.003

Edward, M., & Sahadev, S. (2011). Role of switching costs in the service quality, perceived value, customer satisfaction and customer retention linkage. Asia Pacific Journal of Marketing and Logistics, 23(3), 327–345. doi:10.1108/13555851111143240

Emeterio, I. C. S., Iglesias-Soler, E., Gallardo, L., Rodriguez-Cañamero, S., & García-Unanue, J. (2016). A prediction model of retention in a Spanish fitness centre. Managing Sport and Leisure, 21(5), 300–318. doi:10.1080/23750472.2016.1274675

Ferrand, A., Robinson, L., & Valette-Florence, P. (2010). The intention-to-repurchase paradox: a case of the health and fitness industry. Journal of Sport Management, 24(1), 83–105.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making, 12, 8. doi:10.1186/1472-6947-12-8

Gonçalves, C. R. N. (2012). Retenção de sócios no fitness: estudo do posicionamento, expectativas, bem-estar e satisfação. Obtido de http://www.repository.utl.pt/handle/10400.5/4853

Hall, M. A. (1998). Correlation-based Feature Selection for Machine Learning.

Hurley, T. (2004). Managing Customer Retention in the Health and Fitness Industry: A Case of Neglect. Irish Marketing Review, 17(1/2), 23–29.

Kamakura, W. A., & Wedel, M. (1995). Life-style segmentation with tailored interviewing. Journal of Marketing Research, 308–317.

MacIntosh, E., & Law, B. (2015). Should I stay or should I go? Exploring the decision to join, maintain, or cancel a fitness membership. Managing Sport and Leisure, 20(3), 191–210. doi:10.1080/23750472.2015.1025093

McKinney, W., & others. (2010). Data structures for statistical computing in python. Em Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). SciPy Austin, TX. Obtido de https://pdfs.semanticscholar.org/f6da/c1c52d3b07c993fe52513b8964f86e8fe381.pdf

MIT Technology Review. (2017). Machine Learning: The New Proving Ground for Competitive Advantage. Obtido de https://s3.amazonaws.com/files.technologyreview.com/whitepapers/MITTR_GoogleforWork_Survey.pdf

Moseley, L. G., & Mead, D. M. (2008). Predicting who will drop out of nursing courses: a machine learning exercise. Nurse Education Today, 28(4), 469–475. doi:10.1016/j.nedt.2007.07.012

Pawlowski, T., Breuer, C., Wicker, P., & Poupaux, S. (2009). Travel Time Spending Behaviour in Recreational Sports: An Econometric Approach with Management Implications. European Sport Management Quarterly, 9(3), 215–242. doi:10.1080/16184740903023971

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. Journal of Machine Learning Research, 12(Oct),

2825-2830.

Pridgeon, L., & Grogan, S. (2012). Understanding exercise adherence and dropout: an interpretative phenomenological analysis of men and women's accounts of gym attendance and non-attendance. Qualitative Research in Sport, Exercise and Health, 4(3), 382–399. doi:10.1080/2159676X.2012.712984

Ron Kohavi, & Foster Provost. (1998). Glossary of Terms Journal of Machine Learning. Obtido 25 de Outubro de 2017, de http://ai.stanford.edu/~ronnyk/glossary.html

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354–2364. doi:10.1016/j.eswa.2010.08.023

Walt, S. van der, Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, 13(2), 22–30. doi:10.1109/MCSE.2011.37

Endereço para correspondência:
Pedro Sobreiro
sobreiro@esdrm.ipsantarem.pt