Assessment of Healthcare Services using Models Based on Support Vector Machines

Anatoli Nachev

Abstract*: This article presents a case study that provides assessment of access to the Irish healthcare system and the services it provides. We explore factors related to unmet heath care needs using recent survey data. Our approach is based on using support vector machines for building predictive models that analyse and measure those factors. The proposed methodology is novel for the domain. Following the behavioural model for access to medical care, we group factors into three categories: predisposing, enabling, and needs, and analyse each group. Experimental results show and measure the primary causes of imbalances and inequalities of treatment in the Irish healthcare system today.

***Keywords*: Classification, Healthcare, Machine Learning, Support Vector Machines.**

I. INTRODUCTION

Aimof this paper is to analyse recent survey data related to the living conditions in Ireland with focus on the healthcare system and services it provides to the citizen. Our target is to estimate factors that affect unmet health care needs and to quantify their role. Further to previously published research in the domain, this study addresses some gaps and proposes an alternative method for data analysis based on machine learning and support vector machines.

According to Wolfe et al. [16], unmet healthcare needs could be explained by as weather individuals who needs access to the healthcare services receive them or not and weather the services are provided in the optimal way. Analysing unmet needs allows to estimate the ability of the healthcare system to address imbalances and inequalities of treatment. Previous research in the domain primarily identifies factors related to unmet needs in minorities or groups of citizens in specific countries and regions [13], [14], [15]. Data is and typically analysed by statistical methods. Literature does not report using support vector machines, proposed here. The rest of the paper is structured as follows: Section II discusses the CRISP-DM methodology that outlines the steps we followed to carry out data analysis. It also discusses support vector machines used to build predictive models. Finally, is presents the metrics used to estimate model performance. Section III discusses the data used, provides details about the models created, and presents experimental results. It also interprets and analyses results. Section IV gives conclusions and summarizes contribution.

Manuscript received on 03 November 2022 | Revised Manuscript received on 11 November 2022 | Manuscript Accepted on 15 December 2022 | Manuscript published on 30 December 2022.

***** Correspondence Author (s)

Dr. Anatoli Nachev*, Business Information Systems, University of Galway, Galway, Ireland. Email: anatoli.nachev@universityofgalway.ie.

II. METHODOLOGY

A. CRISP-DM

One of the most common project methodologies in the fields of data mining, data science, and machine leaning is the Cross Industry Standard Process for Data Mining (CRISP-DM) – a framework introduced by the CRISP-DM consortium in late 1990s [1], [2].

As illustrated in Fig. 1, CRISP-DM defines 6 iterative steps, some of which provide reverse links to previous steps, if needed. In summary, the essential responsibilities and activities defined by those steps include:

- *Business understanding* step requires defining the goals and success criteria in both “business” and “technical” terms.
- *Data understanding* step clarifies the data meaning, sources of data, data quality issues, and provides statistical analysis of data.
- *Data preparation* step includes consolidation of data sources and data pre-processing required to prepare it for building models.
- *Modelling* step deals with data partitioning, establishing model testing strategies, training models, tuning their hyper-parameters, and estimating performance.
- *Evaluation* step requires evaluating results towards success criteria, review the previous steps and work accomplished, and determine if deployment is needed.
- *Deployment* is the final step, if any, responsible for model deployment in business environment and its maintenance.

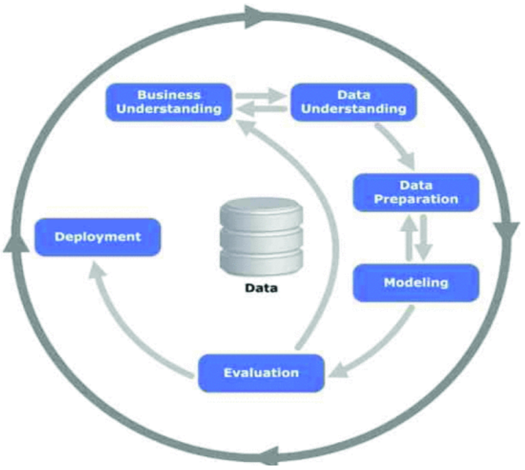


Fig. 1. CRISP-DM life cycle [1].

B. Support Vector Machines

Support vector machines (SVM) is a machine learning technique used for both classification and regression problems, proposed by V. Vapnik [3], [4]. It is based on building a function that maps input patterns to class labels in case of classification or quantitative value in case of regression. Here we consider SVMs used for binary classification. Function that separates between two classes is a linear hyperplane with dimension one less than the original input data. For example, in a two-dimensional input space, function would be a line. Intuitively, the classification function keeps maximal distance to a subset of input points, called support vectors, thus providing maximal separation between the two classes.

Formally, training data can be represented as set of points (1):

$$D = \{(\vec{x}_i, y_i) | \vec{x}_i \in \mathfrak{R}^p, y_i \in \{-1, +1\}\}_{i=1}^{n} \qquad (1)$$

SVM classification function can be represented by (2)

$$f(\vec{x}) = sign(\vec{w}^T\vec{x} + b) \qquad (2)$$

where *w* is weight vector and *b*is a free term.

Over the training, SVM optimizes w, b, and hyper-parameters. This is a quadratic function optimization problem, which in dual form can be represented by finding non-negative Lagrange multipliers α_i that maximize (3):

$$\sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j \qquad (3)$$

Non-zero Lagrange multipliers correspond to data points called support vectors. Also, (2) can be transformed to (4):

$$f(\vec{x}) = sign\left(\sum_i \alpha_i y_i \vec{x}_i^T \vec{x} + b\right) \qquad (4)$$

A technique that enhances SVM performance by making it non-linear is the “kernel trick”. If dot product operationin formulae is replaced by a product of a non-linear function Φ (5), the original input space is transformed to a higher dimensional one and the original class separation hyperplane will be transformed into a higher dimensional one.

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i)\Phi(\vec{x}_j) \qquad (5)$$

As Fig 2. illustrates, in the original 2D input space, class separation by 1D line wouldn’t work well as the two classes are best separable by a curve. If, however, the kernel function applies, a 2D hyperplane built by SVM would separate the classes better in 3D input space.

This study experiments SVM models with five kernel functions: linear $K(\vec{x}_i, \vec{x}_j) = \vec{x}_i^T\vec{x}_j$, polynomial $K(\vec{x}_i, \vec{x}_j) = (\gamma\vec{x}_i^T\vec{x}_j + r)^d$, Gaussian (a.k.a. RBF) $K(\vec{x}_i, \vec{x}_j) = exp(\sigma\|\vec{x}_i - \vec{x}_j\|^2)$, sigmoid $K(\vec{x}_i, \vec{x}_j) = exp(\sigma\|\vec{x}_i - \vec{x}_j\|^2)$, and Laplacian $K(\vec{x}_i, \vec{x}_j) = exp\left(\sigma\|\vec{x}_i - \vec{x}_j\|^2\right)$.

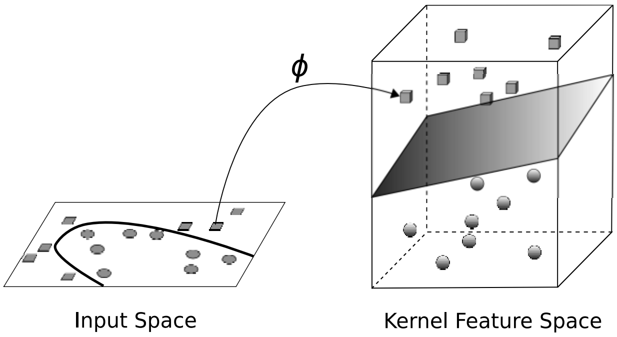
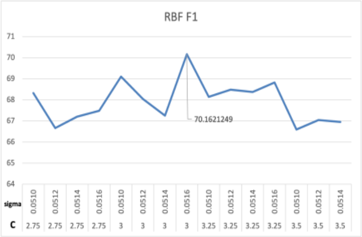
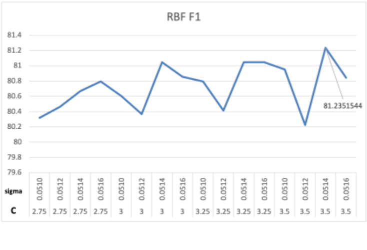
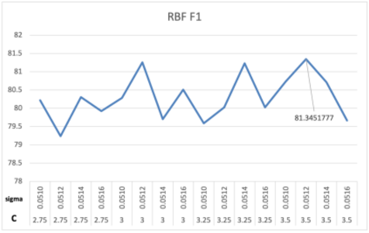


Fig. 2. Kernel function Φ maps input space into a higher dimension, which makes classes better separable by hyperplanes.

In summary, with reference to the CRISP-DM modeling step, we choose SVM for modeling as it is broadly considered now as one of the best machine learning algorithms, outperforming many other [4].

<table>
<tr><th>Predisposing factors</th><th>Enabling factors</th><th>Needs factors</th></tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
</table>

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

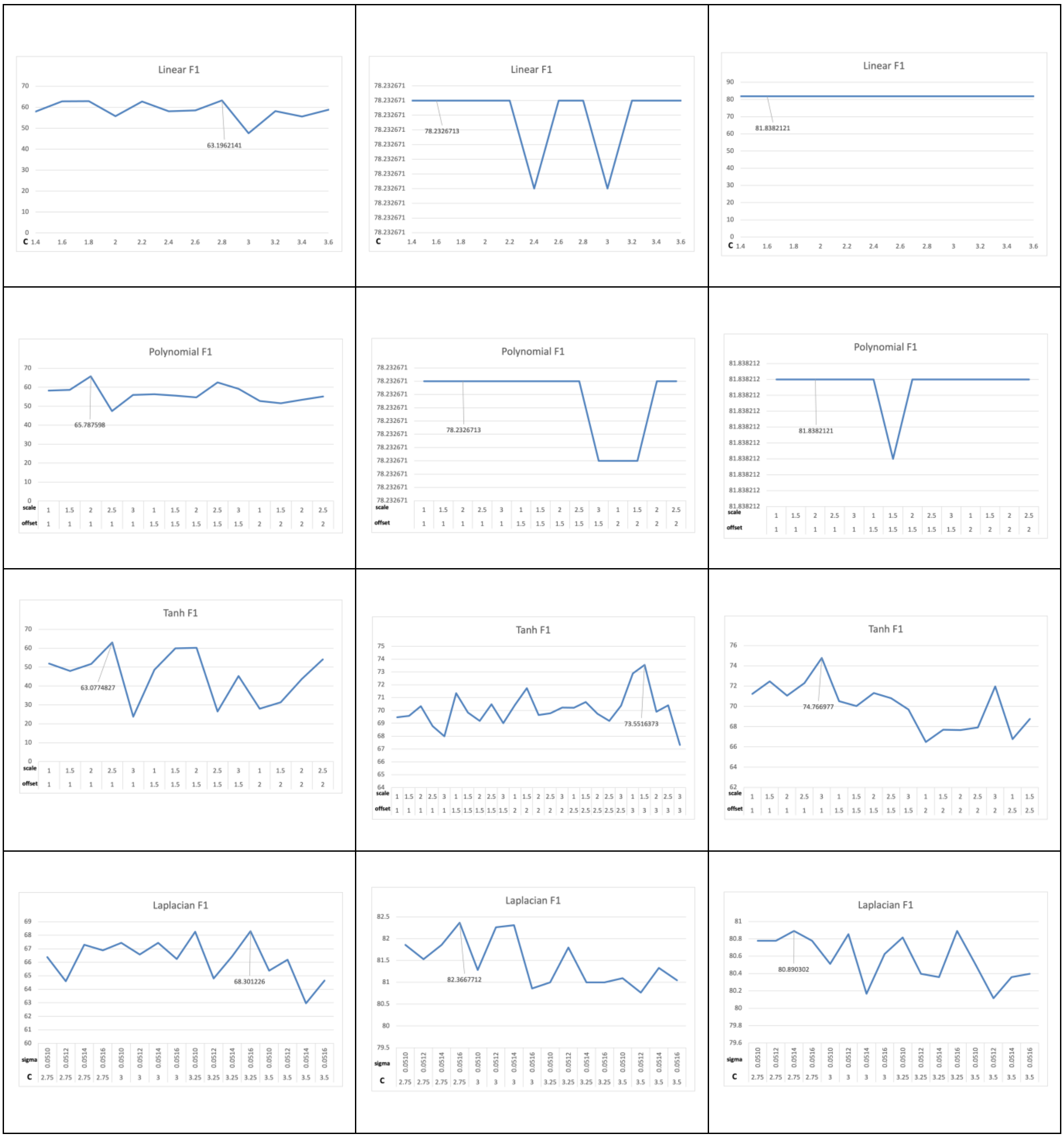


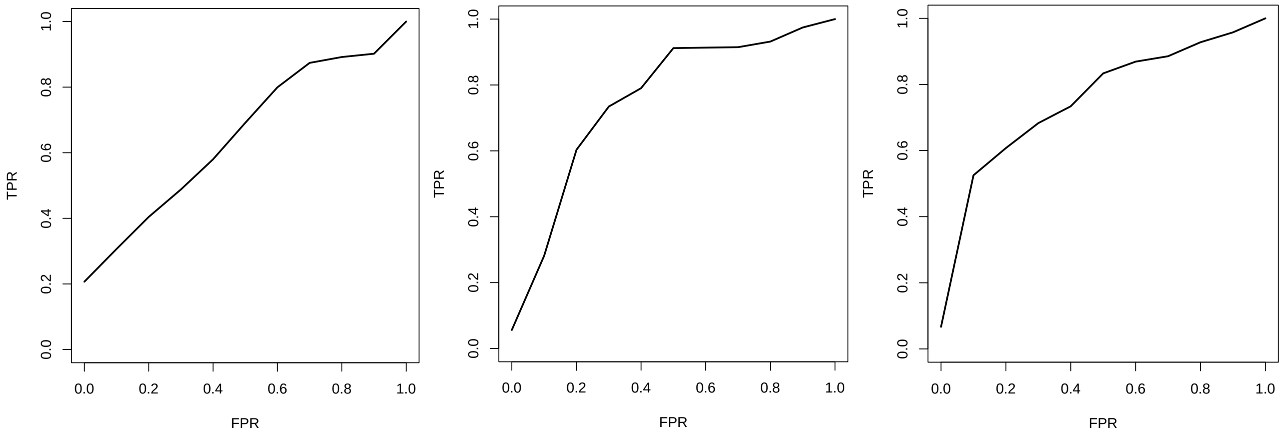
Fig. 3.SVM models performance using RBF, linear, polynimial, hyperbolic tangent, and Laplacian kernels. Primary performance metric is F1.

C. Metrics

Building up a well performing binary classifier requires iterative tuning of SVM hyper-parameters and measuring model performance. By applying test observations having known correct class labels, there are four possible outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions, all summarized in confusion matrix. Model performance can be estimated by various metrics, most common of which is accuracy: *Acc*=(*TP*+*TN*)/(*TP*+*TN*+*FP*+*FP*). It however is not meaningful if dataset is unbalanced, i.e., two classes are unequally presented. As the target variable of our data contains 3.2% “yes” and 96.8% “no” for unmet health care needs, we are in that scenario. Proper estimation of model performance involves other metrics, such as $precision = TP/(TP + FP)$, $recall = TP/(TP + FN)$, and $F1$ that represents a balance between precision and recall as their harmonic mean,

$$F1 = 2 * precision * recall/(precision + recall).$$

We also did ROC analysis on test results by plotting ROC curves and computing area under the curve (AUC) as performance metric [5], [6]. AUC advantage over other metrics is that is provides a single scalar value for model performance over all operational points that map statistical scores to class labels. Among all mentioned metrics we deem F1 as primary for comparing models built on unbalanced data, but also take into consideration AUC, precision, recall, and Acc. The modelling stage also requires to setting test strategy in order to address the performance variance caused by randomness in the modelling process. To avoid overestimation caused by “lucky set composition” we applied 5-fold cross-validation and iterated modelling 10 times for each instance.



(a) (b) (c)

Fig. 4.ROC curves of SVM models trained by:(a) predisposing data with parameters: RBF kernel, C=3, sigma=0.0516; (b) enabling data with parameters: Laplacian kernel, C=2.75, sigma=0.0516; (c) needs data with parameters: linear kernel, C=2.

III. EXPERIMENTS AND DISCUSSION

A. Data

This research explores unmet health care needs and factors related to that. Data is cross-sectional and collected by a national survey, available by the EU-SILC datasets [7], [8]. Combining personal and household data, we constructed and pre-processed a dataset of 34 attributes, grouped into three categories of factors for access to medical care: pre-disposing, enabling, and needs, the meaning of which is outlined in [9]. The dependent variable shows if health care needs are met or not and classes distribution is unbalanced, as discussed above.

B. Modelling

With reference to the modelling stage of CRISP-DM, we built a number of models based on SVM using R packages for machine learning [10], [11]. Three groups of models were created, in accordance with the three categories of factors in data. This approach was preferred in order to explore in detail how each group of factors contributes to the unmet health care needs. Results led to the conclusion that using one kernel for all three categories of factors cannot provide optimal results and selecting specific for each case is a better approach. Using the default mechanism for selecting SVM hyper-parameters, we found that maximum F1 can be achieved by using RBF kernel for the predisposing model, Laplacian for the enabling one and linear kernel for the needs model. A series of experiments aimed to fine-tune each model hyper-parameters and results are outlined in Fig. 3. Best F1 results were obtained as follows: predisposing model with RBF kernel, C=3 and sigma=0.0516; enabling model with Laplacian kernel, C=2.75 and sigma=0.0516; needs model with linear kernel C=2.

Table-I: SVM performance metrics.

Metric:	Healthcare factors		
	Predisposing	Enabling	Needs
Acc	54.90%	70.68%	69.92%
AUC	0.657	0.787	0.741
Precision	98.07%	98.78%	97.97
F1	70.16%	82.37%	81.84%

C. Analysis and Discussion

Analysing the three SVM models we aimed to see how each group of factors contributes to the unmet healthcare needs and how much. Exploring the models performance allows to make conclusions on that. Fig. 4 shows ROC curves for the models and Table-I provides summary of performance metrics. It is evident that the ROC curve of the enabling model shows largest proximity to the top left corner (that represents the best model performance over all operational points), followed in proximity by the needs curve and then by the predisposing curve. This allows to rank the groups of factors, having the enabling factors on top, followed by needs, and lastly by the predisposing ones. This conclusion is also confirmed by the figures in Table-I, especially considering F1 as primary metric. Another series of experiments aimed to explore importance of each factor within the three groups in order to better understand how factors individually unmet health care needs. Applying sensitivity analysis [12] on the variables of each group, we ranked their importance in distinguishing between the two classes. Fig. 5 differentiates Earnings Month with 40% importance as primary contributor among all other predisposing factors, and suggesting that the financial status of the citizen is essential for having access to the healthcare services. Fig. 6 shows that the top three most important enabling factors are Make Ends Meet, Fin Burden Med, and Spend Weekly, having importance between 16% - 22%. These three also suggest that the financial status of the citizen, being related to the access to the paid medical services is most important. Fig. 7 discloses that the general health condition of the individuals is most important needs factor determining the level of health care needs, measured as to 43%. In summary, the data analysis conducted by training SVM models with selection of kernels and hyper-parameters reveals that the most significant factor of having an expected level of and timely access to the Irish healthcare system is affordability to paid services, provided mostly by the private healthcare institutions.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

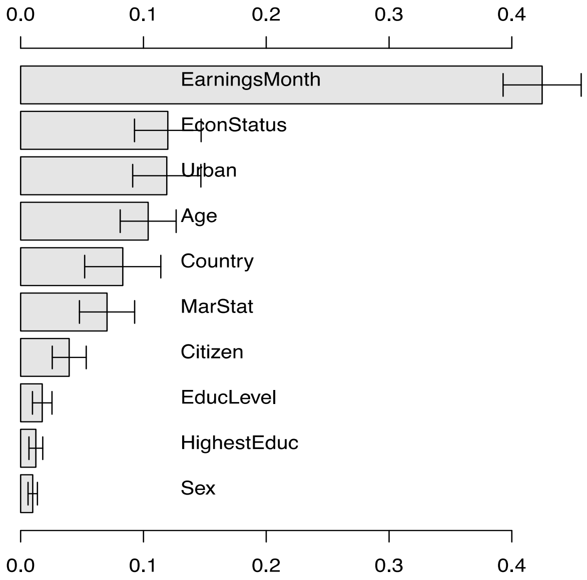


Fig. 5. Importance of predisposing factors.

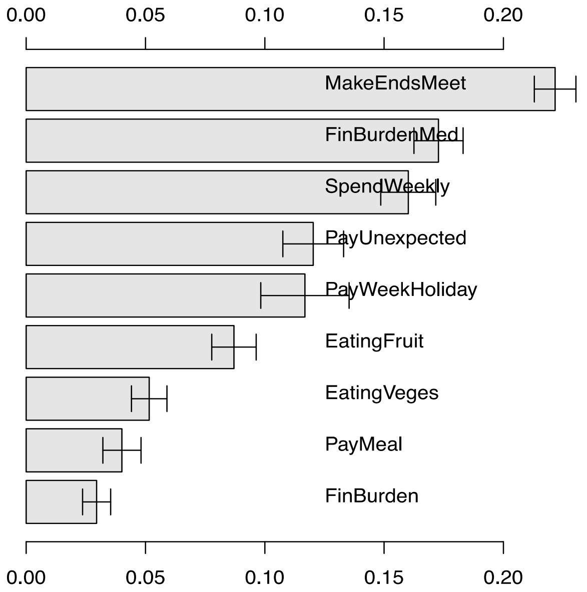


Fig. 6.Importance of enabling factors.

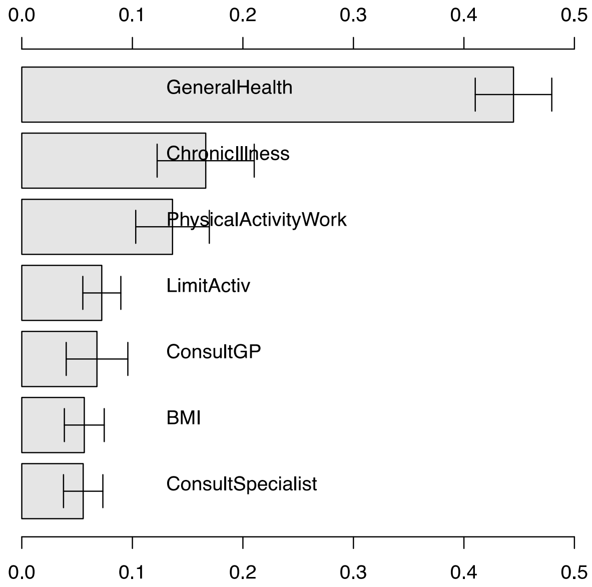


Fig. 7.Importance of needs factors.

IV. CONCLUSION

This study presents analysis of data collected by the EU Statistics on income and living conditions in Ireland. Addressing gaps in methodologies and results reported by previous research in domain, we propose an alternative way for analysis based on SVM models. Exploring factors that affect unmet health care needs we Categorised them into three groups, predisposing, enabling, and needs, as suggested by the behavioral model for access to medical care. Training different SVM models for each group we concluded that they should have specific kernel and selection of hyper-parameters to show optimal performance. Exploring importance of factors that cause unmet health care, we concluded that the Irish healthcare system does not serve satisfactory and timely patients who can't afford paid medical services or having access to private healthcare institutions by voluntary health insurance. A contribution of this study is that the aforementioned conclusions are derived and quantified by a formal method based on data and machine learning techniques.

REFERENCES

1. J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," *2021 IEEE International Conference on Big Data*, 2021, pp. 2337-2344. [CrossRef]
2. C Pete, C Julian, K Randy, K Thomas, R Thomas, S Colin et al., Crisp-DM 1.0, CRISP-DM Consortium, 2000.
3. C. Cortes, V. Vapnik, "Support-vector networks.", Machine Learning, 1995, 20(3), pp. 273-297. [CrossRef]
4. Wang L (Ed.), "Support vector machines: theory and applications.", Springer Science & Business Media, 2005. [CrossRef]
5. Fawcett, T., "An introduction to ROC analysis."*Pattern recognition letters*", 2006, 27(8), pp. 861-874. [CrossRef]
6. Flach, A., "ROC analysis.", *Encyclopedia of machine learning and data mining*, Springer, 2016, 1-8. [CrossRef]
7. European Commission. "EU-SILC User Database Description." Luxembourg: European Commission, 2020.
8. Kotzeva, M. (Ed.), "Eurostat regional yearbook", Publication Office of the EU, 2021, pp. 81-98.
9. Andersen R. "Revisiting the Behavioral Model and Access to Medical Care: Does it Matter?", Journal of Health and Social Behavior, 36:1, 1995. [CrossRef]
10. R Core Team, "R: A language and environment for statistical computing." [Online] Available: http://www.R-project.org/ (URL).
11. Cortez, P., "Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool", ICDM, Springer, LNAI 6171, 2010 pp. 572– 583. [CrossRef]
12. Kewley, R.,Embrechts, M., Breneman, C., "Data strip mining for the virtual design of pharmaceuticals with neural networks," IEEE Transactions on Neural Networks, 11 (3), 2000, pp. 668–679. [CrossRef]
13. S. Connolly, S., Wren, M. "Universal Health Care in Ireland—What Are the Prospects for Reform?", Health Systems & Reform, 2019, 5:2, 94-99. [CrossRef]
14. Connolly, S., Wren, M. "Unmet healthcare needs in Ireland: Analysis using the EU-SILC survey", Health Policy, 2017, vol.121, 434–441. [CrossRef]
15. Popovic N., Terzic-Supic Z., Simic S., Mladenovic B. "Predictors of unmet health care needs in Serbia; Analysis based on EU-SILC data.", PLoS ONE, 2017, 12(11). [CrossRef]
16. Elinson, J. (Ed.) "Health Goals And Health Indicators: Policy, Planning, And Evaluation", 2019. [CrossRef]

AUTHORS PROFILE

Dr. Anatoli Nachev, received his PhD degree at the BAS, Institute of Math and Informatics, section Artificial Intelligence. He received his MSc and BSc degrees in Sofia University, Faculty of Math and Informatics. He is currently a lecturer at Business Information Systems, J. E. Cairnes School of Business & Economics, University of Galway, Ireland. Research interests include business intelligence, business analytics, predictive modeling applied to business processes and business behaviors, machine learning algorithms, data mining focused to business cases, artificial intelligence, etc. He has numerous publications in books, international journals, and conferences in the fields of interest.

RetrievalNumber:100.1/ijeat.B39051212222
DOI: 10.35940/ijeat.B3905.1212222
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)