Texila International Journal of Public Health
Volume 5, Issue 4, Dec 2017

Predicting the Length of Stay among Healthcare Workers in Underserved Communities: A Quantitative Retrospective Cohort Study

Article by Sangiwe Moyo[1,5], Tuan Nguyen Doan[1,2], Jessica A. Yun[3], Ndumiso Tshuma[4,5]
[1]Africa Health Placements, Africa Health Placements, Rosebank, Johannesburg, South Africa
[2]Yale University, New Haven, Connecticut, United States of America
[3]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, South Africa
[4]Regent Business School, Durban, South Africa
[5]School of Public Health, Texila American University, Guyana, South America
E-mail: sangiwemoyo@gmail.com[1]

Abstract

Background: *While prior studies have identified a number of demographic factors related to general health practitioners' decision to stay in public health practice, recruitment agencies have no validated methods to predict how long these health workers will commit to their placement. We aim to use machine learning methods to predict health professional's length of practice in the rural public healthcare sector.*

Methods: *Recruitment and retention data from Africa Health Placements (n=13 698 with 1 838 completers) was used to development machine learning models to predict health workers' length of practice. A cross-validation technique was used to validate the models, to evaluate which model performs better, based on their respective aggregated error rate of prediction. Length of stay was categorised into 4 groups (less than 1 year, less than 2 years, less than 3 years, and more than 3 years). Three machine learning models were trained and used 10-fold cross validation techniques to attain evaluative statistics.*

Results: *The three models attain almost identical results, with negligible difference in accuracy. The 'best'-performing model (Multinomial logistic classifier) achieved a 47.34% [SD 1.63] while the decision tree model achieved an almost comparable 45.82% [SD 1.69]. The three models achieved the average AUC of approximately 0.66 suggesting sufficient predictive signal at the four categorical variables selected.*

Conclusions: *Machine learning models give us an effective tool to predict the recruited health workers' length of practice. These models can be adapted beyond the scope of demographic information such as information about placement location and income. This modelling will also, allow strategic planning and optimization of public health care recruitment.*

Key message

Human resource planning in healthcare can employ machine learning to effectively predict length of stay of recruited health workers who are stationed rural areas.

Introduction

The lack of health workforce is a global crisis which numerous countries have proposed and implemented interventions plans to mitigate this problem [12]. However, there is limited data regarding the impact of these interventions and their sustainability over a long period of time. The loss of health care workers in South Africa cripples the pre-existing delicate health system [3]. As a result, the retention of health workers is essential for the health care system performance. Moreover, the recruitment of health workers should not only focus on nurses and physician, but also on community health workers (CHWs) to help the primary health care systems boost the coverage and address the basic health needs of societies [3].

Health systems in sub-Saharan Africa (SSA) face a serious human resources crisis, with recent estimates pointing to a shortfall of more than half a million nurses and midwives needed to meet the Millennium Development Goals of improving the health and wellbeing of the SSA population by 2015

[4] One of the reasons for this phenomenon is due to brain drain of health professionals [1,5]. Most countries have altered their retirement age in order to extend the working life of their staff. Furthermore, Botswana and South Africa have recruited from other countries within and outside the continent. Currently, there are various frameworks both locally and internationally. However, the effectiveness of these interventions is yet to be seen [6,7]. Another limitation lies in the monitoring and evaluation of these frameworks. The currently available data is fragmented, unreliable, inadequate, and not comparable nationally or internationally, better databases and collection systems still need to be developed[1,2,7].

There is a shortage of health professionals in South Africa hence the need to invest in attracting and retaining staff that stay long. Migration of health workers from low- and middle-income countries (LMICs) to high-income countries is one of the most controversial aspects of globalization, having attracted considerable attention in the health policy discourse at both the technical and political level [1,8]. Recruiting health workers for LMICs decapitates the health systems of those countries and they often spend a significant amount training health care workers only to lose them at the end. To make matters worse there are no alternatives for the population to seek health from the private sector or next health facility as these may be hundreds of kilometres away or too expensive.

Universal access to good quality care and optimal patient safety is the goal of health systems and governments all over the world. Even though developed countries have made significant achievements towards attainment of this goal, many developing countries in Africa lag behind due to financial, material and human resource constraints[9].

Turnover in the health workforce is a concern as it is costly and detrimental to organizational performance and quality of care. Some studies have focused on the influence of individual and organizational factors on an employee's intention to quit [10]. High turnover rates have great implications not only for the quality, consistency and stability of services provided to people in need, but also for the working conditions of the remaining staff such as increased workloads, disrupted team cohesion and decreased morale [11,12]. A variety of individual and organizational factors have been found to impact turnover [10].

Despite the fact that financial incentive is not the only factor, there is no doubt that it plays a vital role in the migration decision-making process [5]. A World Health Organization (WHO) study of four African countries showed the major reasons behind health worker migration are better salary, safer environment, living conditions, lack of facilities, lack of promotion, and heavy workloads [57]. A better life and revenue are the primary source of choices to migrate [513]. One of the mentioned obstacles to migration was language barrier, which was the basis of patient care[14,15]. Patients express their distresses by describing their symptoms and pain, and report changes in health status to professionals who need to comprehend. Nurses or doctors must be able to communicate with one another, other members of the health care team, and patient families. They need the current and technical language fluency to communicate under stress and duress [5].

The misdistribution of health workers between urban and rural areas is a policy concern in virtually all countries. It prevents equitable access to health services, contribute to increased health-care costs and underutilization of health professional skills in urban areas, and is a barrier to universal health coverage. To address this long-standing concern, the WHO has issued global recommendations to improve the rural recruitment and retention of the health workforce [16].

The supply of foreign health practitioners is pivotal to the delivery of health care in rural and remote areas of South Africa. A study has shown that 84% of South African population uses public healthcare, served by only 30% of the trained and certified doctors[17]. In a larger context, Sub-Saharan Africa faces severe lack of healthcare workers, with only 3% of the world's total medical staff while facing 24% of the global burden of disease [7]. The arrival of foreign medical workforce reduces the maldistribution of physicians in South Africa, improving access to healthcare to people in rural areas [7].

To date, greater effort has focused on recruitment, with significantly less attention to medical workforce retention. A challenge to improve health access in rural areas is to maintain high retention rate of the health workforce. Currently, there are few existing empirical studies regarding the factors that influence the length of stay ([1] [2]). Previous attempts to identify these factors mainly focus on health worker satisfaction at medical facilities and retention strategy of staffing agencies. However, from our study we recognize that demographic factors, such as nationality or marital status, could significantly influence the length of practice of foreign practitioners.

This paper aims to develop a predicting tool for the length of practice of foreign health care workers, given their demographic information. Machine learning methods are well-suited for this challenge. Rather than traditionally considering the effect of one demographic variable on the length of practice, machine learning examines all potential predictors simultaneously in an unbiased manner, and identifies pattern of information that are useful to make prediction.

Methods

Study design

A quantitative retrospective cohort study was conducted using secondary data, collected from the Africa Health Placements (AHP)

Study setting

South Africa Health, healthcare worker population in underserved communities and distribution and retention levels. AHP recruits foreign and locally qualified health professionals to be placed in underserved communities in South Africa. Underserved areas like rural areas often face challenges in recruiting and retaining health workers, government has responded with programmes like compulsory community service and rural allowance to address this challenge.

Data acquisition

With a large recruitment and retention dataset from AHP, we programmed three machine learning predictive models using relevant demographic data. We evaluated the model performance by doing 10-fold cross-validation. The aim was to choose a model that performs significantly better in predicting length of practice.

Longitudinal individual health worker records are maintained at AHP. These health workers included professionals from South Africa and the rest of the world seeking employment in underserved facilities in South Africa. Data was collected using two methods (i) customised online portal completed by Health Care Workers (HCW) and (ii) Interviews by Recruitment Officers through email, skype and telephonic conversations. Data were captured onto a database and customer management system called Docwize. The online portal is available at the AHP website as a contact form. Once registered, the HCW receives login details to complete their application on Docwize. This system allowed them to input personal and professional information as well as upload certificates, which would then be verified with the respective regulatory authorities while keeping them informed on the next steps until they secured a job offer. The HCW had an option of completing the application online or supplying the details to the Recruitment Officers who then updated the system. It takes an average of 18 months to complete the recruitment process, 75% of the HCW were discouraged by the regulatory delays resulting in incomplete data. The length of stay was continuously monitored during their employment contract, emails and telephonic contact to establish their last date of employment at a particular facility.

Statistical analysis

Dataset description and manipulation

We took a complete cases approach, using only data from successfully recruited health workers without missing observations. The Africa Health Placements dataset contains 62 variables and 13 698 entries, in which there were 2 079 successfully recruited practitioners. Among these 2 079 professionals, some chose not to provide personal information such as Marital Status or Gender. After data cleaning, there were 1 838 entries with completed fields to meet the requirements of this study.

[1] Here: http://etd.uwc.ac.za/xmlui/handle/11394/3470

[2] Here: https://www.mja.com.au/journal/2002/176/10/workforce-retention-rural-and-remote-australia-determining-factors-influence?inline=true

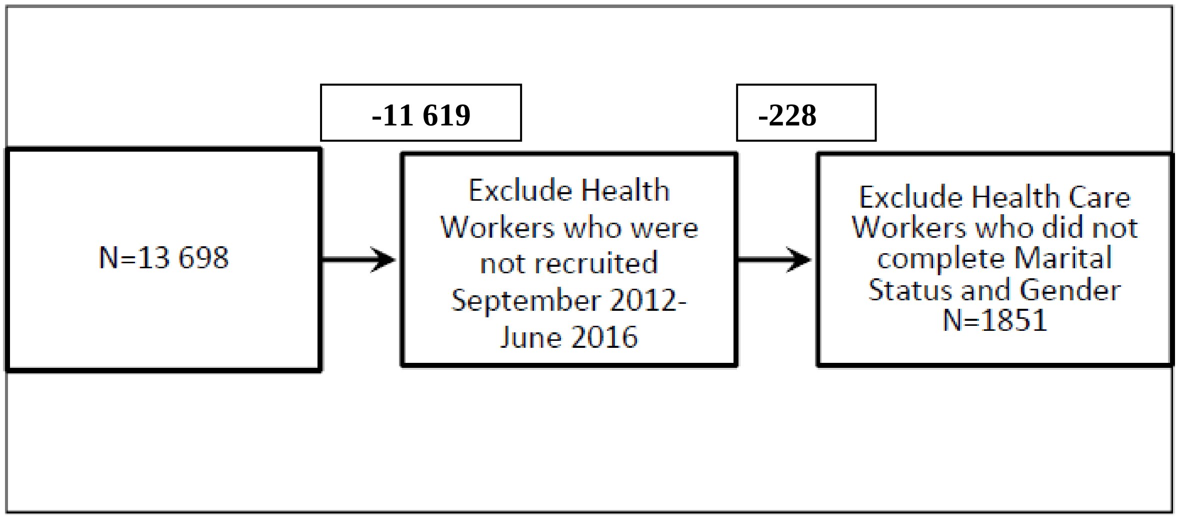
-11 619
-228
N=13 698
Exclude Health Workers who were not recruited September 2012- June 2016
Exclude Health Care Workers who did not complete Marital Status and Gender N=1851

Figure 1. The sample

The variables that are used to develop our machine learning models are chosen based on their availability in the AHP data system. They are Nationality, Profession, Relationship, and Gender. Since there are a lot of missing values in our Age variable dataset, a complete case approach with Age could have further reduced our dataset to merely 914 entries and undermine the ability of the model to learn from existing data. Hence we opted to exclude it from the final analysis. Notably, all of our 4 predictors are categorical variables. A challenge with having categorical variables in machine learning is that to fully represent each variable, we have to use a large number of dummy variables to represent each level within the variable. For example, since our data had records from 145 countries, we needed 144 dummy variables to represent all the existing countries. This method would result in a very sparse dataset and usually not useful in predictive modelling. Hence, we transcribed each variable as follows:

Nationality: categorical data of 145 different countries represented. Instead of recording Nationality as it is, the Nationality variable is transcribed based on World Bank's classification of countries into 4 categories: Low Income, Lower Middle Income, Upper Middle Income, and High Income.

Professions: categorical data of 22 different registered professions, recorded into 3 different categories: doctor, nurse and other

Gender: categorical data of 2 level: male and female

Relationship status: categorical data of 3 levels: married, single or other.

Machine learning model development

As shown on Table 1 three different machine learning classification models were used to train the dataset: Multinomial Logistic Regression, Decision Tree and Naives Bayes Classification. The issue was approached as a classification, rather than a regression problem. The use of regression method was not optimal in this case, due to (i) the lack of quantitative numerical variables in our demographic information, (ii) the wide range of value of the dependent variables (length of practice measured in days), and (iii) the non-continuous nature of the dependent variables. A regression method would require a much larger dataset to arrive at a model of relatively acceptable fit. With our current available dataset, the fit is approximately 18% with high internal sum of squares. Moreover, in strategic workforce planning, a precise prediction of the length of practice in days (or months) is generally not expected. A prediction of whether a specific healthcare worker will stay for 1 year, 2 year or longer is usually acceptable.

Cross-validation

To decide which of the 3 models perform best, we would have to see how well they predict new, unseen data. A challenge to our research was the lack of test data which we could have used for model testing and evaluation. Splitting our existing data into a 80/20 ratio – 80% of the data for training, 20% for testing- was an option, but not optimal as we wanted to use all data available for training.

We constructed and examined our 3 models with repeated 10-fold cross-validations, which partitioned the original sample into 10 disjoint subsets which used nine of those subsets in the training process, made predictions about the remaining subset, and record the misclassification error. To avoid opportune data splits, we average misclassification error across the 10 folds. A comparison between the average misclassification errors of the 3 machine learning models allowed us to decide which model performs the best when facing unseen data.

Ethical approval

Permission to conduct the study was obtained from Africa Health Placements. Data provided did not have any names of the healthcare workers that had been interviewed.

Results

Three machine learning models were trained and used 10-fold cross validation techniques to attain evaluative statistics. The three models attain almost identical results, with negligible difference in accuracy. The 'best'-performing model (Multinomial logistic classifier) achieves a 47.34% [SD 1.63] while the decision tree model achieves an almost comparable 45.82% [SD 1.69].

Multiclass Area Under the Curve (AUC) was computed by building multiple Receiver Operating Characteristic (ROC) curves (one class versus another) and taking the average AUC, as defined by Hand and Till [18]. The three models achieve the average AUC of approximately 0.66 (Multinomial logistic at 0.6652, Decision tree at 0.6635, Naive Bayes at 0.6602), suggesting sufficient predictive signal at the four categorical variables selected.

Overall, the three models had significant accuracy in classifying the length of stay of healthcare workers (p-value< 2.2e-16) see table 1. Additionally, Kappa statistics measures how much better each of the classifiers is performing over the performance of a classifier that simply guesses at random according to the frequency of each class see Table 1. The Cohen's Kappa statistics of the Multinomial Logistics, Decision Tree and Naive Bayes are 0.2658, 0.2649, 0.2521 respectively as shown on Table 1, suggesting a fair (but not substantial) agreement according to Landis and Koch [19]between prediction and response corrected by the agreement expected by chance.

All three models perform reasonably well at identifying those who are likely to stay for less than 1 year as shown in Table 3 The sensitivity of this class was greater than 75% for all three models, showing than that they correctly identify more than ¾ of those who are likely to stay less than 1 year as shown on Tables 3. Specificity of this class is not particularly high (all lower than 65%), so all three models do not do as well in identifying those who are staying for more than one year. However, with negative positive rate as high as 84%, it means that when the model negatively classifies a person out of those who stay for less than 1 year, such classification is likely to be correct.

In contrast, all three models perform poorly at identifying those who are staying between 2 and 3 years (Table 3). With sensitivity goes down as low as 0% (decision tree) and specify goes up to 100%, the three models must have learned to negatively assign a majority (all in decision tree case) out of this class. This is likely the result of imbalanced data sample with considerably too little sample data of this class see Figure 4.

In figure 5, Interesting, this graph looks a little bit different from the previous graph. It points to an observation: we are not very successful in recruiting from some country, but once we do, they tend to stay for a very long time, take Russia for example, however, if we look at the sample size, this indication does not seem to be very reliable. Some countries have very high average length of stay, simply because we have a very small sample size of them.

In table 2, more males 981 (53%) than females 870 (47%) were recruited. Males stay longer than females by 187, 69 days. South Africa had the greatest number of health workers constituting 381 (41%), followed by United Kingdom 361 (39%), Nigeria 106 (11%) and Netherlands 86 (9%). Doctors 1549 (89%) were the most recruited health workers, Nurses 107 (6%) and other professionals were 75 (4%). With Regards to relationship status, single health care workers constituted 61% of the recruited, 31% were married, 8% were cohabiting.

Discussion

This research showed that a majority of foreign qualified health care workers stay at their placement facilities for an average of 2 years. While a constant rate of foreign recruitment per year can "fill the gap" in paper, the low average length of practice signifies a hidden cost of recruiting, relocating and training of new healthcare professionals. Effective workforce planning from government or non-profit organizations, thus, requires a tool to predict the length of practice of incoming health professionals.

The three models attain almost identical results, with negligible difference in accuracy. The 'best'-performing model (Multinomial logistic classifier) achieves a 47.34% [SD 1.63] while the decision tree model achieves an almost comparable 45.82% [SD 1.69]. The three models achieve the average AUC of approximately 0.66 (Multinomial logistic at 0.6652, Decision tree at 0.6635, Naive Bayes at 0.6602), suggesting sufficient predictive signal at the four categorical variables selected. This is an indication that when Human Resource for Health decision makers apply machine learning to their data sets this can effectively enable them to source health care workers who are most likely to stay the longest in underserved communities.

Machine learning must be applied coupled with other qualitative methods like exit interviews so as to give an in-depth understanding of the health care worker perceptions and experiences that relate to their length of stay. A mixed method would have generated a better understanding of why certain gender, countries, age and experience tend to stay longer than others.

Incomplete fields in the data were another issue as many candidates were excluded from the study due to missing information.

Limitations of the study

Incomplete fields in the data were another issue as many candidates were excluded from the study due to missing information. We could not obtain Age as one of the predictors, although we recognized that it could potentially influence health worker long-term plan to stay.

Conclusions

Machine learning models give us an effective tool to predict the recruited health workers' length of practice. These models can be adapted beyond the scope of demographic information (i.e., information about placement location, income, etc.), allowing strategic planning and optimization of public health care recruitment.

Acknowledgements

The authors would like to thank the African Health Placement for providing the dataset used in the study

References

[1]. Alhassan RK, Spieker N, van Ostenberg P, Ogink A, Nketiah-Amponsah E, de Wit TFR. Association between health worker motivation and healthcare quality efforts in Ghana. Hum Resour Health. 2013; 11(1):37. doi: 10.1186/1478-4491-11-37.

[2]. Agyepong IA, Anafi P, Asiamah E, et al. Health worker (internal customer) satisfaction and motivation in the public sector in Ghana. Hum Resour Heal. 2012; 11(247). doi: 10.1186/1472-698X-12-25.

[3]. Buchan J, Couper ID, Tangcharoensathien V, et al. Early implementation of WHO recommendations for the retention of health workers in remote and rural areas. Bull World Health Organ. 2013; 91(11):834-840. doi:10.2471/BLT.13.119008. NDoH. National Health Insurance; 2017.

[4]. Bangdiwala IS, Fonn S, Okoye O, Tollman S. Workforce Resources for Health in Developing Countries. Public Heal Rev. 2010; 32(1):296-318.

[5]. Cometto G, Tulenko K, Muula AS, Krech R. Health Workforce Brain Drain: From Denouncing the Challenge to Solving the Problem. PLoS Med. 2013; 10(9). doi:10.1371/journal.pmed.1001514.

[6]. Dovlo D. The Brain Drain and Retention of Health Professionals in Africa. A case study Prep a Reg Train Conf Improv Tert Educ sub-Saharan Africa Things that Work. 2003:23–25.

[7]. Delobelle P, Rawlinson JL, Ntuli S, Malatsi I, Decock R, Depoorter AM. Job satisfaction and turnover intent of primary healthcare nurses in rural South Africa: A questionnaire survey. J Adv Nurs. 2011;67(2):371-383. doi:10.1111/j.1365-2648.2010.05496.x.

[8]. George G, Gow J, Bachoo S. Understanding the factors influencing health-worker employment decisions in

South Africa. Hum Resour Health. 2013; 11(1):15. doi: 10.1186/1478-4491-11-15.

[9]. Hand DJ. A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems. Mach Learn. 2001:171-186.

[10]. Habte, D., Dussault, G., Dovlo D. Challenges confronting the health workforce in Sub-Saharan Africa. World Hosp Heal Serv. 2004; 40(2):23-26. https://www.researchgate.net/profile/Gilles_Dussault/publication/8373533_Challenges_confronting_the_health_workforce_in_sub-Saharan_Africa/links/0fcfd510c3af1833e7000000.pdf#page=22.

[11]. Hatcher AM, Onah M, Kornik S, Peacocke J, Reid S. Placement, support, and retention of health professionals: national, cross-sectional findings from medical and dental community service officers in South Africa. Hum Resour Health. 2014; 12(1):12:14. Doi: 10.1186/1478-4491-12-14.

[12]. Kok MC, Dieleman M, Taegtmeyer M, et al. Which intervention design factors influence performance of community health workers in low- and middle-income countries? A systematic review. Health Policy Plan. 2014; 30(9):1207-1227. doi:10.1093/heapol/czu126.

[13]. Landis JR, Koch GG. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. Biometrics. 1977; 33(2):363. doi: 10.2307/2529786.

[14]. Labonté R, Sanders D, Mathole T, et al. Health worker migration from South Africa: causes, consequences and policy responses. Hum Resour Health. 2015; 13(1):92. doi: 10.1186/s12960-015-0093-4.

[15]. Rosenthal EL, Brownstein JN, Rush CH, et al. Community health workers: part of the solution. Health Aff (Millwood). 2010; 29(7):1338-1342. doi:10.1377/hlthaff.2010.0081.

[16]. Steinmetz S, Vries DH de, Tijdens KG. Should I stay or should I go? The impact of working time and wages on retention in the health workforce. Hum Resour Health. 2014; 12(1):23. doi: 10.1186/1478-4491-12-23.

[17]. Sieleunou I. Health worker migration and universal health care in Sub-Saharan Africa. Pan Afr Med J. 2011; 10: 55. http://www.ncbi.nlm.nih.gov/pubmed/22384301%5Cn
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3290885.

[18]. Viscomi M, Larkins S, Sen Gupta T. Recruitment and retention of general practitioners in rural Canada and Australia: a review of the literature. Can J Rural Med. 2013; 18(1):13-24.

Appendix

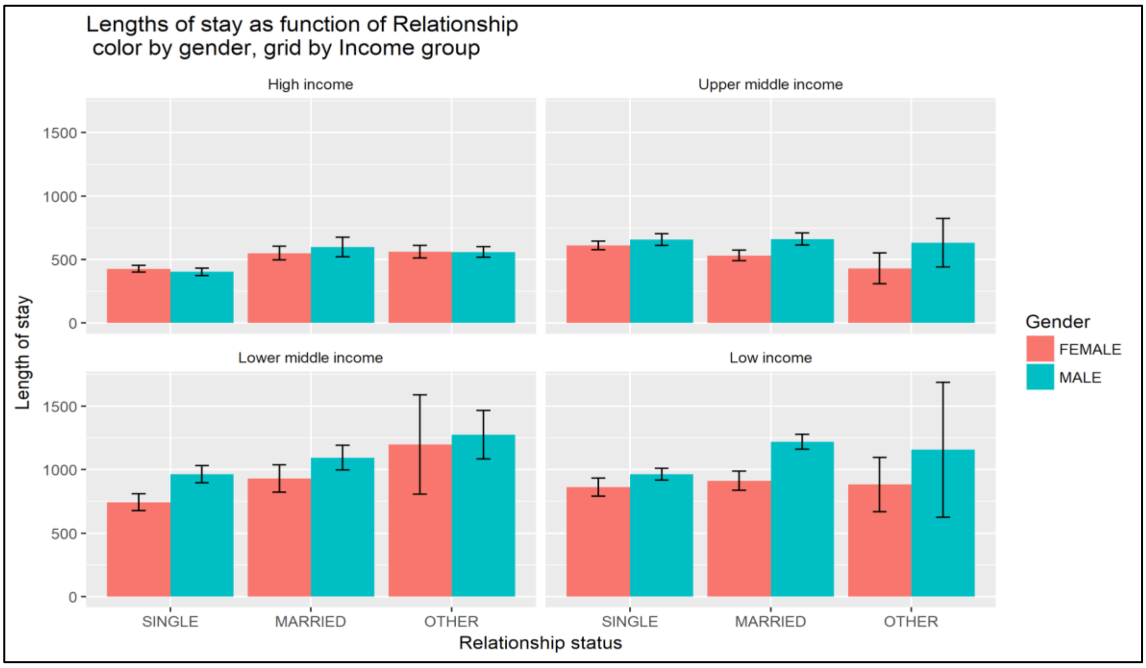

Figure 2. Length of stay as function of relationship colour by gender, grid by income group

DOI: 10.21522/TIJPH.2013.05.04.Art005
ISSN: 2520-3134

Table1. Machine learning results

	Techniques		
	Multinomial logistic	Decision tree	Naive Bayes
Accuracy	47.34% (1.63)	45.82% (1.69)	47.01% (1.62)
95% CI	(46.22, 50.84)	(46.66, 51.28)	(45.19, 49.81)
AUC	0.6652 (need SD)	0.6635	0.6602
No information rate [NIR]	0.376	0.376	0.376
P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16	< 2.2e-16
Cohen’s Kappa	0.2658	0.2649	0.2521

Length of stay

Table 2. Length of stay by gender, nationality, profession and relationship status

	Mean length of stay (days)	**Standard deviation (sd)**	**Sample (n)**	**Percentage (%)**
Gender				
Female	602.67	497.6	870	47%
Male	790.36	630.3	981	53%
Total			**1 851**	**100%**
Nationality (top 4)				
South Africa	548.65	388.1	381	41%
United Kingdom	475.11	373.3	361	39%
Nigeria	1,096.09	719.7	106	11%
Netherlands	753.36	532.7	86	9%
Registered Profession				
Doctor	713.67	587.2	1 549	89%
Nurse	575.38	498.2	107	6%
Support staff	497.51	328.4	75	4%
Total			**1 731**	**100%**
Relationship Status				
Single	623,57	529.3	1 124	61%
Married	869,18	658.2	576	31%
Partner (Cohabitating)	656,80	478.3	145	8%
Divorced	486,40	321.1	5	0%

Table 3. Predictions of length of stay across the three models

	less than 1 year	less than 2 years	less than 3 years	more than 3 years
Multinomial logistic techniques				
Sensitivity	0.7685	0.3248	0.0369	0.5425
Specificity	0.6548	0.8503	0.9766	0.7896
Positive Predictive value	0.5728	0.4533	0.2340	0.3700
Negative Predictive value	0.8244	0.7673	0.8398	0.8834
Balanced accuracy	0.7166	0.5876	0.5068	0.6661
Decision Tree techniques				
Sensitivity	0.7858	0.3740	0.000	0.4897
Specificity	0.6469	0.8075	1.000	0.8150
Positive Predictive value	0.5728	0.4260	NaN	0.3761
Negative Predictive value	0.8337	0.7716	0.8379	0.8751
Balanced accuracy	0.7164	0.5908	0.5000	0.6524
Naives Bayes techniques				
Sensitivity	0.7728	0.2658	0.0403	0.5630
Specificity	0.6391	0.8752	0.9760	0.7675
Positive Predictive value	0.5633	0.4485	0.2449	0.3556
Negative Predictive value	0.8236	0.7573	0.8401	0.8852
Balanced accuracy	0.7059	0.5704	0.5081	0.6653

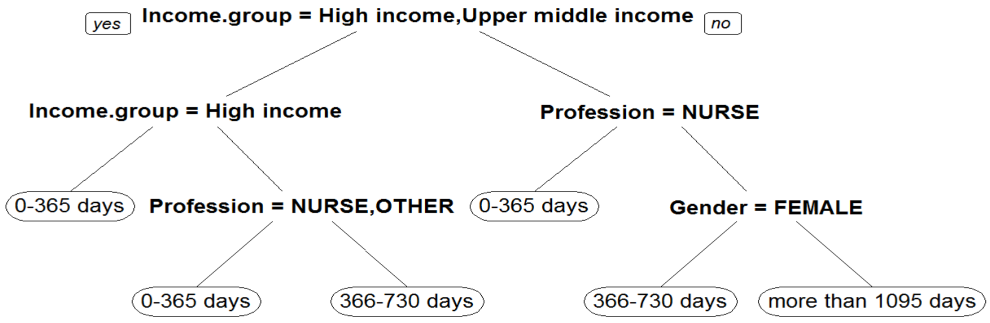

Figure 3. Decision tree based on income

DOI: 10.21522/TIJPH.2013.05.04.Art005
ISSN: 2520-3134

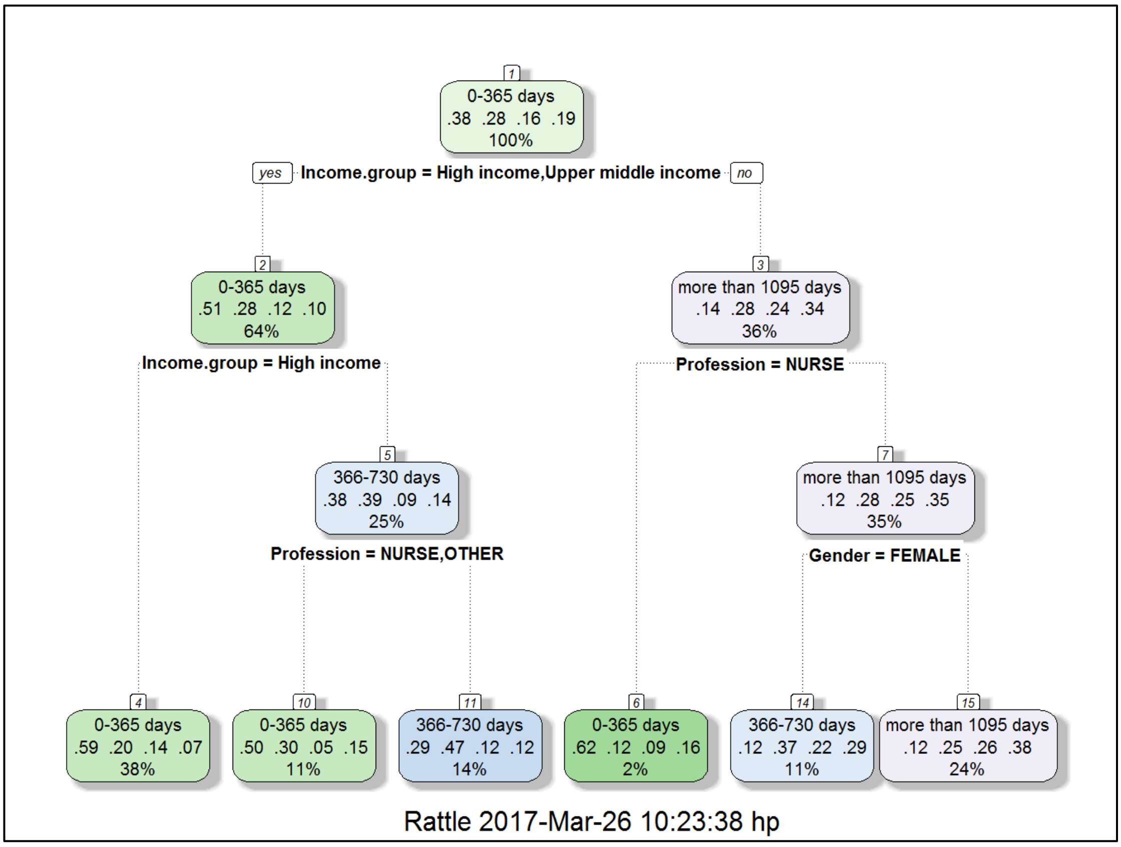

Figure 4. Decision tree on income and gender

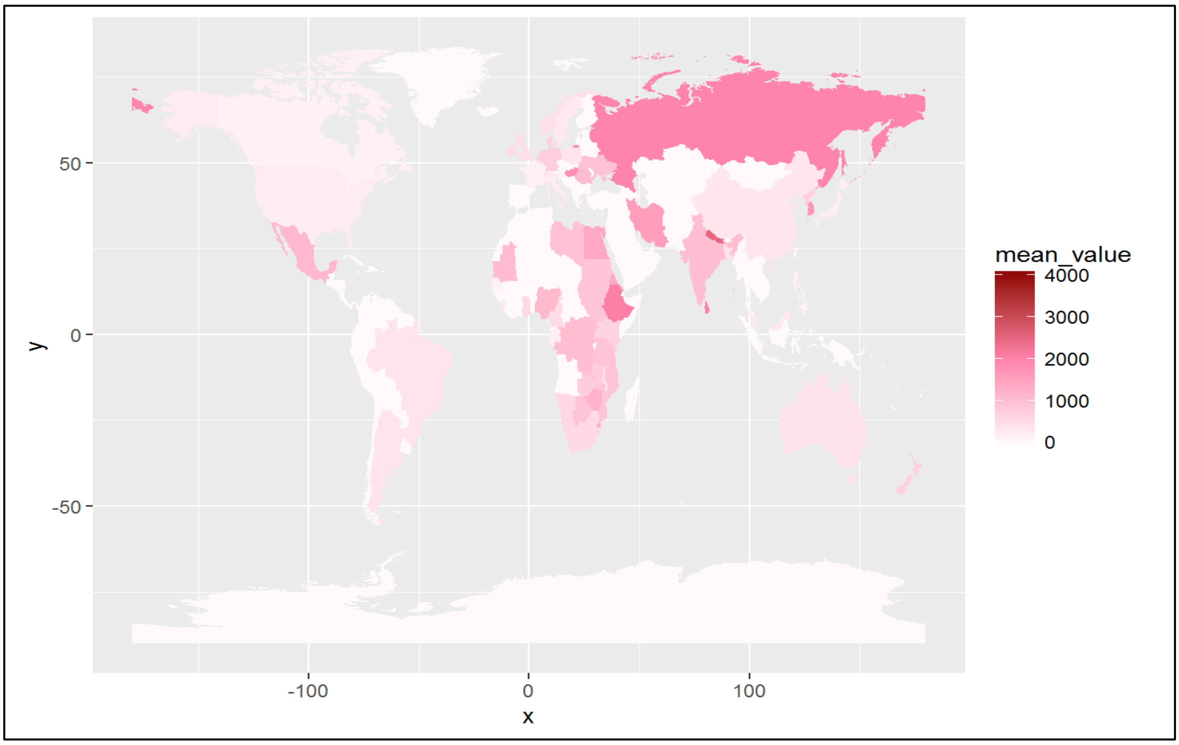

Figure 5: Map showing world distribution of the candidates sourced