**CHECKPOINT- 3 (3/20/2016)**

**Predicting the Success Rate of India Movies**

By this checkpoint, we have an update on our project analysis by data modelling.

**Data Modelling**

Choosing appropriate model for data mining is not an easy task. Many times the data we chose might not give the results we intend to. Data models in a large picture can be divided into two categories

1. Supervised model
2. Unsupervised model

Supervised data model: We use this model when we know our target variable, and when we know the categories of the data. Training data is provided in supervised model.

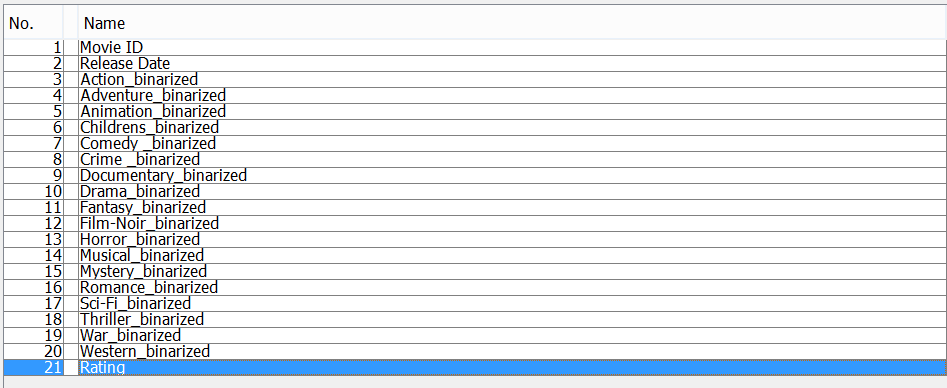
Unsupervised model: This model is used when the target variable is not known or the categories of the data is unknown. Training data is not provided in this model. This is often used to categorize or cluster the target variables.

For our model we know the target variable. We are trying to get the overall rating of a movie based on existing parameters like Actors rating, Director rating, Movie Genre, Release date.

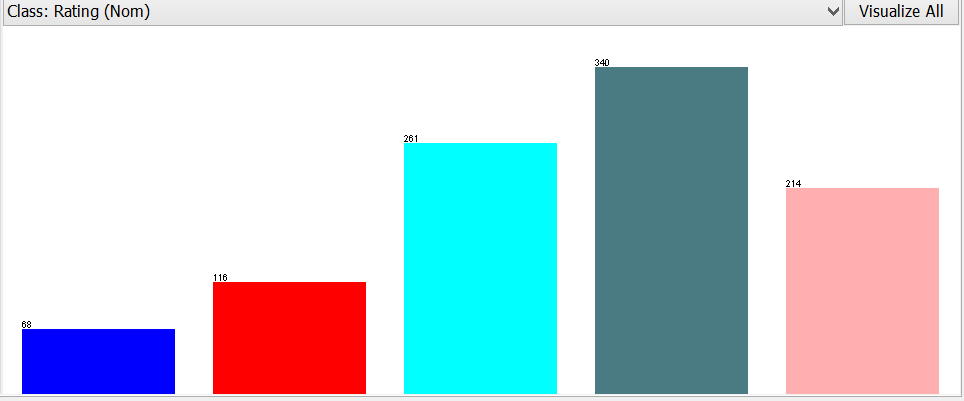
We have used J48 Classification and Naïve bayes models in our project. However initially when we built the data model with the attributes we had the accuracy of the result was very poor.

**Model #1**

Below is the list of data attributes that were used to build our model which resulted in more inaccurate results than accurate results.



**Attributes used for model 1**



**Overall Rating of movies when converted to nominal from numeric**

All the attributes when loaded into Weka were in numeric format. However to support/build our model we had to convert the attributes from numeric to nominal and binary. Binary converts the data into two instances 0 and 1 i.e. low and high. Whereas nominal values are stored as numbers too however these numbers act as indexes into an array of possible attribute values which is very efficient.

The above graph represents the nominal values for overall rating. Rating is between 1 and 5 so we can see five differently colored columns and the values on top of those bars are the number of movies that fall into that rank.

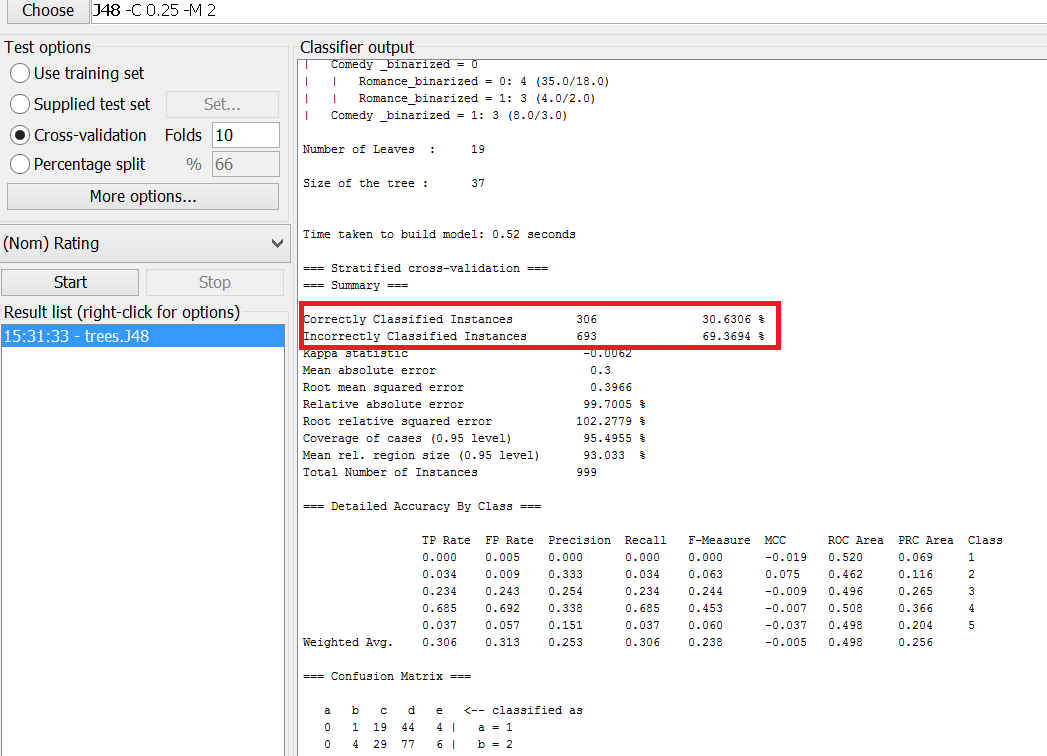
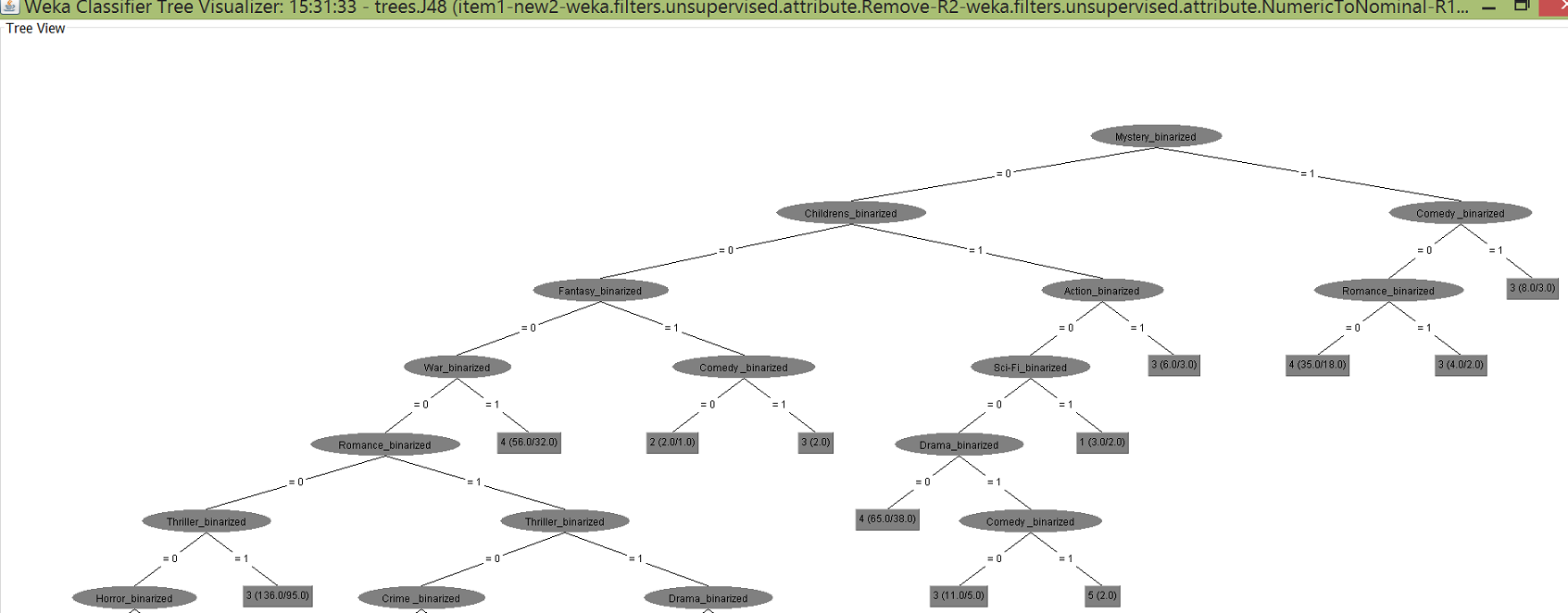


Fig 1.1- **J48 classification result**

**Fig1.2 J48 Result**



As highlighted in fig 1.1 we can see the incorrect instances are higher compared to correct instances. Also if we look at the tree we can see the branches are based on the genre of the movie. The result was not much different from j48 model when we fed the same data to Naïve Bayes model.

We had to go back to the data preparation task and add two attributes manually, this made us reduce the number of data instances from 1000 to 200.

**Model # 2**

After the incorrect model we built using the data set adding two more attributes. J48 Classification model is used to build decision tree. Using the new data set we built the decision tree which produced more accurate results than our first model. We achieved 96% correct instances and 4% incorrect instances. Apart from those two attributes that were added rest all remained same. Below screenshots represent our model which is accurate and our final model for project.



Fig 2.1 **Attributes used for Second model**

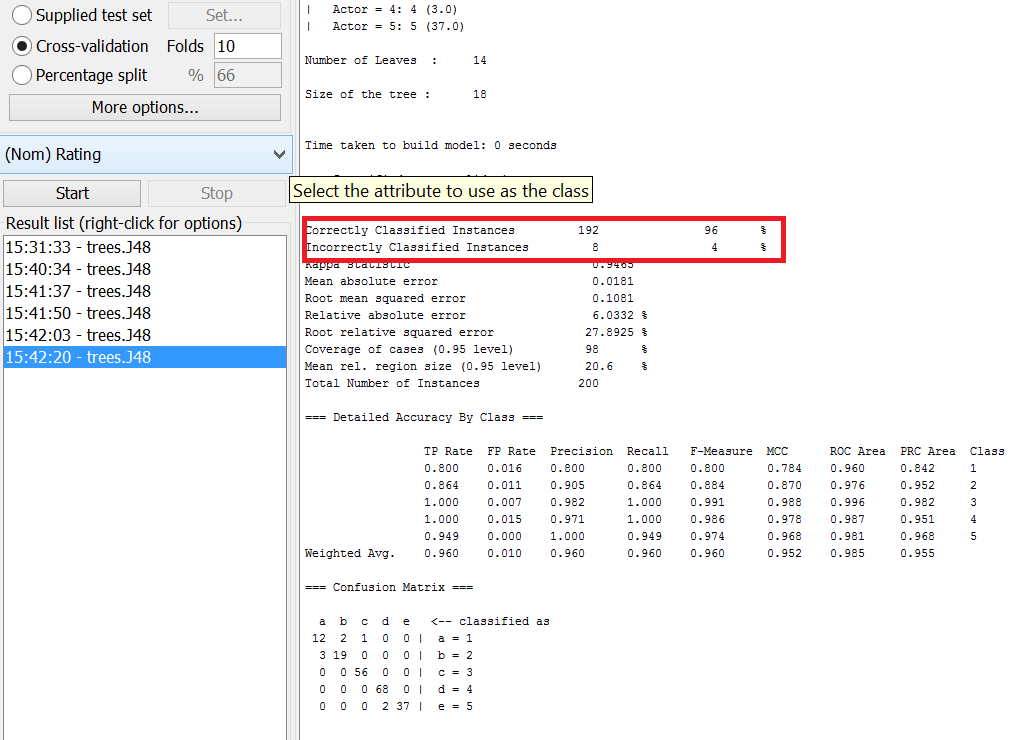


Fig 2.2 Result of J48 (model#2)

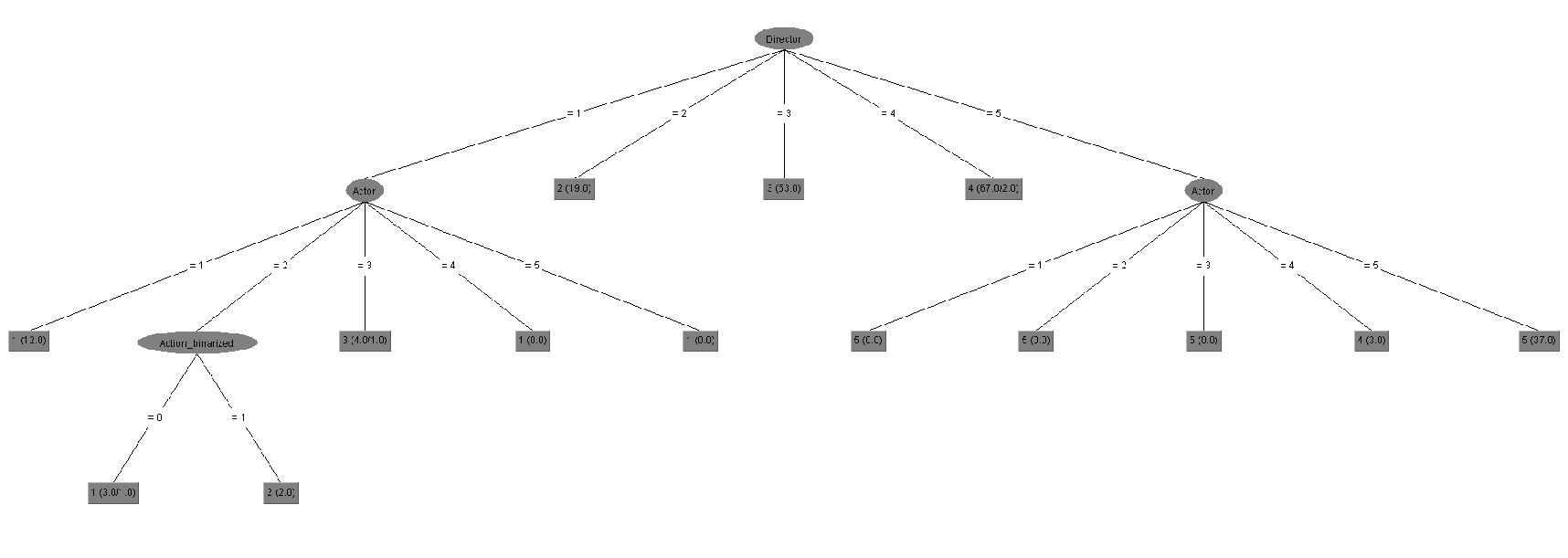
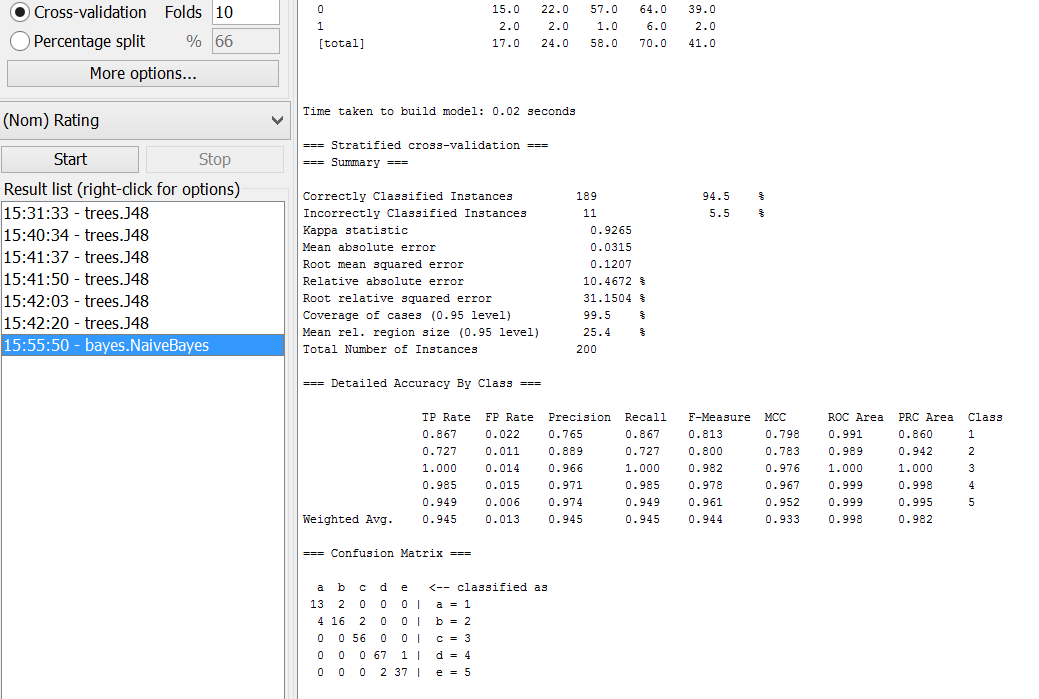


Fig 2.3 J48 Tree

In the second model if we see the tree, the top of the tree is director and the branches are actor rating. This model was more accurate than first model, however it completely ignored the Genre of the movie.

We have also built the Naïve bayes model. The accuracy of the Naïve bayes model was slight less than that of the J48 model. Below is the result of the Naïve bayes model which hosts 94.5% correctly classified instances compared to 96% correctly classified instances of J48 and 5.5% of incorrectly classified instances compared to 4% incorrectly classified instances of J48



**DELIVERABLES AND CHECKPOINTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Checkpoint Date** | **Expected Deliverable** | **Responsible Team Members** | **Checkpoint Results** |
| 02/15/2016 | Project Submission and overview | Mani Karthik | Project has been submitted and is in progress |
| 03/08/2016 | Data collection, roles and responsibilities | Rajender |  |
| 03/20/2016 | Update on project code/analysis | srujan |  |
|  |  |  |  |