International Scientific Conference 20 UNITECH 26

INTERNATIONAL SCIENTIFIC CONFERENCE
19-21 November 2026
GABROVO

ПРОГНОЗИРАНЕ НА НИВА НА РИСК ПОСРЕДСТВОМ SVM, C&R, CHAID, EXHAUSTIVE CHAID, BOOSTED, RANDOM FOREST И NAÏVE BAYES ИНСТРУМЕНТИ

Георги Георгиев[1], Димитриос Парусис[2]

[1]Технически университет - Габрово, ул. „Хаджи Димитър“ № 4, Габрово, България
[2]Технически университет - Габрово, ул. „Хаджи Димитър“ № 4, Габрово, България
givanow@abv.bg

RISK LEVEL PREDICTION BY SVM, C&R, CHAID, EXHAUSTIVE CHAID, BOOSTED, RANDOM FOREST AND NAÏVE BAYES INSTRUMENTS

Georgi Georgiev[1], Dimitrios Parousis[2]

[1]Technical University of Gabrovo, 4 “Hadji Dimitar” Str., Gabrovo, Bulgaria
[2]Technical University of Gabrovo, 4 “Hadji Dimitar” Str., Gabrovo, Bulgaria
givanow@abv.bg

Abstract
The paper proposes a methodology for the investment risk prediction about the potential projects implementation for the next-generation educational platforms integration in ICT resources digitalization for education purposes. The methodology is based on Adaptive Machine Learning methods and algorithms. The design and verification procedures about Support Vector Machine (C-SVM and Nu-SVM classifiers) and k-Nearest Neighbors Methods procedures for different variations of the Kernel functions and metric distances have been conducted. Classification models for multivariate decision making on the basis of Classification & Regression (C&R), CHAID Exhaustive CHAID, Boosted Trees and Random Forest Algorithm were analyzed. The Naïve Bayes algorithm applicability in Gaussian distribution of the input economic metrics with efficiency confirmed over 90.0 % has been evaluated.

Keywords: Risk Level; C-SVM; Nu-SVM; C&R; CHAID; Exhaustive CHAID; Boosted and Random Forest Trees; Naïve Bayes.

1. ВЪВЕДЕНИЕ

Иновациите в образователните технологии се превръщат в основен способ за трансформиране на традиционните начини на усвояване на знания от консервативни и конвенционални до отговарящи на съвременните световни образователни критерии [1, 2]. Платформите за адаптирано обучение от следващо поколение са основна предпоставка за решаването на основните предизвикателства пред образователната сфера. Изграждането на иновативни дигиталните умения и компененции на обучаемите е възможност за адекватност на промените и изискванията в съвременната бизнес среда [3, 4]. Дигиталните средства, имащи отношение към подобряване на учебния процес е необходимо да отразяват иновациите в технологияния напредък на информационно-комуникационните инструменти и приложения, приложими за образователни цели [5]. Като тяхната роля следва да бъде, по-скоро

Print ISSN: 1313-230X, Online ISSN: 3033-1404

допълваща, а не заместваща. Всичко това не може да бъде постигнато без наличие на подходящи и достатъчни инвестиции [6-8].

В разглеждания доклад е синтезиран подход на категоризация на инвестиционния риск във връзка с изпълнение на европейски програми към проекти за внедряване образователни платформи от ново поколение в учебните регионални центрове. Анализът се основава на фактори микроклимата на бизнес средата на специфичната местна/секторна среда, респективно: Market Volatility Index (MVI), Inflation Rate (IR) и Stock Index Change (SIC), по отношение на "Средна" и „Висока" степен на инвестиционен риск. Подходът включва аналитичен инструментариум, представляващ съвкупност от методи и алгоритми на Машинното обучение, сред които са:

- Support Vector Machine;
- CHAID, Exhaustive CHAID, C&R, Boosted и Random Forest;
- Алгоритъм на Naïve Bayes (NB).

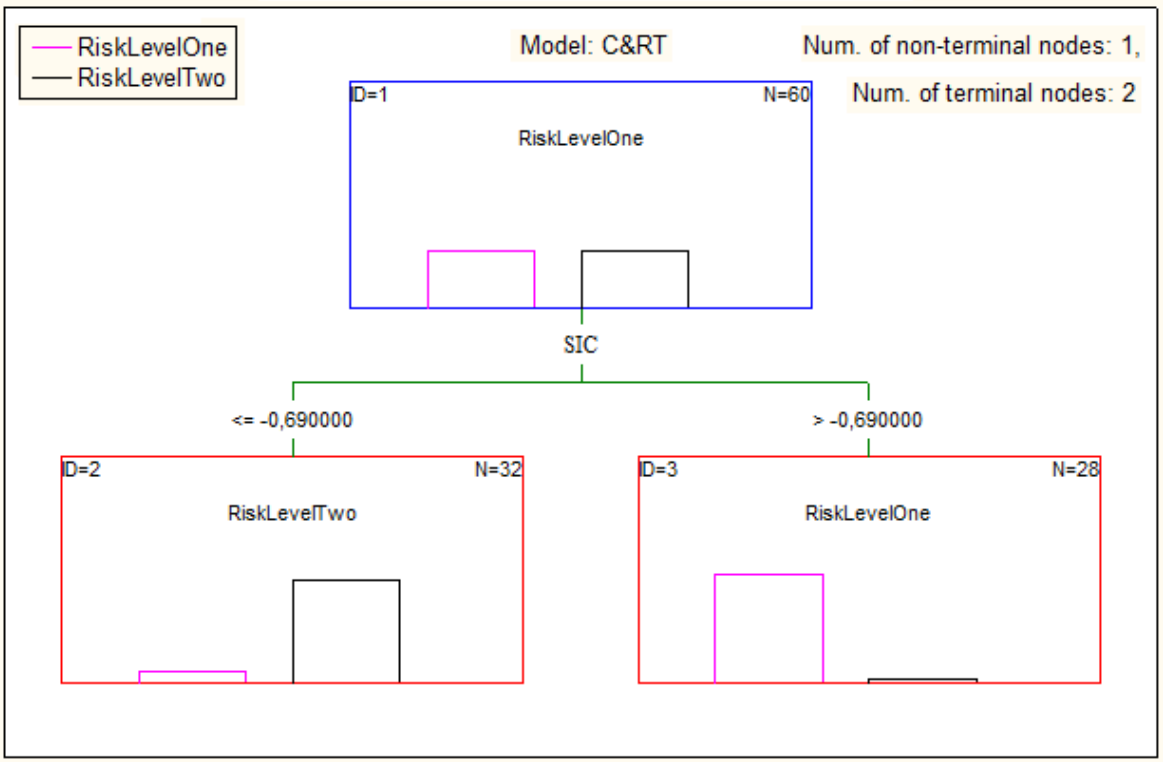
2. АНАЛИЗ И ОЦЕНКА НА РИСКА ЧРЕЗ МЕТОД НА ОПОРНИТЕ ВЕКТОРИ

Във връзка с дефинирания обект на изследване са създадени C-SVM класификатори при задаване на базисни Kernel функции, както следва: Linear (L); Polynomial (P); Radial Basis Function (RBF) и Sigmoid (S). Направена е комплексна оценка на всеки от класификационните модели в програмна среда STATISTICA. Приложената процедура регламентира разделяне на данните във входните обучаващи данни във фиксирани съотношения за обучение и тестване. Изведени са спецификации в хода на анализа, обобщени в структурата на Таблица 1. Според резултатите са получени идентични степени 93.333 % на

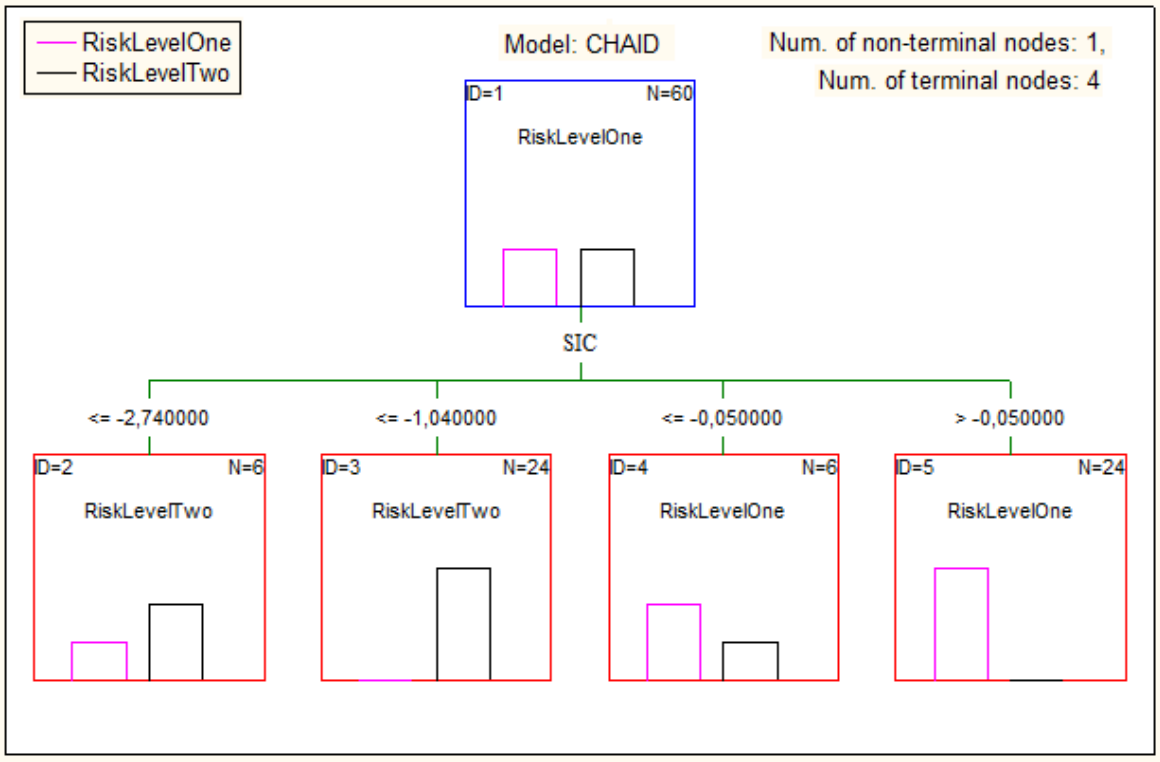
Таблица 1. Спецификация на C-SVM класификационни модели

Kernel тип функция: Linear
a) SVM type: capacity = 2,000 **b)** Number of support vectors: 23 **c)** Cross-validation accuracy (%): 93.333 **d)** Support vectors per class: 11 (class № 1), 12 (class № 2) e) Class. accuracy (%): 93.333 (Train), 93.333 (Test), 93.333 (Overall)
Kernel тип функция: Polynomial
a) SVM type: capacity = 7,000 **b)** Number of support vectors: 33 **c)** Cross-validation accuracy (%): 80.000 **d)** Support vectors per class: 17 (class № 1), 16 (class № 2) e) Class. accuracy (%): 84.444 (Train), 60.000 (Test), 78.333 (Overall)
Kernel тип функция: Radial Basis Function
a) SVM type: capacity = 3,000 **b)** Number of support vectors: 24 **c)** Cross-validation accuracy (%): 93.333 **d)** Support vectors per class: 12 (class № 1), 12 (class № 2) e) Class. accuracy (%): 93.333 (Train), 93.333 (Test), 93.333 (Overall)
Kernel тип функция: Sigmoid
a) SVM type: capacity = 7,000 **b)** Number of support vectors: 23 **c)** Cross-validation accuracy (%): 93.333 **d)** Support vectors per class: 12 (class № 1), 11 (class № 2) e) Class. accuracy (%): 93.333 (Train), 93.333 (Test), 93.333 (Overall)

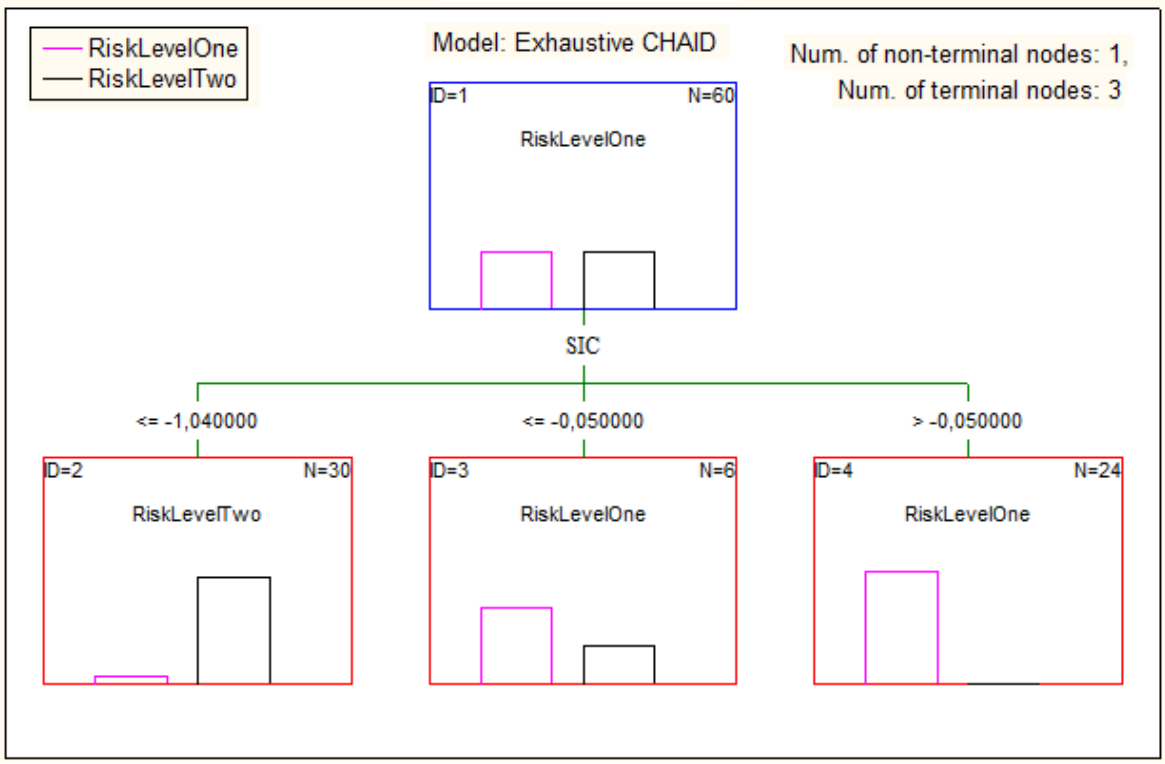
метриките: „Точности при обучение и валидация“ и „Крайна точност при тестова верификация“. Посочените количествени метрики се отнасят за случаите на “L”, “RBF” и “S” Kernel функционални типове. Спад в ефективността на тестовите C-SVM класификатори е отчетен при “Polynomial” тип на базисната Кернел функция. Установеният праг на класификационна точност при обучение



а)



б)



в)

Фиг. 1. *C&R а), CHAID б) и Exhaustive CHAID в) модели*

достига 84.444 %, а този при тестване е сведен до значително редуцирани нива от 60.000 %. Посредством подход V-fold крос-валидране е регистрирано съвпадение 93.333 % при "L", "RBF" и "S", както и праг 80.000 % при "Polynomial" тип във връзка със заложения критерий "Cross-Validation Accuracy".

Според представените спецификации при формиране на хиперповърнината, обособяваща подобластите на принадлежност на изходните класове в N-мерното класификационно пространство за класификаторите, удовлетворяващи нива на ефективност над 93.000 %, са необходими:

- 13 опорни вектора, респективно 11 за клас № 1 и 12 при клас № 2 за "Linear" Kernel функция;
- 14 опорни вектора, съответно 12 за клас № 1 и 12 при клас № 2 за "Radial Basis Function";
- 13 опорни вектора, следователно 12 за клас № 1 и 11 при клас № 2 за "Sigmoid" тип на активационната Кернел функция.

3. АНАЛИЗ И ОЦЕНКА НА РИСКА ЧРЕЗ МНОГОВАРИАНТЕН ИЗБОР НА РЕШЕНИЯ

Проведен е систематичен анализ относно ефективността на инструменти за генериране на модели за многовариантен избор на решения в две основни фази, свързани с обучение по:

Фаза №1: Класически "C&RT", "CHAID" и "Exhaustive" подходи. В хода на изпълнения е направена оценка на значимостта на приложение информативни признаци. Според предварителния процес на ранкиране на входните променливи при обучение са построени "C&RT модел", "CHAID структура" и "Exhaustive CHAID дърво на решенията" на основата на "SIC" базисен икономически индикатор при изключване на останалите два признака. Указаните структури в последователен ред, построени на

	Classification matrix (DTs) Response: Risk Level Model: C&RT			
	Observed	Predicted RiskLevelOne	Predicted RiskLevelTwo	Row Total
Number	RiskLevelOne	27	3	30
Column Percentage		96.43%	9.38%	
Row Percentage		90.00%	10.00%	
Total Percentage		45.00%	5.00%	50.00%
Number	RiskLevelTwo	1	29	30
Column Percentage		3.57%	90.63%	
Row Percentage		3.33%	96.67%	
Total Percentage		1.67%	48.33%	50.00%
Count	All Groups	28	32	60
Total Percent		46.67%	53.33%	

а)

	Classification matrix (DTs) Response: Risk Level Model: CHAID			
	Observed	Predicted RiskLevelOne	Predicted RiskLevelTwo	Row Total
Number	RiskLevelOne	28	2	30
Column Percentage		93.33%	6.67%	
Row Percentage		93.33%	6.67%	
Total Percentage		46.67%	3.33%	50.00%
Number	RiskLevelTwo	2	28	30
Column Percentage		6.67%	93.33%	
Row Percentage		6.67%	93.33%	
Total Percentage		3.33%	46.67%	50.00%
Count	All Groups	30	30	60
Total Percent		50.00%	50.00%	

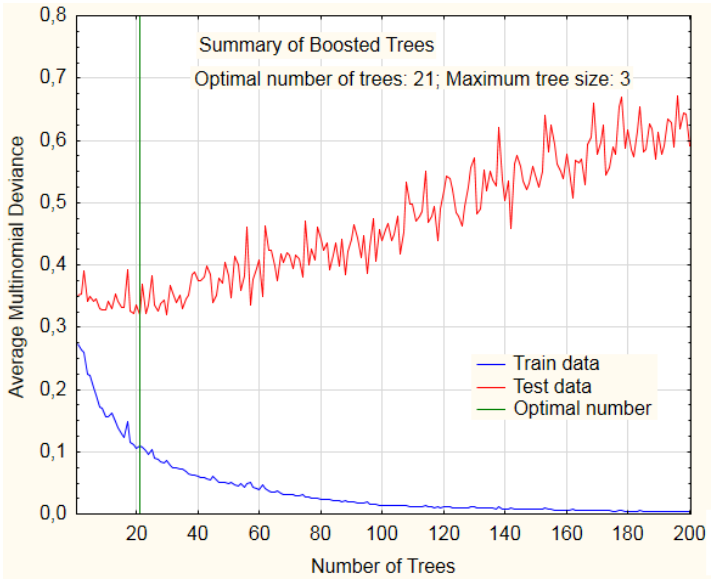
б)

Фиг. 2. *Спецификация на корекно / некоректно класифицираните еталони при а) C&RT, б) CHAID и Exhaustive CHAID модели*

основата на „1 non-terminal и 2 terminal структурни възела за C&RT структура“, „1 non-terminal и 4 terminal структурни възела при CHAID модел“ и „1 non-terminal и 3 terminal структурни възела относно Exhaustive CHAID класификатор“, са показани на Фиг. 1. Обобщени спецификации на изходните матрици, отразяващи разпределението на опитните наблюдения при верификация, са дадени на Фиг. 2. Установените нива на класификационна точност за първа и изходна група при тестовия C&RT модел се равняват, респективно на 90.00 % и 96.67 % (Фиг. 2.а)).

Отчетени са идентични нива на посочената метрика за качество на класификаторите при прогнозна „средна степен на риск“ и „висока степен на риск“ при “CHAID” и “Exhaustive CHAID” класификатори, съответно 93.33 % (Фиг. 2.б)). Налице е минимално предимство на модела за многовариантен избор на решения на базата “C&R” пред “CHAID” и “Exhaustive CHAID” подходи, където калкулираната крайна точност е 93.335 %

<u>Фаза №2:</u> “Boosted” подход и “Random Forest” метод. Създадени са серии от Boosted и Random Forest модели за многовариантен избор на решения, експериментално ограничени до максимално количество до „60“ на Фиг. 3.а) и „200“ структури на Фиг. 4.а). Приети са два базистни критерии за оценка на качеството на класификация, респективно “Average Multinomial Deviance (AMD)” при Boosted подход и “Misclassification Rate (MR)” за моделите по Random Forest метод. Поведението на посочените критерии е изследвано в хода на извършените и паралелно представено за обучаващите и тестовите процедури. Наблюдавани са по-добри вариации на



а)

Classification matrix (DTs)
Response: Risk Level
Analysis sample;Number of trees: 21

	Observed	Predicted RiskLevelOne	Predicted RiskLevelTwo	Row Total
Number	RiskLevelOne	22	2	24
Column Percentage		100.00%	10.00%	
Row Percentage		91.67%	8.33%	
Total Percentage		52.38%	4.76%	57.14%
Number	RiskLevelTwo		18	18
Column Percentage		0.00%	90.00%	
Row Percentage		0.00%	100.00%	
Total Percentage		0.00%	42.86%	42.86%
Count	All Groups	22	20	42
Total Percent		52.38%	47.62%	

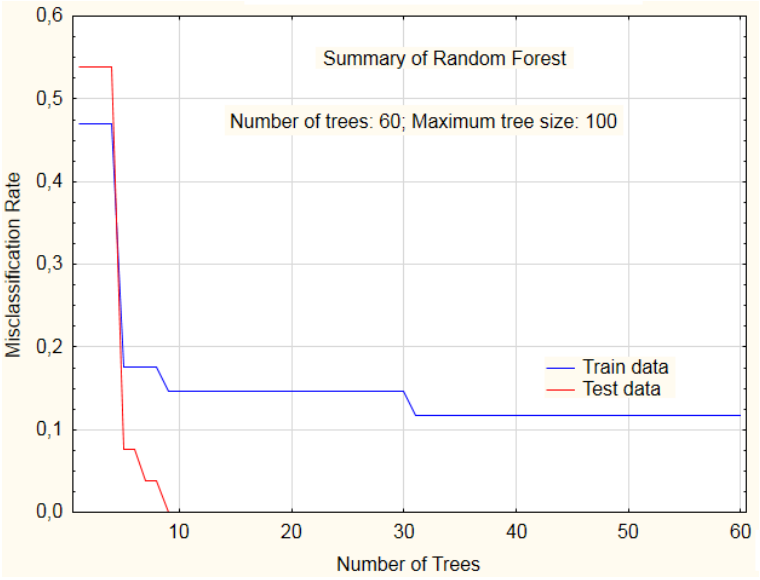
б)

Фиг. 3. *Изменение на Average Multinomial Deviance а) и спецификация на коректно / некоректно класифицираните еталони при б) при Boosted метод*

“AMD” и “MR” индекси при оперирane с данните с обучение спрямо тестовите процеси. Относно “Average Multinomial Deviance” е наблюдавана редукция на показателя при обучение от приблизително начално „0.28“ до „ниски нива, близки до нулев еталон“, докато при тестване е отчетен нарастващ вариационен характер със съизмерими пикове и спадове от „0.35“ при начална до „0.59“ при крайна тестова структура. Установено е оптимално количество структури – „21“, и максимална структурна големина, ограничена до „3“, за изследваните 200 “Boosted” структури. По отношение на “Misclassification Rate” са регистрирани приблизителни начални нива при обучение „0.470“ и тестване „0.535“ при базисна структура, намаляващи до установени нива „0.12“ и „ниски нива, близки до нулева стойност“ при краен “Random Forest” модел (Фиг. 4.а). Съгласно представените матрични спецификации на Фиг. 3.б) и Фиг. 4.б) могат да бъдат калулирани следните обобщени точности на класификация –
95.835 % за “Boosted” и 87.845 % при “Random Forest” метод, определящо поголямата пригодност на “Boosted” подход.

4. АНАЛИЗ И ОЦЕНКА НА РИСКА ПОСРЕСТВОМ NAÏVE BAYES АЛГОРИТЪМ

По отношение на обекта на изследване е създаден класификационен модел



а)

Classification matrix (DTs)
Response: Risk Level
Training set sample; Number of trees: 60

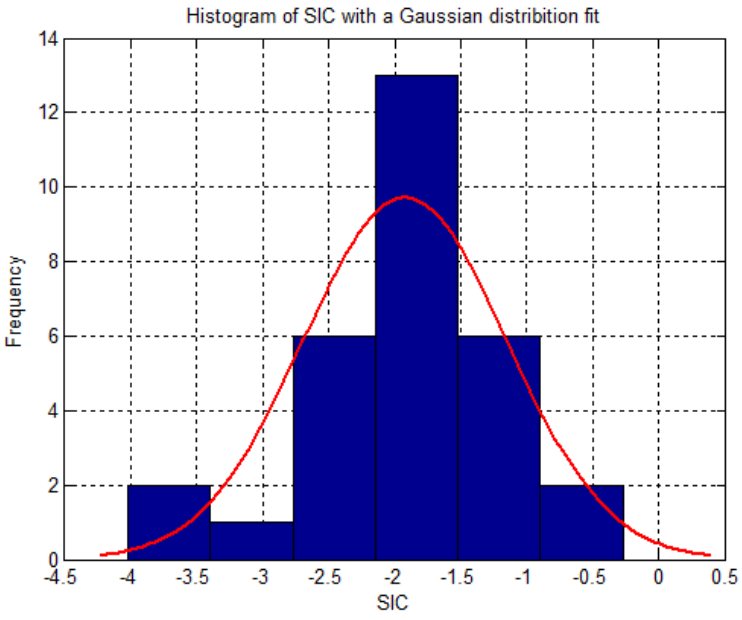
	Observed	Class Predicted RiskLevelOne	Class Predicted RiskLevelTwo	Row Total
Number	RiskLevelOne	13	3	16
Column Percentage		92.86%	15.00%	
Row Percentage		81.25%	18.75%	
Total Percentage		38.24%	8.82%	47.06%
Number	RiskLevelTwo	1	17	18
Column Percentage		7.14%	85.00%	
Row Percentage		5.56%	94.44%	
Total Percentage		2.94%	50.00%	52.94%
Count	All Groups	14	20	34
Total Percent		41.18%	58.82%	

б)

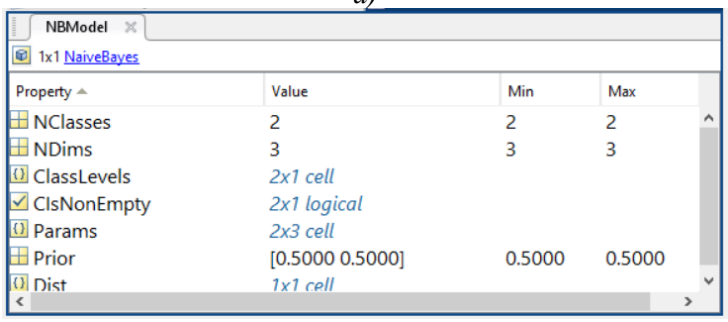
Фиг. 4. *Изменение на Misclassification rate а) и спецификация на корекно / некоректно класифицираните еталони при б) при Random Forest алгоритъм*

посредством Алгоритъма на Naïve Bayes. Процедурата е изпълнена при инициране на описателна функция на плътност на вероятностите на входните обучаващи от Гаусов тип. Във връзка с това на Фиг. 5.а) е дадена интерпретация при приложено Гаусово разпределение на информативен признак “Stock Index Change”. Генерираната хистограма показва сравнително добро приближение между изходната характеристика и теоретичната крива, описваща Закона за Нормално разпределение. По този начин е потвърдена коректността на процедурата по адаптиране на NB алгоритъм за прогнозиране на на инвестиционния риск относно изпълнението на проекти за иновационно обновяване на обучителните ресурси с платформи от ново поколение.

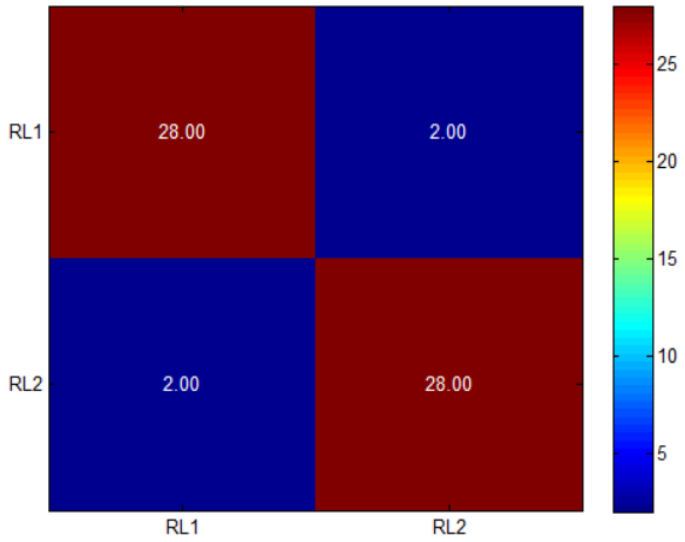
Съвкупните изходни променливи при процесите на синтез на Naïve Bayes класификационен модел, изпълнени в програмна среда MATLAB, са представени на Фигура 5.б). Променливите отразяват набора от информативни признаци, целеви изходни групи, калкулирамата априорна вероятност на класова принадлежност и набор от други специфични показатели. Направена е крос-валидания на създадения класификатор при фиксирано процентно съптношение 75 % : 25 % за обучение и тестване на крос-



а)



б)



в)

Фиг. 4. *"SIC" Гаусово разпределение а), спецификация на изходните променливи б) и класификационна матрица в) при Naïve Bayes алгоритъм*

валидиращите модели. Според построената матрица на коректни и некоректни класификации на Фиг. 4.в), констатираните с правилно асоциирана принадлежност са 28 наблюдения от съставна първа и втора изходна група. Общата калкулирана класификационна точност попада в поръдък 93.3 % - еталон за висока ефективност при разпознаване и категоризация на инвестиционен риск.

5. ЗАКЛЮЧЕНИЕ

- констатираната точност 80.000 % при“Polynomial” Кернел функция определя C-SVM модел като недостатъчно ефективен;
- Random Forest структура се асоциира с най-ниска степен на адекватност спрямо C&R, CHAID, Exhaustive CHAID и Boosted;
- Точността при NB (93.33 %) следва да бъде подобрена чрез замяна на типа на разпределение на входните данни с “Kernel”, “Gamma”, “Lognormal” или “Poisson”;
- най-висока ефективност 95.835 % е постигната при Boosted подход. Приложими за целите на класификационната задача са C&R Tree, последван от CHAID, Exhaustive CHAID, C-SVM при “L”, “RBF” и “S” Kernel функции и Naïve Bayes.

БЛАГОДАРНОСТИ: *Публикацията е във връзка с проект №BG16RFPR002-1.014-0013-C01 „Дигитализация на икономиката в среда на Големи данни– Втори етап“ (ДИГД2), финансиран по програма „Научни изследвания, иновации и дигитализация за интелигентна трансформация“ 2021 – 2027 и съфинансиран от Европейския съюз.*

ЛИТЕРАТУРА

[1] Issah M, Antwi S. Role of macroeconomic variables on firms’ performance: Evidence from the UK. Cogent Economics & Finance 2017;1(1):1-18.

[2] Kumar V. Predictive analytics: A review of trends and techniques. International Journal of Computer Applications 2018;182(1):31-37.

[3] Constantin A, Madalina-Gabriela A, Capusneanu S. Econometric model used for GDP correlation analysis and economic aggregates. Economic Computation and Economic Cybernetics Studies and Research 2019;53(1):183-197.

[4] Schiltz F, Masci Ch, Agasisti T, Horn D. Using regression tree ensembles to model interaction effects: A graphical approach. Applied Economics 2018;50(58):6341-6354.

[5]Tyrychtr J. Analytical system with decision tree for economic benefit. Agris on line Papers in Economics and Informatics 2017;9(4):123-129.

[6] Ozhegov E, Ozhegova A. Regression tree model for prediction of demand with heterogeneity and censorship. Journal of Forecasting 2019;39(3):1-6.

[7] Woloshco N. Adaptive trees: A new approach to economic forecasting. Journal of Organization for Economic Co-operation and Development 2020;1(1):1-43.

[8] Lohar P, Chowdhury K, Afli H, Hasanuzzaman M, Way A. A multinomial Naive Bayes classification approach for customer feedback analysis task. In: Proceedings of the eighth international joint conference on natural language processing, vol. I, 2017, p. 161-169.