ITCS 6162 - KNOWLEDGE DISCOVERY IN DATABASES

PROJECT REPORT

Team Members

Gaurav Pareek (801020037)

Nipun Gupta (801016923)

Rahul Bairathi (801019746)

TABLE OF CONTENTS

ABSTRACT

A data researcher has acute interest in obtaining valuable information stored in the database. The data is termed as raw facts and figures and it requires some sophisticated processing to output useful knowledge and information. As a student of Knowledge Discovery in Database, we are interested in finding patterns and strategies which may alleviate our awareness about the world.

In this research project, we are performing experiments on the world ranking, with an aim to understand the impacts of various indicators on the reputation of a state. We also aim for understanding the actions on various indicators and how they affect the reputation as a whole.

The data is taken from Fragile States Index and it is fed to various classifiers and action rule miners. The results of classification and generated action rules are discussed at the end of the report.

INTRODUCTION

Due to explosive growth in the amount of data, it has become difficult to comprehend it. The strategies in the area of Knowledge Discovery in Databases aims to identify patterns in data and extract knowledge. It is done by employing several data mining algorithms with desired thresholds along with some necessary pre-processing. In this project, we are implementing two of the chief strategies in data mining: Classification Rules and Actions Rules mining. The dataset on which these strategies are practiced, is taken from “Fragile States Index” [1].

PROBLEM DESCRIPTION

This project consists of two tasks, namely; data classification and action rules mining. Here, we need to employ Machine Learning classification algorithms on the picked dataset from “Fragile States Indexes”. One of the major concerns in the Machine Learning is the accuracy of the model. For an attempt to improve the accuracy, we have added new features and performed discretization of attributes.

The second task consists of generating Action Rules to transition from the Alert state to the Sustainable state. This task was accomplished by employing “AC4ft-Miner”.

TOOLS USED

The experiments in this project are done using following two tools: WEKA [2] and Lisp Miner [3].

PROJECT PHASES

DATA SELECTION

The data selected for this experiment is from Fragile States Index (FSI) and Social Progress Index dataset (SPI).

The Fragile States Index (FSI) produced by The Fund for Peace (FFP), is a critical tool in highlighting not only the normal pressures that all states experience, but also in identifying when those pressures are outweighing a state's capacity to manage those pressures. By highlighting pertinent vulnerabilities which contribute to the risk of state fragility, the Index — and the social science framework and the data analysis tools upon which it is built — makes political risk assessment and early warning of conflict accessible to policy-makers and the public at large.

The attributes of selected data are as follows:

1. **Country:** Name of the country

2. **Year:** Year of the data
3. **Rank:** Rank of the country based on the "Total" attribute. The country with lowest total is ranked 1st.
4. **Total:** Total of all indicators. The indicators are in the range 0 – 10.
5. **C1: Security Apparatus:** The Security Apparatus indicator score is assigned by considering security threats to the country. The threats include terrorism, attacks, rebel movements, mutinies, etc.
6. **C2: Factionalized Elites:** The Factionalized Elites indicator score is assigned by considering the fragments of a country like religion, ethnicity, class, clan, etc. It measures the political transition, struggle and competition.
7. **C3: Group Grievance:** The Group Grievance considers political and social division among various groups of the society. This indicator scores a country based on the access to various services and resources to these group.
8. **E1: Economy:** This indicator consider the economic decline for the state. It considers factors such as per-capita income, Gross National Product, Inflation, Productivity, Unemployment Rate, debt, business failures etc.
9. **E2: Economic Inequality:** The economic inequality indicator scores a country on the basis of its economic inequality across various dimensions. For example, identity groups (race, religion, ethnicity), economic status, education, region etc.
10. **E3: Human Flight and Brain Drain:** This considers the impact of human displacement on the economy of the country and country's development. This may involve emigration of workers, labors or other middle working class people.
11. **P1: State Legitimacy:** This indicator measures a country's performance on the basis of representativeness and openness of the governing bodies and its relationship with the citizens. The indicators looks at the level of confidence among the citizens, processes and state institutions.
12. **P2: Public Services:** The public services indicator scores the basic state functions for serving people. This includes but is not limited to, provision of essential services such as transport, electricity, education, health, water and sanitization.
13. **P3: Human Rights:** This indicator ranks on the basis of observation of fundamental human rights of the people. It looks at the abuse of legal social rights or people groups and institutions.
14. **S1: Demographic Pressures:** This indicator measures the pressures on the population related to access for various services like health, food supply, safe-water and other life sustaining resources.
15. **S2: Refugees and IDPs:** It measure the pressure on a state by displacement of large communities due to political, environmental or social reasons. It also considers the refugee inflow.
16. **X1: External Intervention:** It considers the impact of external entities in the functioning of the state's governance.

One of the chief task in our experiment is to devise ways to improve the accuracy of our classification and action rule model. For this purpose, the FSI dataset is augmented with additional attributes. The augmented dataset is taken from **"Social Progress Index"**[4], which is an aggregation of social and environmental indicators for the countries of the world. The data span across three dimensions: Human Needs, Wellbeing and Opportunities.

There were several indicators available for augmenting the dataset. However, the social progress indicators gives information about the attributes that really matter. It is not something very generic like income, population, literacy rate, etc. However, it discusses the capacity of a state to strengthen their basic building block and allow its citizen to lead and sustain a quality life.

The attributes that were selected for augmentation are as follows:

1. **Personal Safety:** The dimensions evaluated by Personal Safety are: Homicide Rate, Level of Crime, Traffic Death, Criminality and Political Terror.
2. **Environmental Quality:** The Environmental Quality attribute measures state index based on factors like emission of greenhouse gases, biodiversity and habitat, wastewater treatment and outdoor air pollution.
3. **Tolerance and Inclusion:** This attribute include ranks for tolerance of immigrants, homosexuals, discrimination against minorities and community safety.
4. **Press Freedom Index :** This indicator is an evaluation of freedom of expression in private discussion, academia or culture related backgrounds.
5. **Tolerance for immigrants:** It is based on a survey asking individuals whether the area in which they live is safe for immigrants or not.
6. **Discrimination and violence against minorities:** It is an indicator representing group grievances and violence on the grounds of culture, religion and community.

PRE-PROCESSING

DECISION ASSIGNMENT

In the dataset, the decision attribute is "Total", which only contains continuous numerical values. However, the algorithms for classification and action rules generation require nominal values for the decision. Therefore, as the foremost pre-processing steps, we have classified the "Total" using percentile ranks. The selected classes are:

Decision Value	Percentile Range
Alert	75 - 100 percentile
Warning	50 - 74.9 percentile
Stable	25 - 49.9 percentile
Sustainable	0 - 24.9 percentile

The classes are filled using Microsoft Excel, by the use of PERCENTRANK function on the "Total" and "New Total" fields and the "Decision" and "New Decision" are the decision columns.

DISCRETIZATION OF ATTRIBUTE VALUES

The second pre-processing step was "Discretization". During classification, operating on numerical or continuous attributes has very limited suitability of suitability for new records. This Discretization process resolves this problem by reducing the size of the domain. It is done for all numerical attributes before generating classification rules to improve accuracy. As an example, the accuracy of the Naive Bayes classification model was raised from 53% to 59% for the year 2017 dataset, after discretization.

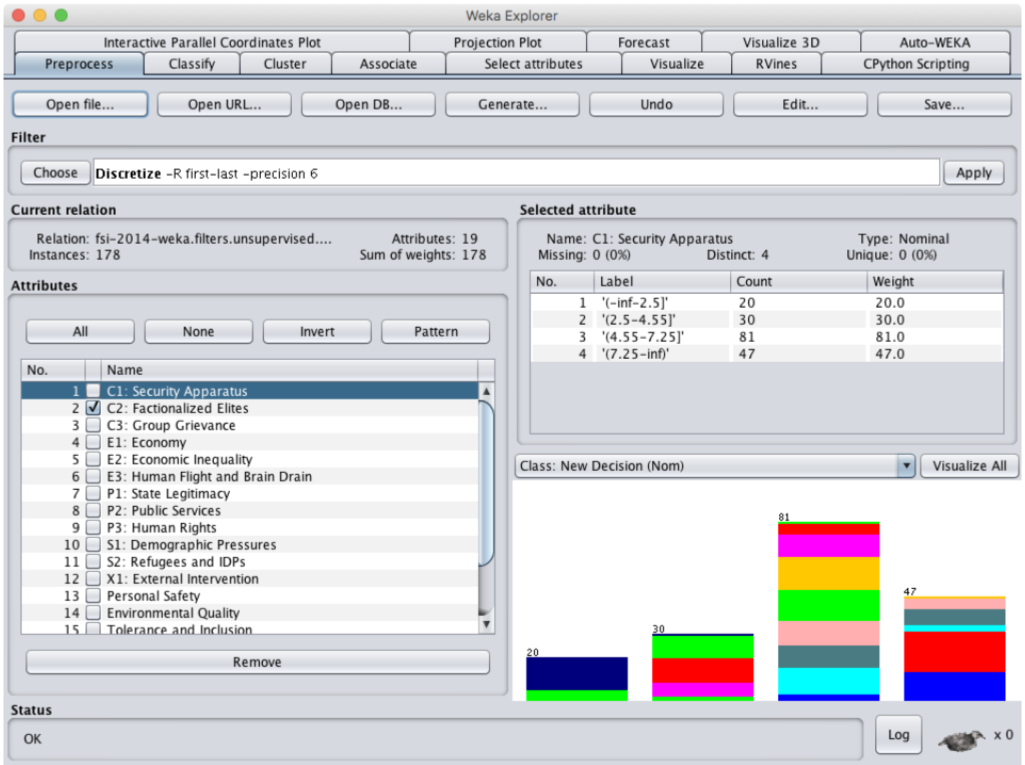

FIGURE 1: DISCRETIZED ATTRIBUTES IN WEKA

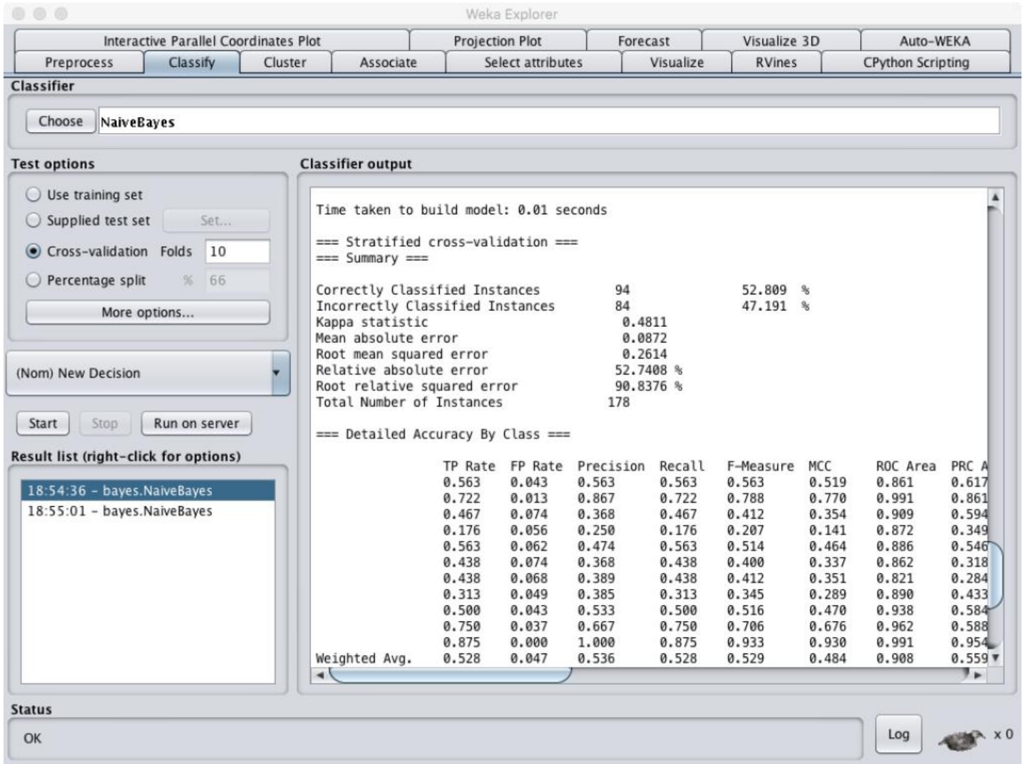

FIGURE 2: CLASSIFICATION RESULT WITHOUT DISCRETIZATION

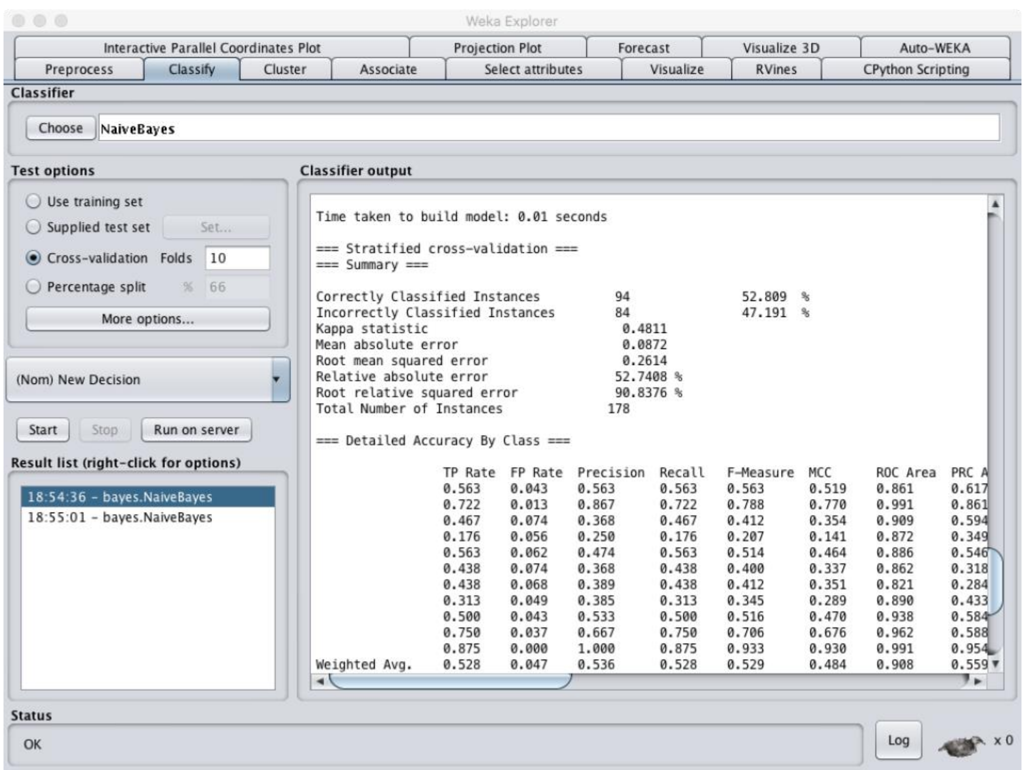

FIGURE 3:CLASSIFICATION PERFORMACE AFTER DISCRETIZATION

CLASSIFICATION OF DATA

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

NAIVE BAYES CLASSIFIER

Naive Bayes[5] is a very popular classification technique implementing using Bayes' probability theorem. It is based on the assumption that each feature of a particular class is independent of other features. This assumption simplifies the probability calculation method as if the attributes are independent of each other, then the complete probability can be calculated by mere multiplication of probability of individual attributes.

The probability relation of the Bayes' theorem is as follows:

$$P(t) = \frac{P(c).P(c)}{P(t)}$$

In the above relation, "c" represents class and "t" represents a tuple. P(c|t) is the posterior probability, which means the probability of class "c" given the data tuple "t". P(t|c) is the likelihood probability, which means that probability of occurrence of "t" given the class is "c".

For a dataset with multiple attributes, the Naïve Bayes' classifier replaces the likelihood probability in the above relation with the product of likelihood probability of each attribute given the class "c". The probability is calculate for every class and the class with highest probability is assigned to the testing set under consideration.

RESULTS

The classification results using Naïve Bayes' classifier for the dataset of years 2014 to 2017 are as follows:

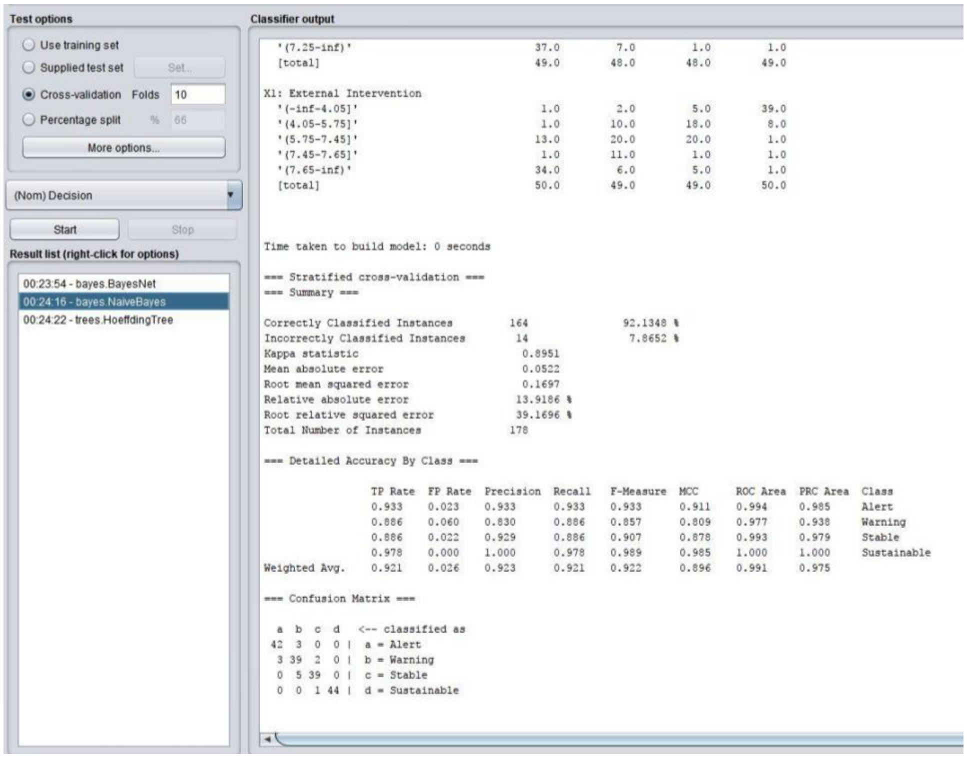

FIGURE 4: NAIVE BAYES - 2014 DATASET WITHOUT ADDITIONAL FEATURES

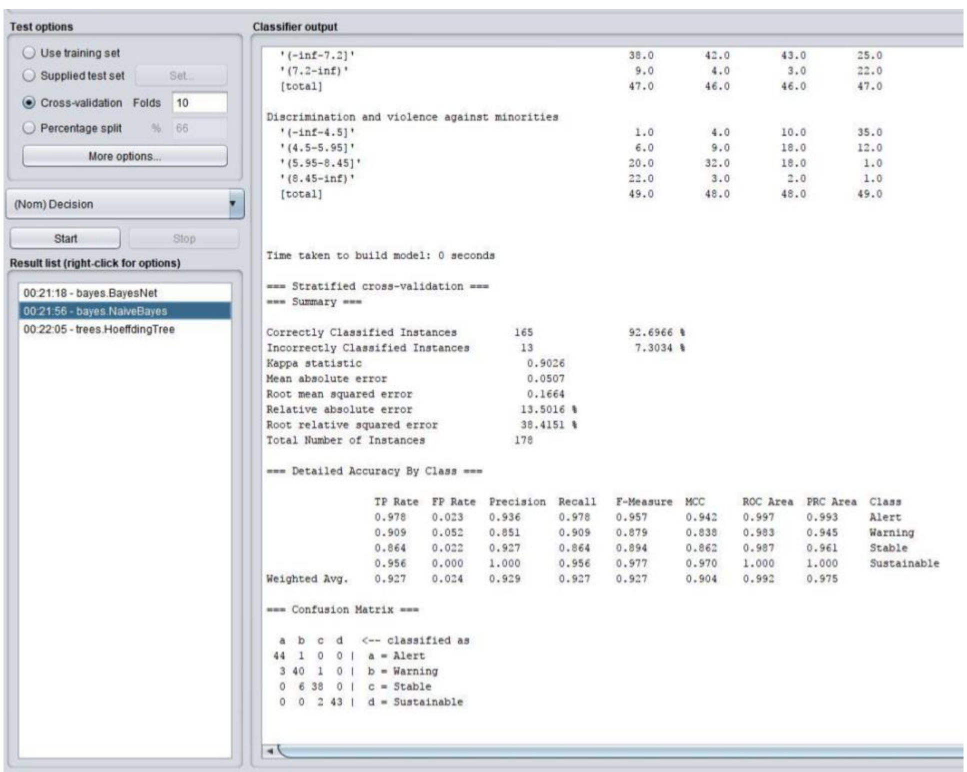

FIGURE 5: 2014 DATASET WITH ADDITIONAL FEATURES

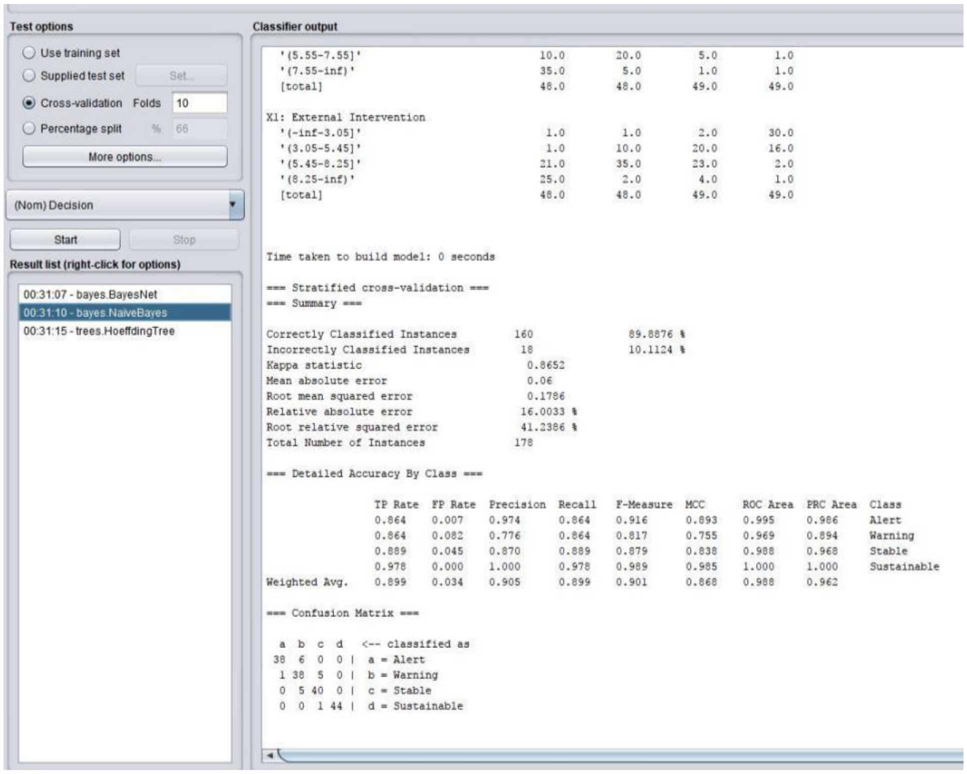

FIGURE 6: NAIVE BAYES - 2015 DATASET WITHOUT ADDITIONAL FEATURES

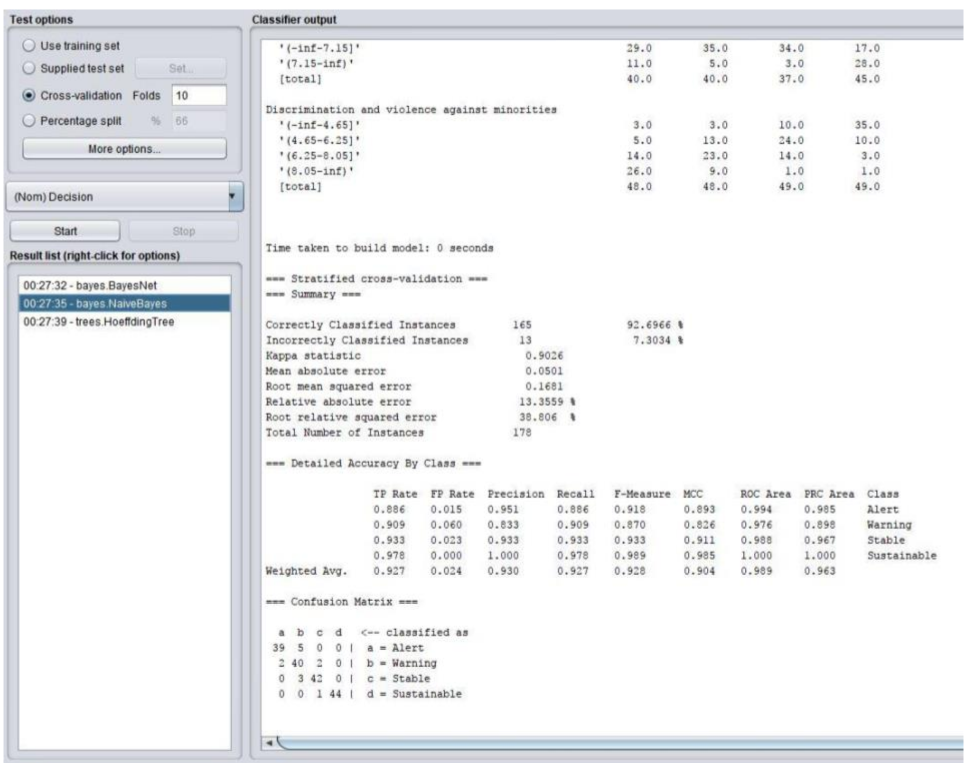

FIGURE 7: NAIVE BAYES - 2015 DATASET WITH ADDITIONAL FEATURES

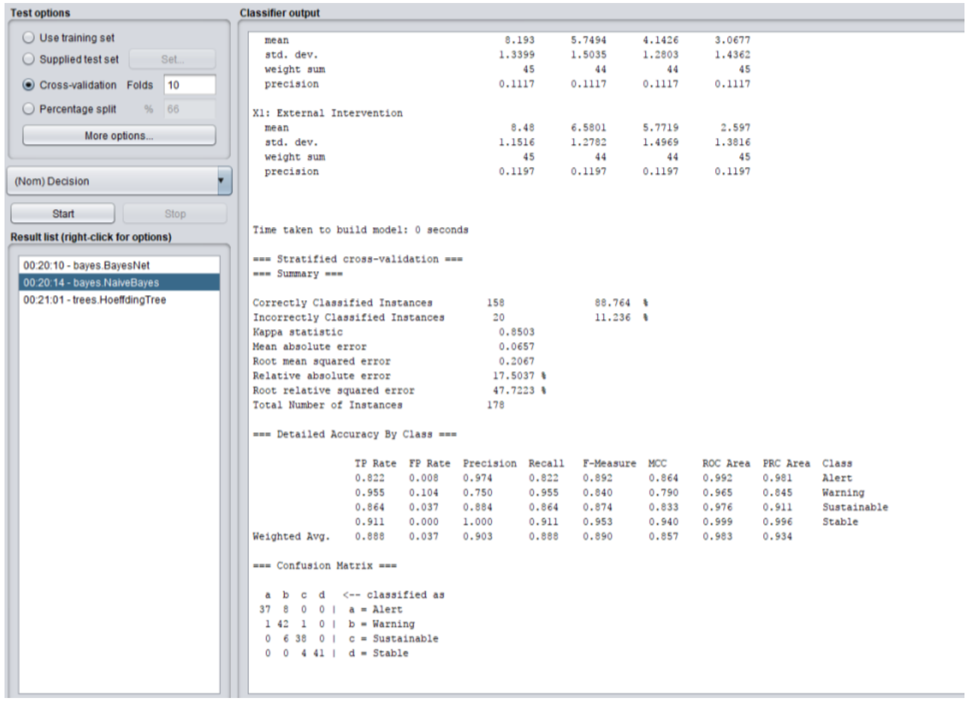

FIGURE 8: NAIVE BAYES - 2016 DATASET WITHOUT ADDITIONAL FEATURES

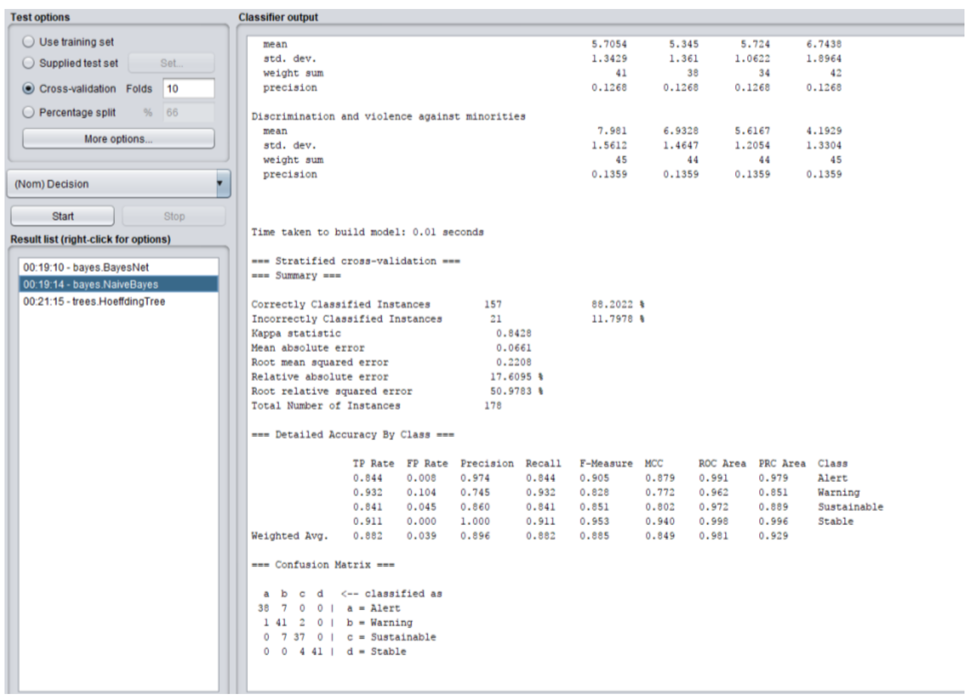

FIGURE 9: NAIVE BAYES - 2016 DATASET WITH ADDITIONAL FEATURES

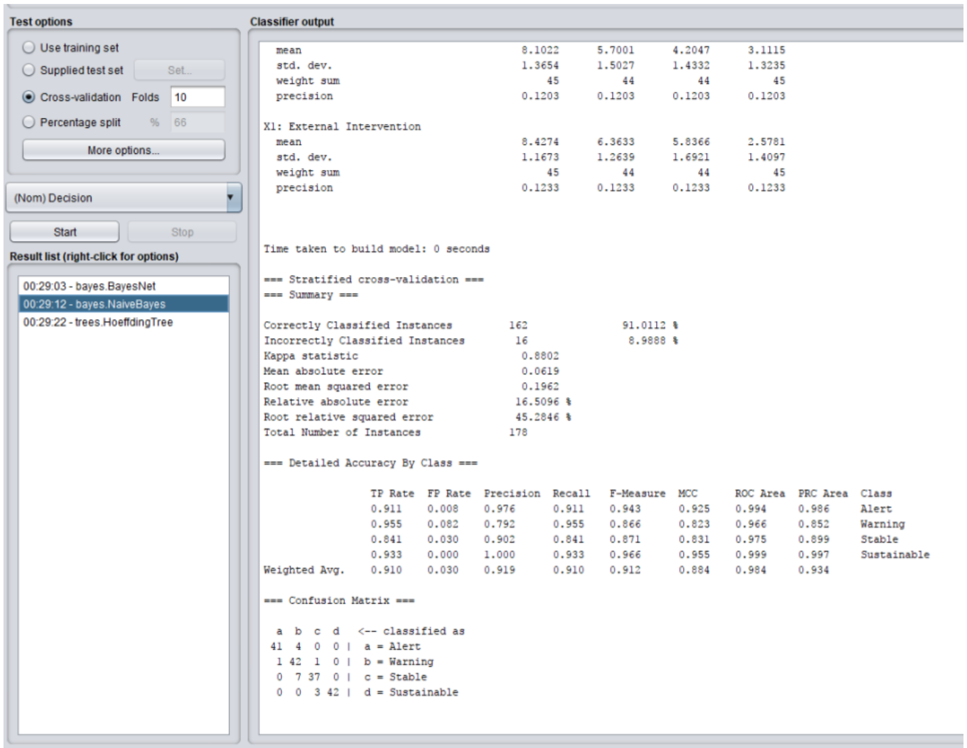

FIGURE 10: NAIVE BAYES - 2017 DATASET WITHOUT ADDITIONAL FEATURES

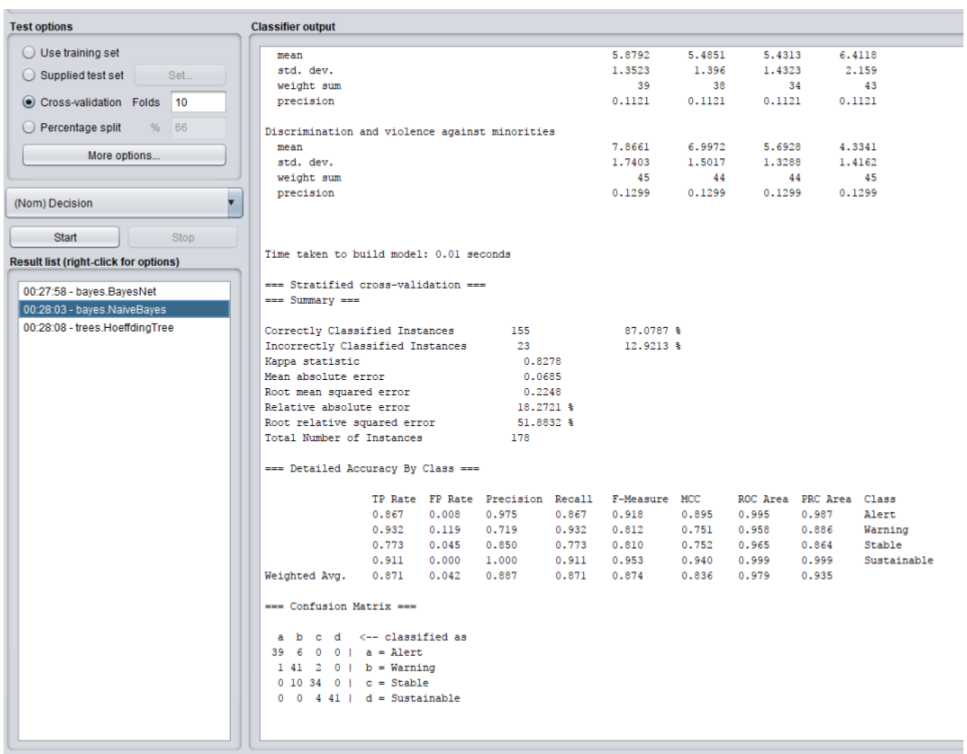

FIGURE 11: NAIVE BAYES - 2017 DATASET WITH ADDITIONAL FEATURES

BAYESNET CLASSIFIER

The most prominent feature of the Naive Bayes model is also its drawback. It seems impractical to ignore the independence relations among the features in many cases. This problem is tackled by the BayesNet model[6]. This model prepares a dependency graphs which are directed acyclic graphs. Each node of the graph represents a probability distribution over a set of random variables. The BayesNet network encodes conditional independence among the variables. In the network, each node is conditionally independent of is non-descendants, given the states of its parent nodes. This independence relation is utilized to reduce the number of parameters needed to characterize the probability distribution. Thus, the calculation of posterior probability becomes more efficient.

RESULTS

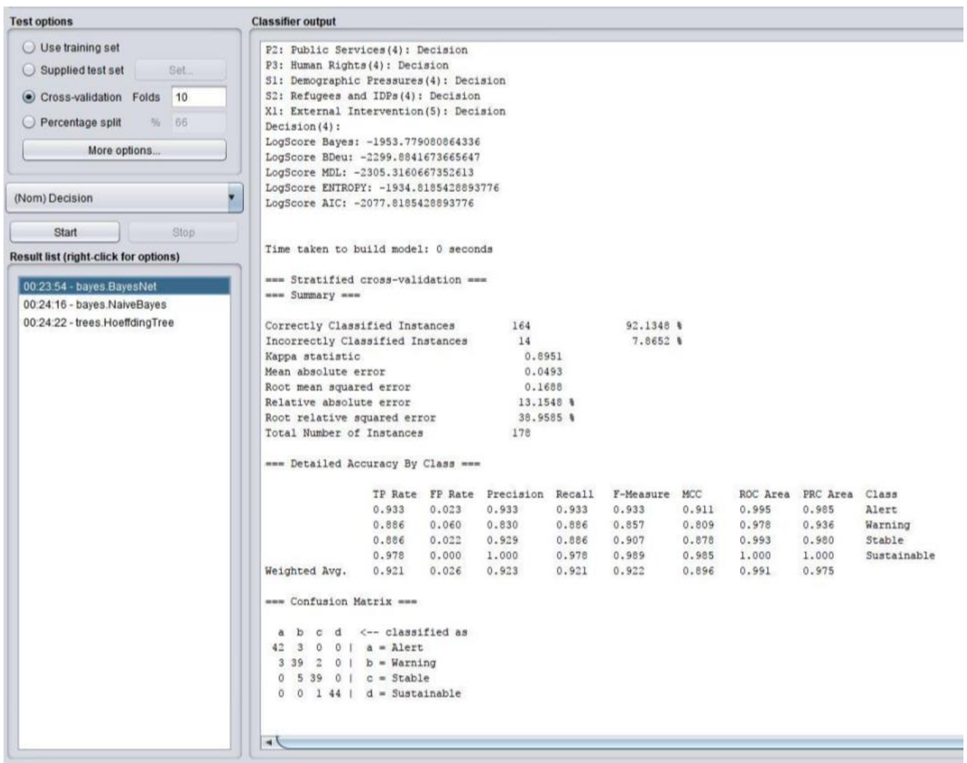

FIGURE 12: : BAYES NET - 2014 DATASET WITHOUT ADDITIONAL ATTRIBUTES

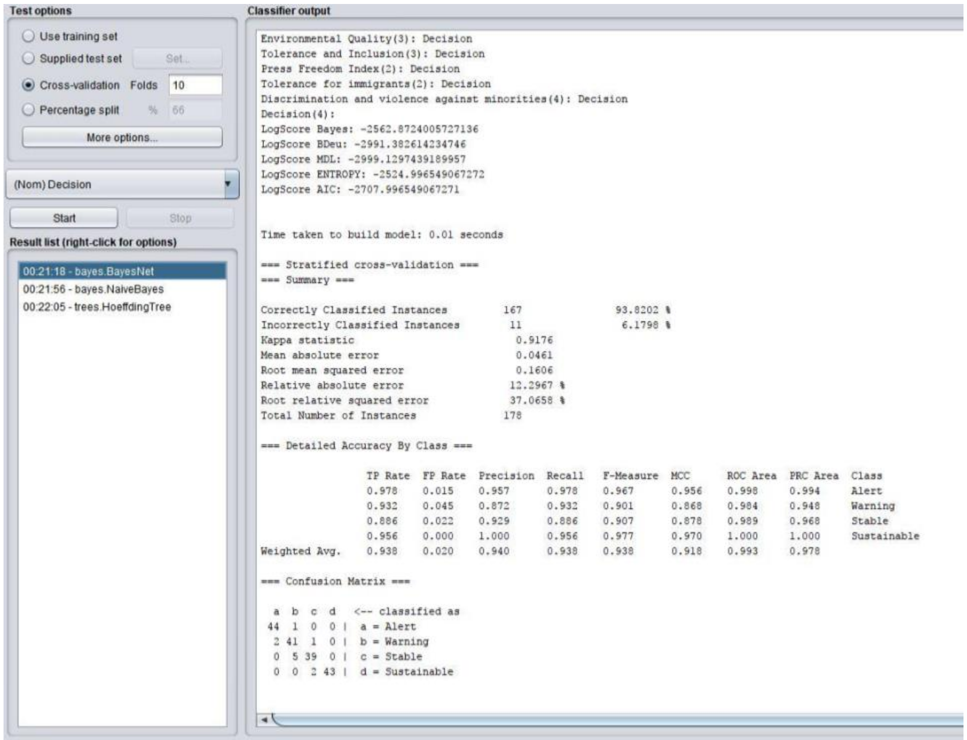

FIGURE 13: : BAYES NET - 2014 DATASET WITH ADDITIONAL ATTRIBUTES

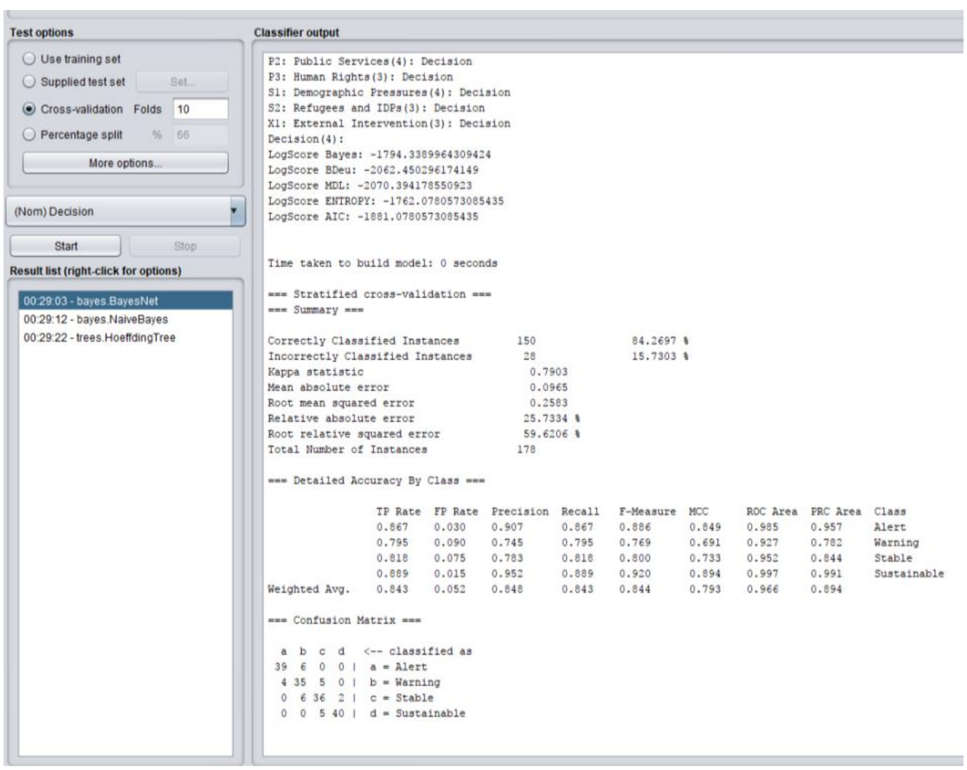

FIGURE 14: BAYES NET - 2014 DATASET WITHOUT ADDITIONAL ATTRIBUTES

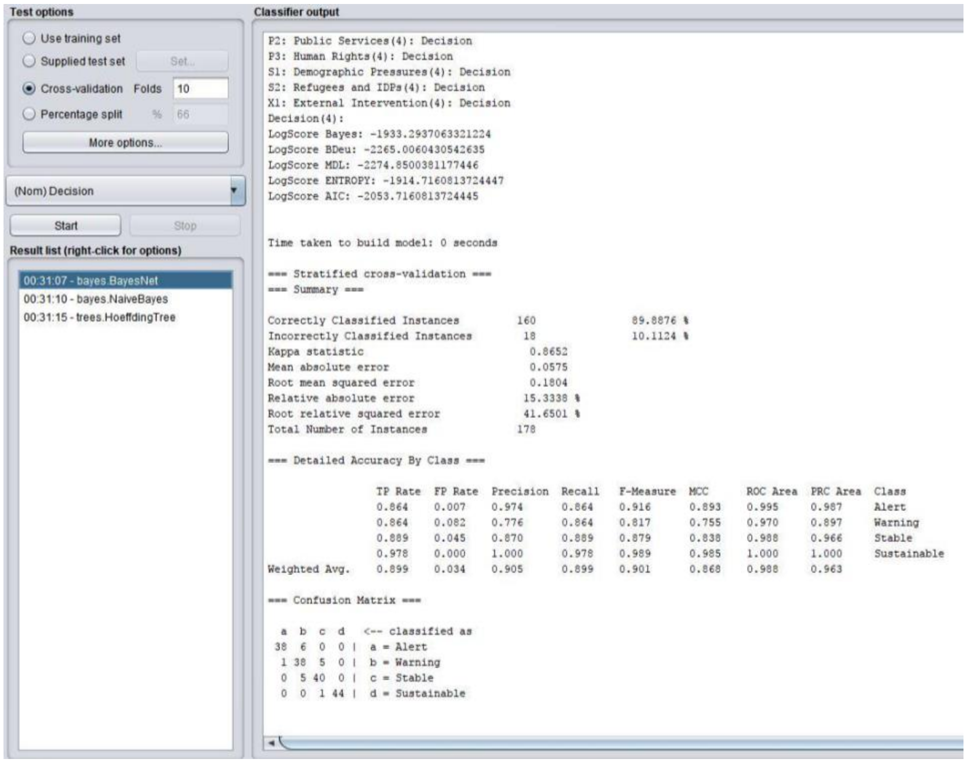

FIGURE 15: BAYES NET - 2015 DATASET WITHOUT ADDITIONAL ATTRIBUTES

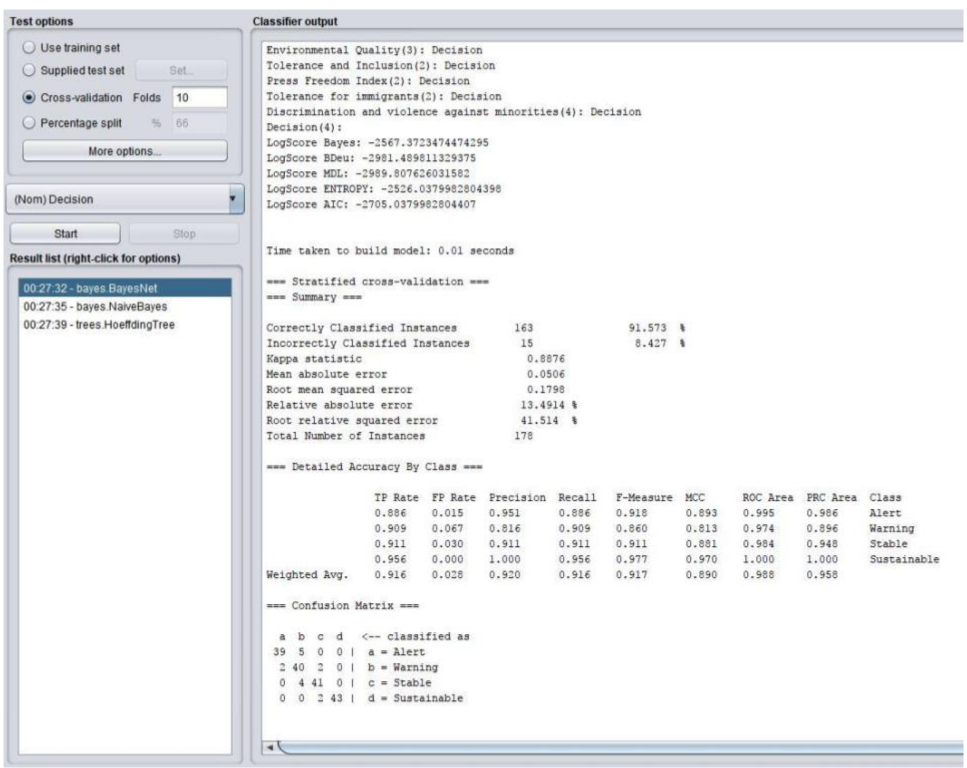

FIGURE 16: BAYES NET - 2015 DATASET WITH ADDITIONAL ATTRIBUTES

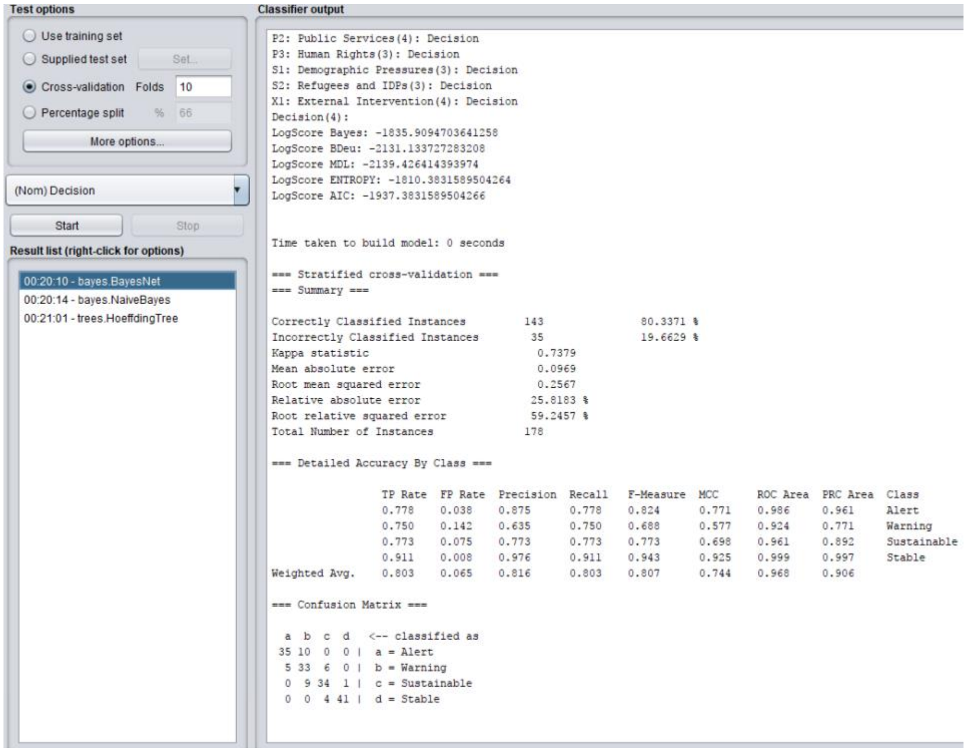

FIGURE 17: BAYES NET - 2016 DATASET WITHOUT ADDITIONAL ATTRIBUTES

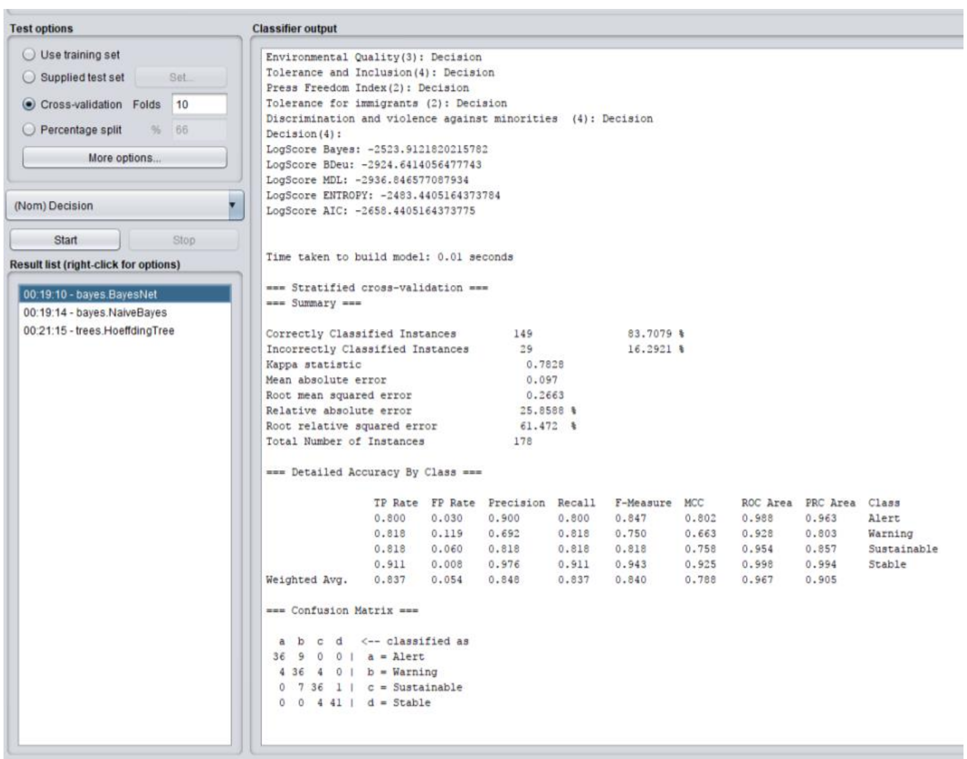

FIGURE 18: BAYES NET - 2016 DATASET WITH ADDITIONAL ATTRIBUTES

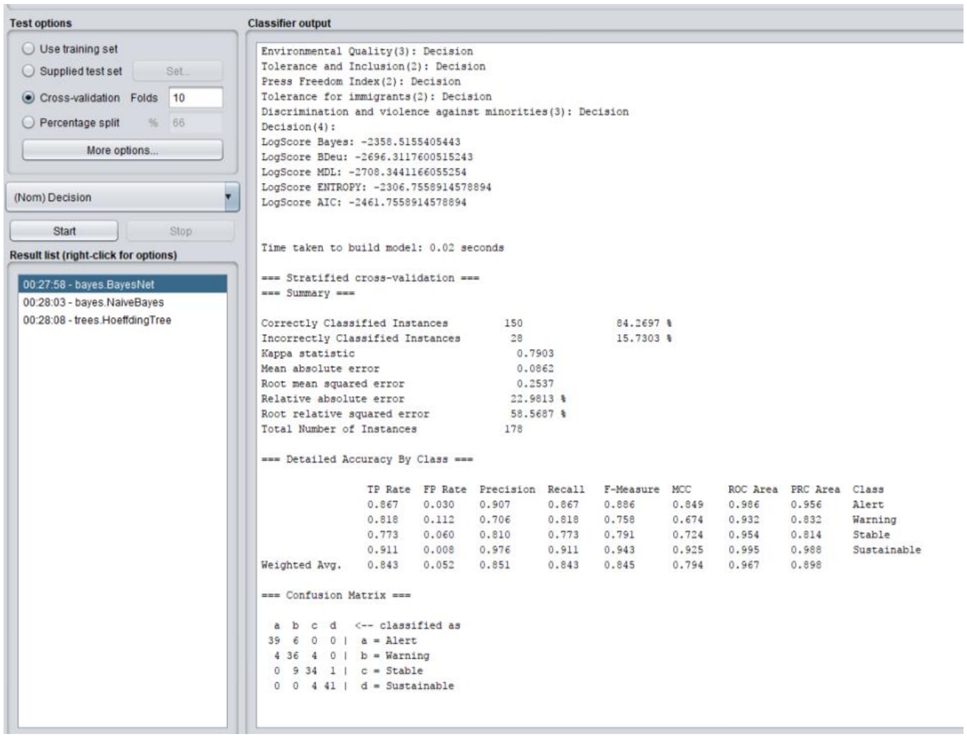

FIGURE 19: BAYES NET - 2017 DATASET WITH ADDITIONAL ATTRIBUTES

HOEFFDING TREE CLASSIFIER

A Hoeffding tree[7] (VFDT – Very fast decision tree) is an induction algorithm capable of learning from large datasets. It is based on the assumption that the examples which generate distribution of nodes does not change over time. This idea is leveraged with Hoeffiding bound, which is a quantification of the observations. This generates some statistics with a predefined precision, which is also known as the “goodness of an attribute” for classification. Hoeffding tree is a high performing algorithm both in terms of classification performance and accuracy.

RESULTS

The results of classification using Hoeffding tree classification is as follows:

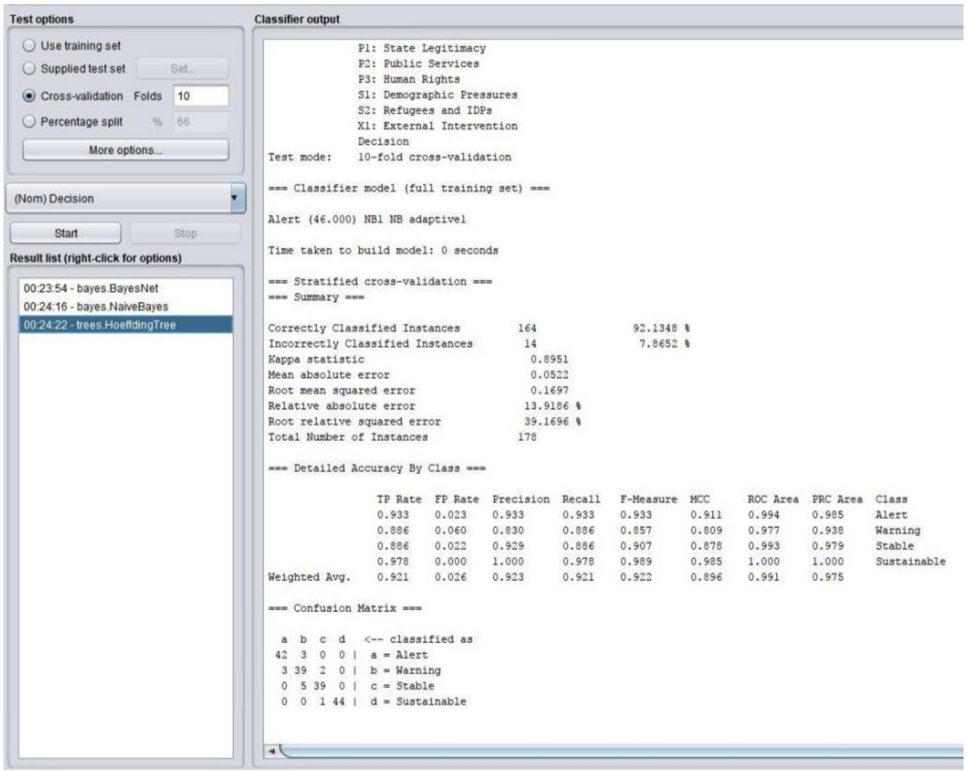

FIGURE 20: HOEFFDING TREE - 2014 DATASET WITHOUT ADDITIONAL ATTRIBUTES

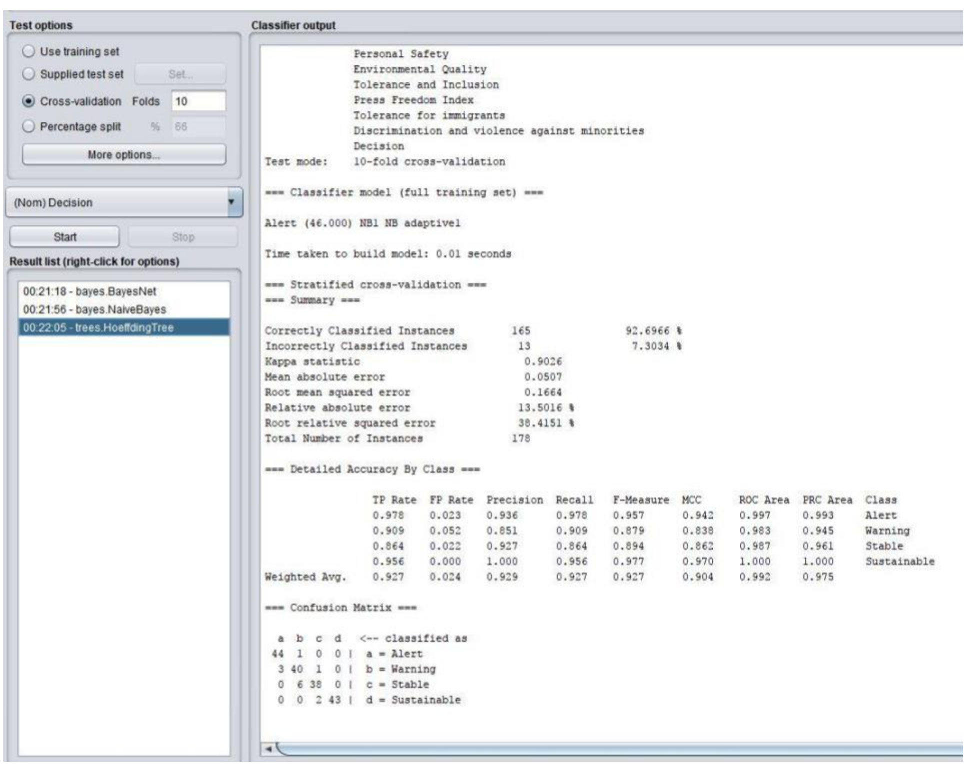

FIGURE 21: HOEFFDING TREE - 2014 DATASET WITH ADDITIONAL ATTRIBUTES

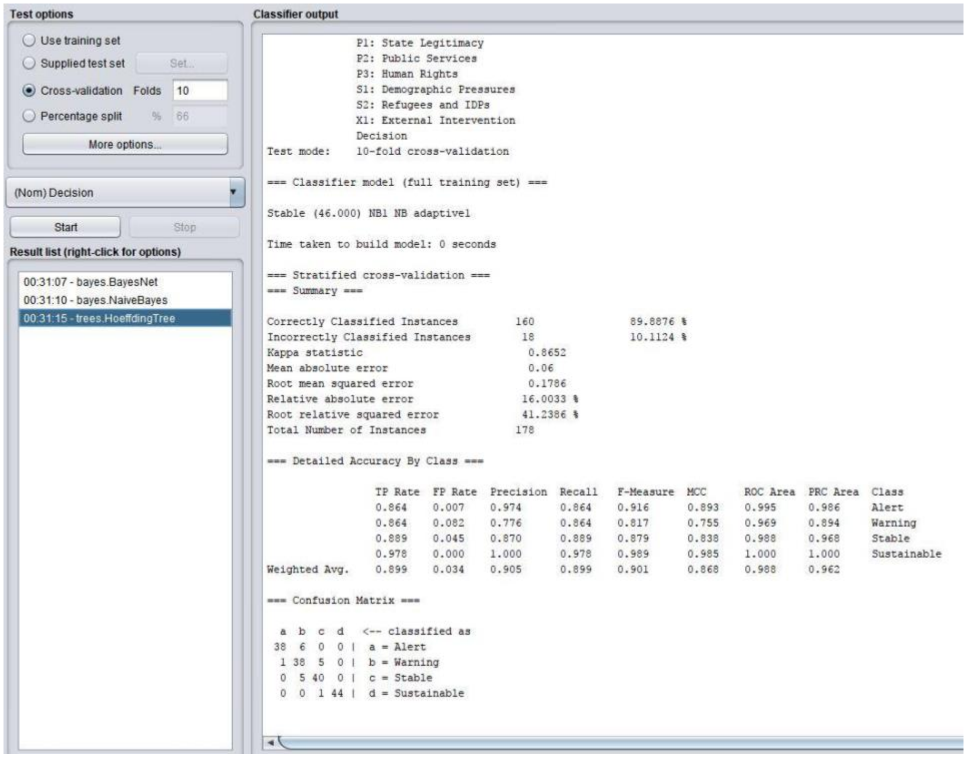

FIGURE 22: HOEFFDING TREE - 2015 DATASET WITHOUT ADDITIONAL ATTRIBUTES

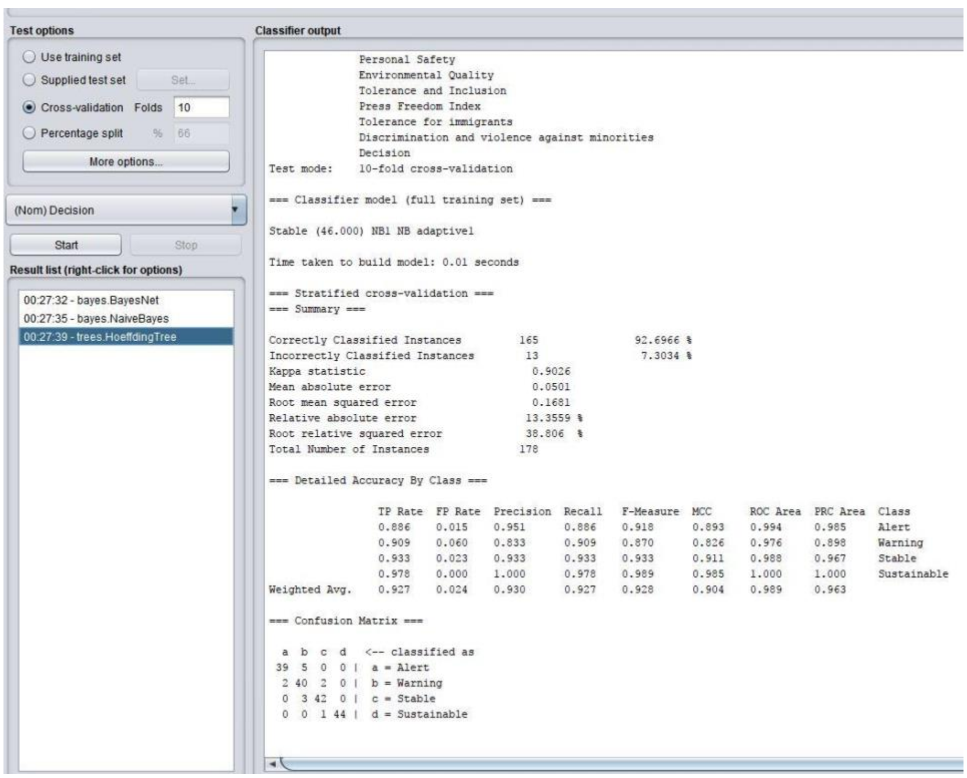

FIGURE 23: HOEFFDING TREE - 2015 DATASET WITH ADDITIONAL ATTRIBUTES

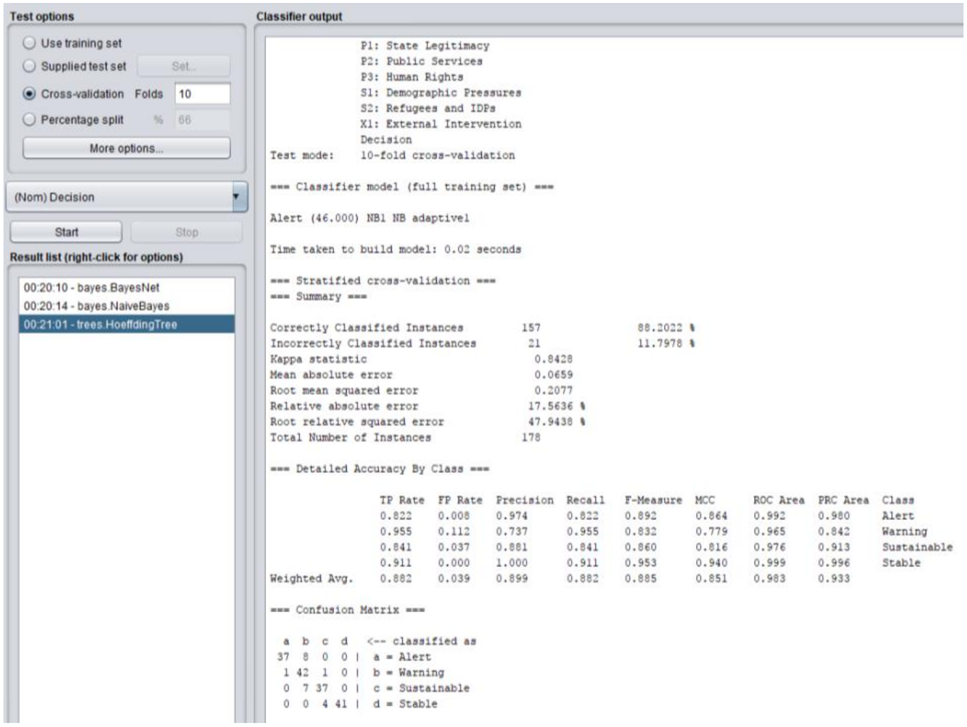

FIGURE 24: HOEFFDING TREE - 2016 DATASET WITHOUT ADDITIONAL ATTRIBUTES

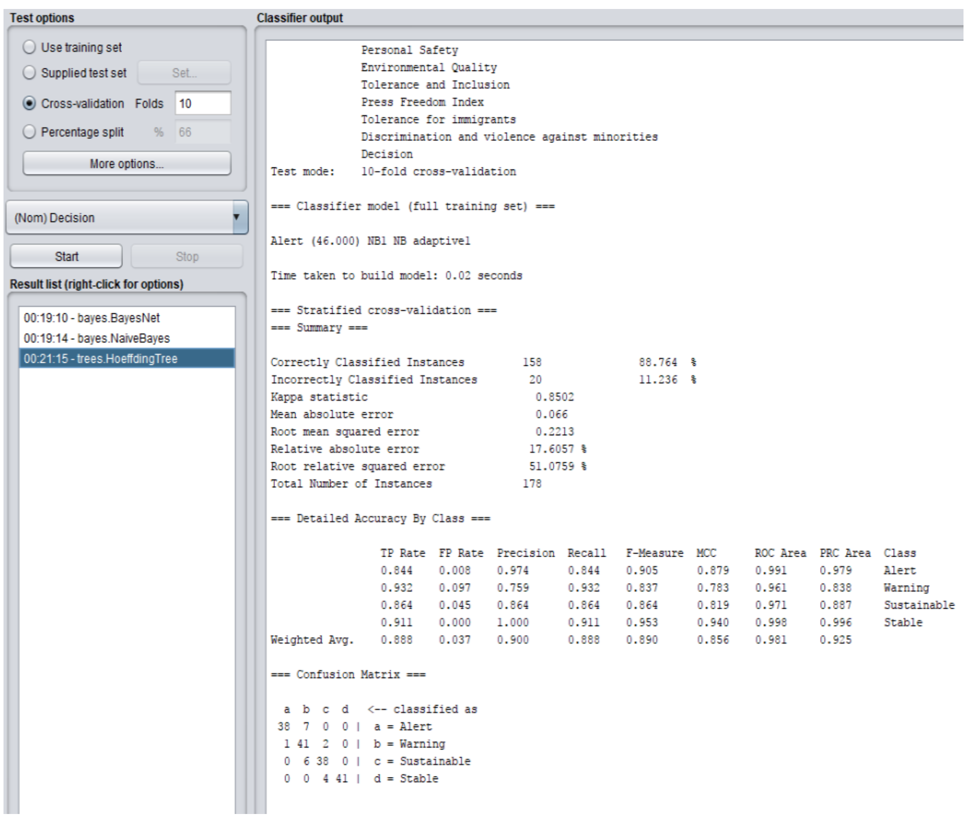

FIGURE 25: HOEFFDING TREE - 2016 DATASET WITH ADDITIONAL ATTRIBUTES

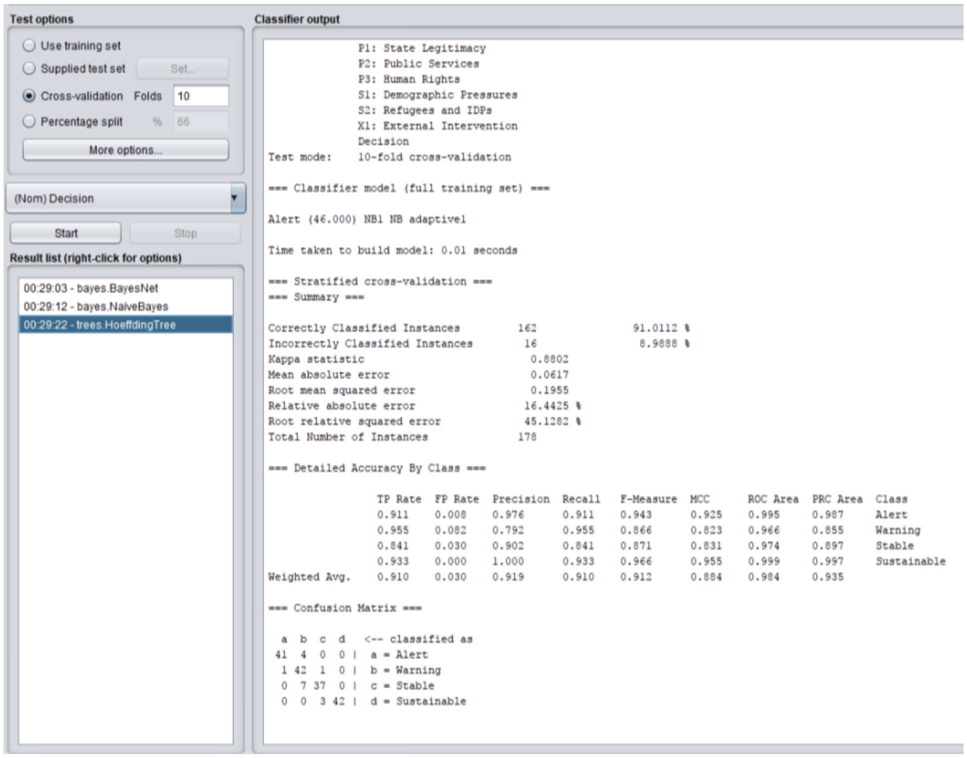

FIGURE 26: HOEFFDING TREE - 2017 DATASET WITHOUT ADDITIONAL ATTRIBUTES

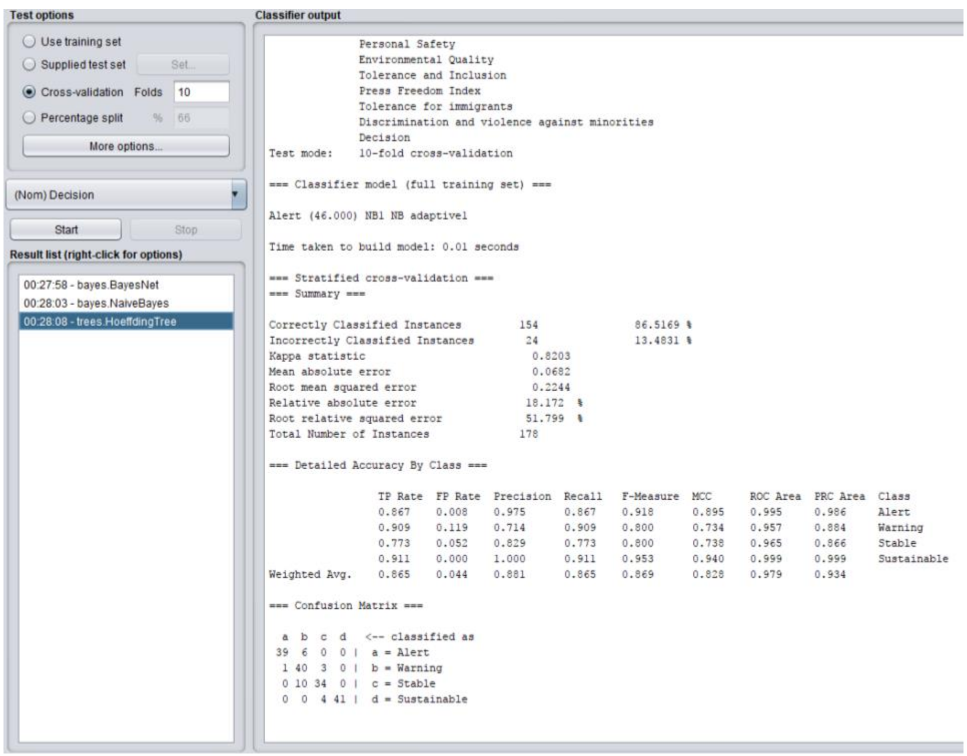

FIGURE 27: HOEFFDING TREE - 2017 DATASET WITH ADDITIONAL ATTRIBUTES

RESULTS COMPARISON

The above screenshots indicate the performance of Naive Bayes classifier on our dataset. Naive Bayes, despite of being very simple, performs extremely well. We were able to achieve up to 93% accuracy and the precision ranges from 85% - 95% for all the classes.

The Bayes' Net classifier, being more considerate about the relationship between the attributes of dataset has performed little better than Naive Bayes. The accuracy is consistently between 85 - 92%. An additional feature of this classifier was the execution performance. It took less time than Naive Bayes for classification of our dataset. Hoeffding Tree is a more sophisticated classifier. It has given us the accuracy ranging between 85 - 90%.

From these results, it is easy to judge that Bayes' net is the best classifier. It is better in terms of both execution performance and the resultant accuracy.

ACTION RULES MINING

STABLE / FLEXIBLE ATTRIBUTE SELECTION

The experiment of Action Rules mining is performed in 4 stages, each stage representing different sets of stable and flexible attributes. The stages are as follows:

STAGE 1

Stable Attributes: Factionalized Elites
Flexible Attributes: All remaining attributes

HYPOTHESIS:

In this stage, we wish to analyze the impact of Factionalized Elites on other attributes. Factionalized Elites indicates fragmentation of state institutions along ethnic, class, clan, racial or religious lines. It also factors the use of nationalistic political rhetoric by ruling elites, often in terms of nationalism, xenophobia and communal irredentism. Example, when we keep Factionalized Elites as a stable attribute, Security Apparatus of the country goes from a more fragile state to a less fragile state. Security Apparatus indicates security threats to a state, such as bombings, attacks and battle-related deaths, rebel movements, mutinies, coups, or terrorism.

YEAR: 2014

C2__Factionalized_Elites(<8;9)) : (C1__Security_Apparatus(<8;9)) ->
C1__Security_Apparatus(<6;7))) >÷< (Decision(Alert) -> Decision(Warning))

YEAR: 2015

C2__Factionalized_Elites(<8;9)) : (C1__Security_Apparatus(<7;8)) & X1__External_Intervention(<8;9)) -> C1__Security_Apparatus(<6;7)) & X1__External_Intervention(<6;7))) >÷< (Decision(Alert) -> Decision(Warning))

YEAR: 2016

C2__Factionalized_Elites(<8;9)) : (C1__Security_Apparatus(<8;9)) -> C1__Security_Apparatus(<6;7))) >÷< (Decision(Alert) -> Decision(Warning))

Year: 2017

C2__Factionalized_Elites(<8;9)) : (C1__Security_Apparatus(<8;9)) -> C1__Security_Apparatus(<6;7))) >÷< (Decision(Alert) -> Decision(Warning))

CONCLUSION:

In this stage, we have considered the attribute Factionalized Elites as stable, which is an indicator of various fragments of the society. As we can see from the generated rules that this attribute has direct impact on the Security Apparatus, External Intervention, Public Services and Tolerance and Inclusion. Therefore, if a state works on the latter aspects while preservice its elite fragments, it can move towards a more sustainable state.

STAGE 2

Stable Attributes: Demographic Pressure, Refugees and internally displaced persons (IDPs), Group grievance and Human flight and brain drain
Flexible Attributes: All remaining attributes

HYPOTHESIS:

In this stage, we wish to analyze the impact of Social Indicators. Social indicators indicates pressures upon the state deriving from the population itself or the environment around it such as population pressures related to food supply, access to safe water, and other life-sustaining resources, or health, such as prevalence of disease and epidemics. It also includes pressure upon states caused by the forced displacement of large communities as a result of social, political, environmental or other causes, measuring displacement within countries, as well as refugee flows into others. Existence of tension or violence between groups, which can undermine the state's provision of security. Pressures related to discrimination, ethnic violence, communal violence, sectarian violence, and religious violence. Example, when we consider Social Indicator as a stable attribute State Legitimacy goes from a more fragile state to a less fragile state. State Legitimacy indicates corruption and lack of representativeness, as citizens can lose confidence in state institutions and processes.

Year: 2014

C3__GROUP_GRIEVANCE (01)(<8;9)) : (P1__STATE_LEGITIMACY (01)(<9;10>) -> P1__STATE_LEGITIMACY (01)(<7;8))) >÷< (DECISION(ALERT) -> DECISION(WARNING))

Year: 2015

S1__DEMOGRAPHIC_PRESSURES (01)(<8;9)) : (PERSONAL_SAFETY (01)(<4;5)) -> PERSONAL_SAFETY (01)(<3;4))) >÷< (DECISION(ALERT) -> DECISION(WARNING))

Year: 2016

E3__HUMAN_FLIGHT_AND_BRAIN_DRAIN (01)(<7.3;8.3)) : (PERSONAL_SAFETY(<4.2;5.2)) -> PERSONAL_SAFETY(<3.2;4.2))) >÷< (DECISION(ALERT) -> DECISION(WARNING))

Year: 2017

E3__HUMAN_FLIGHT_AND_BRAIN_DRAIN(<7.3;8.3)) : (PERSONAL_SAFETY(<1;2)) & X1__EXTERNAL_INTERVENTION(<8;9)) -> PERSONAL_SAFETY(<2;3)) & X1__EXTERNAL_INTERVENTION(<7;8))) >÷< (DECISION(ALERT) -> DECISION(WARNING))

CONCLUSION:

In this stage of experiment, we have kept Demographic Pressure, Refugees and internally displaced persons (IDPs), Group grievance and Human flight and brain drain under consideration, to study their impact on other attributes. The results indicate major effect on State Legitimacy and Personal Safety. The State Legitimacy is an indicator of confidence on the governing bodies and personal safety is directly related to the sense of security in the population. Clearly, these two attributes affect the mindset of citizens which in turn leads to grievances, migration and pressure on the state governance. Therefore, as a suggestion from this experiment, to bring a country to a more sustainable state, the governance should work elevating the confidence among its citizens.

STAGE 3

Stable Attributes: State legitimacy, Public services, Human rights, Security apparatus, Factionalized elites, External intervention

Flexible Attributes: All remaining attributes

HYPOTHESIS:

In this stage, we wish to analyze the impact of Political Indicators. Political indicators indicates corruption by ruling elites, resistance to transparency, level of democracy, illicit economy, and protests and demonstrations. Disappearance of essential services such as healthcare, education, sanitation, public transportation, police, and infrastructure.Violation or uneven protection of basic rights such as press freedom and civil liberties can challenge the security apparatus of a state which can lead to use of force, weakening the social contract, conflict, riots and protests. Example, when we keep the Political Indicator as a stable attribute Press Freedom Index which indicates evaluation of freedom of expression in private discussion, academia or culture related backgrounds goes from a more fragile state to a less fragile state.

YEAR: 2014

P1__State_Legitimacy (01)(<7;8)) : (Press_Freedom_Index (01)(<2;3)) -> Press_Freedom_Index (01)(<3;4))) >÷< (Decision(Alert) -> Decision(Warning))

YEAR: 2015

X1__External_Intervention (01)(<7;8)) : (Press_Freedom_Index (01)(<2;3)) -> Press_Freedom_Index (01)(<3;4))) >÷< (Decision(Alert) -> Decision(Warning))

YEAR: 2016

P3__Human_Rights(<7.2;8.2)) : (Press_Freedom_Index (02)(<2.3;3.3)) -> Press_Freedom_Index (02)(<3.3;4.3))) >÷< (Decision(Alert) -> Decision(Warning))

YEAR: 2017

P2__Public_Services (01)(<8;9)) : (Press_Freedom_Index (01)(<3.3;4.3)) -> Press_Freedom_Index (01)(<2.3;3.3))) >÷< (Decision(Alert) -> Decision(Warning))

CONCLUSION:

The attributes which were taken as stable are mostly related to political and basic human needs. It is understandable that they should have direct impact on the personal safety, human flight and brain drain and group grievances. This is what has been indicated by the generated action rules. From these rules we can conclude that if a state wants to improve the situation of its population, then it must alleviate issues related to availability of basic resources and politics.

STAGE 4

Stable Attributes: Economy, Economic Inequality, Human Flight and Brain Drain

Flexible Attributes: All remaining attributes

HYPOTHESIS:

In this stage, we wish to analyze the impact of Economic Indicators. Economic indicators indicates economic inequality in education, economic status , poverty, fairness of housing, food. In terms of progressive economic decline includes unemployment, per capita income, inflation, business failures. It also includes failure of state to pay salary of government employe, armed forces and failure to meet the responsibilities to citizens like pension payments. Example, when we consider Economic Indicators as a stable attribute, Demographic Pressures goes from a more fragile state to a less fragile state. Demographic Pressures indicates high population volume relative to food supply and seeking for life sustaining resources which is always a difficult task for government to protect citizens.

YEAR: 2014

E2__Economic_Inequality (01)(<5;6)) : (S1__Demographic_Pressures (01)(<5;6)) -> S1__Demographic_Pressures (01)(<4;5))) >÷< (Decision(Stable) -> Decision(Sustainable))

YEAR: 2015

E1__Economy (01)(<4;5)) : (S1__Demographic_Pressures (01)(<2;3)) -> S1__Demographic_Pressures (01)(<5;6))) >÷< (Decision(Sustainable) -> Decision(Stable))

YEAR: 2016

E2__Economic_Inequality(<6.2;7.2)) : (S1__Demographic_Pressures (01)(<5.2;6.2)) -> S1__Demographic_Pressures (01)(<6.2;7.2))) >÷< (Decision(Warning) -> Decision(Sustainable))

YEAR: 2017

E3__Human_Flight_and_Brain_Drain (02)(<7.2;8.2)) : (S1__Demographic_Pressures (02)(<8;9)) -> S1__Demographic_Pressures (02)(<7;8))) >÷< (Decision(Alert) -> Decision(Warning))

CONCLUSION:

This experiment mainly focus on economic aspects of a state. The generated rules indicated that the economic aspects have direct impact on demographic pressures, state legitimacy, public services and safety. It is a known fact that economy is the base for all. Therefore, if the state has to alleviate issues related to public safety, provide better services to the citizens, then it should clearly work towards economic stability.

REFERENCES:

1. http://fundforpeace.org/fsi/indicators/
2. https://www.cs.waikato.ac.nz/ml/weka/documentation.html
3. http://lispminer.vse.cz/tutorial/index.html
4. https://www.socialprogressindex.com/
5. http://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/naive-bayes-classifier.pdf
6. http://www.cs.technion.ac.il/~dang/journal_papers/friedman1997Bayesian.pdf
7. https://samoa.incubator.apache.org/documentation/Vertical-Hoeffding-Tree-Classifier.html
8. https://en.wikipedia.org/wiki/Fragile_States_Index