A Comparison of Machine Learning Algorithms to Predict Social Determinants of Health

Hello every one my name is Obed osei Akoto and im to present to you my masters project on. Before this project I didn't know a lot of factors affect a person on what kind of health treatment they get.

Abstract

The environments in which people are born, live, learn, work, play, worship, and age have an impact on a variety of health, functional, and quality-of-life outcomes and risks. In this study use data mining, machine learning (ML), and neural network (NN) technologies to predict socio economic determinants of health in this study. The database contains data on 673 patients, their non-unique class traits, and BARRIERS, which are the target attributes. To build prediction models for our categorization using SDH data, several different machine learning methods were investigated. The results showed that the best machine learning approach varies depending on the classification algorithm, as assessed by predictive precision, recall, and F-measure score, showing that a method for constructing predictive models from SDH data is beneficial.

Introduction

The dataset contains 673 patients, each with their own set of seven distinct characteristics. Table 1 summarizes the dataset's characteristics. Age, Language, Birth, Zip Code, Nearest cancer center, Kilometers to nearest cancer center, Duration to nearest cancer center, and Outcomes are the seven characteristics used to predict SDH. The seven class attributes are considered independent/feature variables, while the 'outcomes' attribute is a dependent or target variable. The SDH target attribute 'outcomes' has 93 labels, each of which is a binary value with a value of 0 indicating no label and 1 indicating a label.

Preprocessing of data

Preprocessing assists in the transformation of data in order to produce a more accurate machine learning model. Preprocessing performs a range of tasks to improve data quality, including missing value filling, data normalization/standardization, and feature selection. There are 673 samples in the collection.

Selection of Features

Pearson's correlation method is a well-known method for determining the most important characteristics or qualities. The correlation coefficient, which is related to the output and input quality, is calculated using this method. The coefficient's value remains between 0 and 1. A value over or below 0.5 indicates a substantial correlation, whereas a value of zero indicates no correlation.

MultiLabelBinarizer

Because your target property was in array form, we used a Multilabelbinarizer to transform the labels into distinct columns.

Normalization

Normalization is a data preparation technique for machine learning that is frequently used. Normalization is the process of changing the values of numeric columns in a dataset to a similar scale without distorting the ranges of values or losing information. Some algorithms, in order to model the data accurately, require normalization as well.

More Explanation for Normalization (example)

Assume your input dataset has two columns, one with values ranging from 0 to 1 and the other with values ranging from 10,000 to 100,000. When attempting to integrate the values as features during modeling, the large disparity in scale of the numbers may cause complications.

Method for training and testing data sets/Evaluation Strategy

After being cleaned and preprocessed, the dataset is ready to train and test. We used K-fold cross-validation and the 80 percent /20 percent train/test splitting strategy separately to examine the performance of the various machine learning models. The train/split method randomly splits the dataset into training and testing sets. In the K cross-validation method, the data is divided into K folds. One fold is used for validation/testing, while the remaining K-folds are used for training. The procedure will be continued until each K fold has been tested. To evaluate performance, the average of all recorded Kth test results is used.

Creating and implementing a classification system

Problem Transformation, Adapted Algorithms, and Ensemble Approaches are the three main strategies for solving a multi-label classification problem. To analyze the SDH, we applied the Problem Transformation method and various ML classification techniques such as BinaryRelevance, ClassifierChain, OneVsRestClassifier, and neural network (CNN). We used Kth value = 10 for the CNN algorithm. The proposed model diagram is shown in Figure 7 and 8.

Implementation of the BinaryRelevance model in 2.4.1

This is the simplest basic technique, which treats each label as a separate single-class classification problem. Let's have a look at an example, which is provided below. The data set is as follows: X is the independent variable, while Y is the target variable. In binary relevance, this topic is separated into four independent single class classification tasks.

Implementation of the ClassifierChain model

Each succeeding classifier is trained on the input space as well as all prior classifiers in the chain, with the starting classifier being trained exclusively on the input data. [Fig3]

Implementation of the OneVsRestClassifier model

One-vs-all is a strategy that involves matching only one classifier per class. For each classifier, the class is compared to all other classes.

NN

With varying hidden layer levels, we generated two independent neural network models. The neural network was tested with hidden layers 1 and 2, three different thresholds, and 512 epochs. The weighted total of input is processed by the activation function of CNN's hidden layer. We used sigmoid and RELU activation functions in our study. The Keras and TensorFlow libraries were used to generate the neural network models. From the Keras library, a Sequential class was used. The target variables are stored in the 'Outcomes' attribute. The optimizer is required during the backpropagation method of CNN to reduce the output error. We utilized rmsprop as an optimizer (Root Mean Square Propagation).

This can be done `7! = **5040**` ways

(input image [7,1]) convolution (128 Filters, size [1,5])= 128 Different Output having size (7,1)

Conclusion

Utilizing Python, we preprocessed the data. In the CITI Program dataset. On our dataset, we employed multiple machine learning methods to predict SDH, including BinaryRelevance(GaussianNB), BinaryRelevance((KNeighborsClassifier), BinaryRelevance((RandomForestClassifier), ClassifierChain(GaussianNB), ClassifierChain(KNeighborsClassifier), ClassifierChain(RandomForestClassifier), OneVsRestClassifier(GaussianNB), OneVsRestClassifier(KNeighborsClassifier), OneVsRestClassifier(RandomForestClassifier), and OneVsRestClassifier(SVC), and assessed the performance on various measures. precision, recall, and F-measure, all models produce positive outcomes. We also used the NN model to predict SDH in the CITY Program dataset. With varied thresholds and 512 epochs, we employed the 1, 2 hidden layers in the neural network model, which had a good score among our constructed models for SDH. The NN with one hidden layer and BinaryRelevance(GaussianNB), are the most efficient and promising of all the presented models for evaluating SDH, with an accuracy rate greater than 50 for most of the barrier

What are social determinants of health?

Social determinants of health (SDOH) are the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks.

Why did we Normalization:

Similarly, the goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values

Question to answer:

Which is the best classifier for each barrier? Why (statistics)?

ANOVA tells you if there are any statistical differences between the means of three or more independent groups.

BR_NB: It gave the score and it statistically significant from the others.

The ones that are not statistically significant from the best: Are they viable? how much worse are they?
From the anova doc we can determine which ones are worse from the best classifier

Why do you think a given classifier is better than the rest? What are the characteristics of the data and the classifier that allow this?
The score for precision and recall was small compare to other classifiers