Influence of Historical Meteorological Data Processing in a Mobile Application of Weather Prediction, based on Data Mining

Herwin Alayn Huillcen Baca and Flor de Luz Palomino Valdivia

School of Systems Engineering, National University José María Arguedas, Andahuaylas, Peru
hhuillcen@gmail.com, fdeluz3@gmail.com

Abstract

In the city of Andahuaylas Peru, agricultural region, the conditions of the weather and especially the temperature is decisive for the success or failure of the agricultural campaigns, on the other hand the information of UV radiation conditions the prevention of health and life in general, this information is shown by free internet services, but in a wrong way, because it comes from a remote location of the city of Andahuaylas and also does not provide detailed predictions for proper decision making, in this way the present investigation had the purpose of evaluating the influence of historical meteorological data processing on the efficiency of a mobile application of prediction of temperature and UV radiation, which provides real, updated and predicted information of the weather of the city of Andahuaylas, for it was used the historical and current data of the meteorological station of the National University José María Arguedas of Andahuaylas, which were analyzed using data mining to obtain efficient prediction models and were subsequently implemented in the mobile application. To measure the efficiency of the prediction, we compared the mean absolute errors of the models used by the National Service of Meteorology and Hydrology of Peru, of the cities of Arequipa, Iquitos and Lima, with values of 1.24, 2.66 and 1.485 respectively and the mean absolute error of the prediction model of the mobile application with a value of 1.18, which verifies the efficiency of the proposed model and is the expected result.

1 Introduction

In our planet, the elements of the weather are fundamental for the development of the life in general, they are periodic natural phenomena and depends on factors like the geographical location. In the case of Peru, specifically in the city of Andahuaylas, whose major source of economic development is agriculture, the temperature associated with UV radiation is vital and essential, the temperature variations condition the success or failure of a campaign Agricultural, and on the other hand, the levels of UV radiation directly affect the health of the villagers. In this way, predictive and real-time information on temperature and UV radiation is of great importance.

Under this approach, a mobile application for prediction the temperature and UV radiation was implemented, using the historical meteorological information of a station located in the city of Andahuaylas, which, when processed using data mining, generates prediction models that receive input data the current weather information and generates numerical prediction of temperature and UV radiation. Subsequently, the degree of approximation of the actual temperature and UV radiation levels against the prediction was evaluated.

This paper is a contribution as a generalization of the proposed predicted model and as a real source of information of the climate for the city of Andahuaylas.

2 Related works

Weather prediction is an important application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century, so many related works have been realized with data mining techniques and machine learning.

Olaiya and Barnabas (Folorunsho Olaiya, 2012), propose the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed. This was carried out

using Artificial Neural Network and Decision Tree algorithms and meteorological data collected between 2000 and 2009 from the city of Ibadan, Nigeria. A data model for the meteorological data was developed and this was used to train the classifier algorithms. The performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. A predictive Neural Network model was also developed for the weather prediction program and the results compared with actual weather data for the predicted periods. The results show that given enough case data, Data Mining techniques can be used for weather forecasting and climate change studies.

Khan, Muqeem and Javed (Sara khan, April 2016), propose that data mining is a technique that helps in extracting relevant and meaningful information from the set of data. It can be further described as knowledge discovery process that can be applied on any set of data. Data mining techniques when applied on relational databases can be used to search certain trends or patterns. This paper provides a survey of different data mining techniques being used in weather prediction or forecasting. It also reviews and compares various techniques being used in a tabular format.

Bartok, Habala, Bednar, Gazak and Hluchý (Juraj Bartok, 2010), present the methods and technologies for integration of the input data, distributed on different vendors' servers. The meteorological detection and prediction methods are based on statistical and climatological methods combined with knowledge discovery-data mining of meteorological data (messages, weather radar imagery, "raw" meteorological data from stations, satellite imagery and results of common meteorological prediction models).

3 Methodological Approach

3.1 Problem Approach

In the city of Andahuaylas there is a problem with temperature information, as free internet services provide wrong information, usually between three and five degrees less than the actual measurement, this causes confusion and uncertainty among the population.

The cause is that the free internet services takes as a source of data the meteorological information of the airport of Andahuaylas, which is located 17.5 km from the city, with an altitude of 11706 feet or 3568 meters (Corpac SA, 2015), against the altitude of the city of Andahuaylas of 2901 meters. (Regional Government of Apurimac - DIRCETUR, 2015), this difference of 667 meters makes both places present different climates, in addition to belonging to different natural regions.

The city of Andahuaylas and its environs have as main productive activity to the agriculture, which depends almost entirely on the elements of the weather, therefore it is important to have meteorological information predictive of the present day and later for an appropriate decision making, the problem arises because the prediction information is not correct. It is known that in southern Peru, UV radiation has the highest rates in the world, this problem directly affects the health of Peruvians through diseases of skin cancer and eye disease, the human exposure to UV radiation, in the case of Andahuaylas the problem is greater, because there is no source of information regarding the UV indices, both current and predictive, so that people adopt preventive measures during the hours that they will be exposed to the sun.

3.2 Research Method

The objective is to evaluate the influence of historical meteorological data on the efficiency of a mobile application prediction of temperature and UV radiation, based on data mining, in such a way that in case of obtaining a mean absolute error (MAE) (Pablo Cortes Achedad, 2010) acceptable predictive models and research in general, for the case of the temperature must have an average absolute error lower than 2.0 and for the case of UV radiation, less than 1.0.

Finally we propose a generalization of this approach, for the application of prediction models with an optimum amount of data and an efficient algorithm.

4 Analysis of meteorological data

In order to generate optimal models of prediction of temperature and UV radiation, based on the analysis of the behavior of classification algorithms of the WEKA tool (Waikato, 2010), taking as input, the meteorological data of the meteorological station of the National University José María Arguedas of Andahuaylas, whose records of temperature, UV radiation, humidity, wind speed and rain are intervals of 5 minutes. Specifically, prediction model algorithms were chosen for tem-

perature prediction after 24 hours, temperature prediction after 48 hours, prediction of UV radiation after 24 hours and prediction of UV radiation after 48 hours.

4.1 Pre-processing of data

The processed data are registered in a mySQL database of the web server of the National University José María Arguedas of Andahuaylas, these data correspond to all elements of the climate recorded every 5 minutes, for the research was used data of 6 months of registration.

4.2 Selection of variables or characteristics

The variables or characteristics correspond to the attributes taken into account in the generation of input files, these variables are hour, minute, temperature, absolute humidity, wind speed, rainfall and UV radiation. The objective was to evaluate which elements are correlated with temperature and UV radiation. Figures 1 and 2 show correlations of variables.

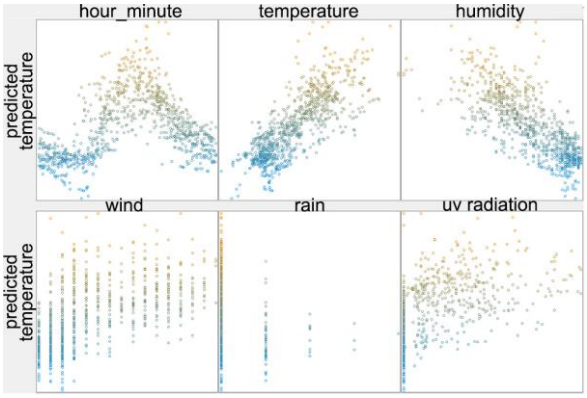

Figure 1: Correlation of variables for temperature prediction

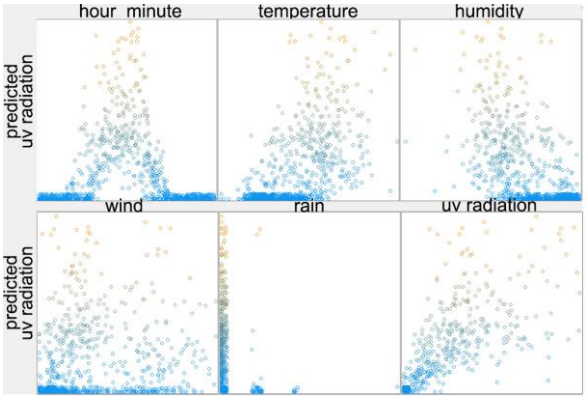

Figure 2: Correlation of variables for UV radiation prediction

4.3 Prediction algorithms

For the selection of candidate algorithms to generate the prediction models, we used as reference the investigations that were used as antecedents (José M. Molina, 2010); however, they can not be used in their entirety, since it depends very much on the nature of the Attributes and class attributes, so we have the following classification algorithms for generating prediction models: reptree, m5p, kstar, linear regression, additive regression, bagging, decision table, conjunctive rule, simple linear regression. In fact, the WEKA tool (Waikato, 2010) does not allow the use of other available classification algorithms. There are other algorithms that were not taken into account by the high mean absolute error (MAE) that results from the data analysis.

4.4 Extraction of knowledge

The input files are 28 corresponding to each prediction taking data of 60, 30, 15, 7, 3 and 2 days prior to data collection, counting input files and candidate classification algorithms, proceeded to train the respective algorithms for each prediction model, in such a way to be compared among them by means of the absolute average error.

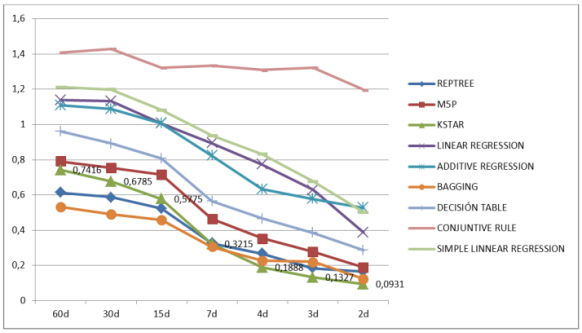

Figure 3: Results of algorithm training for temperature prediction models at 24 hours

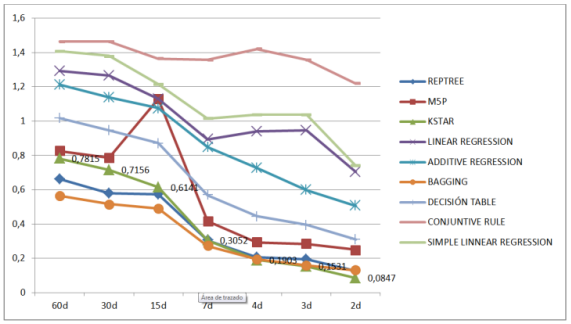

Figure 4: Results of algorithm training for temperature prediction models at 48 hours.

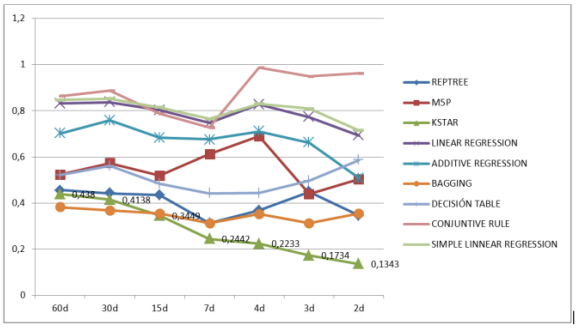

Figure 5: Results of algorithm training for UV radiation prediction models at 24 hours.

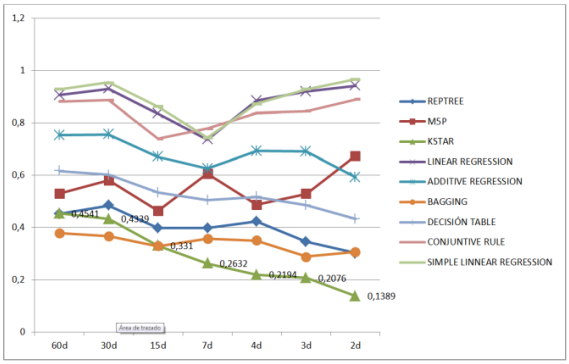

Figure 6: Results of algorithm training for UV radiation prediction models at 48 hours.

4.5 Interpretation and Evaluation

From Figure 3, it is observed that as the size of historical meteorological data decreases, the mean absolute error (MAE) of temperature predictions to 24 hours is similar to the results of the training of algorithms for prediction models of Temperature to 48 hours, UV radiation to 24 hours and UV radiation to 48 hours. For the evaluation of the choice of the prediction algorithm, it can be seen that in Figures 3, 4, 5 and 6, the algorithm that obtains the minimum value of the mean absolute error is the KSTAR algorithms, with a value of 0.0931 For models of prediction of temperature at 24 hours, of UV radiation at 24 hours and of UV radiation at 48 hours yields values of 0.0847, 0.1343 and 0.1349 respectively, all of them through the algorithm KSTAR. Finally **we conclude for the generation of prediction models we used the KSTAR algorithm and the data size are 2 days before the date and time of prediction**.

5 Construction of Mobile Application.

5.1 Mobile Application

The mobile application was developed for platforms Android (Google, 2015), works from version 4.0 API Level 14, the development tool used was Android Studio Beta 0.8.6. The choice of the platform and the development tool obey the non-functional requirement to provide free information and easy access.

Currently the mobile application is available for download in the "Google Play" repository (Google, 2015), under INFORAD's name, it is free and available since July 2015. Figure 7 shows the main interface.

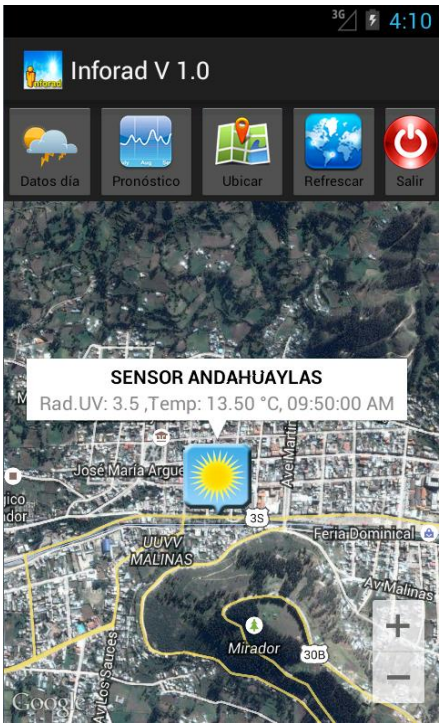

Figure 7: Screenshot of the main interface of the INFORAD mobile application

5.2 General Scheme of the application

The INFORAD mobile application has the following functional requirements:

- Show the temperature prediction for the current day for each hour of the day.
- Show the temperature prediction for the next day for each hour of the day.
- Show the prediction of UV radiation for the current day for every 30 minutes of the day, from 6:00 a.m. to 6:00 p.m.

- Show the prediction of UV radiation for the next day for every 30 minutes of the day, from 6:00 a.m. to 6:00 p.m.
- Display real-time information on temperature and UV radiation levels, updated every 5 minutes.
- Show health recommendations according to UV radiation levels.

The only non-functional requirement is that the INFORAD mobile application must provide free and easily accessible information. To satisfy these requirements, several components are required to work synchronously, Figure 8 shows the component diagram.

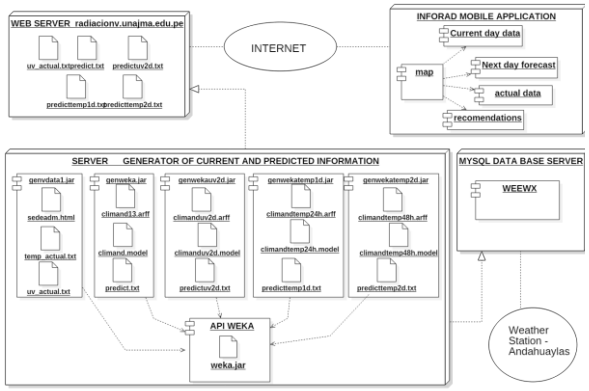

Figure 8: Component diagram of the INFORAD mobile application

5.3 Meteorological Station

The meteorological station used in the present investigation is the DAVIS INSTRUMENTS CORPORATION[1], model WIRELESS VANTAGE PRO2, this equipment was acquired in September of 2014 and installed in November of 2014.

5.4 Database Server

The disadvantage of the meteorological station is the lack of database connection, so we used the mySQL server of the university, whose platform is Debian GNU / Linux[2], version 7.0, however another service was required to connect to the weather station, Extracts data every 5 minutes and registers them in the mySQL database, then the WEEWX[3] service is installed and configured, which is free and meets the requirements.

1 http://www.davisnet.com/
2 http://www.debian.org/
3 http://www.weewx.com/

5.5 Generation of current and predicted information.

The data mining process for knowledge extraction is a process involving hardware resource consumption and resource time, which can not be loaded to a mobile device, for reasons of hardware and processing limitations, then developed programs that generate current and predicted weather information, executed on the general purpose server of the Universidad Nacional José María Arguedas, the resulting information is uploaded to subdomain of the university (http://radiacionv.unajma.edu.pe), for later use by the INFORAD mobile application. Figure 9 shows the prediction interface through intuitive and easy-to-read graphs.

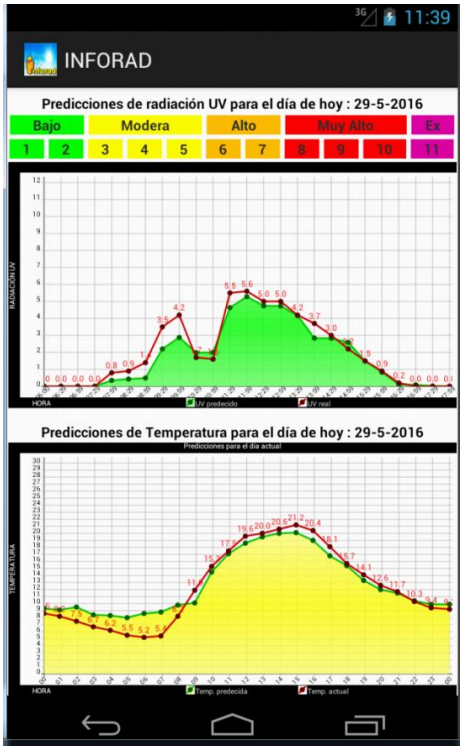

Figure 9: Screenshot of the prediction interface of INFORAD mobile application

5.6 Usage statistics.

According to figures 10 and 11, it is observed that there are 706 installations per user, from July 2015 to July 2017, likewise has an average rating of 3,778.

Figure 10: Component diagram of the INFORAD mobile application

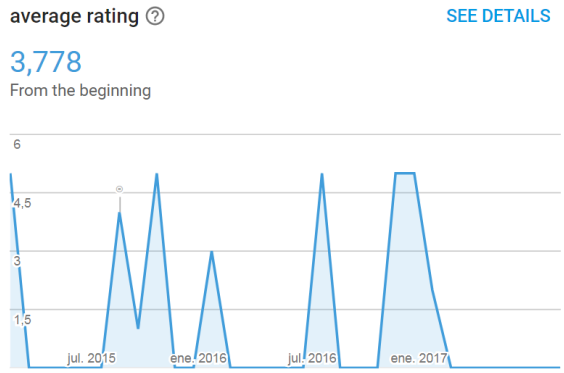

Figure 11: Component diagram of the INFORAD mobile application

6 Result Obtained

6.1 Result of Errors Obtained

To evaluate this efficacy, we used the mean absolute error (MAE) of each prediction versus the real value and then the mean of all predictions to obtain the mean absolute error (MAE), which refers to the efficacy of the prediction. However, it is also important to calculate the mean absolute percentage error (MAPE) (Pablo Cortes Achedad, 2010). Five consecutive days of evaluation of the predictions were chosen, reaching the following results:

	Temperature Prediction		UV Radiation Prediction	
	At 24 hours	At 48 hours	At 24 hours	At 48 hours
MAE	1.18	1.45	0.98	0.87
MAPE	9.32%	12.00%	38.62%	40.77%

Table 1: MAE and MAPE values of predictions

6.2 Results of prediction effectiveness

The type of numerical prediction addressed in the present research corresponds to a regional model, so to evaluate the effectiveness of the results of the predictions, a comparison of the mean absolute error (MAE) of the proposed model with Other prediction models of the region, in this case there are no predictive models for the city of Andahuaylas, but there is information about the effectiveness of the ETA model (Vergaray, GJ (2010), SENAMHI, 2013) of the National Service of Meteorology and Hydrology of Peru, for some Cities of Peru, as follows:

Model	City	MAE
ETA / SENAMHI	Arequipa	1.24
ETA / SENAMHI	Iquitos	2.66
ETA / SENAMHI	Lima	1.485
KSTAR/INFORAD	Andahuaylas	1.18

Table 2: Comparison of mean absolute errors (MAE) of temperature prediction to 24 hours for ETA models and the one raised

7 Conclusions and Future Work

The analysis and evaluation of historical meteorological data, through data mining, have an optimal influence on the efficiency of the mobile application of prediction of temperature and UV radiation, since smaller errors have been obtained than the ETA model (Vergaray, GJ (2010) , Currently used by the National Service of Meteorology and Hydrology of Peru, for the case of temperature prediction at 24 hours, the present investigation has an average absolute error of 1.17 compared to a value of 1.80 generated by the ETA model.

The KSTAR classification algorithm is the most suitable for the generation of prediction models of temperature and UV radiation, for the city of Andahuaylas.

The optimum data size for general prediction models of temperature and UV radiation for the city of Andahuaylas is 2 days ago, as it is proven that the prediction is more accurate when taking near-occurrence data.

The degree of certainty or approximation of temperature predictions is better predicted for the next day than for the subsequent day, because the mean absolute percentage error (MAPE) is 11.46% and 12.0% respectively.

The degree of certainty or approximation of predictions of UV radiation is better predicted for the next day than for the subsequent day, because the mean absolute percentage error (MAPE) is 38.62% and 40.77% respectively.

Predicting the temperature generates errors less than predicting UV radiation, because the temperature has more stable values with respect to UV radiation.

It is possible to implement prediction models and later prediction applications for other elements of the climate that are also important as information, such as rainfall, humidity, atmospheric pressure, wind speed and direction.

It is known that the weather in general is periodic, it has an annual repetition cycle, so it would be interesting to analyze the historical meteorological data of at least 2 years ago to generate predictions for the whole following year, day by day, even hour per hour.

8 References

Corpac SA (10 of 05 of 2015). Andahuaylas airport. Retrieved on September 10, 2015, Andahuaylas Airport: http://www.corpac.gob.pe/Docs/Aeropuertos/AdmCorpac/ANDAHUAYLAS.pdf

Folorunsho Olaiya, A. B. (2012). Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. *I.J. Information Engineering and Electronic Business, 1*, 51-59.

Google. (2015). Android. Retrieved 09 2015, https://www.android.com/

José M. Molina, J. G. (2010). Técnicas de Minería de Datos basadas en Aprendizaje Automático. Recuperado el 10 de 09 de 2015, de Técnicas de Minería de Datos basadas en Aprendizaje Automático: https://santiagozapatakdd.files.wordpress.com/2011/03/curso-kdd-full-cap-3.pdf

Juraj Bartok, O. H. (2010). Data mining and integration for predicting significant meteorological phenomena. *Procedia Computer Science, Volume 1, Issue 1, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2010.04.006.*, Pages 37-46.

Keffer, T. (2010). WeeWX: weather Linux software. Retrieved on September 10, 2015, of WeeWX: Linux weather software: http://www.weewx.com/

Pablo Cortes Achedad, LO (2010). Organization Engineering. As Pablo Cortes Achedad, Engineering Organization (pp. 349, 350). Madrid, Spain: Díaz de Santos SA

Regional Government of Apurimac - DIRCETUR. (2015). Sub Regional Directorate of Foreign Trade and Tourism - Andahuaylas. Retrieved on September 10, 2015, Sub Regional Directorate of Foreign Trade and Tourism - Andahuaylas:http://dirceturandahuaylas.gob.pe/Principales-Atractivos-Turisticos.php

Sara khan, M. M. (April 2016). A Critical Review of Data Mining Techniques in Weather Forecasting. IJARCCE - International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4.

SENAMHI. (2013). The weather forecast, cap 13. Retrieved September 10, 2015, of the weather forecast: http://200.58.146.28/nimbus/weather/pdf/cap13.pdf

Vergaray, GJ (2010). Verification of the temperature of the ETA - SENAMHI model. Retrieved 09 2015, http://ftp.cptec.inpe.br/etamdl/WorkEtaIV/Estudo_de_Caso/Estudo_de_Caso_WorkETA_4_Gerardo_Vergaray.pdf

Waikato. (2010). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Retrieved on September 10, 2015, Weka 3 - Data Mining with Open Source Machine Learning Software in Java: http://www.cs.waikato.ac.nz/ml/weka/

Wikispaces. (2015). WEKA - ARFF stable version. Retrieved 09 2015, https://weka.wikispaces.com/ARFF+(stable+version).