Personalized Health Monitoring using Predictive Analytics

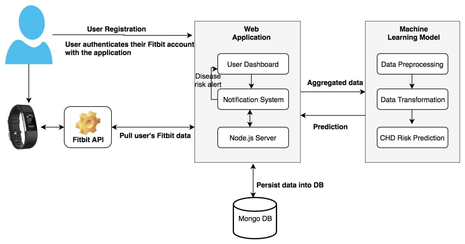
Project Advisor: Dr.Wencen Wu

Computer Engineering Department

Amin, Poojitha (MS Software Engineering)
Anikireddypally, Nikhitha (MS Software Engineering)
Khurana, Suraj (MS Software Engineering)
Vadakkemadathil, Sneha (MS Software Engineering)

Introduction

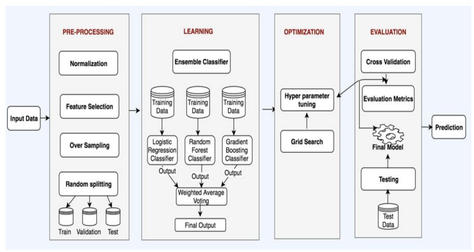
Disease identification and diagnosis of ailments are at the forefront of machine learning research in healthcare. Physical health monitoring using wearable sensors plays a vital role in early detection of abnormalities in the physical health of people. According to a study by Centers for Disease Control and Prevention, about 9% of the population now have diabetes, and one in four are unaware of it. Today, the increasingly sedentary nature of many forms of recreation time and increasing urbanization results in decreased physical activity and thereby leads to a rise in health problems.



A lot of wearables available today can provide important cues to people, however, these devices are not able to perform advanced predictions from the collected data about a disease condition. In this project, we propose a real-time analytics approach on data from activity trackers, to monitor the vital signs of a person, like heart rate and notify the user in case of significant changes. The project aims to develop a model to perform predictions based on the captured physiological parameters. The outcome will be a personalized healthcare service which can significantly improve diagnostic accuracy, healthcare quality, and persons' quality of life.

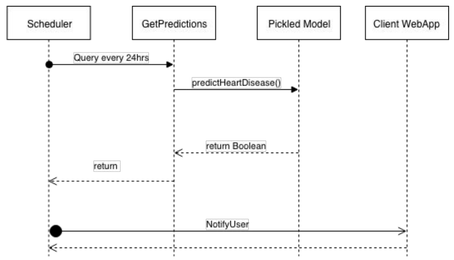
Methodology

Our ML design includes all the ML phases - acquiring, processing and modeling data, model optimization and evaluation. ML routines include experimenting in a loop and testing. Fine tuning is also performed as part of ML routine. These kind of routines helps in improving the algorithms and thereby, achieve better results. The model is then prepared for deployment and the results are consumed by the users to know about their heart disease.



The input to the system consists of real-time data collected from Fitbit device as well as lifestyle-related information about the person. Real-time data collected from the Fitbit device form a continuous streaming data. Collecting the lifestyle-related information is a one-time process during the user registration phase. The system will have an interface to fetch Fitbit data upon user authentication and authorization.

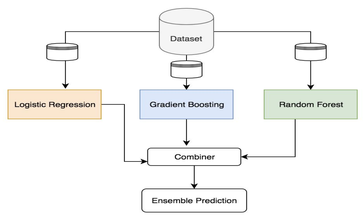
This final phase deals with sending notifications to the user in case of any abnormalities in his/her vitals. Challenges in implementing this module include filtering out false alarms during the period in which a user is exercising or playing games. This module will be triggered only when the predicted value from the machine learning model indicates a risk of heart disease. In such a scenario, the application will send an alert to the user with an appropriate message. This phase also deals with defining notification types, scheduling the notifications, tracking them for delivery and responding to user-selected actions.



Machine Learning Model

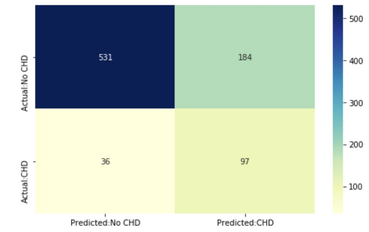
An ensemble machine learning model combines different base models into one predictive model to reduce bias, variance, or improve predictions. We built an ensemble of models and used the weighted average voting to predict the heart disease risk category class.

- Random Forest Classifier - Ensemble way of classification by constructing multiple decision trees.
- Gradient Boosting Classifier - Ensemble of prediction models with an objective of minimizing the loss of the model through gradient descent.
- Logistic Regression Classifier - Classification model is used to predict the probability of the classes. In logistic regression, the output variable is a binary variable encoded as either 1 or 0.



Analysis and Results

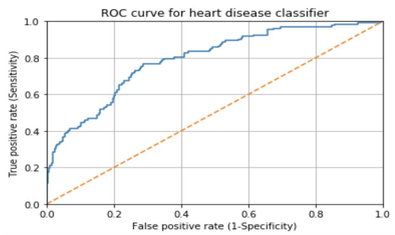
The performance of a classification model is determined by using a confusion matrix. We trained our model using a training data set consisting of 3392 records and a test data set of 848 records. Our result from the ensemble approach in the confusion matrix form is as below-



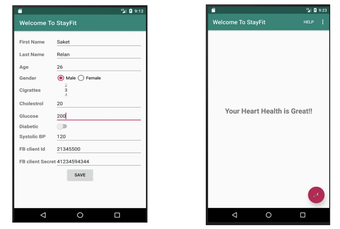
We achieved an accuracy of 87% with the Logistic Regression model. However, further analysis of data identifies majority of the data belongs to one class and hence the events of interest have been misclassified.
Intuitively we know that, the identification of the negative data points in our problem is not helpful and we should focus on identifying the positive heart disease cases. The best metric to maximize in our scenario is the Recall. Recall or Sensitivity is the ability to identify a heart disease condition in a population of individuals affected with heart disease.
For our data set, we have 848 records, of which 133 (16%) actually are likely to get heart diseases in the future.

Sensitivity =TP/TP+FN = 97/97 + 36 = 73%
Specificity =TN/TN+FP = 531/531 + 184 = 74%

In other words, our model will identify 73% of people with the heart disease, but 26% of healthy people will incorrectly be assumed as a heart disease condition. Our measure of accuracy will be a combination of good sensitivity and specificity.
ROC curve is another way of visualizing the performance of a binary classifier. The area under the resulting ROC curve (AUC) gives a complete picture of the performance of a trained classifier on a given dataset. TP (sensitivity) can then be plotted against FP (1 – specificity) For each threshold the True Positive rate is plotted against the False Positive rate. The output is a Receiver Operating Characteristic curve as given below:



STAY FIT



Summary/Conclusions

We obtained an accuracy of 87% with the Logistic Regression model. However, relying on accuracy alone in an imbalanced class scenario can be misleading. Hence, we use the resampling technique to overcome the class imbalance issue and built an ensemble model to improve the recall values. Our ensemble model using the combination of logistic regression, gradient boosting, and random forest provides an improvement of 40%-70% improvement in the recall values. We get an AUC score of 80%. The model evaluation results of our ensemble classifier show that our approach performs better than many other existing techniques.

Key References

H. He, E.A. Garcia (2009). "Learning from imbalanced data", IEEE Transactions on Knowledge and Data Engineering, 21 (9), p. 1263-1284

BioLINCC: Framingham Heart Study-Cohort (FHS-Cohort). (n.d.). Retrieved from https://biolincc.nhlbi.nih.gov/studies/framcohort/

Palaniappan, S. & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. International Journal of Computer Science and Network Security 8(8) 343-350.

Radha, P., & Divya, R. (2016). Multiple time series clinical data with frequency measurement and feature selection. Advances in Computer Applications (ICACA), IEEE International Conference, 250-254. DOI:10.1109/ICACA.2016.7887960

Liang, Z., Martell, M., & Nishimura, T. (2016, October). A Personalized Approach for Detecting Unusual Sleep from Time Series Sleep-Tracking Data. Healthcare Informatics (ICHI), 2016 IEEE International Conference , 18-23, DOI: 10.1109/ICHI.2016.99

Acknowledgements

The authors are deeply indebted to their project advisor Dr. Wencen Wu and Prof. Dan Harkey, Director of the Software Engineering Department.