MODELLING A FOOD INDUSTRY PROCESS BY DECISION TREE

Brankica Sobota-Šalamon[1], Jasminka Dobša[2], Vladimir Mihoković[3]

[1]Natura agro d.o.o. Đurđevac
48 350 Đurđevac, Basaričekova 16, Croatia
[2]University of Zagreb, Faculty of organization and informatics
42 000 Varaždin, Pavlinska 2, Croatia
[3]Ekotours Co.
10 000 Zagreb, Cernička 20, Croatia

jasminka.dobsa@foi.hr (Jasminka Dobša)

Abstract

Food safety is not only an issue concerning human health and business profit, but it is a political issue as well. Classical retroactive approach to food safety based on control of final product and governmental inspection survey is being changed to producers and distributors' self-control approach. In that respect, there are a number of new tools that are in force: ISO, HACCP, IFS, BRC, Hygieneomic. The main obstacle of technology production are variations in different steps of the process. Our aim is to develop a mathematical model which would contribute to a better understanding of technology process in ought to decrease risks present in it.
Subject of our paper is fresh cheese - product made from very unstable raw material: cow milk and starter cultures (Genus Lactobacillus). Variables measured during the cheese production are first analyzed by principal component analysis to detect the most relevant components in the technology process. The quality of the fresh cheese is measured by two variables: quantity of cfu/ml of Escherichia coli and quantity of cfu/ml of Yeasts and Moulds in the final product. These quantities are regulated by Ministry of helth. The technology of cheese production is modelled by decision tree which classifies final products in two classes according to their quality.
Modelling of a technological process by decision tree classifier, information about the most relevant variables in cheese production can be revealed and boundary values for some of the observed variables dividing positive from negative examples can be obtained. These boundary values could be compared against existing standards in cheese production technology.

Keywords: Technology of cheese production, standards in food technology, principal components, decision tree classifier.

Presenting Author's biography

Jasminka Dobša. Jasminka Dobša teaches statistical courses at the Faculty of organization and informatics, University of Zagreb. She received her bachelor's degree and master's degree at University of Zagreb in Mathematics and PhD at University of Zagreb in Computer Science. Her fields of interest are data and text mining.

1 Introduction

The main obstacle of technology production are variations in different steps of the process. Our aim is to develop a mathematical model for a specific technology process which would contribute to a better understanding of technological process in ought to decrease risks present in it.

Subject of our work is fresh cheese - product made from very unstable raw material: cow milk and starter cultures (Genus Lactobacillus).

In the second section we describe the process of cheese production and highlight the points in which particular variables are monitored. Originally, we measured 21 variables, but in final analysis we discarded some of them on the basis of estimation done by an expert in domain, according to their relevance to dependent (or control) variables: quantity of cfu/ml of pathogen Escherichia coli (EC) and quantity of cfu/ml of spoiling micro organisms Yeasts and Moulds (Y&M) in the final product. These two criteria for cheese quality are analysed separately. They are regulated by the Ministry of health: less then 100 cfu/ml of EC and of Y&M is allowed in a product of satisfactory quality. We used these criteria to separate products into two classes according to their quality.

In the third section we give preliminary analysis of measured variables for two classes: the products of desirable quality (DQ) or positive examples and product of undesirable quality (UQ) or negative examples. We conducted two experiments: the first is analysis according to dependant variable of cfu/ml of pathogen EC and the second is analysis according to dependant variable of cfu/ml of Y&M. The analysis according to conjunction of criteria was not interesting because the second criteria is much more restrictive and conjunction of criteria gives very similar division of examples as the second criteria gives alone. Measured variables generally are very dispersed, the assumptions on normal distribution are not met and correlations between variables are weak (if they are significant). That is why traditional statistic methods failed (regression analysis, linear discriminate analysis) and that is one of the reasons why decision tree is chosen for modelling of technology process of cheese production. The other reason is the fact that decision tree gives us exactly the information that we wanted to obtain by modelling: its hierarchical nature provides inside in relevance of specific variables in cheese production and split values on the branches of the tree gives us boundary values for some of the observed variables which separate positive from negative examples. These boundary values could be compared against existing standards in cheese production technology which might be subsequently reconsidered.

Decision trees [1, 2] are one of the most widely used and practical methods for inductive inference. They are used to predict membership of examples in the classes of categorical dependent variable from their measurements on predictor variables. Advantages of decision tree learning is that they are robust to noisy data, they make no statistical assumptions and can handle data that are represented on different measurement scales. Nevertheless, if statistical assumptions are met, it is recommending to use more traditional statistical methods first. In this paper we use the algorithm of QUEST (Quick, Unbiased, Efficient Statistical Trees) [3] implemented by Statistica 7.1. Bertelli and coauthors already have used QUEST for research in the food industry. In their paper [4] they classify honey from different floral sources using the method of principal components for selection of variables in the way to use only those variables which have big loans in principal components. In this paper we also use principal components, but not with a view to select variables by it, but rather to detect the most relevant components in technology process and their correlation with dependent variables. Beside application in food industry QUEST has already bin used in the field of medicine [5], social sciences [6,7,8], meteorology [9,10], and so on.

In the forth section we present the results of process modelling by employing a decision tree. We succeeded in construction of interpretative tree only in the first experiment in which the dependent variable is cfu/ml of EC. However, in the case of the second experiment (dependent variable cfu/ml of Y&M) we did not succeeded in construction of an interpretative tree and the reasons for that are explained in the third section which gives preliminary analysis of measured variables. Generally speaking, variables used in the second experiment are very dispersed and noisy. The most important aim for us was to constructed tree that is interpretative, and that is why in the process of construction we were faced with the problem of trade off between simple and interpretative tree and accuracy of classification achieved by it.

The last section gives conclusions and directions for further work.

2 Description of technology process

The technology process of cheese production is proceeding in the following fazes [11,12]:

1. Transport of milk from the farms – receiving and quick cooling in sheet chiller to 4-6°C.

2. .Storage in refrigerate tanks at 4-6^{o}C for 2 – 24 hours and microbiological monitoring of:

2.1 Aerobic Mesophilic Bacteria (AMB) in raw milk (variable AMB-RM) – Method of detection: Media tryptic glucose yeast agar 72 h on 30 ^{o}C,

2.2 Enterobacteriaceae in raw milk (variable E-RM) – Method of detection: Media violet red bile glucose agar 24 – 48 h on 37°C,
2.3 Escherichia coli (EC) in raw milk (variable EC-RM) – Method of detection: Media endo agar 24 – 48 h on 37 °C and confirmation test.

3. Transport of milk by pumps to pasteurizer and heating to 42-45°C for skimming in separator. Standardization of milk fat by addition of cream.
4. Pasteurization in sheet pasteur on 83-85°C for 20 sec and quick chilling to 30-32°C.
5. Transport of milk by pumps to Schulemberg bath for fresh cheese. Microbiological monitoring of:
5.1 equipment cleanness is controlled by slide method (sanibact procedure : AMB (variable AMB-EC) and Enterobacteriaceae (variable E-EC),
5.2 AMB in pasteurized milk (variable AMB-PM) in ml – Method of detection: Media tryptic glucose Yeast agar 72 hours on 30 °C,
5.3 EC in pasteurized milk (variable EC-PM) in ml – Method of detection: Media nutrient broth 14 – 16 hours on 37 °C - Brilliant green bile broth 2 % 24 hours on 37 °C - Endo agar 24 – 48 h on 37 °C and confirmation test,
5.4 Physicochemical monitoring: percentage of milk fat , acidity of milk and temperature.
6. Addition of frozen starter culture (Streptococcus sp. and Lactobacillus sp.,10^7 cfu/ml) to milk.
7. Addition of enzyme rennin – clotting of milk. Measuring of milk acidity.
8. Cutting of cheese at the end of progression of starter culture and at the optimal pH point product pressing for elimination of whey. Here is measured time of fermentation (time needed that pH achieve optimal value).
9. Discharge of cheese by special tools. Pack by machine to plastic can (0.25 , 0.5 and 1 kg). Product is monitored to:
9.1 EC in final product (variable EC-FP)
9.2 Y&M in final product (variable YM-FP) Method of detection: Media saburaud dextrose agar 25° for 3-5 days.
10. Chilling in cold-storage till 4-6°C and distribution.

During the process we measured 21 variables, but during the analysis we selected relevant variables for each of two experiments. For he first experiment relevant variables are: AMB in raw milk (AMB-RM), Enterobacteriaceae in raw milk (E-RM), EC in raw milk (EC-RM), AMB in pasteurized milk (AMB-PM), EC in pasteurized milk (EC-PM), equipment cleanness on AMB (AMB-EC), equipment cleanness on Enterobacteriaceae (E-EC), milk pH at cutting point of cloth (pH-CPC) and cloth pH before pressing (pH-CBP).
For the second experiment relevant variables are: AMB in raw milk (AMB-RM), AMB /ml of rinsing water (AMB-RW), Y&M /ml of rinsing water (YM-RW), AMB in pasteurized milk (AMB-PM), equipment cleanness on AMB (AMB-EC), AMB in air (AMB-A) and Y&M in air (YM-A).

3 Preliminary analysis of variables

3.1 Descriptive statistics

In the first step we will compare variables observed in conducted experiments by their descriptive statistics. The intension is to show that generally variables are dispersed and in the most variables there is no significant difference between mean values of observed variables for two defined classes: class of products of undesirable quality (UQ) and that of desirable quality (DQ). From Tab. 1 it can be seen that coefficients of variation (the forth and fifth column) are big for all variables and classes except for two last variables, which is indication of big dispersion of data. Significantly better values of means for measured variables (the second and the third column) are shown bolded. It can be seen that values of EC in raw milk and EC in pasteurized milk are significantly better for the class of products of desirable quality. There in no significant difference between the means of the rest measured variables analyzed in the first experiment.
From Tab. 2 it can be seen that in the second experiment coefficients of variation (the forth and the fifth column) are big for all observed variables. The smallest variation of data there is in variables of AMB in raw milk and AMB in air. Significantly different values of mean (the second and the third column) for two observed classes are shown bolded. The variable of AMB in air is significantly smaller for class of products of desirable quality (which is expected) whereas the variable of AMB in /ml of rinsing water is significantly smaller for the class of products of undesirable quality (which is unexpected).

3.2 Principal components analysis

As a preliminary analysis of observed variables we have also conducted principal component analysis [13,14] of variables observed in both experiments.
A principal components analysis provides an insight into the most relevant factors round which variables are grouped. It allows expressing large proportion of total variance of data with smaller number of variables in directions of maximal variation of data. The first principal component is the linear combination of original variables with maximal variance, the second principal component is linear combination with

Tab. 1 Descriptive statistics for relevant variables in the first experiment. In the second and the third column there are means of observed variables for products of undesirable quality (UQ) and products of desirable quality (DQ) respectively. Significantly different values of means are shown bolded (more desirable value is shown bolded). In the forth and the fifth column there are coefficients of variation for class UQ and DQ respectively.

Variable	Mean UQ	Mean DQ	CV UQ (%)	CV DQ (%)
AMB - RM	5379487	4846847	39.75	40.52
E - RM	565641	452459	68.17	54.34
EC - RM	303590	**290243**	53.83	59.26
AMB - PM	112.05	113.46	150.41	92.78
EC - PM	1.72	**0.66**	159.10	188.47
AMB - EC	1.87	1.46	67.33	85.57
E - EC	0.74	0.48	110.03	164.30
pH - CPC	5.83	5.79	3.27	4.73
pH - CBP	4.52	4.52	0.92	1.03

Tab. 2 Descriptive statistics for relevant variables in the second experiment. In the second and the third column there are mean values of measured variables for products of undesirable quality (UQ) and products of desirable quality (DQ) respectively. Significantly different values of means are shown bolded (more desirable value is shown bolded). In the forth and the fifth column there are coefficient of variation for class UQ and DQ respectively.

Variable	Mean UQ	Mean DQ	CV UQ (%)	CV DQ (%)
AMB - RM	5379487	4846847	39.75	40.52
AMB - RW	**15.21**	24.57	89.32	56.43
YM - RW	0.64	0.50	110.23	130.59
AMB - PM	112.05	113.46	150.41	92.78
AMB - EC	1.87	1.46	67.33	85.57
AMB - A	44.23	**37.87**	36.90	40.27
YM - A	30.13	24.90	51.74	57.44

maximal variance orthogonal to the first component, and so on.

Tab. 3 shows loans of original variables for the first experiment (dependent variable EC in final product or EC-FP) in the first three principle components which account for 57.12% of the variation of the data. Greatest loans in the first component (which accounts for 21.67% of the variation) correspond to variables which describe cleanness of equipment (AMB–EC and E-EC) in the positive direction and variables of acidity of product in two different points of production (variables pH-CPC and pH-CBP) in the negative direction of the component. Greatest loans in the second component (which accounts for 20.21% of the variation) correspond to variables that describe quality of raw milk. The third component (that accounts for 15.44% of the variation) contrasts quality of pasteurized milk (variables AMB-PM and EC-PM) with acidity of cheese before pressing (variable pH-CBP). The number of principal components was chosen following with the criterion that eigenvalue of principal component is greater than one.

Tab. 3 Loans of variables in the first three principal components (the first experiment). In the last row there are correlations of dependent variable with principal components.

Variable	PC 1	PC 2	PC 3
AMB – RM	-0.1354	**-0.5724**	0.3966
E – RM	-0.2447	**-0.7893**	-0.0936
EC – RM	-0.4095	**-0.7349**	0.0791
AMB – PM	0.1167	0.1164	**0.4994**
EC – PM	-0.1568	0.0686	**0.6932**
AMB – EC	**0.6155**	-0.4018	-0.3283
E – EC	**0.6920**	-0.3678	-0.1182
pH – CPC	**-0.7021**	0.0143	-0.1032
pH – CBP	**-0.5451**	0.1145	**-0.5961**
EC-FP	0.0165	-0.1163	0.2014

Tab. 4 Loans of variables in the first three principal components (the second experiment). In the last row there are correlations of dependent variable with principal components.

Variable	PC 1	PC 2	PC 3
AMB – RM	0.2873	**-0.4975**	**-0.4833**
AMB – RW	-0.1035	-0.4059	**0.8566**
YM – RW	**0.5724**	0.0940	**0.4536**
AMB – PM	-0.3529	**-0.7721**	-0.0095
AMB – EC	0.1303	**-0.5835**	-0.1181
AMB – A	**-0.7021**	0.3344	-0.0231
YM – A	**-0.8922**	-0.0957	0.0406
YM-FP	0.2015	0.1287	-0.1781

Tab. 4 shows loans of original variables for the second experiment (dependent variable of Y&M in final product) in the first tree principal components which accounts for 64.6% of the variation. Greatest loans in the first component (which accounts for 26.45% of the variation) correspond to variables that measure AMB and Y&M of the air in the negative direction to the contrary of variable of Y&M in /ml of rinsing water. In the second component greatest loans correspond to variables of AMB in raw milk and air, and variable of equipment cleanness on AMB, while the third component contrasts quality of rinsing water to variable of AMB in raw milk.

The number of principal components was chosen following the criterion that eigenvalue of principal component is greater then one.

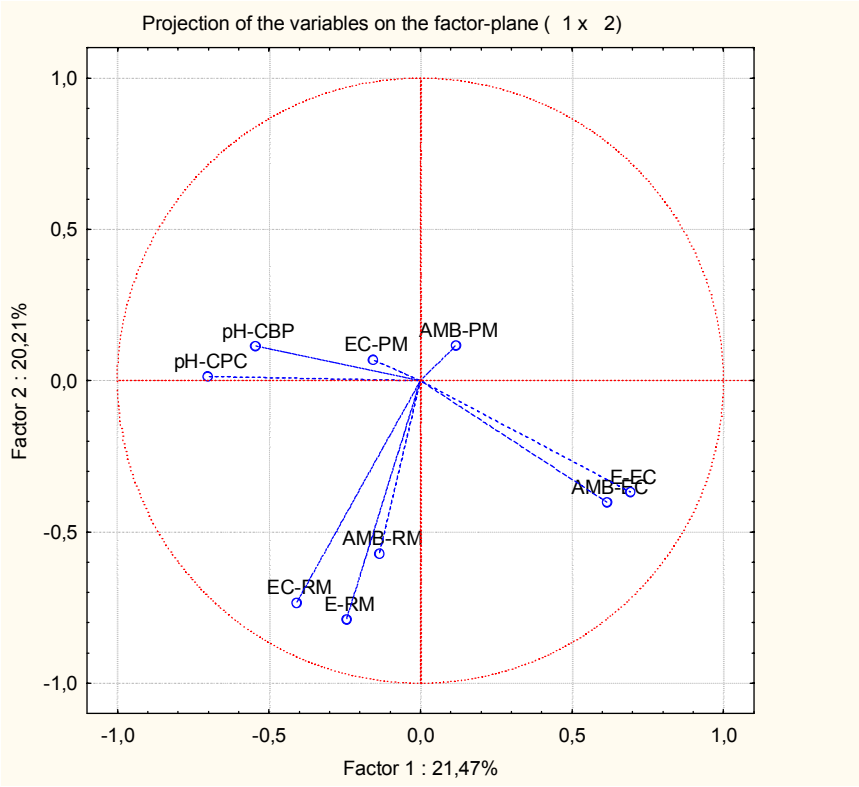

Fig. 1 Projection of variables on the first two principal components (the first experiment). Biggest positive correlations with the first principal component have variables that measure cleanness of equipment (AMB-EC and E-EC), while the biggest negative correlations with the first component have variables of acidity of milk and clot (pH-CBP and pH-CPC). The biggest loans in the second principal component have variables that measure quality of raw milk (AMB-RW, E-RM and EC-RM).

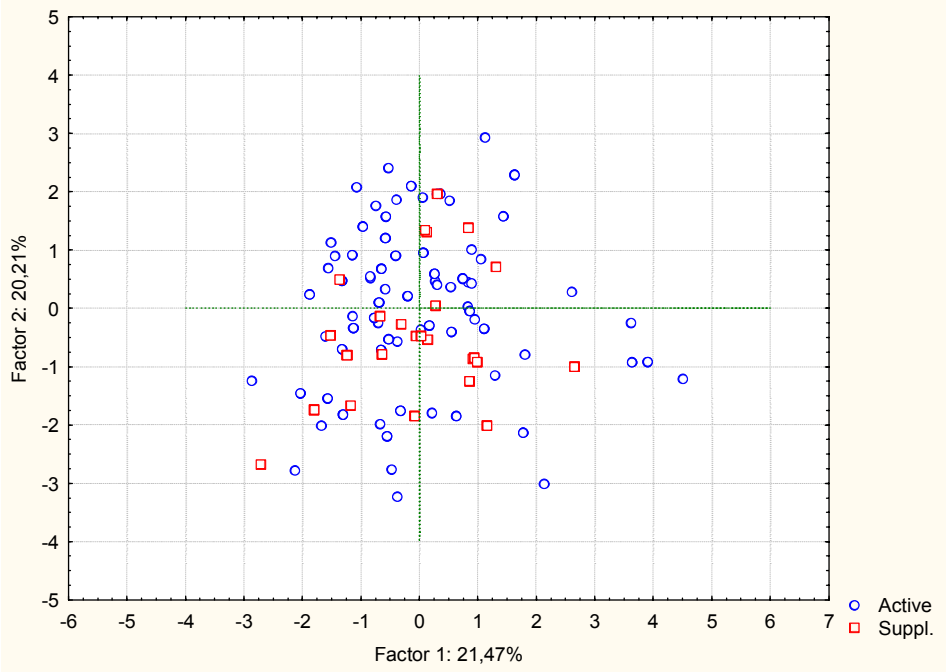

Fig. 2 Projection of examples on the first two principal components (the first experiment). The principal components are computed on the basis of only positive examples (active examples or cases), while negative examples are supplementary. It can be seen that negative examples tend to be located in the negative direction of the second principal component. From the Fig. 1 it is evident that the largest negative loans in the second principal component have variables of quality of raw milk. That indicates that cheese of undesirable quality tends to be made from the raw milk of inferior quality.

ISBN 978-3-901608-32-2

In the last rows in Tab. 3 and 4 there are correlations of dependent variables in the first and in the second experiment respectively with principal components. The correlations are fairly low, and in the case of the second experiment they are not even interpretable. Generally speaking, we find the first experiment, in which control variable is quantity of EC in the final product, much more interpretative.
On the Fig. 1 projection of the variables in the first experiment on the first two principal components is shown. On the Fig. 2 projection of examples on the first two principal components is shown. The principal components are computed on the basis of only positive examples (they are called active examples or cases), while negative examples are supplementary. It can be seen that negative examples (cheeses of undesirable quality) tend to be located in the negative direction of the second principal component (below x axis). From the Fig. 1 it is clear that the largest negative loans in the second principal component have variables of quality of raw milk. That indicates that cheese of undesirable quality tend to be made from raw milk of inferior quality. On the other hand, below the x axis there are lots of positive examples or projections of products of the good quality. This means that it is possible to produce good quality cheese from raw material or relatively inferior quality.

4 Design of a decision tree classifier

4.1 Method

A decision tree is a rule for predicting the class of an example from the values of its predictor variables. It classifies examples by sorting them down the tree from the root to some leaf node, which provides the classification of the example. Each node in the tree specifies a test of some variable and each branch descending from that node corresponds to condition set on the variable. As we deal here with ordered variables, the condition has a form $X \leq c$, or it represents some split of the variable. If example satisfies this condition it is send to left subnode, otherwise it is send to right subnode.
The algorithm of QUEST is response to its precursor: FACT algorithm [13], which in every its step selects variable which alone gives the best classification by using the variable with the largest *F*-statistics and employs linear discriminant analysis for every class to select split constant *c* of the chosen variable. QUEST uses ANOVA *F*-statistics and Levene *F*-statistics to select the variable. It selects a point of split *c* by employing a 2-means clustering algorithm which gives as a result two superclasses of initial classes. Then it uses quadratic discriminate analysis on the superclasses to get constant *c*.
One characteristic of classification trees is that if no limit is placed on the number of splits that are performed, eventually "pure" classification will be achieved, with each terminal node containing only one class of examples. However, "pure" classification is usually unrealistic, especially in the cases of noisy data which we are facing in this research. One of the options for controlling when splitting stops is to allow splitting to continue until all terminal nodes are pure or contain no more examples than a specified minimum fraction of the sizes of the classes. This is so called FACT style direct stopping criteria and here we have used threshold of 20% for fraction of class sizes.

4.2 Results

We have used 150 examples, 100 out of which were used as training examples and the rest were used as test examples. Our intension was to get interpretable tree which could give us information about importance of measured variables for the quality of the final product (those which are located in the highest nodes) and boundary values which separate positive from negative examples (splits on the branches of the tree).

We did not succeeded in that task in the case of the second experiment where dependent variable is quantity of Y&M in the final product, but we were successful in the case of the first experiment where dependent variable is quantity of EC in the final product. The reasons for that can be found in the third section where preliminary analysis of used variables in both experiments is given. The variables used in the second experiment where dependent variable is the quantity of Y&M in the final product are more dispersed (noisy), a variable of AMB in raw milk has much lower value for the class of products of undesirable quality (which is unexpected), and so on. In the opinion of experts in domain the most relevant variables for quantity of Y&M in the final product are variables of AMB and Y&M in the air. These variables have highest negative loans in the first principal component, but control variable of Y&M in the final product is not correlated with the first principal component with negative correlation as it would be expected.
The designed tree for the first experiment has seven terminal nodes and six splits. It gives accuracy of 71% of correctly classified training examples and 68% of correctly classified test examples. Similar accuracy values on the training and test set indicates that designed tree is not overfitted. A tree with better classification results could have been designed, but we opted for interpretability above accuracy. According to the tree the most relevant variable in cheese production in quality of pasteurized milk or, more specifically, quantity of EC in the pasteurized milk, while the variables of quality of raw milk (quantity of Enterobacteriaceae and AMB in raw milk) come second in the order of importance.
The selected variable in the first and

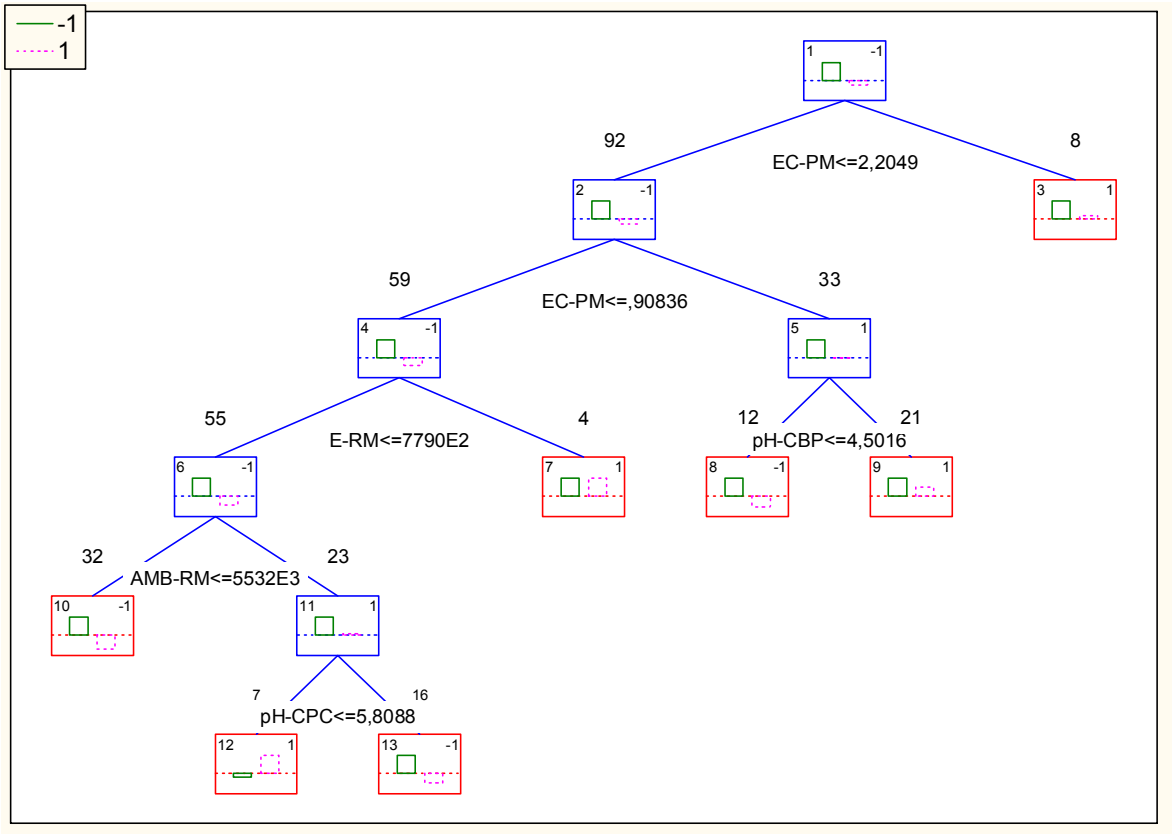

Fig. 3. Classification tree for the first experiment in which dependent variable is quantity of EC in the final product. It has seven terminal nodes and six splits and it classifies training examples with accuracy of 71% and test examples with accuracy of 68%. In right corner of node there is label of predicted class (1 is DQ, -1 is UQ) while in left corner there is node number.

in the second node is quantity of EC in pasteurized milk. In the first node splitting constant is 2.20, while in the second node splitting constant is more strict, being 0.90. On the right branch of the second node about two thirds of negative examples are classified, while on the left branch of that node less then 20% of negative examples are classified. That is why, according to estimation done by an expert on the basis of this research, the standard regarding EC in pasteurized milk should be 1.00 cfu/ml, which is currently the internal standard. The splitting value in the forth node on variable of quantity of Enterobacteriaceae in raw milk is $7790 \cdot 10^2$ cfu/ml (or 5.89 log cfu/ml), while splitting value in the sixth node on variable of quantity of AMB in raw milk is $5532 \cdot 10^3$ cfu/ml (or 6.74 log cfu/ml). Examples which are classified on the left branch in the forth and in the sixth node (corresponding to variables that measure quality of raw milk) are classified as a positive examples or products of the good quality. That is why splitting constants for variables of AMB in raw milk and Enterobacteriaceae in raw milk will be suggested as a standards in the observed technology process.

5 Discussion and further work

From the preliminary analysis of measured variables for the second experiment where the dependent variable is the quantity of Y&M in the final product it can be seen that there is a lot of noise and inconsistency in these data. We did not succeed in designing an interpretative decision tree in this experiment.

For the first experiment, from principal components analysis and the designed decision tree, it can be concluded that the most relevant elements in cheese production, according to criteria of allowed quantity of cfu/ml of EC in the final product, is primarily the quality of milk after the pasteurization and secondarily the quality of raw milk. By this research the existing standard for the allowed quantity of EC in pasteurized milk is confirmed, while new internal standards for quality of raw milk (quantity of allowed Enterobacteriaceae and AMB) are to be established. From the projection of examples on the plane spread by first two principal components it can be seen that products of the bad quality mainly are made from the raw milk of inferior quality, but also there are products of the good quality made of raw milk of

ISBN 978-3-901608-32-2

inferior quality. It indicates that it is possible to get good product from bad raw material and high standardization and technological discipline would enable to get good products from existing raw milk. Namely, the quality of raw milk can not be improved significantly in a short period of time.

In further work we intend to introduce new scored variables which will depend on behavior of the employees during the technological process, their personal hygiene, cross-contamination, cleanness of the production room, ventilation in the production room, and so on. The perception of experts in the domain is that the quality of the final product depends strongly on numerous factors which can not be measured exactly.

6 References

[1] T.M. Mitchell. Machine Learning. McGraw-Hill. 1997.

[2] L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and Regression Trees. Wedsworth, Belmont. 1984.

[3] W. –Y. Loh, and Y.-S Shih. Split selection methods for classification trees, *Statistica Sinica*, 7: 815-840,1997.

[4] D. Bertelli, M. Plessi, A.G. Sabatini, M. Lollo, and F. Grillenzoni. Classification of Italian honeys by mid-infrared diffuse reflectance spectroscopy (DRIFTS), *Food Chemistry*, 101: 1582-1587, 2007.

[5] M.O. Hoque, Q.H. Feng, P. Toure, A. Dem, C.W: Critchlow, S.E. Hawes, T. Wood, C. Jeronimo, E. Rosenbaum, J. Stern, M.J. Yu, B. Trink, N.B. Kiviat, and D. Sidransky. Detection of aberrant methylation of four genes in plasma DNA for the detection of breast cancer, *Journal of Clinical Oncology*, 24: 4262-4269, 2006.

[6] K. Stahl. Influence of hydroclimatology and socioeconomic conditions on water-related international relations, *Water International*, 30:270-282, 2005.

[7] S. Kannebley, G.S. Porto, and E.T. Pazello. Characteristics of Brazilian innovative firms: An empirical analysis based on PINTEC - industrial research on technological innovation, *Research Policy*, 34: 515-527, 2005.

[8] M. Pal, and P.M. Mather. An assessment of the effectiveness of decision tree methods for land cover classification, *Remote Sensing of Environment*, 86: 554-565, 2003.

[9] J. Elsner, G. Lehmiller, and T. Kimberlain. Objective classification of Atlantic hurricanes, *Journal of Climate*, 9:2880-2889, 1999.

[10] M. Carter, and J. Elsner. A statistical method for forecasting rainfall over Puerto Rico, *Weather and Forecasting*, 12:515-525, 1997.

[11] S. Miletić. Mlijeko i mliječni proizvodi. Hrvatsko mljekarsko društvo , Zagreb. 1994.

[12] D. Sabadoš. Kontrola i ocjenjivanje kakvoće mlijeka i mliječnih proizvoda , 2. dopunjeno izdanje. Hrvatsko mljekarsko društvo , Zagreb. 1996.

[13] B.F.J. Manly. Multivariate Statistical Methods: A primer. Chapman and Hall. 1986.

[14] A. C. Rencher. Methods of Multivariate Analysis, second edition. Wiley-Interscience. 2002.

[15] W.-Y. Loh, N. Vanichestakul. Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of American Statistical Association*, 83:715-728, 1988.

