UNIVERSITÀ
DI SIENA
1240

Department of Information Engineering &
Mathematics

MSc in Engineering Management

Business Intelligence course – Project work

“Effects of classification methods on Dow Jones Index Stocks (weekly data)”

Matteo Vadi
University ID : 092663
matteo.vadi@student.unisi.it

Siena, 27/07/2020

1. ABSTRACT

This project work concerns the effects of two different classification approaches on a dataset composed by different stock price data, collected over the first six months of the 2011. The stocks field this work deals with are heterogenous, since they belong to the Dow Jones Index, the collection of the 30 more representative NYSE – New York Stock Exchange stock prices.

In the collection-prediction phases there is a time lag of one unit (in this work, a week), since the data collected can't be used for the prediction in the same time period.

The two approaches for this supervised learning try to predict if a stock is good to be purchased in the following week, by looking at the % of change in the price for the following week.

In this direction, with respect to the original dataset, a categorical attribute has been identified as the output variable.

The first method is a Classification Tree approach, while the second one concern the Naïve Bayes, done on two different train & test sets for the comparison.

2. INTRODUCTION & BACKGROUND

The usage of Machine Learning (ML) techniques in forecasting the financial index prices has become more and more frequent during years, especially from the last few decades. This is due to the expansion of the various approaches and the different computational capabilities of the modern technologies, with respect to the development and the more easier access to the financial world, from everywhere and every time.

The massive approaching to the financial world has opened new frontiers in research field, especially for what concern the possibility of predicting the variability of a stock index prices during time, in order to help the financial operator in making decision of Buy and Sell.

This work goes in this direction, since it tries to make some kind of classification approaches for prediction an output variable which models the decision the operators have to take (buy / sell), according to the information they have.

The dataset chosen is the one already treated in (M. S. Brown, M. Pelosi & H. Dirska - 2013), descripted in an accurate way in the specific paragraph.

Going into details, first a decision rule has been identified. In particular, the one reported in the project work is the easiest, that is:

$$\text{if \% change price in the following week} \geq 0 \text{ then ``purchase or not sell''}$$

With this kind of logical relationship is possible to split the output variable into two different levels, that are "purchase or not sell" (purchase if you don't have it, don't sell if you already have it in your financial portfolio) and "sell or not purchase" in the opposite situation (when % change price in the following week is < 0).

This decision rule does not consider the possibility (actually, this happens, with different entity, every time) of fees and other taxes methods that can affect its quality.

Just think about a stock index which will produce a small increment (in %) in the price in the following week (so it is reasonable to buy it today and gain in difference in terms of price that I expected in the next week) which can't overcome the fee that I pay to buy it. Is possible to review the rule, according to this kind of considerations (e.g. buy / not sell if %change ≥ 0.5, sell / not buy if %change ≤ - 0.5, not buy / not sell if - 0.5 < %change < 0.5).

Second, with two very simplified ways with respect to the various more complicated algorithms there are in literature, the project work focus on a Classification Trees method and on a Naïve Bayes one.

It is possible to find variety of methods that use Genetic Algorithms (GAs) and Genetic Programming (GP) to predict stock and security movements before this work. These methods take different approaches; some researches use GAs and GPs to develop classification rules, while others use GAs in hybrid ways.

The most influent research on this work is the one realized by (M. S. Brown, M. Pelosi & H. Dirska - 2013); the dataset I choose came from their job, where they developed an NGA (Niche Genetic Algorithm) to derive a set of classification rules based on a train set inside it, that latter they applied to another set of data (test set). The algorithm they developed is named Dynamic-radius Species-conserving Genetic Algorithm (DSGA); in their work the aim is the prediction of a set of decision rules, taken into account by a Tabu List, summarized in a sort of purchase indicator, that can be utilized by the financial operator in making decision on buy / sell.

3. DATASET

For what concerns the dataset, as already introduced, it deals with the Dow Jones Industrial Index for the first two quarters of the 2011. It is the most known index of the NYS Exchange which is computed by looking at the price of the 30 "blue chips" (the most capitalized companies in the USA), which may change from year to year.

The DJ index for the 2011 involved companies indices from different industrial sectors, very heterogeneous from each other; we can find companies which belong to the industrial field, but also others from the financial and the chemical-pharmaceutical one. In Table 1 is reported a list of all the 30 blue chips which composed the DJ index in that year, together with the stock identification code:

Table 1. *Financial indices of the DJ Index in 2011.*

3M (MMM)	DuPont (DD)	Intel (INTC)
American Express (AXP)	ExxonMobil (XOM)	IBM (IBM)
Alcoa (AA)	General Electric (GE)	Johnson & Johnson (JNJ)
AT&T (T)	Hewlett-Packard (HPQ)	JPMorgan Chase (JPM)
Bank of America (BAC)	The Home Depot (HD)	Kraft (KRFT)
Boeing (BA)	Travelers (TRV)	Mc Donald's (MCD)
Caterpillar (CAT)	United Technologies (UTX)	Merck (MRK)
Chevron (CVX)	Verizon (VZ)	Microsoft (MSFT)
Cisco Systems (CSCO)	Wal-Mart (WMT)	Pfizer (PFE)
Coca-cola (KO)	Walt-Disney (DIS)	Procter & Gamble (PG)

The dataset, treated for the first time in (M. S. Brown, M. Pelosi & H. Dirska - 2013), is developed around different entries. In this work, each record (row) is data for a week. There are 750 observation of 16 variables which composed the dataset, but most of the attributes are then not used for the classification phase. The only ones that are useful for the aim of this work are:

Table 2. *Useful attributes for the classification.*

Attribute	*Description*
Quarter	The quarter the record belongs to. In this dataset there is a collection of data from January to June (1st and 2nd quarters).
Stock	The code for the identification of the company stock.
% price changed	The percent change in the stock price for week *x*, which is the week prior to the week that the algorithm attempts to predict the stock for.
% volume changed	The percent change in volume during week *x* compared to week *x - 1*. Volume is the number of shares of a stock sold.
Days to next dividend	The number of days until the next dividend.
% return of next dividend	The percent return based upon the stock price of week *x* of the amount of the dividend.
% price changed next week	The percent change in the stock price for week *x+1*.

Is important to notice how the collected data are connected with each other; in predicting stock prices, the available data in a certain time period (in this case, the week *x*) give an information only after the end that time period. So it is possible to use them only for the prediction in the following time period; this is why in the original dataset there are two numerical attributes that represent the same variable but referring to different time buckets.
In this direction, an output variable has been identified, based on the "*% price changed next week*" numerical attribute, according to the decision rule already explained before. Finally, in the dataset there were some NA values connected to the "*% volume change*" with respect to the previous week for the first record of the year for each stock (that is, the records with date = "1/1/2011" for each stock have no previous date for the computation of "*% volume change*"). The problem connected to the removal of NA values from the dataset could have been in the "*stock*" attribute; the risk in fact was connected to the possibility of removal too much record of a specific stock with respect to the others. But, fortunately, each NA values refers only to a specific stock index, so there were no reason in keeping them (e.g. by imputation with the mean value for example). The dataset without the NA values is composed by 720 observations.

4. METHODS

In order to explore deeply the classification phase and algorithms, a “Preprocessing & data preparation” phase and a preliminary statistical analysis have been performed.

4.1 DATA VISUALIZATION

Splitting the useful attributes into two different categories according to their nature, in Figure 1 are reported the histograms for the numerical attributes, while in Figure 2 the pie chart for “*quarter*” and the bar chart for the output variable (“*decision*”).

Figure 1. *Histograms for the numerical attributes.*

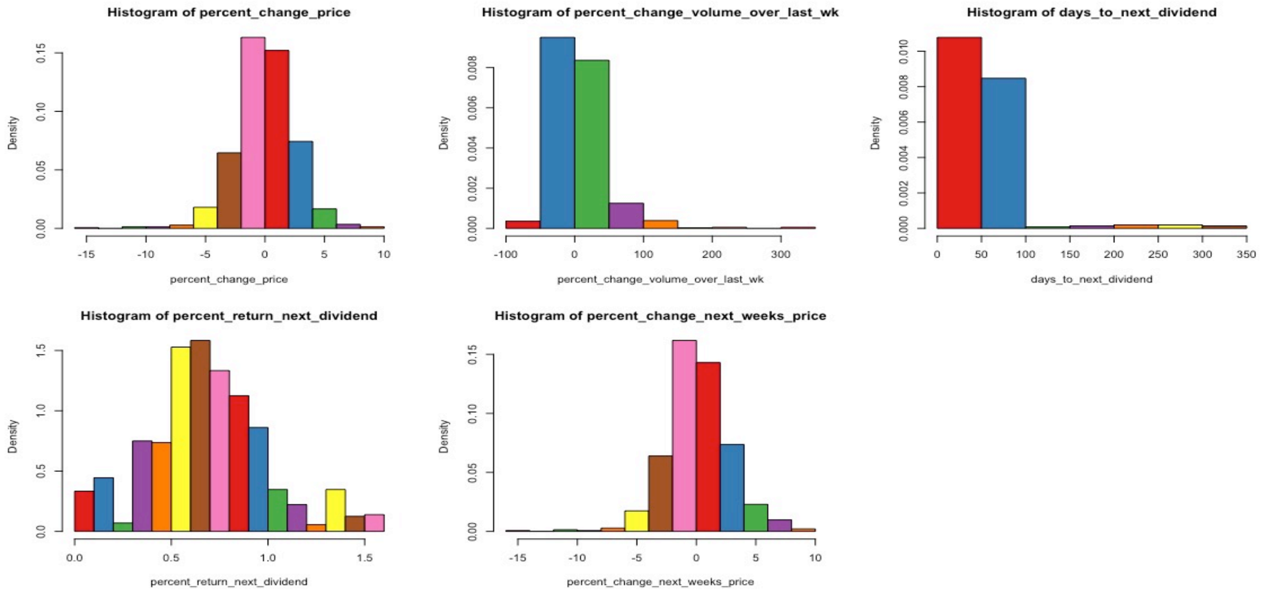
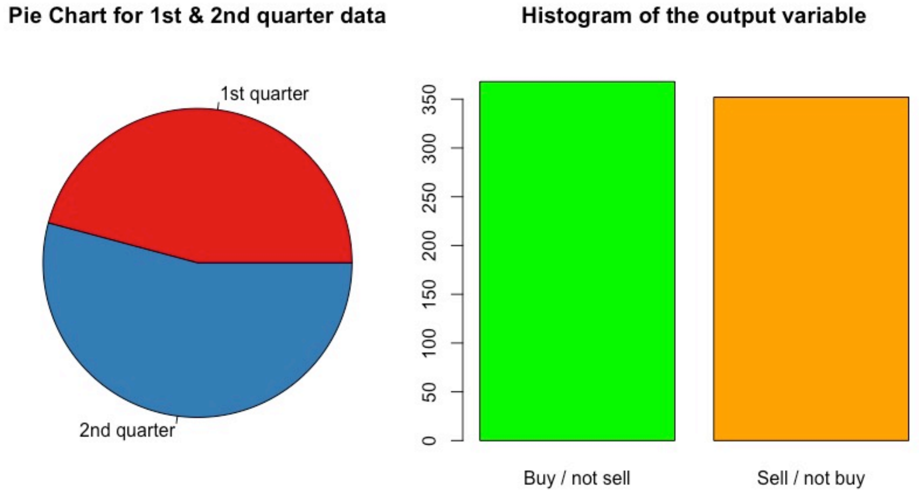


Figure 2. *Pie chart for “quarter” & bar chart for “decision”.*



As we can see in the pie chart, there are less records for the 1st quarter than for the 2nd; this can affect the classification phase results if the dataset, according to what has been proposed in (M. S. Brown, M. Pelosi & H. Dirska - 2013) will be split into train & test sets accordingly. Later on this aspect will be consider more in details.

4.2 DATA EXPLORATION

To keep track of the summary statistics of the distribution of each attribute, a specific matrix ("*ss_matrix*") has been realized (Table 3).

Table 3. *Summary statistics matrix for the numerical attributes.*

Summary Statistics	*Mean*	*Median*	*Mode*	*MAD*	*Var*	*Std.dev*	*CV*
percent_change_price	0.03	0.00	0.0	1.88	6.3e+00	2.50	8311
percent_change_volume_over_last_wk	5.59	0.513	1.4	27.72	1.6e+03	40.54	725
days_to_next_dividend	52.26	47.00	54.0	29.49	2.1e+03	45.88	88
percent_return_next_dividend	0.69	0.680	1.1	0.23	9.3e-02	0.31	44
percent_change_next_weeks_price	0.19	0.036	0.0	2.00	7.1e+00	2.66	1378

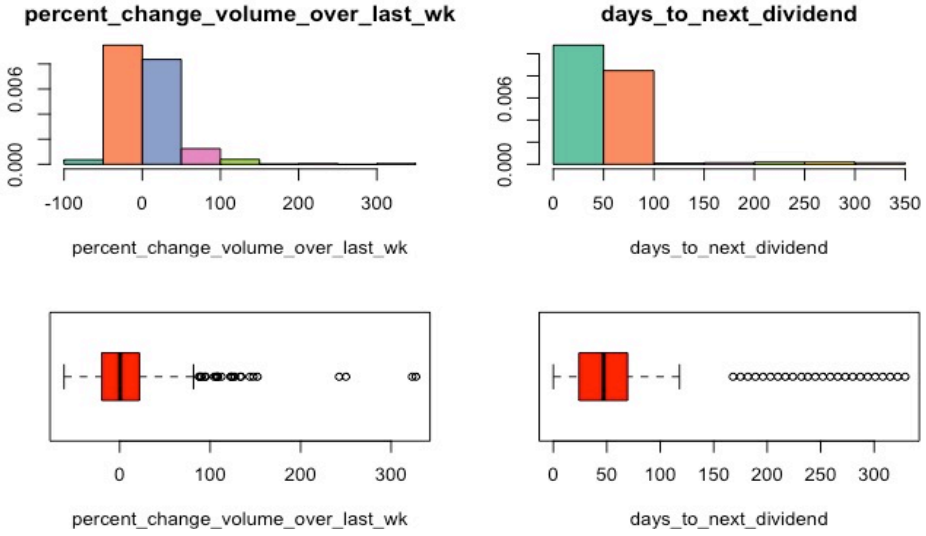
From the previous histograms is possible to see that some numerical attributes are really touched by the outliers (especially for "*days to next dividend*" attribute) so different kind of measures of location, based on the quantiles have been identified (Table 4).

Table 4. *Mid Mean, Trimmed Mean and 10% Winsorized Mean.*

Quantile mean measures	*Mid mean*	*Trimmed mean*	*Winsorized mean*
percent_change_price	0.0571	0.0379	0.0301
percent_change_volume_over_last_wk	0.3777	5.2388	5.5936
days_to_next_dividend	46.9773	52.3848	52.2597
percent_return_next_dividend	0.6794	0.6912	0.6916
percent_change_next_weeks_price	0.1188	0.2016	0.1933

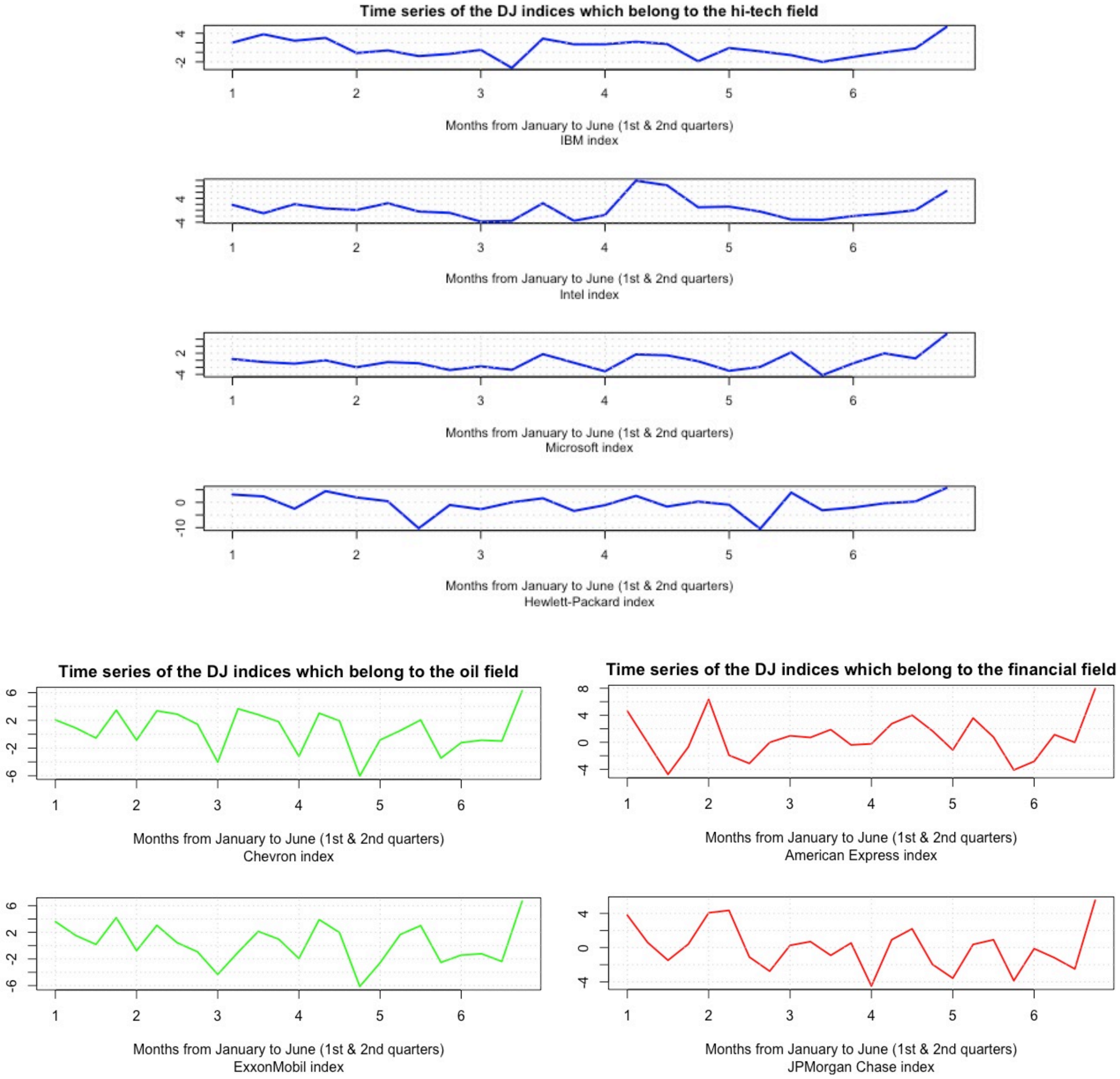
By the comparison between the real mean and the Mid Mean is possible to see how they are different from each other, especially for what concerns the "*days to next dividend*" and the "*% change volume over last week*" attributes, the two variables more influenced by the outliers (Figure 3). This aspect should be taken into account in the classification phase results.

Figure 3. *Outliers for "days to next dividend" and "% change volume over last week".*



In order to see the evolution of the % change in the prices of the stock indices a time series representation has been realized for the stocks that belong to the hi-tech, financial, oil and chemical-pharmaceutical fields (Figure 4.)

Figure 4. *Time series evolution of % change in price week by week from January to June for the hi-tech, financial and oil sectors.*



The strength of the relationship between the numerical attributes has been described into a correlation matrix. From those statistical measures is possible to see that they are very little correlated each other. This is mainly due to the presence of the outliers in their distributions, since the correlation index is computed by looking at the mean values and the mean values are not so consistent with respect to the values which follow the distribution in a less precise way.

4.3 CLASSIFICATION

For the prediction phase, the heterogeneity in the values of the numerical attributes and in the nature of the variables force to act an a-priori standardization on the dataset; in this direction the focus is on the factorization of the categorical attributes ("*stock*" and "*decision*", with the particular case of the "*quarter*" attribute which is integer – can assume only values equal to 1 or 2 – but that has been considered as a factor variable with two levels – "I" or "II") and in a min-max normalization between [0,1] for the numerical ones. In this way the standardize dataset presents only factor or numerical variables, with scaled values, not so much different from each other, that can be used in the classification phase in order to reduce the computational effort and improve the accuracy.

In both the classification approaches treated in this work, there is the need of distinguishing the original dataset into two different subsets, called train set and test set. The aim is to use the train set in the algorithm (Classification Trees and Naïve Bayes) to identify a prediction rule which will give indication on the output variable ("*decision*") that has to be used to predict the test set. To get the accuracy of the model is possible to count how many times the realization of the output variable predicted is equal to its real value for each record in the test set.

As already done in (M. S. Brown, M. Pelosi & H. Dirska - 2013), the subdivision realized is the one between the records of the standardize dataset which belong to the 1st quarter of the year (i.e. the rows with the variable "*quarter*" = "I") for the test set and to the 2nd for the test set. In this way a train set with 330 observations and a test set with 390 observations have been created. The original reason of (M. S. Brown, M. Pelosi & H. Dirska - 2013) in choosing this kind of train & test sets can be found in the nature of the problem: collect data for a certain time of period and use them for the following (in terms of quarter of years).

Then another subdivision has been performed, in order to compare the two situations. The choice this time follows the proportion of the dataset, without considering which stock indices belong to one of the first two quarters of the year. In fact, with an hold out method has been proposed a situation with 2/3 of the records for the train set and 1/3 for the test set.

4.3.1 CLASSIFICATION TREES

The first approach for the prediction is the one realized following the euristic "divide & conquer". By using specific classification function in software (like R) is possible to build a classification tree which involves inside the rule (the path along the tree) that has to be used in the prediction phase. The building of the tree traces the concept of "information gain", that is the amount of information that can be achieved if a specific splitting rule is followed, by the comparison of impurity indices.

The splitting criterion used in the project work takes into account "*deviance*" and "*Gini*" indices.

A score function is then used to convert the value obtained in the classification phase into a target class, by the application of the "majority voting" criterion.

In order to have a classification tree which does not grows too much (with possible effects on the accuracy of the model) a post-pruning following the misclassification criterion on the tree has been realized.

4.3.2 NAÏVE BAYES

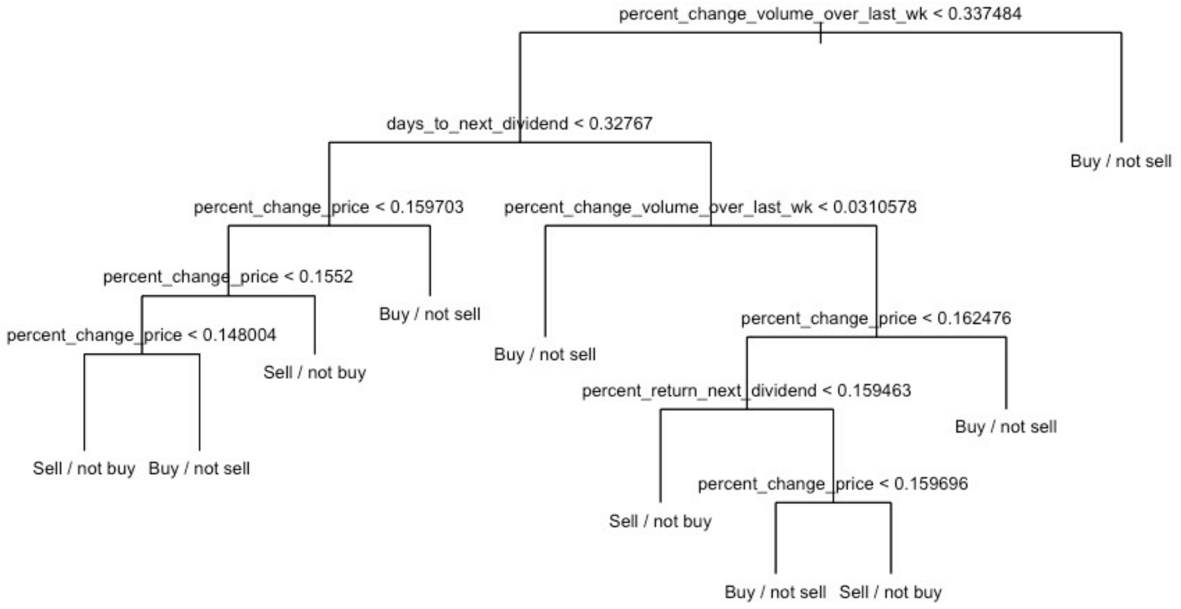
The second approach belongs to the Baesian Methods for the prediction phase, based on the concept of conditional probabilities (a-posteriori).
The probability that the classified values belongs to the target class can be computed in this direction.
By running a specific algorithm the mean and the standard deviation of the fitted Gaussian distribution of the categorical output variable given the numerical predictor attributes, using the Bayes rule, has been computed.
Then the values obtained in this way have been used for the prediction phase.

5. RESULTS

In showing the results of the project work is important to pay attention on two aspects first. The choose of the rule for the creation of the output variable can bring much more different results, since different rules from the simplest one adopted in this project can bring to situation in which the best thing is neither to buy nor to sell. The factor output variable created in this way would have three different levels instead of two. Moreover, the presence of lot of outliers in the distributions of some predictor variables may affect the results presented above.

Case #1 – Results in terms of accuracy of the classification models with train & test sets according to the quarter and "deviance" as splitting criterion for the CTs.

Figure 5. *Classification tree post-pruning with train & test set according to "quarter" and "deviance" splitting criterion.*



percent_change_volume_over_last_wk < 0.337484
days_to_next_dividend < 0.32767
Buy / not sell
percent_change_price < 0.159703
percent_change_volume_over_last_wk < 0.0310578
percent_change_price < 0.1552
Buy / not sell
percent_change_price < 0.148004
Sell / not buy
Buy / not sell
percent_change_price < 0.162476
percent_return_next_dividend < 0.159463
Buy / not sell
Sell / not buy
Buy / not sell
percent_change_price < 0.159696
Sell / not buy
Buy / not sell
Sell / not buy

Figure 6. *Results of the prediction after running the algorithm in case#1.*

```
[1] "The Classification Trees approach is more accurate than the Naive Bayes one"

------------------------          --------
 **accuracy_class_tree**    0.5205

   **accuracy_naive**        0.5026
------------------------          --------
```

Case #2 – Results in terms of accuracy of the classification models with train & test sets according to the quarter and "Gini" as splitting criterion for the CTs.

Figure 7. *Classification tree post-pruning with train & test set according to "quarter" and "Gini" splitting criterion.*

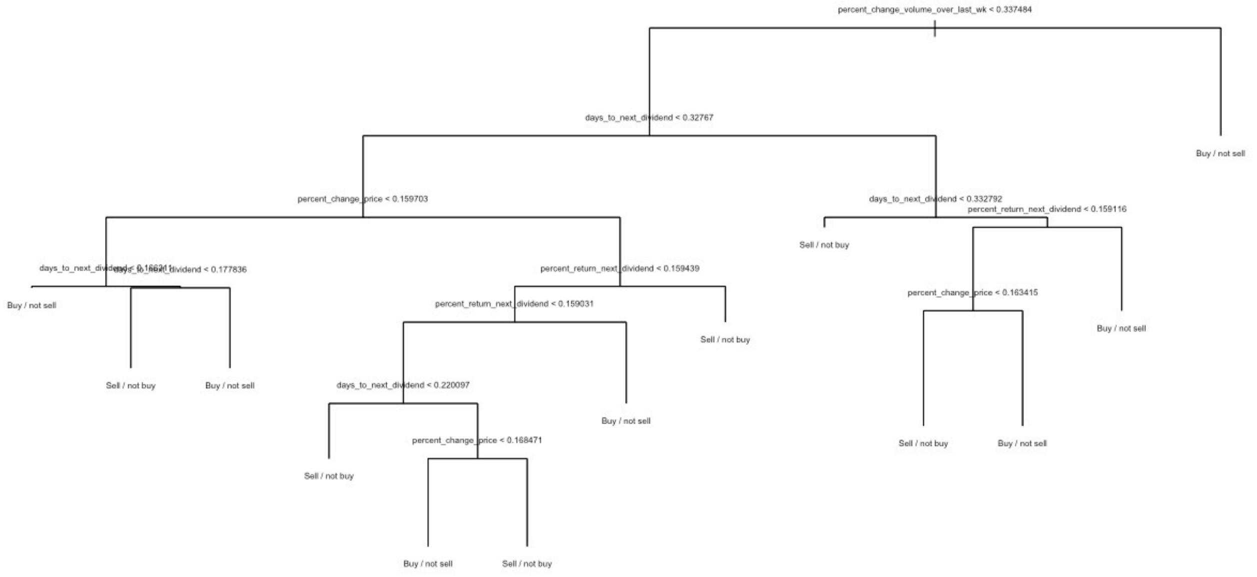

percent_change_volume_over_last_wk < 0.337484
days_to_next_dividend < 0.32767
Buy / not sell
percent_change_price < 0.159703
days_to_next_dividend < 0.332792
percent_return_next_dividend < 0.159116
Sell / not buy
days_to_next_dividend < 0.166211
days_to_next_dividend < 0.177836
percent_return_next_dividend < 0.159439
Buy / not sell
percent_return_next_dividend < 0.159031
percent_change_price < 0.163415
Sell / not buy
Buy / not sell
Sell / not buy
Buy / not sell
days_to_next_dividend < 0.220097
Buy / not sell
Sell / not buy
Buy / not sell
percent_change_price < 0.168471
Sell / not buy
Buy / not sell
Sell / not buy

Figure 8. *Results of the prediction after running the algorithm in case#2.*

```
[1] "The Naive Bayes approach is more accurate than the Classification Trees one"

------------------------          --------
 **accuracy_class_tree**    0.4846

   **accuracy_naive**        0.5026
------------------------          --------
```

Case #3 – Results in terms of accuracy of the classification models with train & test sets sampled randomly and "deviance" as splitting criterion for the CTs.

Figure 9. *Classification tree post-pruning with train & test set sampled randomly and "deviance" splitting criterion.*

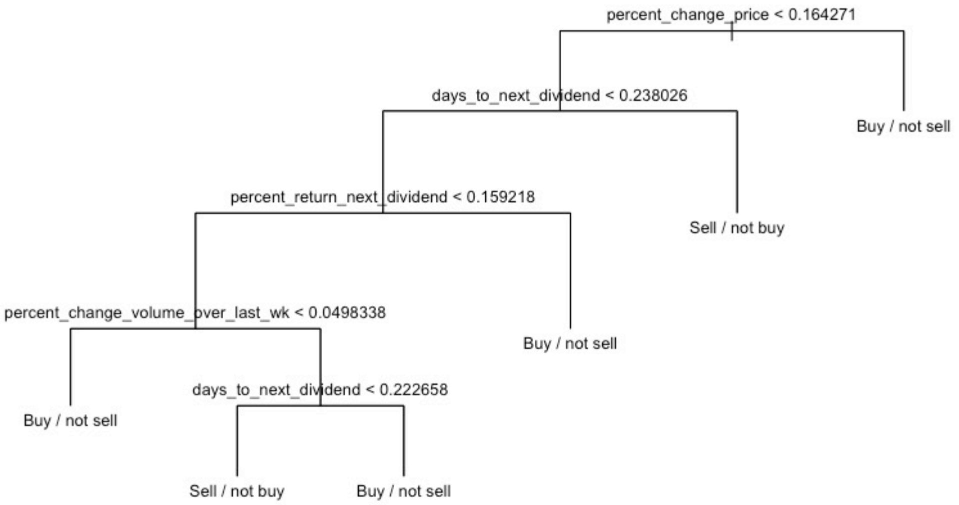

percent_change_price < 0.164271
days_to_next_dividend < 0.238026
Buy / not sell
percent_return_next_dividend < 0.159218
Sell / not buy
percent_change_volume_over_last_wk < 0.0498338
Buy / not sell
days_to_next_dividend < 0.222658
Buy / not sell
Sell / not buy
Buy / not sell

Figure 10. *Results of the prediction after running the algorithm in case#3.*

```
[1] "The Classification Trees approach is more accurate than the Naive Bayes one"

------------------------          --------
 **accuracy_class_tree**    0.5125

   **accuracy_naive**        0.5
------------------------          --------
```

Case #4 – Results in terms of accuracy of the classification models with train & test sets sampled randomly and "Gini" as splitting criterion for the CTs.

Figure 11. *Classification tree post-pruning with train & test set sampled randomly and "Gini" splitting criterion.*

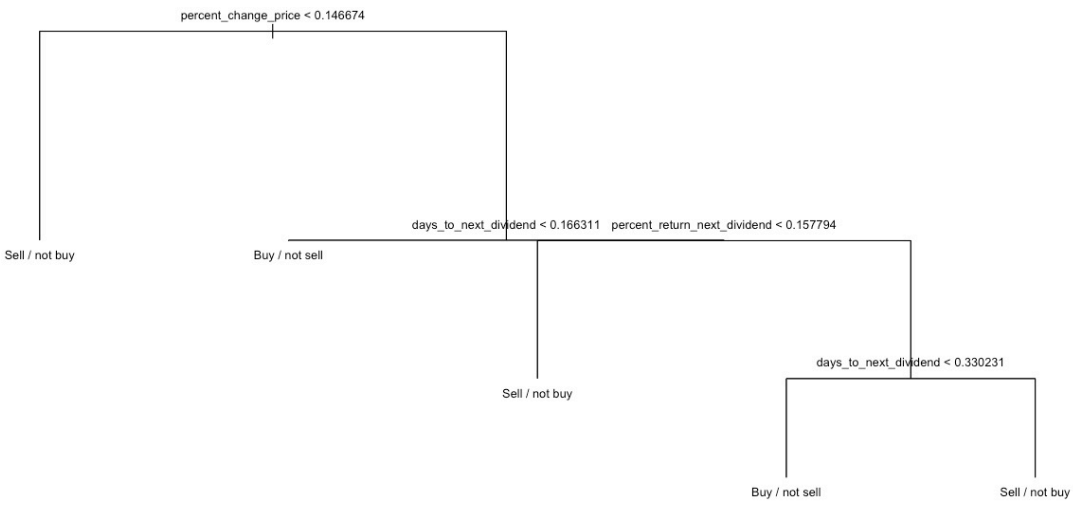

percent_change_price < 0.146674
Sell / not buy
Buy / not sell
days_to_next_dividend < 0.166311
percent_return_next_dividend < 0.157794
Sell / not buy
days_to_next_dividend < 0.330231
Buy / not sell
Sell / not buy

Figure 12. *Results of the prediction after running the algorithm in case#4.*

```
[1] "The Classification Trees approach is more accurate than the Naive Bayes one"

------------------------          --------
 **accuracy_class_tree**    0.5375

   **accuracy_naive**          0.4875
------------------------          --------
```

Before any kind of considerations is important to focus on the fact that the only kind of correct comparison is between case #1 and #2. This is due to the fact that the train & test sets are the same, while in case #3 and #4 they change at every iteration of the algorithm.

In order to keep track of this situation, is possible to run the algorithm a certain number of time and to use standard measure such as the mean of the value obtained to get to the accuracy index.

Even if the two results are not so consistent, from the obtained results is possible to say that there seems to be greater accuracy with the Classification Tree approach. This is probably due to the fact the this kind of classification method is more robust than the other with respect to the outliers in the predictor numerical attributes distribution.

Finally, is possible to point out that the tree dimension in CTs approach are smaller using a Gini index as splitting criterion than the “deviance” one.