Report on developing a Machine Learning Predictive Model on Flower Dataset

Background: As more and more researchers are turning to big data for new opportunities, machine learning models, as the backbone of big data analysis, are used universally in every sector widely. However, owing to the inherent complexity of machine learning methods, they are prone to misuse. Because of the flexibility in specifying machine learning models, the results are often insufficiently reported in research articles, hindering reliable assessment of model validity and consistent interpretation of model outputs.

Objective: To attain a set of guidelines on the use of machine learning predictive models within clinical settings to make sure the models are correctly applied and sufficiently reported so that true discoveries can be distinguished from random coincidence.

Results: The process produced a set of guidelines that consists of

(1) A set of practical sequential steps for developing predictive models.

Conclusions: A set of guidelines was generated to enable correct application of machine learning models and consistent reporting of model specifications and results. We believe that such guidelines will accelerate the adoption of big data analysis, particularly with machine learning methods.

Introduction:

ig data is changing every industry. Medicine is no exception.

With rapidly growing volume and diversity of data in health

care and biomedical research, traditional statistical methods

often are inadequate. By looking into other industries where

modern machine learning techniques play central roles in dealing

with big data, many health and biomedical researchers have

started applying machine learning to extract valuable insights

from ever-growing biomedical databases, in particular with

predictive models [1,2]. The flexibility and prowess of machine

learning models also enable us to leverage novel but extremely

valuable sources of information, such as wearable device data

and electronic health record data [3].

Despite its popularity, it is difficult to find a universally

agreed-upon definition for machine learning. Arguably, many

machine learning methods can be traced back as far as 30 years

ago [4]. However, machine learning started making a broad

impact only in the last 10 years. The reviews by Jordan and

Mitchell [5] and Ghahramani [6] provide accessible overviews

for machine learning. In this paper, we focus on machine

learning predictive methods and models. These include random

forest, support vector machines, and other methods listed in

Multimedia Appendix 1. They all share an important difference

from the traditional statistical methods such as logistic regression

or analysis of variance—the ability to make accurate predictions

on unseen data. To optimize the prediction accuracy, often the

methods do not attempt to produce interpretable models. This

also allows them to handle a large number of variables common

in most big data problems

ig data is changing every industry. Medicine is no exception.

With rapidly growing volume and diversity of data in health

care and biomedical research, traditional statistical methods

often are inadequate. By looking into other industries where

modern machine learning techniques play central roles in dealing

with big data, many health and biomedical researchers have

started applying machine learning to extract valuable insights

from ever-growing biomedical databases, in particular with

predictive models [1,2]. The flexibility and prowess of machine

learning models also enable us to leverage novel but extremely

valuable sources of information, such as wearable device data

and electronic health record data [3].

Despite its popularity, it is difficult to find a universally

agreed-upon definition for machine learning. Arguably, many

machine learning methods can be traced back as far as 30 years

ago [4]. However, machine learning started making a broad

impact only in the last 10 years. The reviews by Jordan and

Mitchell [5] and Ghahramani [6] provide accessible overviews

for machine learning. In this paper, we focus on machine

learning predictive methods and models. These include random

forest, support vector machines, and other methods listed in

Multimedia Appendix 1. They all share an important difference

from the traditional statistical methods such as logistic regression

or analysis of variance—the ability to make accurate predictions

on unseen data. To optimize the prediction accuracy, often the

methods do not attempt to produce interpretable models. This

also allows them to handle a large number of variables common

in most big data problems

ig data is changing every industry. Medicine is no exception.

With rapidly growing volume and diversity of data in health

care and biomedical research, traditional statistical methods

often are inadequate. By looking into other industries where

modern machine learning techniques play central roles in dealing

with big data, many health and biomedical researchers have

started applying machine learning to extract valuable insights

from ever-growing biomedical databases, in particular with

predictive models [1,2]. The flexibility and prowess of machine

learning models also enable us to leverage novel but extremely

valuable sources of information, such as wearable device data

and electronic health record data [3].

Despite its popularity, it is difficult to find a universally

agreed-upon definition for machine learning. Arguably, many

machine learning methods can be traced back as far as 30 years

ago [4]. However, machine learning started making a broad

impact only in the last 10 years. The reviews by Jordan and

Mitchell [5] and Ghahramani [6] provide accessible overviews

for machine learning. In this paper, we focus on machine

learning predictive methods and models. These include random

forest, support vector machines, and other methods listed in

Multimedia Appendix 1. They all share an important difference

from the traditional statistical methods such as logistic regression

or analysis of variance—the ability to make accurate predictions

on unseen data. To optimize the prediction accuracy, often the

methods do not attempt to produce interpretable models. This

also allows them to handle a large number of variables common

in most big data problems

The Iris dataset was used in R.A. Fisher's classic 1936 paper, [The Use of Multiple Measurements in Taxonomic Problems](http://rcs.chemometrics.ru/Tutorials/classification/Fisher.pdf),

It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

The columns in this dataset are:

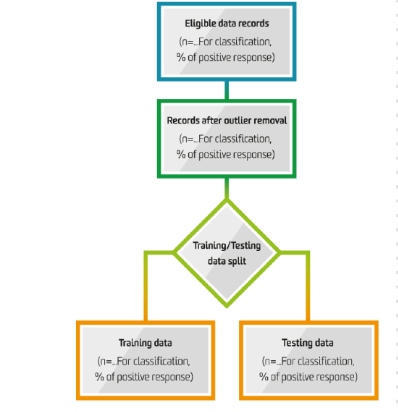
* Id
* SepalLengthCm
* SepalWidthCm
* PetalLengthCm
* PetalWidthCm
* Species

Our Aim is to study and do data engineering on the dataseta and build a predictive model which will identify depending on the feature the species of the flower.

We will be applying logistic regression to build this model as this being a classification task.

The Outcome of the Model will be prediction of flower belonging to one of the 3 category mentioned below:

1. Iris-setosa
2. Iris-versicolor
3. Iris-virginica



Defining the problem and prediction from available data:

When performing ML analyses, 2 assumptions are made:

(1) The desired outputs of the data can be generated given the input data

(2) Available data contain the necessary information to learn the desired output.

It is important to keep these assumptions in mind when considering the input data, ML method, and overall analysis architecture that will be used to address the research question. Classifying the research question and analysis by problem type e

Eg, binary/multiclass classification versus time series or sequential analysis versus exploration or categorization) is important as these factors determine appropriate ML methods given the data available and the desired result. For example, supervised techniques can only be used if an appropriately labeled response is available.

Data Collection

Data sources used for ML analysis are usually large (many data points) and complex (many different types of data) and may, therefore, be difficult or impossible to review manually .Description of the data to be used is critical to assess their quality, reliability, suitability to produce the desired output, potential accuracy of any findings, and especially reproducibility (. Data should be described in detail with respect to its source eg, study, contributor,

### Data Preprocessing, Model Development

Preprocessing steps may include additional statistical and ML methods that supplement the overall analysis. In cases where ML methods are used during preprocessing, reporting guidelines discussed in the Model Training and Evaluation, Model Configuration, Optimization, and Generalization, as well as Validation sections should be included. In cases where specific features have been engineered via domain expertise or data-driven methods, details specific for how those features were defined and validated as well as their similarity to other data sets should be reported.