International Journal of Information Sciences and Application (IJISA). ISSN 0974-2255, Vol.11, No.1, 2019, (Special Issue)

DATA ANALYSIS, VISUALIZATION AND PREDICTIVE MODELING

Sachin Gupta ,Manvendra Singh, Shubham Chaurasia, and Vibhav Kumar

Department of Computer Science and Engineering, IMS ENGINEERING COLLEGE, GHAZIABAD.

Abstract

The main aim of this project is to develop an assistance system that help Data Analysts to directly jump to the data analyzing process rather than to first write codes to read and clean data. This is a software that visualizes and illustrates large numerical data by plotting high quality and colorful graphs, pie charts, histograms, scatter plots, box plots. It will illustrate data in effective ways so as one can study and analyze data more efficiently. This project will help the data scientist to do data analysis and visualization in a very easy manner. It is used to build different classification models to predict output from provided inputs. This will save time and will be much more reliable. It is developed using java components such as java swing and java ML. We have done our best in making graphical user interface as easy as possible It will help Data Analysts to directly jump to the data analysing process rather than to first write codes to read and clean data. It will illustrate data in effective ways so as one can study and analyses data more efficiently.

Introduction:

This project is meant to improve the efficiencies of data scientist by reducing the manual work done by them .This project aims to automate the data mining and cleansing process along with the predictive modelling and visualization. This project makes use of the of Java ML feature which accounts for the predictive modelling and clustering of the data .It uses .arff file format which is an industry standard in terms of data analysis. It also does predictive modelling and uses the concept of confusion matrix. Predictive analytics is used in marketing, financial services, insurance, telecommunications, retail, travel, mobility, healthcare, child protection, pharmaceuticals, capacity planning, social networking and other fields. Although it may be tempting to think that big data makes predictive models more accurate, statistical theorems show that, after a certain point, feeding more data into a predictive analytics model does not improve accuracy. Analyzing representative portions of the available information sampling can help speed development time on models and enable them to be deployed more quickly. This ensures that the efficiencies of the data analysts are improved for good. Once data scientists gather this sample data, they must select the right model. Linear regressions are among the simplest types of predictive models. Linear models essentially take two variables that are correlated -- one independent and the other dependent -- and plot one on the x-axis and one on the y-axis. The model applies a best fit line to the resulting data points. Data scientists can use this to predict future occurrences of the dependent variable.

Steps in working:

- Import your data
- Select the target variable
- Build classification models
- Explore top models with high accuracy
- Deploy the best model

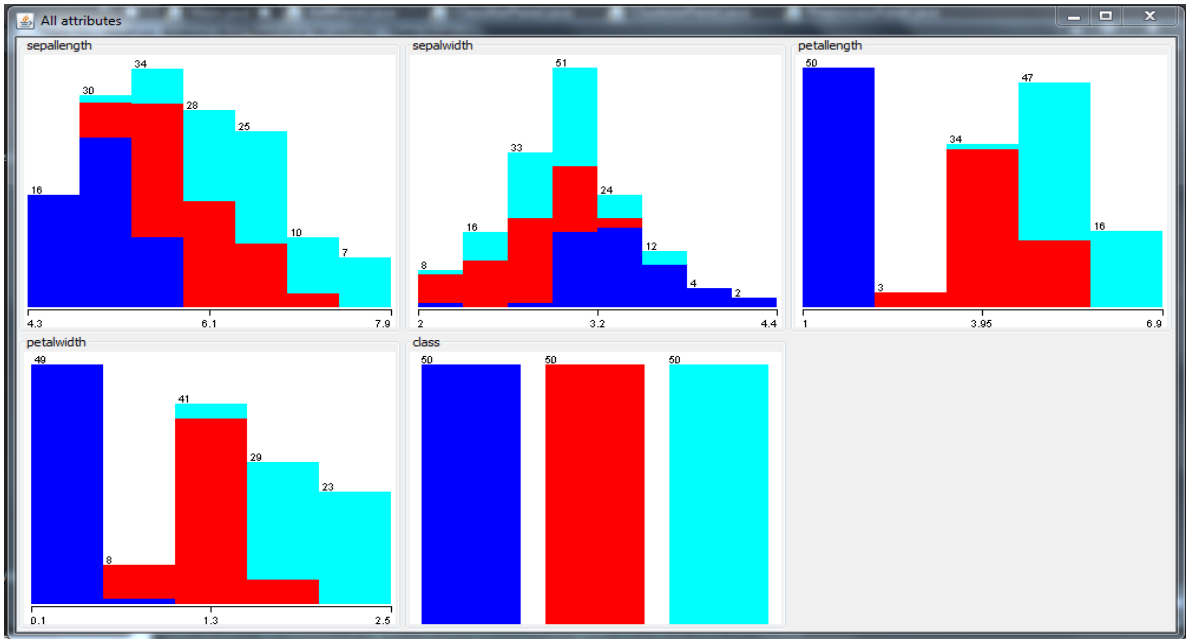
Data import:

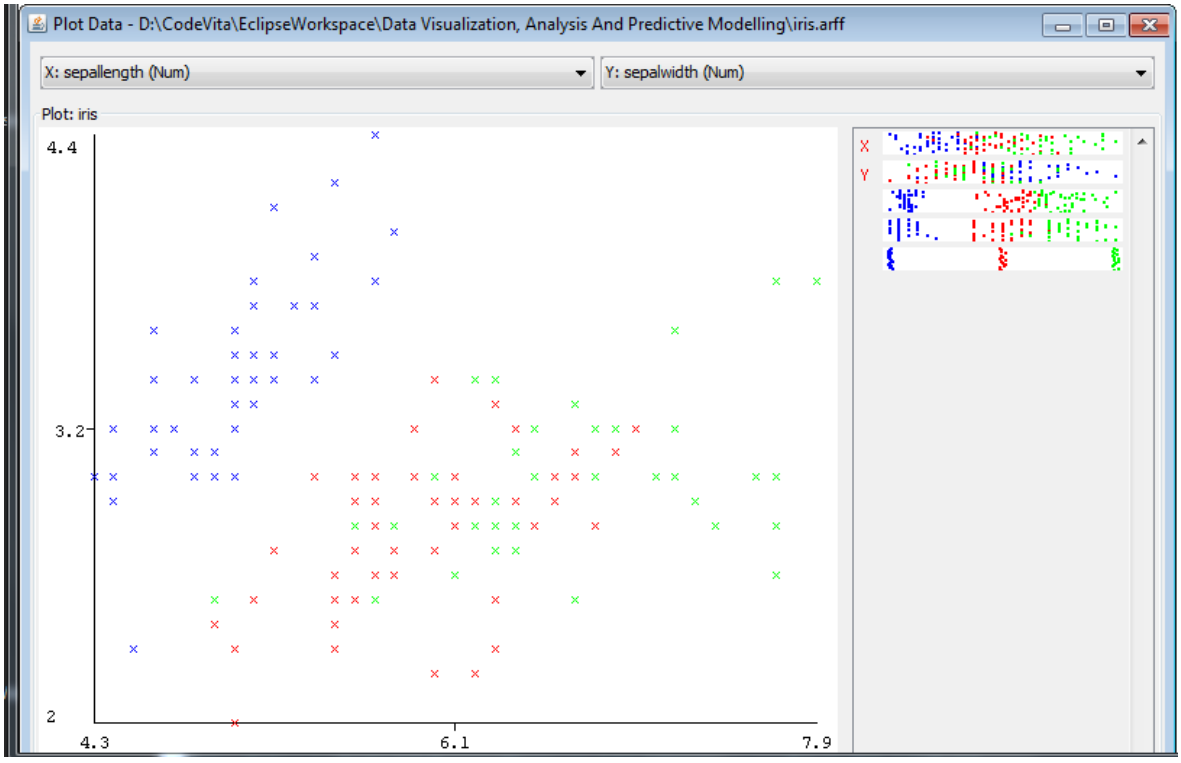
The data is imported in .arff file format .It can include long lists and classified data as well. ARFF is an acronym that stands for attribute relation file format. It is an extension of .csv file format .Here a header is used to provide metadata about the data types in the column. A java class named arffreader is used to create data import possible in this scenario. As soon as data is input to the software it analyses all the data and creates some more data with respect to the user requirement.

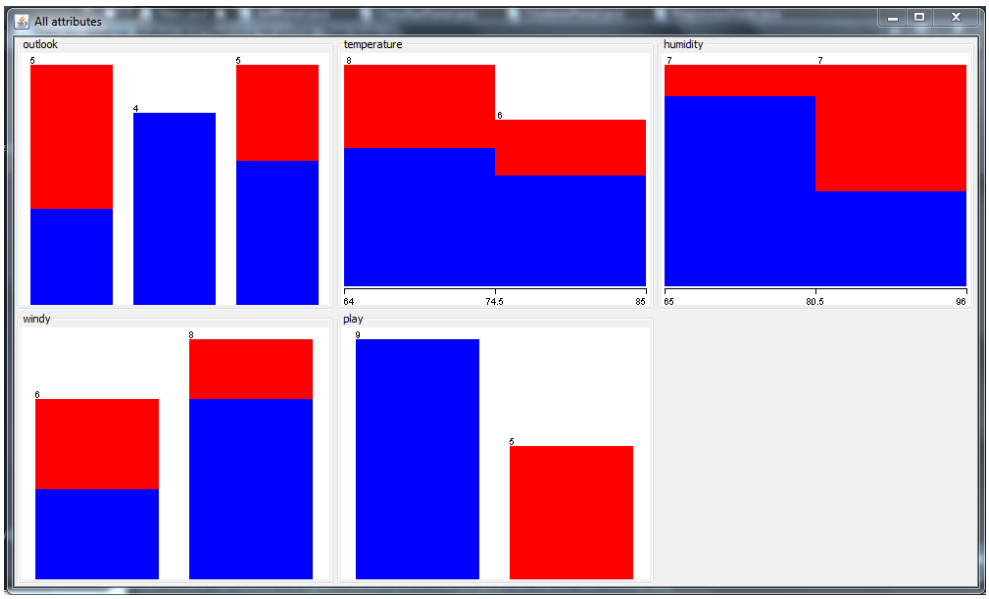
Data Visualization:

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

The data visualisation is very much extended throughout the program. Here are some screenshots regarding the choice of data visualisation.







Modelling methods:

Predictive analytics software relies on methodologies including the following:

Logistic regression:

A statistical analysis method used to predict a data value based on prior observations of a data set.

Time series analysis:

An illustrations of data points at successive time intervals.

Decision trees:

A graph that uses a branching method to illustrate every outcome of a decision.

All these methods uses a concept of confusion matrix.

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Challenges and limitations:

The oversimplification of data is one of the biggest draws of visualization is its ability to take big swaths of data and simplify them to more basic, understandable terms. However, it's easy to go too far with this; trying to take millions of data points and confine their conclusions to a handful of pectoral representations could lead to unfounded conclusions, or completely neglect certain significant modifiers that could completely change the assumptions you walk away with.

The human limitations of algorithms is the biggest potential problem, and also the most complicated. Any algorithm used to reduce data to visual illustrations is based on human inputs, and human inputs can be fundamentally flawed. For example, a human developing an algorithm may highlight different pieces of data that are "most" important to consider, and throw out other pieces entirely; this doesn't account for all companies or all situations, especially if there are data outliers or unique situations that demand an alternative approach. The problem is compounded by the fact that most data visualization systems are rolled out on a national scale; they evolve to become one-size-fits-all algorithms, and fail to address the specific needs of individuals.

Already, there are dozens of tools available to help us understand complex data sets with visual diagrams, charts, and illustrations, and data visualization is too popular to ever go away. We're on a fast course to visualization taking over in multiple areas, and there's no real going back at this point. To some, this may not seem like a problem, but consider some of the effects—companies racing to develop visualization products, and consumers only seeking products that offer visualization.

Conclusion:

The tools like machine learning is not only limited to use for developers and data scientists. Now anyone with the need of it can build and design it Just with use of this automated data analysis tool anyone with basic knowledge can use complex machine learning algorithms with simplicity. It will also help data scientists to skip the steps of cleaning and processing the data and directly jump to the analysis part. This

will not only save their time but also they can work more on analysing the data with results produced by automated tools.

Acknowledgement:

I would like to appreciate the work done in the data robot software. We have taken a lot of inspiration from the working of data robot software. The working stages and the mechanics of the software were useful as a reference for the framework of this project.

References:

[1]- "Java ML Library" - http://java-ml.sourceforge.net/content/java-machine-learning-library-0.1.7-released
[2]- Wikipedia "Data Analysis"- https://en.wikipedia.org/wiki/Data_analysis

[3]- Wikipedia "Data Visualization"- https://en.wikipedia.org/wiki/Predictive_analytics

[4]- Wikipedia "Predictive Analytics"- https://en.wikipedia.org/wiki/Predictive_analytics

[5]-Data Robot "Automated Machine Learning"- https://www.datarobot.com/product/automated-machine-learning/