Predicting Thyroid Dysfunction with Machine Learning

[1]Opetunde Adepoju, [1]Ayodeji Ajayi, [2]Olamilekan Wahab,
[1]Ladoke Akintola University of Technology
[2]Obafemi Awolowo University
ohadepoju@student.lautech.edu.ng

Abstract

Machine learning algorithms have caused great advancements in disease prediction due to the level of accuracy they possess in making predictions. In this research, we apply machine learning to predict thyroid dysfunction of 3,774 patients from the levels of free triiodothyronine (FT3), thyroid stimulating hormone (TSH), triiodothyronine (T3) and thyroxine (T4). Also features like pregnancy status, illness, psychological state and use of any antithyroid drugs at the time the data was collected were used in building the model. We apply supervised learning algorithms such as gradient boosting, random forest and logistic regression and use the accuracy, precision and recall to measure the level of exactness of the classification. We then compare the level of exactness and conclude that the best algorithm for thyroid dysfunction prediction is Logistic Regression. The motivation for this research is to ensure that qualitative healthcare is affordable and accessible to the marginalized in Africa. We believe that adopting machine learning models to predict dysfunction will ensure qualitative healthcare and enhance affordability and accessibility in Africa because it is less expensive to build technological models, which are more accessible than humans

Keywords: Machine Learning, predictive analytics, Thyroid dysfunction, precision medicine

Introduction

The epidemiology of thyroid dysfunctions all over the world is alarming due to its prevalence worldwide. In Africa, thyroid dysfunction is prevalent due to lack of quality nutrition and population isolation. There is therefore a prevalence rate of 1.2 to 9.9% of thyroid dysfunction. Over 27.9% of thyroid dysfunctions are undiagnosed because of the financial status of patients affected, thereby leading to more complications and eventual death. The effects of thyroid dysfunctions are not only on the physical well-being of patients but also on their mental well-being, as proven by research. Thyroid dysfunction causes diseases of the blood cells such as anemia, erythrocytosis leukopenia, and thrombocytopenia. Patients with depression and mental disorders such as difficulty in concentration, over-anxiety, lack of mental alertness, short-term memory lapses, lack of interest, and mood swings display abnormal levels of thyroid hormones, usually caused by thyroid disorders. Other effects such as heart attack, slow metabolism and goiter are threatening.

Based on the understanding of the problem, we proposed a structured analytical method, also known as machine learning, to build a predictive system that accurately predicts thyroid dysfunction before its symptoms are obvious. The idea of the machine learning algorithms is to develop and validate a model that satisfies the need of the solution by providing accuracy in prediction of thyroid dysfunction. We used the guidance of other academic literatures to get an understanding of similar models that have been built and we compare the performance of these models with our model.

The paper is structured to first explain the works of other researchers in building a predictive model to see what methodologies have been found useful in understanding the problem and building an accurate model. We discuss the data used in our study. Next, we discuss the methodologies of the various machine learning algorithms we used and the mathematical concepts behind their implementation. Lastly, we present our results, discuss our conclusions and other possible research areas we discovered in the course of this research and how other researchers can extend in this research moving forward.

With thyroid dysfunctions being one of the major leading disorders of the endocrine system, as well as its life-threatening consequences, building a predictive technology for thyroid dysfunctions will cause advancements in precision medicine, allowing access to an affordable healthcare while reducing the workload of health practitioners.

Literature Review

The early works on predicting diseases with machine learning have integrated many machine learning models. Rasitha et al. used a data mining technique called DBSCAN to conclude that TSH and iodine levels in the body are useful for predicting thyroid dysfunction. (Rasitha et al., 2014). While they did not identify their source of data or the results of their analysis in their paper, they reinforced the fact that the level of TSH in the body is an indicator of thyroid dysfunction.

A. Lui and A. Pappas concluded in their paper that a decision tree model with FTI, TT4 and TSH as features give a good accuracy (Lui and Pappas, 2015). S. Umadevi worked on applying classification algorithms to predict thyroid diseases and concluded that fuzzy based artificial neural network on 21 features gave an accuracy of 90%. (Umadevi et al., 2017). Bahel worked on predicting blood donation using machine learning algorithms and concluded that the non – clustered 5 –fold validated clustering performed best in predicting blood donors, which will perform well for predicting other diseases (Bahel et al., 2017). Most of the work we found built predictive models based on the levels of FTI, TT4 and TSH measurement in patients. However, none of these works considered whether the patients are on an antithyroid or a thyroid-stimulating drug. The studies just discussed are summarized in Table 1 below, which are a fraction of the numerous studies that have been carried out in building predictive models for diseases. We found that none of the studies reviewed have considered psychological effects of thyroid dysfunction as a feature to be considered when building a predictive model.

Table 1: The previous works on disease prediction with machine learning models and their level of accuracy.

Authors	Algorithms	Data	Features	Results
Rasitha et al., 2014	Density-based spatial Clustering of applications with noise (DBSCAN)	Unidentified	TSH levels	Unidentified
Lui and Pappas 2015	CART, Logistic Regression, KNN, SVM	7679 samples, 7 features	TSH, FTI and TT4 levels, age thyroxine level, tumor, surgery	CART (0.01589 gen. error), SVM (0.02621 gen. error), KNN (0.02939 gen. error), Logistic regression (0.03415 gen. error)
Umadevi et al., 2017	KNN, ANN, Fuzzy ANN	7200 samples, 16 features	T3 level, T4 level and 14 other unspecified features	KNN (80% accuracy), ANN (85% accuracy), fuzzy ANN (90% accuracy)
Bahel et al., 2017	ANN, C5 Logistic regression, CART, Random forest	UCI Machine learning repository data	Social influence, months since last blood donation, months since first donation	ANN (83% accuracy), C5 logistic regression (88%), CART (79%), Random forest (75%)

Data

In order to build a machine learning algorithm for accurate prediction, we collected dataset from a local hospital of 3774 patients containing their FTI, TSH, TT4, T4U levels, age, health status at the time of test (sick or not sick), gender and pregnancy status of the women. The dataset mostly contained binary annotations such as presence of pregnancy, goiter and diseases that can affect thyroid stimulation in the body. There were many missing values in the dataset and we performed data cleaning to either replace the missing values with median or drop the missing values. Also from the correlation plot, we realized there were some features that had insignificant correlation to the target variable. An example is the TGB, referral source and 131treatment. We dropped these features to enable an optimum performance of the model. We used 80% and 20% for the training and test splits.

Methodology

Tools Used

For building this model, we use python, Scikit learn, Matplotlib, Numpy and Pandas. Python is a programming language that is multipurpose. We streamline its usefulness to building a predictive model. Matplotlib is a python library used for visualization. We use it to visualize the correlation between the features (such as age, sex, pregnancy status etc) and the target (thyroid dysfunction). Numpy is a python library for mathematical computation. Here, we use it for computing the mean, median and standard deviations of the numerical features contained in the dataset. Scikit learn is the library we used to build the model. Here we used it to build the logistic regression, gradient boosting and random forest.

Machine Learning Algorithms

For this kind of predictive analytics, there are two major kinds of learning algorithms we can apply namely: Supervised learning and Unsupervised learning. This work focuses on supervised learning.

Supervised learning: Supervised learning is a type of machine learning that makes future prediction based on an input and output labels from previous data of a similar occurrence. In supervised learning, learning stops when the algorithms achieves an acceptable performance. There are two major types of supervised learning, which are classification and regression. Classification is a mode of machine learning which classifies an output variable in a binary or a two –way means. Some examples of classification include Decision tree, neural networks, logistic regression, Support vector machine, Naïve Bayes and K-Nearest neighbor. This work is an example of a classification-based model and it focuses on using basic classification algorithms such as decision tree and logistic regression. Regression is a machine learning which is most suitable for predicting a continuous output variable. Many different supervised learning models are used for a regression-based problem. An example is a simple linear regression

Unsupervised learning: Unsupervised learning is a form of machine learning that learns from an unlabeled data, which is neither classified nor categorized. The machine learning algorithms used for unsupervised learning are clustering algorithms. A popular type of clustering algorithm is K-Means clustering. This work does not focus on unsupervised learning because we are working with a categorized dataset.

An explanation of the algorithms used

In this work we use logistic regression, decision trees and gradient boosting. Here is a brief explanation of how these models work to make predictions

Logistic Regression

binary dependent variable. It does this by estimating the parameters of a logistic model. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables. Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types, ”0” and ”1” (which may represent, for example, ”dead” vs. ”alive” or ”win” vs. ”loss”). Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., ”disease A” vs. ”disease B” vs. ”disease C”) that are not ordered. Ordinal logistic regression deals with dependent variables that are ordered. In binary logistic regression, the outcome is usually coded as ”0” or ”1”, as this leads to the most straightforward interpretation. If a particular observed outcome for the dependent variable is the noteworthy possible outcome (referred to as a ”success” or a ”case”) it is usually coded as ”1” and the contrary outcome (referred to as a ”failure” or a ”non-case”) as ”0”. Binary logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a no-case. In this work, we explored the binomial logistic regression.

Table 2: Model Results

Model	Accuracy	Precision	Recall
Logistic regression	0.981	0.827	0.727
XGBoost	0.923	0.791	0.715
Decision trees	0.846	0.754	0.703

Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. XGboosting is a form of gradient boosting. A decision tree is a flowchart-like structure in which each internal node represents a ”test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

Results and Discussion

We realized that age had a correlation with the thyroid functioning in the body. We also realized from the exploratory data analysis that we carried out that the FTI, TSH, TT4 and T4U levels of different healthy age groups were similar while the levels were far diverse in unhealthy people. We did an exploratory data analysis that best explained this outcome.

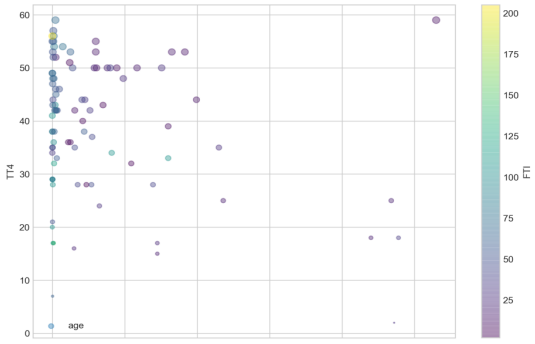

Figure 1: Age in relation to the FTI and TT4 levels of the body

The relationship of age and the level of wellness of the thyroid gland explains the previous work which discovered that the older people tend to have a lower rate of performance of thyroid gland which in turn affects the level of thyroid hormones in the body. This effect plays out on their susceptibility to thyroid dysfunction at a later age in life.

Conclusion

From the work carried out and the previous related work, we conclude that machine learning algorithms will perform better than human experts at predicting thyroid dysfunction will. Other research areas related to this work might include the use of more advanced techniques called deep learning to build predictive models. Deep learning technique even provides a higher accuracy than machine learning techniques and provides the ability for the data to learn by itself from an unfamiliar and unknown data. This concept is otherwise known as Reinforcement Learning.

References

1. Dr.G.Rasitha Banu, Baviya , "predicting Thyroid disease using Data Mining Technique", IJMTER journal, Volume -2,Issue -3,page no- (666-670),March 2015.

2. Rasitha et al., 2015 A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease

3. Imbus JR, Randle RW, Pitt SC, Sippel RS, Schneider DF. Machine learning to identify multigland disease in primary hyperparathyroidism. J Surg Res. 2017 Nov;219:173-179. doi: 10.1016/j.jss.2017.05.117. Epub 2017 Jun 29. PMID: 29078878; PMCID: PMC5661967.

4. A. Begum and A. Parkavi, "Prediction of thyroid Disease Using Data Mining Techniques," 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), 2019, pp. 342-345, doi: 10.1109/ICACCS.2019.8728320.

5. K. Rajam and R. Jemina Priyadarsini"A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques" ,IJCSMC, Vol. 5, Issue. 5, May 2016, pg.354–358.

6. Umadevi S , Dr .Jeen Marseline K.S,"Applying Classification Algorithms to Predict Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 10, September 2017

7.]UCI Machine learning repository (patient's data) (online).Available:http://archive.ics.uci.edu/ml/machinelearning-databases/ thyroid-disease/hypothyroid

8. Khushboo Taneja, Parveen Sehgal, Prerana "Predictive Data Mining because of Diagnosis over Thyroid Disease the use of Neural Network" International Journal regarding Research of Management, Science amp; Technology (E-ISSN: 2321- 3264) Vol. 3, No. 2, April 2016

9. Rao, N., Razia, S.: Machine learning techniques for thyroid disease diagnosis. Indian J. Sci. Technol. 9(28) (2016). https://doi.org/10.17485/ijst/2016/v9i28/9370

10. Albert Y. Lui Alexandra M. Pappas., Thyroid Dysfunction: Prediction and Diagnostics., December 11, 2015

11. "Free T4." : The Test. American Association for Clinical Chemistry, 29 Oct. 2015. Web. 12 Dec. 2015.

12. Faix, James D., MD, and Linda M. Thienpont, PhD. "Thyroid-Stimulating Hormone." - AACC.org. American Association for Clinical Chemistry, 1 May 2013. Web. 12 Dec. 2015.

13. "TSH." : The Test. American Association for Clinical Chemistry, 29 Oct. 2015. Web. 12 Dec. 2015.

14. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. New York: Springer Science+Business Media, 2013. Electronic.