Т Е Х Н И Ч Е С К И У Н И В Е Р С И Т Е Т – В А Р Н А

Мая Петрова Тодорова

ИЗСЛЕДВАНЕ НА МЕТОДИ НА МАШИННО ОБУЧЕНИЕ ЗА АНАЛИЗ НА ОНКОЛОГИЧНИ ЗАБОЛЯВАНИЯ

А В Т О Р Е Ф Е Р А Т

на дисертация за получаване на образователната и научна степен „ДОКТОР”

по докторска програма: Автоматизирани системи за обработка на информация и управление

професионално направление: 5.3. Комуникационна и компютърна техника

Научен ръководител: доц. д-р инж. НЕДЯЛКО НИКОЛОВ

Рецензенти:

1.

2.

Варна, 2021 г.

Дисертационният труд е обсъден на 06.07.2021г. в катедра „СИТ“ на катедрен съвет.

Автор: Мая Петрова Тодорова

Заглавие: Изследване на методи на машинно обучение за анализ на онкологични заболявания

Т Е Х Н И Ч Е С К И У Н И В Е Р С И Т Е Т – В А Р Н А

Мая Петрова Тодорова

ИЗСЛЕДВАНЕ НА МЕТОДИ НА МАШИННО ОБУЧЕНИЕ ЗА АНАЛИЗ НА ОНКОЛОГИЧНИ ЗАБОЛЯВАНИЯ

А В Т О Р Е Ф Е Р А Т

на дисертация за получаване на образователната и научна степен „ДОКТОР”

Варна, 2021 г.

Дисертационният труд съдържа 144 страници, включително 42 фигури, 58 таблици, 44 формули, оформени в 4 глави, приноси на дисертационния труд, списък с публикациите на автора по темата на дисертационния труд и списък на използваната литература от 146 заглавия, от които 49 на кирилица и 97 на латиница.

Защитата на дисертационния труд ще се състои на ..... г. от ..... ч. в ..... на открито заседание на жури сформирано със заповед на Ректора № ...../..... г.

Материалите по защитата (дисертацията, рецензиите и становищата) са на разположение на интересуващите се в Докторантски център, стая 318 НУК.

ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Актуалност на проблема

Основна характеристика на съвременната информационна епоха е динамичното и бързо нарастване на масивите с данни и обмена на информационни потоци.

Методите на машинното обучение се използва за решаването на широк клас задачи в медицината и в частност онкологията. В медицинската онкология методите на машинното обучение основно се прилагат в процеса на идентификация и диагностика на заболявания, като подпомагат медицинският екип при вземане на решения относно: подобряване на начина на живот на пациента, предписание на терапия и медикаменти, ранна оценка на критичните състояния. Машинното обучение може да подобри класификацията и категоризацията на съхранената документацията за пациент, да подпомогне процеси, свързани с наблюдение и прогнозиране на хода на заболяването, базирани на големи обеми от данни.

Разработката и модификацията на методи и алгоритми за определяне на вида на туморното образувание, определяне на преживяемостта на пациент и определяне стадия на заболяване са основни изследователски и приложни задачи. Не съществува цялостен универсален метод или алгоритъм за решаване на конкретна задача, което поражда необходимост от изследване на съществуващите подходи, като се прилагат различни методи и алгоритми за класификация и регресия с цел откриване на най-точния и ефективен класификатор.

2. Цел и задачи на изследването

Целта на настоящият дисертационен труд е изследване и оценка на приложимост и ефективност на методи на машинното обучение върху голям обем от данни за пациенти с онкологични заболявания. Направените изследвания да бъдат полезни както за специалистите в областта, така и за пациентите със злокачествени образувания.

За изпълнение на целта са формулирани следните **задачи**:

1. Анализ и изследване на методи и алгоритми за класификация и регресия в машинното обучение.
2. Генериране на модели за предсказване на стадия на онкологично заболяване.
3. Класификация на онкологични лечебни заведения на база преживяемост на пациент и топографски код на заболяване.
4. Избор на метрики за оценка на точност и ефективност на създадените класификационни и регресионни модели.

3. Обект и предмет на изследване

Обект на изследване са методи и алгоритми на машинното обучение за създаване на регресионни и класификационни модели.

Предмет на изследване са данни в областта на медицинската онкология, методи и алгоритми за класификация и регресия.

4. Методи на изследване

В дисертационния труд са използвани методи и алгоритми на машинното обучение за определяне стадия на онкологично заболяване и оценка на лечебно заведение. Използвани са статистически и аналитични методи за анализ на данни и извличане на знания. Приложени са научно-изследователски методи като системен анализ и моделиране. За апробиране на резултатите от научното изследване са използвани техники за моделиране и сравнителен анализ.

5. Място на изследване

Изследванията са проведени в лабораториите на катедра СИТ при ТУ – Варна.

6. Научна новост на изследването

Направени са регресионни модели за предсказване стадия на онкологично заболяване. Предимството на създадените модели е, че дават възможност за автоматизиране на процеса по стадиране. Разработена е методика за класифициране на лечебно заведение. Обучени и тествани са класификатори, който дават оценка на лечебно заведени, на база предложената методика.

7. Практическа ценност на изследването

Експерименталните изследвания са проведени с реални данни, за регистрирани пациенти с онкологични заболявания и лечебни заведения приложили лечение на пациенти, които представляват извадка от Националния Раков Регистър.

Към резултатите с приложна насоченост могат да се посочат получени експериментални данни, които предоставя възможност на пациентите да направят информиран избор за медицинско здравно заведение, където да им бъде приложено лечение при поставяне на тежката диагноза - находка на злокачествено заболяване.

8. Апробация на изследването

Основните етапи от разработването на дисертационния труд и основните резултати от изследванията са докладвани и публикувани в следните научни форуми и издания:

Конференции:

- VIII International Scientific Conference “Engineering. Technologies. Education. Safety”, 08-11.06.2020 , Borovets – 2 доклада;
- XXIX International Scientific Conference Electronics (ET). IEEE, 2020, September 16 - 18, 2020, Sozopol – 1 доклада;
- International Conference “Automatics and Informatics 2020” (ICAI 20), IEEE, 2020, October 1-3, Varna – 1 доклад;

Списания:

- Компютърни науки и технологии, Година XVIII, Брой 1/2020, стр. 111-117 и стр. 141-146 – 2 статии

Списъкът на публикациите е приложен в края на автореферата.

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. ОБЗОР НА МЕТОДИ НА МАШИННОТО ОБУЧЕНИЕ И DATA MINING ЗА АНАЛИЗ НА ДАННИ ЗА ПАЦИЕНТИ С ОНКОЛОГИЧНИ ЗАБОЛЯВАНИЯ

В Първа глава на дисертационния труд е представена актуалността на темата. Направено е въведение в Data Mining областта. Класифицирани са видовете задачи за извличане на знания от големи обеми от данни. Формулирани са задачите за класификация и регресия. Осъществен е обзор на публикации със сравнителен характер относно методи и алгоритми на машинното обучение и тяхното приложение в медицинската онкология.

ИЗВОДИ КЪМ ГЛАВА ПЪРВА

- В дисертационния труд Data Mining се дефинира като процес на анализ, изследване и моделиране на големи масиви от данни с цел откриване на закономерности и взаимовръзки, за предсказване на събития и тенденции, чрез използване на алгоритми от областта на машинното обучение;
- Класифицирани са задачите, които са обект на изследване в дисертационния труд. В зависимост от методите за извличане на знания се решават задачи за предсказване. Съгласно методите за анализ на данни се решават задачи за регресия и класификация;
- Основни методи на машинното обучение като: Дърво на решения, Опорни вектори, К-най-близък съсед, Наивен Бейсов класификатор и Ансамблови алгоритми намират широко приложение в областта на медицината и в частност медицинската онкология.

- Технологиите за извличане на данни позволяват откриването на модели в медицинските данни;
- Анализът на публикуваните научни изследвания показва, че няма цялостен универсален метод или алгоритъм за решаване на конкретна задача. Всеки метод или алгоритъм зависи от спецификата на задачата и от използваните данни;
- Качеството на използваните данни и обемът на извадката оказва съществено влияние на получените резултати;
- Информацията, която се съхранява в Националния Раков Регистър за всеки регистриран пациент с онкологично заболяване позволява да се изследва и анализира приложимостта и ефективността на методите на машинното обучение.

ГЛАВА 2. ОПИСАНИЕ НА МЕТОДИ И АЛГОРИТМИ НА МАШИННОТО ОБУЧЕНИЕ ЗА АНАЛИЗ НА ПАРАМЕТРИТЕ НА ОНКОЛОГИЧНИ ЗАБОЛЯВАНИЯ

Във Втора глава са проведени теоретични изследвания на методи и алгоритми за извличане на знания от данни чрез обучение на класификатор – K - най-близък съсед, Опорни вектори, Наивен Бейсов класификатор и Дърво на решения. За всеки метод са посочени предимства и недостатъци. Направен е сравнителен анализ на четири алгоритъма за създаване на дърво на решения – CART, ID3, C4.5, C5.0. Представени са метрики за оценка на качество и сравнение на модели.

В Таблица 2.2. са съпоставени четири метода на машинно обучение на база пет показателя.

Таблица 2.2. Методи на машинно обучение - сравнителен анализ

Метод	Подходящ обем на извадка	Клас задачи	Разбираемост / интерпретация	Изчислителна мощ	Точност
k-NN	Малък Среден	Класификация	Разбираем и лесен за интерпретация	Голям изчислителен ресурс	Качеството на класификация зависи от правилния избор на метрика за разстояние и стойността на k.
NBC	Малък Среден	Класификация	Разбираем и лесен за интерпретация	Малък изчислителен ресурс	Ниско качество на класификация при силно колерирани функции.
SVM	Малък Среден Голям	Класификация Регресия	Сложен за разбиране	Голям изчислителен ресурс	Високо качество на класификация и регресия.
DT	Малък Среден Голям	Класификация Регресия	Разбираем и лесен за интерпретация	Малък изчислителен ресурс	Високо качество на класификация и регресия.

ИЗВОДИ КЪМ ГЛАВА ВТОРА

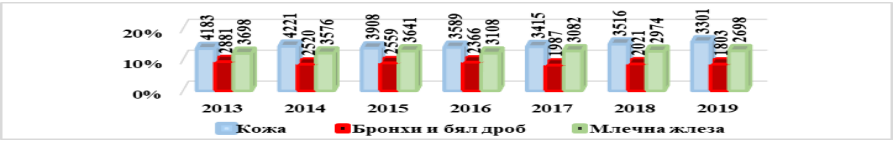
- Моделите създадени с DT, SVM, NBC и k-NN могат да подпомогнат процеса по поставяне на диагноза въз основа на медицински показатели на пациенти, избор на подходящо лечение и определяне на преживяемост на пациент. Чрез тези методи може да се предскаже резултат от проведен или не проведен курс на лечение, стадия на онкологично заболяване, да се открие обща тенденция на заболеваемост сред населението въз основа на дадено онкологично заболяване, възрастова група, пол, регион и фамилна обремененост и да се създаде оценъчна (рейтингова) система на лечебните заведения.
- Анализирайки предимства, недостатъци и отчитайки параметричните характеристики на алгоритмите CART, ID3, C4.5 и C5.0 за целите и задачите на дисертационния труд ще бъдат изследвани и приложени два алгоритъма за реализация на метода дърво на решения - CART и C4.5.
- За сравнение и оценка на класификационните модели ще бъдат използвани следните метрики - точност на класификатор, прецизност, пълнота, F-мярка, класификационна грешка, време за обучение, ROC криви и точност на предсказване.
- Регресионните модели ще бъдат съпоставени и оценени на база коефициент на детерминация, RMSE, MSE, време за обучение и точност на предсказване.
- Оценката на генерираните модели се реализира чрез прилагане на статистически метод за валидиране – k-кратно кръстосано валидиране.

ГЛАВА 3. ИЗСЛЕДВАНЕ НА МОДЕЛИ ЗА ПРЕДСКАЗВАНЕ СТАДИЯ НА ОНКОЛОГИЧНО ЗАБОЛЯВАНЕ ЧРЕЗ МЕТОДИТЕ НА МАШИННОТО ОБУЧЕНИЕ

В Трета глава на дисертационния труд е зададена методиката на експерименталните изследвания. Осъществен е подбор, анализ, логически контрол за коректност, подготовка и обработка на данни за пациенти с регистрирани онкологични заболявания в периода от 2013г. до 2019 година. Създадени и обучени са регресионни модели за предсказване стадия на онкологично заболяване чрез прилагане на алгоритми реализиращи три метода на машинно обучение – Дърво на решения, Опорни вектори и Ансамблов. Направена е оценка на моделите. Проведени са тестове. Представени и анализирани са експерименталните резултати.

Подбор на данни от избраната приложна област

Според данни от Националния Раков Регистър се наблюдава висока заболеваемост от злокачествени новообразувания на: млечна жлеза, кожа, бронхи и бял дроб. Изследвана е тази тенденция върху данни от генералната извадка (Фиг.3.4.).



Фиг. 3.4. Регистрирани пациенти в периода 2013 – 2019г.

Анализ и подготовка на данни

Основани етапи оказващи съществено влияние върху качеството на получените резултати е анализът на база логически контрол на данните, подготовката и обработката. Подготовката и обработката на данни за изследване включва: определяне на структура, премахване на грешки, преобразуване, създаване и редуциране на променливи.

- **Логически контрол на данните**

Всяко злокачествено заболяване се характеризира с топография и морфология. Топографията и морфологията се кодират съгласно Международната класификация на болестите (МКБ) [8].

Топография

Топографският код описва локализацията на първичните неоплазми. Топографските категории имат четиризначни кодове от C00.0 до C80.9. Структурата на топографски код е представена на Фиг.3.5.

Фиг. 3.5. Топографски код

Морфология

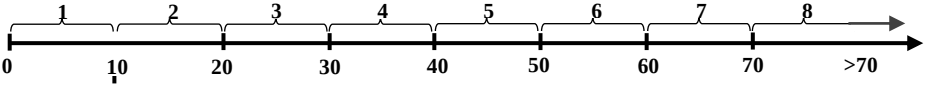
Морфологичният код описва характеристиките на тумора. На Фиг. 3.6. е представена структура на морфологичен код.

Фиг. 3.6. Морфологичен код

Световен стандарт за стадиране на рак е Международната TNM (Tumor – Node - Metstasis) система. Стадиранетото е начин да се опише локализацията на тумор и неговото разпространение.

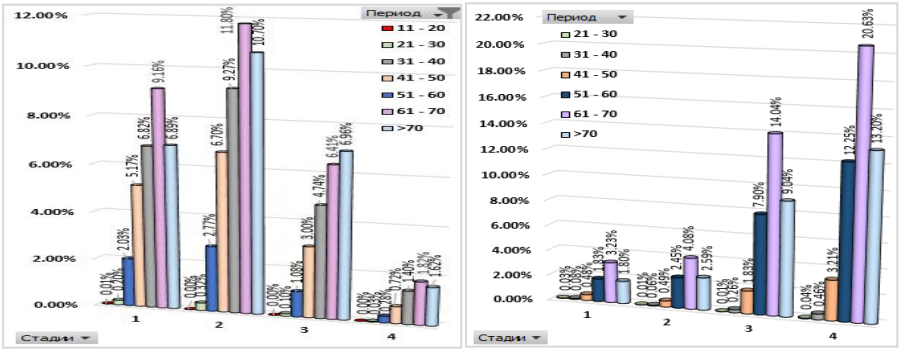
- **Подготовка и обработка на данни**

На база информация за възрастта на пациента при диагностициране са формирани осем възрастови групи (Фиг. 3.7.).



Фиг. 3.7. Възрастови групи

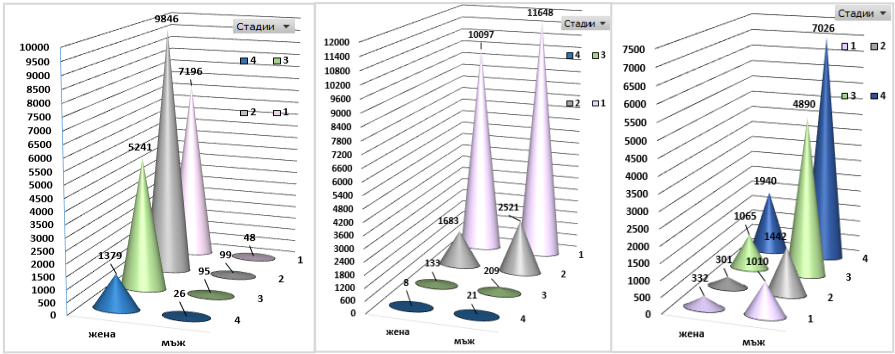
Изследвани са данните за стадирането на пациенти с онкологични заболявания на гърда (МКБ С50), кожа (МКБ44), бронхи и бял дроб (МКБ С34) и за всички МКБ групи. Изчислен е процентът на заболелите пациентите с определен стадий на заболяване по възрастови групи (Фиг. 3.8.).



а) МКБ С50 б) МКБ С34

Фиг. 3.8. Заболеваемост на пациенти по стадий и възрастова група

Изследва се зависимостта между пол на пациент и стадии на заболяване. Броят на пациентите разделени по пол с онкологични заболявания на млечна жлеза, кожа, бронхи и бял дроб с определен стадии на заболяване са представени на Фиг.3.9.



а) МКБ С50 б) МКБ С44 в) МКБ С34

Фиг. 3.9 Полово стадиране на пациенти

Приложен е корелационен анализ за описание на силата и посоката на зависимост между променливите величини Големина на тумор (T), Състояние на лимфни възли (N), Метастази (M) и Стадий на заболяване (Stage). Изчислени са коефициентите на корелация за три независими извадки, съдържащи данни за пациенти с регистрирани злокачествени заболявания с МКБ-та C50, C34 и C44 чрез Формула (3.1).

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (3.1)$$

Стойностите на коефициентите са представени в Таблица 3.12.

Таблица 3.12. Коефициенти на корелация за МКБ C50, C34 и C44

МКВ	T	N	M	
C50	0.7251	0.6410	0.5590	**Stage**
C34	0.3350	0.4036	0.8143	
C44	0.9345	0.3796	0.3706	

Получените стойности на коефициентите на корелация, изчислени за три МКБ групи, показват наличие на връзка между данните и необходимост от определяне на степента на зависимост между Стадий и Големина на тумор, Състояние на лимфни възли и Метастази за цялостната извадка обхващаща всички МКБ групи. Изчислените коефициентите на корелация са представени в Таблица 3.13.

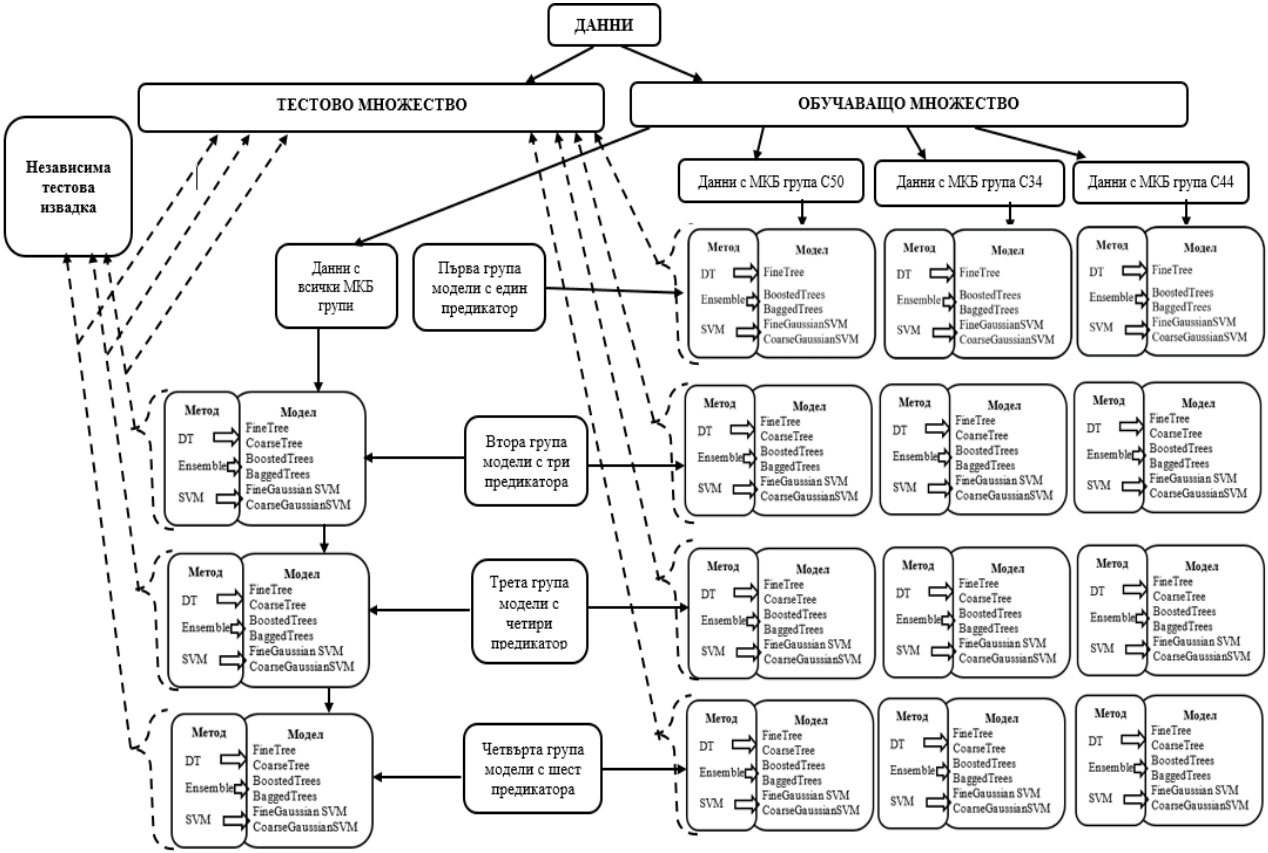
Таблица 3.13. Корелационна матрица

	T	N	M	Stage
T	1			
N	0.3934	1		
M	0.3612	0.3282	1	
Stage	**0.7597**	**0.5709**	**0.6923**	1

3.6. Създаване, обучение, оценка и тестване на модел чрез методи на машинно обучение. Експериментални резултати.

От генералната извадка, която съдържа 169109 записа, след обработка се формират три извадки. Обемът на първата извадка е 24 441 записа, съдържаща данни за пациенти със злокачествени заболявания на млечна жлеза. Втората извадка е с обем 27 882 записа и включва данни за пациенти със злокачествени образувания на кожа. Третият набор от данни съдържа 20 632 записа и обхваща пациенти с онкологично заболяване на бронхи или бял дроб. Всяка извадка се разделя на обучаващо и тестово множество. Обучаващото множество съдържа 80% от входните данни, а тестовото –20%.

Схема на експерименталните изследвания е представена на Фиг. 3.11.



Фиг. 3.11. Схема на експериментални изследвания

Първа група модели

Създават се четиридесет и пет регресионни модела за трите извадки. Изследвано е влиянието на всяка предикаторна променлива върху прогнозната. Използвана метрика за оценка на създадените модели е Root Mean Squared Error (Таблица 3.15).

Таблица 3.15. Root Mean Squared Error

Метод	Модел	Предикаторна променлива	C50	C44	C34
Дърво на решения	Fine Tree	Пол	0.87	0.41	0.92
		Възрастова група	0.87	0.41	0.92
		МКБ	0.86	0.40	0.91
Ансамблов	Boosted Trees	Пол	0.87	0.42	0.93
		Възрастова група	0.87	0.42	0.93
		МКБ	0.86	0.41	0.92
	Baggeded Trees	Пол	0.88	0.41	0.92
		Възрастова група	0.88	0.41	0.92
		МКБ	0.86	0.40	0.91
Опорни вектори	Fine Gaussian SVM	Пол	0.88	0.42	0.94
		Възрастова група	0.88	0.42	0.94
		МКБ	0.87	0.41	0.93
	Coarse Gaussian SVM	Пол	0.89	0.43	0.95
		Възрастова група	0.89	0.43	0.95
		МКБ	0.87	0.41	0.93

Моделите се прилагат и върху тестови данни. Изчислява се точността им при предсказване стадия на онкологично заболяване (Таблица 3.16.).

Таблица 3.16. Точност на предсказване

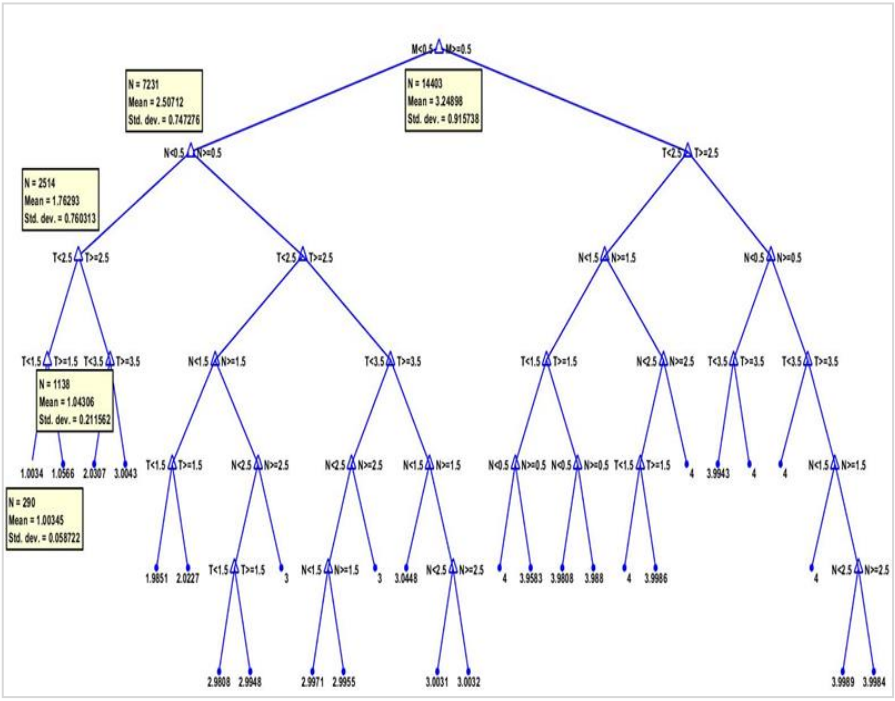
Метод	Модел	Предикатор			C50 %	C44 %	C34 %
		Пол	Възрастова група	МКБ			
Дърво на решения	Fine Tree	1			41.31	75.36	34.03
			1		41.31	75.36	34.03
				1	**41.32**	**75.38**	**37.05**
Ансамблови алгоритми	Boosted Trees	1			41.31	75.36	34.02
			1		41.31	75.36	34.02
				1	41.32	75.36	34.03
	Bagged Trees	1			41.30	75.36	34.03
			1		41.30	75.36	34.03
				1	**41.32**	**75.38**	**37.05**
Опорни Вектори	Fine Gaussian SVM	1			41.30	75.36	34.01
			1		41.30	75.36	34.01
				1	41.31	75.36	34.02
	Coarse Gaussian SVM	1			41.29	75.27	33.98
			1		41.29	75.27	33.98
				1	41.31	75.36	34.02

Моделите създадени на база една предикаторна променлива не са подходящи за определяне стадия на злокачествено заболяване, защото стойността на RMSE е голяма, а точността на предсказване е малка.

Втора група модели

Създадени и обучени са модели за предсказване на стадия на заболяване на база T, N и M. Използвани са две метрики за оценка - RMSE и R-Squared.

Моделите създадени с метода дърво на решения са конструирани с алгоритъм CART. На Фиг. 3.12 е представено дървото на решение на създаден модел на база извадка съдържаща данни за пациенти със злокачествени заболявания на бронхи и бял дроб. Въз основа на предикаторни променливи - Големина на тумор, Състояние на лимфни възли и Метастази – моделът трябва да определя стадият на заболяването.



Фиг. 3.12. Дърво на решения

Различните дървовидни модели са създадени въз основа на зададени стойности на параметрите:

- Минимален брой наблюдения в краен възел;
- Максималната дълбочина на дървото.

Моделите са съпоставени по коефициент на детерминация и средна квадратична грешка (Таблица 3. 18).

Таблица 3.18. RMSE и R-Squared на модели

Метод	Модел	C50		C44		C34	
		RMES	R^2	RMES	R^2	RMES	R^2
Дърво на решения	Fine Tree	**0.09**	**0.99**	**0.10**	**0.93**	**0.11**	**0.99**
	Coarse Tree	0.12	0.98	0.11	0.92	**0.11**	**0.99**
Ансамблов	Boosted Trees	0.13	0.98	0.12	0.92	0.18	0.96
	Baggeded Trees	**0.72**	**0.32**	**0.40**	**0.03**	**0.53**	**0.66**
Опорни вектори	Fine Gaussian SVM	0.16	0.96	0.12	0.91	0.12	0.98
	Coarse Gaussian SVM	0.19	0.95	0.12	0.91	0.18	0.96

Коефициентът на детерминация е статистическа метрика, която ни дава информация какъв процент от дисперсията на целевата променлива се прогнозира чрез стойностите на независимите променливи. Изчислена е точността на предсказване.

Таблица 3.19. Точност на модели създадени на база TNM системата за стадиране

Метод	Модел	C50	C44	C34
Дърво на решения	Fine Tree	**99.32%**	**98.78%**	**99.00%**
	Coarse Tree	99.14%	98.61%	**99.00%**
Ансамблов	Boosted Trees	**99.32%**	98.67%	**99.00%**
	Bagged Trees	**47.49%**	**75.42%**	**83.46%**
Опорни Вектори	Fine Gaussian SVM	**99.32%**	**98.78%**	**99.00%**
	Coarse Gaussian SVM	94.91%	**98.78%**	**99.00%**

Моделите Fine Tree и Boosted Trees са устойчиви. При тези модели се наблюдава тенденция за запазване на висока точност, както при обучение, така и при тестване с независим набор от данни. Fine Gaussian SVM и Coarse Gaussian SVM моделите, създадени с метода на опорните вектори, отчита по-ниска точност при обучение, но по-висока точност при прилагане върху тестов набор от данни. Най-ниска е точността на Bagged Trees моделите.

Съпоставяйки оценъчните параметри на първата група модели с втората група модели, създадени въз основа на TNM системата за стадиране и точността, която постигат при тестови множества от данни, възниква необходимост от изследване на нови модели, при които да се увеличи броя на предикаторните променливи, като се запазят трите основни предикатора - T, N и M.

С цел увеличаване на точността при предсказване стадия на злокачествено заболяване са създадени модели чрез нови комбинации от предикаторни променливи. Формирани са трета и четвърта група модели. Третата група модели определя стадия отчитайки връзката между МКБ на заболяване, Големина на тумор, Състояние на лимфни възли и Метастази. В четвъртата група модели се търси зависимост между Пол на пациент, Възрастова група и предикаторните променливи от трета група. Таблица 3.20 и Таблица 3.21 съдържат точността на моделите в процес на обучение и точността на моделите при предсказване.

Таблица 3.20. Точност на модели при обучение

Метод	Модел	Група	C50	C44	C34
Дърво на решения	Fine Tree	Трета	**0.99**	**0.93**	0.98
		Четвърта	**0.99**	**0.93**	0.98
	Coarse Tree	Трета	0.98	0.92	**0.99**
		Четвърта	0.98	0.92	**0.99**
Ансамблов	Boosted Trees	Трета	0.98	0.92	0.96
		Четвърта	0.98	0.92	0.96
	Bagged Trees	Трета	**0.77**	**0.91**	**0.84**
		Четвърта	**0.63**	**0.61**	**0.72**
Опорни вектори	Fine Gaussian SVM	Трета	0.96	**0.91**	0.98
		Четвърта	0.95	0.92	0.98
	Coarse Gaussian SVM	Трета	0.94	**0.91**	0.95
		Четвърта	0.94	0.91	0.93

Таблица 3.21. Точност на предсказване на модели от трета и четвърта група

Метод	Модел	Група	C50	C44	C34
Дърво на решения	Fine Tree	Трета	**99.32%**	98.72%	**99.01%**
		Четвърта	**99.29%**	**98.71%**	**99.01%**
	Coarse Tree	Трета	99.14%	98.60%	99.00%
		Четвърта	99.12%	98.61%	99.00%
Ансамблов	Boosted Trees	Трета	99.31%	98.23%	99.00%
	Boosted Trees	Четвърта	99.30%	98.69%	99.00%
	Bagged Trees	Трета	**72.32%**	**72.51%**	**87.31%**
	Bagged Trees	Четвърта	**63.76%**	**86.76%**	**78.58%**
Опорни вектори	Fine Gaussian SVM	Трета	99.25%	98.65%	98.99%
		Четвърта	97.07%	98.27%	93.66%
	Coarse Gaussian SVM	Трета	96.25%	**98.78%**	99.00%
		Четвърта	96.18%	97.15%	94.04%

Експерименталните резултати показват устойчивост и висока ефективност на създадените модели, както и достоверност на получените резултати, което е предпоставка за създаване на обобщен модел обучен с данни от обучаващото множество, съдържащо информация за всички регистрирани пациенти със злокачествени заболявания в периода 2013г. – 2019г., обхващащ всички МКБ групи. Използвани метрики за оценка на създадените модели са MSE, RMSE, R-Squared и време за обучение. Резултатите от експерименталните изследвания са представени в три отделни таблици.

Таблица 3.22. Оценъчни параметри на модели с предикатори T, N, M

Метод	Модел	MSE	RMES	R-Squared
Дърво на решения	Fine Tree	**0.14**	**0.37**	**0.89**
	Coarse Tree	**0.14**	**0.37**	**0.89**
Ансамблов	Boosted Trees	0.15	0.39	0.88
	Bagged Trees	**0.64**	**0.79**	**0.49**
Опорни вектори	Fine Gaussian SVM	0.16	0.40	0.87
	Coarse Gaussian SVM	0.15	0.38	0.88

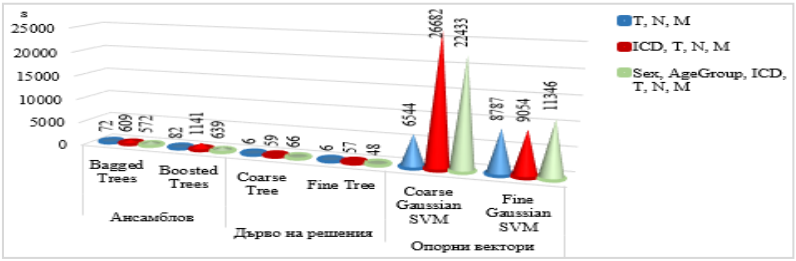
Таблица 3.23. Оценъчни параметри на модели с предикатори МКБ, T, N, M

Метод	Модел	MSE	RMES	R-Squared
Дърво на решения	Fine Tree	**0.03**	**0.17**	**0.98**
	Coarse Tree	**0.03**	**0.17**	**0.98**
Ансамблов	Boosted Trees	0.06	0.25	0.95
	Bagged Trees	**0.13**	**0.36**	**0.89**
Опорни вектори	Fine Gaussian SVM	0.05	0.23	0.96
	Coarse Gaussian SVM	0.07	0.26	0.95

Таблица 3.24. Оценъчни параметри на модели с предикатори Пол, Възрастова група, МКВ, T,N,M

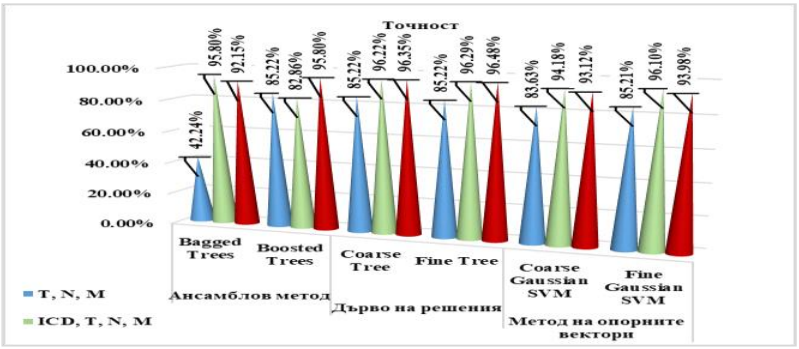
Метод	Модел	MSE	RMES	R-Squared
Дърво на решения	Fine Tree	**0.03**	**0.17**	**0.98**
	Coarse Tree	**0.03**	**0.17**	**0.98**
Ансамблов	Boosted Trees	0.06	0.25	0.95
	Bagged Trees	0.07	0.27	0.94
Опорни вектори	Fine Gaussian SVM	0.07	0.26	0.95
	Coarse Gaussian SVM	**0.09**	**0.28**	**0.93**

Сравнителна оценка на модели на база необходимо време за обучение (Фиг. 3.16.).



Фиг. 3.16. Време за обучение на модели

Сравнителна оценка на модели на база точност на предсказване (Фиг. 3.17.).



Фиг. 3.17. Тестова точност на модел

За определяне на приложимостта на създадените модели е използвана независима извадка, съдържаща данни за пациенти регистрирани през 2010г., 2011г. и 2012 година. Времевият интервал е различен от периода, от който са използваните данни за обучение и тестване. Резултатите от проведените тестове са представени в Таблица 3.25.

Таблица 3.25. Точност на модел при независима извадка

Метод	Модел	Предикаторни променливи		
		T, N, M	ICD, T, N, M	Sex, AgeGroup, ICD, T, N, M
Дърво на решения	Fine Tree	**84.13%**	**95.00%**	**95.13%**
	Coarse Tree	84.02%	94.94%	95.03%
Ансамблов	Boosted Trees	84.12%	93.84%	93.86%
	Bagged Trees	42.92%	82.15%	91.40%
Опорни вектори	Fine Gaussian SVM	84.13%	92.88%	94.79%
	Coarse Gaussian SVM	81.99%	92.37%	93.27%

ИЗВОДИ КЪМ ГЛАВА ТРЕТА

- След направен анализ на данните, за онкологично болните пациенти, обхващащ опознаване на източниците, от които са интегрирани данните, разбиране на значението и допустимите стойности на съдържащите се в тях променливи възниква необходимост от осъществяване на логически контрол на данните и редуциране на променливи с цел:
  - премахване на тези, които не представляват интерес за конкретните цели на експерименталните изследвания;
  - избор на най-информативните, които да участват при обучението на моделите.
- Качеството на получените модели зависи от анализът, структурата и предварителната подготовка и обработка на данните, което изисква преобразуване на стойности, създаване на нови променливи и отстраняване на допуснати грешки в данните (Таблица 3.6., Таблица 3.7., Таблица 3.8., Таблица 3.9.).
- Реализиран е статистически анализ на данните, в резултат на който се доказва наличие на силна корелационна зависимост между големина на тумор, състояние на лимфни възли, метастази и стадия на злокачествено заболяване (Таблица 3.13.);
- Изследвано е влиянието на всяка входна променлива върху предсказаната, с цел откриване на фактори, които в най-голяма степен влияят върху стадирането. Създадени са четири групи регресионни модели с три метода на машинното обучение - дърво на решения, опорни вектори и ансамблов - за три независими съвкупности от

данни. Първата група модели определя стадия на база една предикаторна променлива, втората група модели използва три предикаторни променливи, третата – четири и четвъртата група шест предикаторни променливи. Оценка на качеството на моделите е реализирана въз основа на следните метрики:

- RMSE;
- MSE;
- R-Squared;
- време за обучение;
- точност на предсказване.

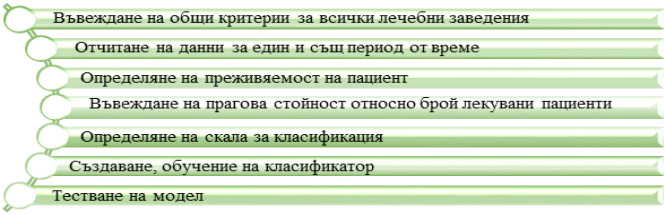
- Оценени и сравнени са 99 модела в процес на обучение и в процес на тестване. В резултат на проведените експерименти и получените резултати за приложените метрики, за оценка се достига до извода, че при създадените и обучени модели от втора, трета и четвърта група се наблюдава устойчивост, висока ефективност и достоверност, което позволява създаването на обобщен модел, които да се обучават с данни от множество, съдържащо информация за всички регистрирани пациенти в седем годишен период, независимо от локализацията и подлокализацията на злокачественото заболяване. Неподходящи са моделите от първа група, защото те регистрират ниска точност на предсказване (Таблица 3.16.) и високи стойности за RMSE (Таблица 3.15.).
- Обобщените модели са създадени на база три предикаторни променливи (Големина на тумор, Състояние на лимфни възли и Метастази), четири (МКБ, Големина на тумор, Състояние на лимфни възли и Метастази) и шест (Пол, Възрастова група, МКБ, Големина на тумор, Състояние на лимфни възли и Метастази). Оценка на моделите е направена в процес на обучение, в процес на тестване с тестово множество и с независима извадка.
- От създадените 18 модела с три метода на машинно обучение, моделът Fine Tree създаден с метода дърво на решения чрез алгоритъм CART е най-подходящ за определяне стадия на злокачествено заболяване. Моделът е обучен да определи стадия на база следните данни за пациент – пол, възрастова група, МКБ на заболяване, големина на тумор, състояние на лимфни възли и метастази. Този извод се обосновава от данните получени в резултат на проведените експерименти, получените стойности на използваните метрики за оценка на създадените модели и тяхното съпоставяне. За модел Fine Tree се отчита средноквадратична грешка 0.03 и коефициент на детерминация 0.98 (Таблица 3.24.). Необходимото време за обучение

на модела е 48s (Фигура 3.16.). При тестване на модела с данни от тестовото множество се постига най-висок резултат от 96.48% (Фигура 3.17.). Моделът постига най-голяма точност (95.13%) и при тестване с данни от независима извадка (Таблица 3.25.). Моделът е устойчив.

ГЛАВА 4. КЛАСИФИКАЦИЯ НА ОНКОЛОГИЧНИ ЛЕЧЕБНИ ЗАВЕДЕНИЯ НА БАЗА ПРЕЖИВЯЕМОСТ И ТОПОГРАФСКИ КОД НА ЗАБОЛЯВАНЕ

В Четвърта глава на дисертационния труд е зададена последователността на експерименталните изследвания за решаване на класификационната задача. На база анализ, обработка и извличане на знания от големи масиви с данни е предложена методика за решаване на формулираната задача. Въз основа на създадената методика са разработени класификационни модели за оценка на лечебни заведения. Приложени са метрики за оценка на създадените класификатори. Представени са експерименталните изследвания и резултати.

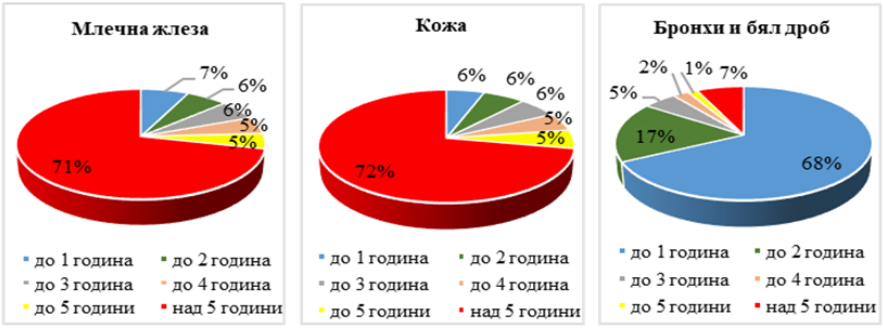
Процесът по създаване на класификационна система обхваща седем етапа (Фиг. 4.1.).



Фиг. 4.1. Етапи

Определя се преживяемостта на пациентите, като показател за относителният дял на пациентите с дадено заболяване, които са живи след определен момент от поставяне на диагнозата, за дванадесет МКБ групи.

На Фиг. 4.2. са представени графики за преживяемостта на пациентите с онкологични заболявания на млечна жлеза, кожа, бронхи и бял дроб, пикочен мехур, черен дроб и хранопровод.



Фиг. 4.2. Преживяемост на пациенти с определена МКБ група

Установена е зависимост между топография на злокачествено заболяване и преживяемост на пациент. С преживяемостта под една година са над 50% от всички пациенти регистрирани със злокачествени заболявания локализирани в областта на черен дроб (C22), хранопровод (C15), панкреас (C25), стомах (C16), главен мозък (C71), бронхи или бял дроб (C34). Над 50% от всички регистрирани пациенти с онкологично заболяване на кожа (C44), млечна жлеза (C50), женски полови органи (C51_58), мъжки полови органи (C60_63), устна (C00) и пикочен мехур (C67) имат преживяемост над 5 години.

При класифициране на лечебно заведение се отчита два основни критерия - преживяемостта на всеки пациент и локализацията на злокачественото заболяване в тялото. На база преживяемост пациентът поставя оценка на лечебното заведение. Създава се категориална променлива именувана от докторанта ИндексенСтатус. Променливата получава една от две възможни стойности - висок или нисък.

Класификация на лечебните заведения се формира за всяка основна група злокачествени заболявания кодирани съгласно Международната класификация на болестите. Определен е броят на пациентите получили лечение в медицинско здравно заведение за всяка МКБ група. За да бъдат коректно съпоставими лечебните заведения се въвежда прагова стойност относно брой обслужени пациенти, като се отчита и обемът на извадката (Таблица 4.2.).

Таблица 4.2. Прагова стойност за извадка

Брой записи	Прагова стойност	Извадка	Брой лечебни заведения
[500 , 1000]	10	C15	14
		C00	15
(1000, 2000]	30	C22	10
		C71	14
(2000, 4000]	40	C25	16
(4000, 6000]	50	C16	26
>6000	100	C34	19
		C44	24
		C50	27
		C51_58	33
		C60_63	32
		C67	18

Детайлизират се получените високи и ниски оценки на база локализация и подлокализация на заболяване.

За формиране на цялостна оценка на дадено лечебно заведение се определя зависимост между съотношението на високите и ниските оценки получени от пациентите. За всяка от дванадесетте МКБ групи е изчислен процентът на високите и ниските оценки. Получените резултати са представени в две отделни таблици, като извадките в тях са групирани на база въведения критерии за определяне на стойността на Индексния Статус на лечебно заведение.

Таблица 4.5. Процентна стойност за Индексен Статус на база преживяемост над една година.

МКБ	Локализация	Високи оценки	Ниски оценки
C15	Хранопровод	22.30%	77.70%
C34	Бронхи и бял дроб	22.52%	77.48%
C22	Черен дроб	27.78%	72.22%
C16	Стомах	36.32%	63.68%
C25	Панкреас	22.63%	77.37%
C71	Главен мозък	43.99%	56.01%
	Усреднен %	**29.26%**	**70.74%**

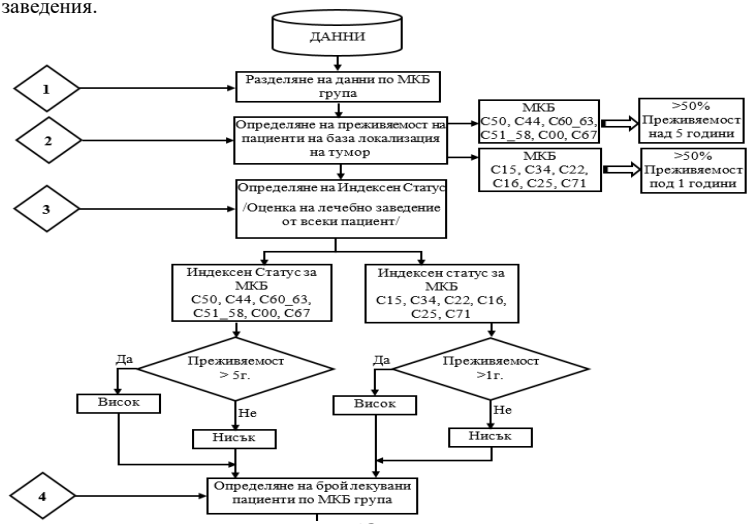
Таблица 4.6. Процентна стойност на Индексен Статус на база преживяемост над пет години.

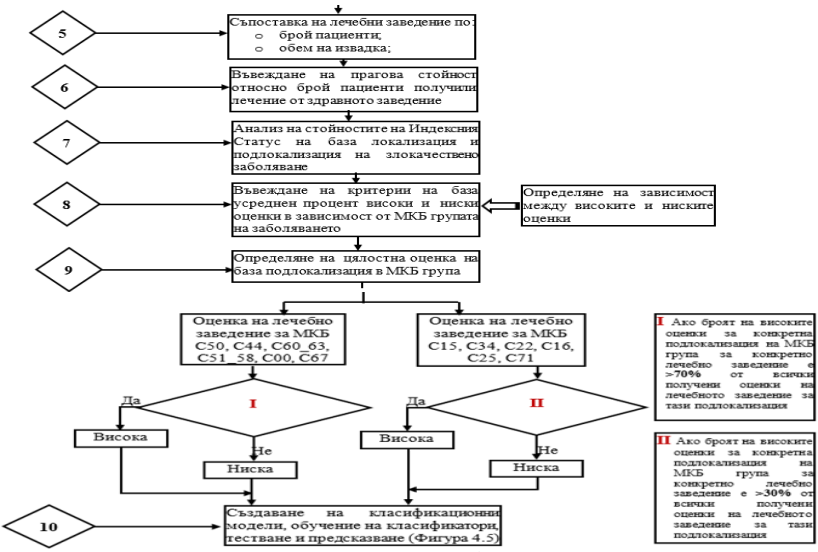
МКБ	Локализация	Високи оценки	Ниски оценки
C00	Устна	73.82%	26.18%
C60_63	Мъжки полови органи	61.52%	38.48%
C51_58	Женски полови органи	59.32%	40.68%
C44	Кожа	72.18%	27.82%
C50	Млечна жлеза	72.06%	27.94%
C67	Пикочен мехур	54.70%	45.30%
	Усреднен %	**65.60%**	**34.40%**

Основен критерии за определяне на цялостна оценка на едно лечебно заведение за конкретна група МКБ е усредненият процент.

4.3. Класификационни модели. Експериментални резултати

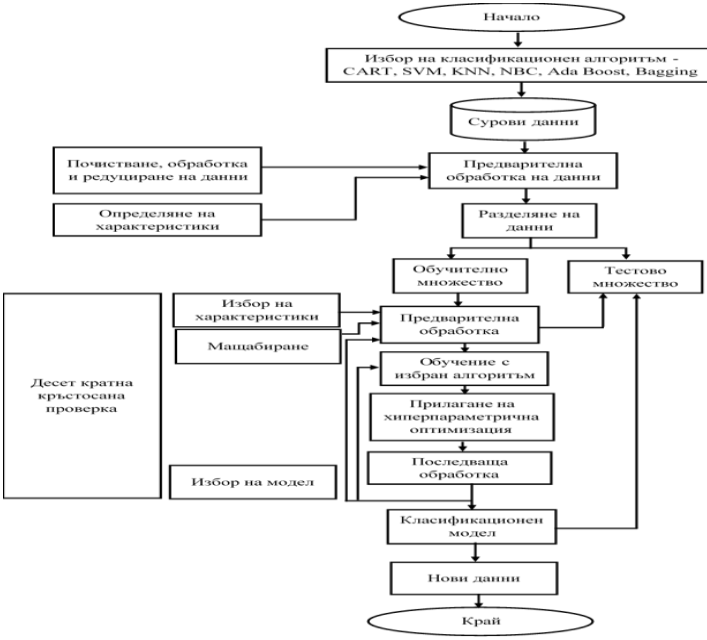
На Фиг. 4.4. е представена схема на предложената методиката за класифициране на лечебни заведения.





Фиг. 4.4. Методика за класифициране

Процедурата по изграждане на класификационен модел е зададена на Фиг. 4.5.



Фиг. 4.5. Работна процедура за създаване, тестване и приложение на класификатор

За създаване на класификационни модели са използвани пет метода на машинно обучение. Класификационните модели имат за цел да предскажат Оценката на лечебно заведение на база три предикаторни променливи – МКБ, Лечебно заведение и Индексен статус. За предпазване на моделите от преобучение е приложена десет кратна кръстосана валидация, съгласно която обучаващото множество се разделя на десет части и алгоритъмът се изпълнява десет пъти, като всеки път девет от десетте части се използват като обучаваща извадка, а останалата една част се използва за тестова. Пълна ефективност на модел се постига чрез усредняване на оценката на грешката от всички десет изпитания. Намаляване на пристрастията и на дисперсията се дължи на факта, че всяка точка от данни се среща само веднъж в множеството за валидиране и девет пъти в тренировъчната извадка. При този подход всички наблюдения се използват както за обучение, така и за валидиране.

Приложени са два алгоритъма за създаване на дървовидните модели – CART и C4.5. Реализирани са по три модела за всяка от дванадесетте извадки с алгоритъмът KNN и SVM, дванадесет модела с Наивен Бейсов класификатор и по два модела за всяка извадка с ансамблови алгоритми.

Използвани метрики за оценка на качество и сравнение на създадените модели са:

- Точност (Accuracy);
- Класификационна грешка;
- F-мярка;
- Прецизност (Precision)
- Пълнота (Recall);
- Receiver Operating Characteristic (ROC) криви;
- Време за обучение

Създадени и обучени са дванадесет класификатора с алгоритъм C4.5. В Таблица 4.11. е представена точност и класификационна грешка на моделите.

Таблица 4.11. Оценъчни параметри на модели TreeC4.5

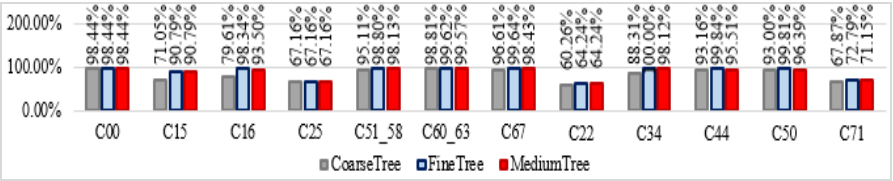
МКБ група	Точност на класификатор	Класификационна грешка
C34	98.38%	0.02
C50	99.64%	0.00
C44	99.58%	0.00
C51_58	97.85%	0.02
C60_63	99.50%	0.00
C71	98.28%	0.02
C16	99.47%	0.01
C25	99.40%	0.01
C67	99.03%	0.01
C00	98.42%	0.02
C15	**89.50%**	0.10
C22	**98.84%**	0.01

Таблица 4.10. съдържа изчислените оценъчни параметри на тридесет и шест класификатора създадени с алгоритъм CART.

Таблица 4.10. Оценъчни параметри на модели FineTree, MediumTree и CoarseTree

МКБ	Класификатор	Точност	Прецизност	Пълнота	F-мярка	Кл. грешка
C34	FineTree	**99.97%**	**1.00**	**1.00**	**1.00**	**0.00**
	MediumTree	97.59%	0.99	0.97	0.98	0.02
	CoarseTree	88.45%	0.87	0.95	0.91	0.12
C50	FineTree	**99.81%**	**1.00**	**1.00**	**1.00**	**0.00**
	MediumTree	97.25%	0.97	0.98	0.98	0.03
	CoarseTree	93.09%	0.95	0.93	0.94	0.07
C44	FineTree	**99.93%**	**1.00**	**1.00**	**1.00**	**0.00**
	MediumTree	96.39%	0.97	0.97	0.97	0.04
	CoarseTree	93.69%	0.96	0.94	0.95	0.06
C51_58	FineTree	**99.96%**	**0.99**	**0.98**	**0.98**	**0.01**
	MediumTree	98.54%	0.98	0.96	0.97	0.01
	CoarseTree	98.00%	0.97	0.94	0.96	0.02
C60_63	FineTree	**99.63%**	**0.99**	**0.99**	**0.99**	**0.00**
	MediumTree	99.62%	0.99	0.99	0.99	0.00
	CoarseTree	98.80%	0.96	0.97	0.96	0.01
C71	FineTree	**99.18%**	**0.99**	**0.99**	**0.99**	**0.01**
	MediumTree	94.75%	0.97	0.95	0.96	0.05
	CoarseTree	81.28%	0.86	0.83	0.84	0.19
C16	FineTree	**98.52%**	**0.98**	**0.99**	**0.99**	**0.01**
	MediumTree	91.82%	0.91	0.96	0.94	0.08
	CoarseTree	80.28%	0.82	0.87	0.84	0.20
C25	FineTree	**99.53%**	**1.00**	**0.99**	**0.99**	**0.00**
	MediumTree	**99.53%**	**1.00**	**0.99**	**0.99**	**0.00**
	CoarseTree	97.79%	1.00	0.93	0.96	0.02
C67	FineTree	**99.33%**	**0.98**	**0.96**	**0.97**	**0.01**
	MediumTree	98.52%	0.97	0.90	0.93	0.01
	CoarseTree	96.06%	0.82	0.86	0.84	0.04
C00	FineTree	**99.21%**	**0.99**	**0.98**	**0.99**	**0.01**
	MediumTree	**99.21%**	**0.99**	**0.98**	**0.99**	**0.01**
	CoarseTree	99.21%	0.99	0.98	0.99	0.01
C15	FineTree	**88.70%**	**0.85**	**0.78**	**0.81**	**0.11**
	MediumTree	**89.37%**	**0.87**	**0.78**	**0.82**	**0.11**
	CoarseTree	81.73%	0.72	0.68	0.70	0.18
C22	FineTree	**98.51%**	**0.98**	**0.98**	**0.98**	**0.02**
	MediumTree	**98.51%**	**0.98**	**0.98**	**0.98**	**0.02**
	CoarseTree	98.68%	0.98	0.99	0.98	0.01

Точността, която се постига при предсказване на оценката на лечебно заведение със създадените класификационни модели е представена на Фиг. 4.10.



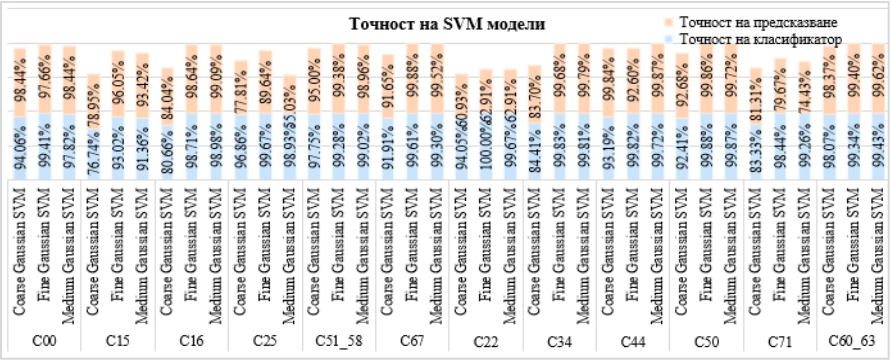
Фиг. 4.10. Точност на предсказване

Съпоставяйки оценъчните параметри на създадените дървовидни модели, с два различни алгоритъма, се достига до извода, че стойностите на показателите са близки, но в десет от дванадесет случая моделите FineTree създадени с алгоритъм CART реализират най-голяма точност и допускат най-малка класификационна грешка. Двете обучителни извадки (C15, C22), за които класификаторите TreeC4.5 реализират по-висока точност, съдържат малък обем данни.

Създадени са тридесет и шест класификатора с метода К-най-близък съсед и тридесет и шест с метода на опорните вектори. Стойностите на оценъчните параметри са представени в Таблица 4.12. и Фиг. 4.11.

Таблица 4.12. Оценъчни параметри на модели Fine KNN, Medium KNN и Coarse KNN

МКБ група	FineKNN			MediumKNN			CoarseKNN		
	Точност	Клас. грешка	Тестова точност	Точност	Клас. грешка	Тестова точност	Точност	Клас. грешка	Тестова точност
C34	**99.89%**	0.00	**99.68%**	99.03%	0.01	98.50%	89.97%	0.10	88.63%
C50	**99.84%**	0.00	**99.82%**	98.76%	0.01	99.49%	93.27%	0.07	95.00%
C44	**99.81%**	0.00	**99.87%**	99.10%	0.01	95.14%	95.19%	0.05	95.55%
C51_58	**99.14%**	0.01	**99.06%**	98.30%	0.02	98.33%	96.82%	0.03	93.55%
C60_63	**99.58%**	0.00	**99.46%**	98.96%	0.01	98.92%	97.53%	0.03	97.18%
C71	**99.18%**	0.01	**72.79%**	94.75%	0.05	71.15%	81.28%	0.19	67.87%
C16	**98.52%**	0.01	**98.34%**	91.82%	0.08	93.50%	80.28%	0.20	79.61%
C25	**99.53%**	0.00	**67.16%**	99.53%	0.00	67.16%	97.79%	0.02	67.16%
C67	**99.64%**	0.00	**99.64%**	98.36%	0.02	98.43%	92.61%	0.07	93.58%
C00	**99.21%**	0.01	**98.44%**	99.21%	0.01	98.44%	99.21%	0.01	98.44%
C15	**88.70%**	0.11	**90.79%**	**89.37%**	0.11	**90.79%**	81.73%	0.18	71.05%
C22	**99.67%**	0.00	**63.58%**	97.36%	0.03	61.59%	77.19%	0.23	59.60%



Фиг. 4.11. Точност на SVM модели

Две групи класификационни модели са създадени с ансамбловия метод и една група с Найвен Бейсов класификатор. При конструирането на ансамблови модели са използвани две технологии. Първата технология е Бустинг, при която всички класификатори се реализират последователно. Втората технология е Багинг, където се използва паралелно създаване на

композиция от класификатори. Експерименталните резултатите са представени в Таблица 4.13. и Таблица 4.14.

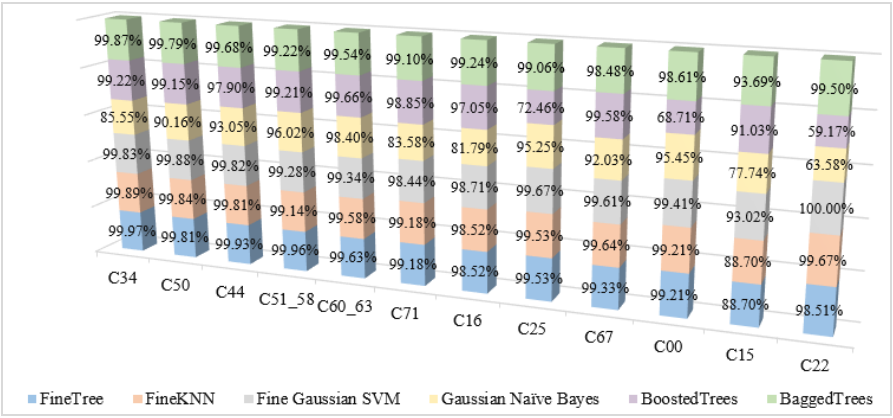
Таблица 4.13. Оценъчни параметри на модели BoostedTrees и BaggedTrees

МКБ група	Класификатор	Точност на класификатор	F-мярка	Класификационна грешка	Тестова точност
C34	BoostedTrees	99.22%	0.99	0.01	98.39%
	BaggedTrees	99.87%	1.00	0.00	99.84%
C50	BoostedTrees	99.15%	0.99	0.01	99.03%
	BaggedTrees	99.79%	1.00	0.00	99.63%
C44	BoostedTrees	97.90%	0.98	0.02	96.56%
	BaggedTrees	99.68%	1.00	0.00	99.21%
C51_58	BoostedTrees	99.21%	0.98	0.01	98.96%
	BaggedTrees	99.22%	0.98	0.01	99.06%
C60_63	BoostedTrees	99.66%	0.99	0.00	99.52%
	BaggedTrees	99.54%	0.99	0.01	99.46%
C71	BoostedTrees	98.85%	0.99	0.01	72.46%
	BaggedTrees	99.10%	0.99	0.01	73.44%
C16	BoostedTrees	97.05%	0.98	0.03	95.92%
	BaggedTrees	99.24%	0.99	0.01	98.94%
C25	BoostedTrees	72.46%	0.18	0.28	78.11%
	BaggedTrees	99.06%	0.98	0.01	67.16%
C67	BoostedTrees	99.58%	0.99	0.00	43.71%
	BaggedTrees	98.48%	0.99	0.00	62.25%
C00	BoostedTrees	68.71%	0.00	0.31	64.84%
	BaggedTrees	98.61%	0.98	0.01	99.22%
C15	BoostedTrees	91.03%	0.86	0.09	100.00%
	BaggedTrees	93.69%	0.90	0.06	94.74%
C22	BoostedTrees	59.17%	0.00	0.41	43.71%
	BaggedTrees	99.50%	0.99	0.00	62.25%

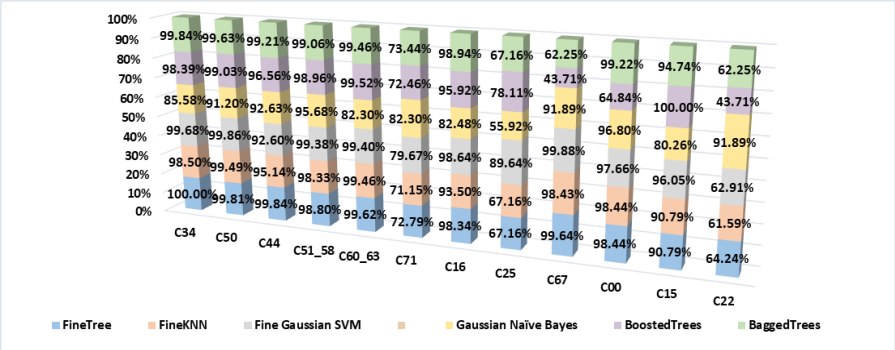
Таблица 4.14. Оценъчни параметри на модел Gaussian Naïve Bayes

МКБ група	Точност на класификатор	F-мярка	Класификационна грешка	Тестова точност
C34	85.55%	0.88	0.14	85.58%
C50	90.16%	0.92	0.10	91.20%
C44	93.05%	0.95	0.07	92.63%
C51_58	96.02%	0.91	0.04	95.68%
C60_63	98.40%	0.95	0.02	98.26%
C71	83.58%	0.87	0.16	82.30%
C16	81.79%	0.86	0.18	82.48%
C25	95.25%	0.92	0.05	55.92%
C67	92.03%	0.55	0.08	91.89%
C00	95.45%	0.92	0.05	96.80%
C15	77.74%	0.59	0.22	80.26%
C22	63.58%	0.94	0.05	91.89%

Класификационните модели създадени с пет метода на машинното обучение са съпоставени по два основни критерия - точността на класификатор (Фиг. 4.12.) и точност на предсказване (Фиг. 4.13.).



Фиг. 4.12. Точност на класификационен модел



Фиг. 4.13. Точност на предсказване на класификационен модел

Въз основа на получените резултати за обучителна и тестова точност на създадените модели са формулирани следните изводи:

- Класификационните модели създадени с метода дърво на решения (CART), обучени с данни от големи тестови извадки, постигат най-голяма точност в три от пет случая. Големите тестови извадки съдържат над 6000 наблюдения. Точността на FineTree класификаторите, които определят оценката на лечебно заведение приложило лечение на пациенти със злокачествени заболявания локализирани в областта на бронхи и бял дроб (C34), млечна жлеза (C50), кожа (C44), женски полови органи (C51_58) или мъжки полови органи (C60_63) е съответно 99.97%, 99.81%, 99.93%, 99.96%, 99.63%. При големи обучителни извадки с най-малка точност са моделите създадени с Найвен Бейсов класификатор. Точността на тези модели варира в граници от 85.58% до

96.80%. Точността на моделите като цяло е висока, над 85%. FineTree и FineGaussianSVM класификаторите отчитат най-висока точност при предсказване. За всички моделите създадени с ансамбловия метод прогнозната точност е по-малка от обучителната.

- Класификаторите създадени с метода на опорните вектори, обучени с данни от малки извадки (до 2000 наблюдения), регистрират най-голяма точност при три от пет обучителни множества. Обучителната точност на FineGaussianSVM моделите определящи оценката на лечебни заведения приложили лечение на пациенти със злокачествени заболявания локализирани в областта на устна, черен дроб и интрахепаталните жлъчни пътища, панкреас, хранопровод или главния мозък е съответно 99.41%, 100.00%, 89.24%, 93.02%, 79.67%. Четири от класификаторите създадени с алгоритми CART, KNN и SVM отчитат по-малка точност при прилагане върху тестови набор данни в сравнение с точността в процес на обучение.
- Точността на класификаторите създадени с методите дърво на решение, К-най-близък съсед, опорни вектори и ансамблови алгоритми обучени с данни от средни по обем извадки - C16 (злокачествени образования на стомах) и C67 (злокачествени образования на пикочен мехур) е над 97.00%. Най-слабо е представянето на метода Найвен Бейсов класификатор. При тестови набор данни най-малка точност (43.71%, 62.25%) се отчита за класификаторите създадени с ансамбловия метод.
- На база получени резултати от направените експериментални изследвания се достига до извода, че може да бъде създаден обобщен класификатор, който да определи рейтинга на шестдесет и три лечебни заведение, които са приложили лечение на пациента, за пет годишен период от време, със заболявания обозначени със сто и девет МКБ-та.

Създадени и обучени са седем класификатора с пет метода на машинното обучение. Моделите са съпоставени по точност на класификатор, класификационна грешка, време за обучение, Receiver Operating Characteristic криви и точност при предсказване.

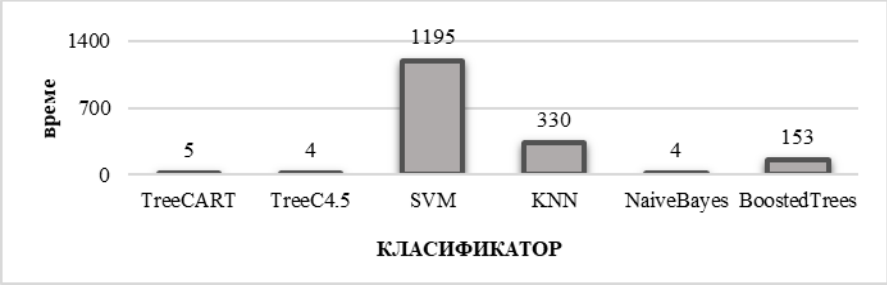
Представени са експериментални резултати.

Таблица 4.15. Точност и класификационна грешка на модел

Метод	Модел	Точност	Класификационна грешка
Дърво на решения	TreeCART	**99.51%**	**0.49%**
	TreeC4.5	99.29%	0.71%
Опорни вектори	SVM	**99.51%**	**0.49%**
К-най-близък съсед	KNN	99.25%	0.75%
Наивен Бейсов класификатор	Naive Bayes	**81.07%**	**18.93%**
Ансамблов	BoostedTrees	95.21%	4.79%
	BaggedTrees	**99.57%**	**0.43%**

Точност над 99.5% и класификационна грешка под 0.5% се регистрира за три от моделите създадени с ансамблов метод, дърво на решение и опорни вектори. Разликата в точността на тези класификатори е 0.06%. С най-малка точност (81.07%) и най-голяма грешка (18.93%) е моделът създаден с метода Наивен Бейсов класификатор. Под 96% точност регистрира класификаторът създаден с алгоритъма AdaBoost.

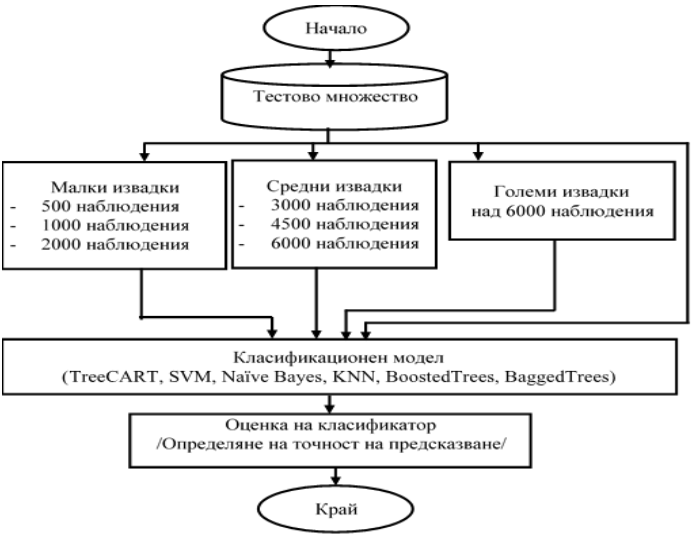
Класификаторите са съпоставени по параметър време за обучение (Фиг. 4.14.).



Фиг. 4.14. Време за обучение на класификационен модел

Времето за обучение на класификатор е от значение при работа с голям обем данни. Моделите TreeCART, TreeC4.5 и NaiveBayes изискват най-малко време. Моделът SVM създаден с метода на опорните вектори изисква много повече време за обучение в сравнение с останалите пет модела.

Моделите са тествани с данни от тестовото множество. Формирани са различни по обем извадки съдържащи различен брой наблюдения с цел изследване на поведението на класификаторите. Алгоритмичната процедура за тестване на класификатор е представена на Фиг. 4.16.



Фиг. 4.16. Тестване на класификатор

Изчислена е усреднена точност на предсказване, на класификатор за малки, средни и големи по обем извадки (Таблица 4.16., Таблица 4.17., Таблица 4.18.).

Таблица 4.16. Усреднена точност на предсказване на класификатори при малки по обем извадки.

Модел	Точност на предсказване / Средна стойност
TreeCART	**97.02%**
SVM	96.31%
KNN	**97.01%**
NaiveBayes	**72.62%**
BoostedTrees	91.67%
BaggedTrees	96.92%

Таблица 4.17. Усреднена точност на предсказване на класификатор при средни по обем извадки.

Модел	Точност на предсказване / Средна стойност
TreeCART	**98.42%**
SVM	**99.09%**
KNN	96.25%
NaiveBayes	**82.51%**
BoostedTrees	94.78%
BaggedTrees	97.94%

Таблица 4.18. Усреднена точност на предсказване на класификатор при голяма по обем извадка.

Модел	Точност на предсказване / Средна стойност
TreeCART	**98.03%**
SVM	**98.63%**
KNN	97.15%
NaiveBayes	**82.44%**
BoostedTrees	94.26%
BaggedTrees	97.62%

Точността на предсказване на класификаторите TreeCART, SVM, KNN и BaggedTrees, при тестване с данни от малки извадки е голяма и варира в граници от 82.07% до 100%. Усредената точност е висока, като за пет от моделите е над 96.00%. Най-голяма точност постигат класификаторите TreeCART и KNN. Два от класификаторите TreeCART и SVM при тестване със средни и големи по обем извадки отчитат над 98% точност на предсказване. Разликата в точността на предсказване при тези два класификатора не е голяма - 0.67% и 0.60%. Експериментални изследвания проведени с класификатор BoostedTrees, използвайки данни от двадесет и девет тестови извадки показват, че само в три от случаите се регистрира точност от над 96%. Тестовете са направени с извадки които съдържат данни само от един клас и обемът им е до 3000 наблюдения, в останалите случаи точността му е от 82.30% до 95.71%. Най-малка точност на предсказване отчита NaiveBayes модела, за всяка от извадките, независимо от обема й. Регистрираната точност на предсказване варира в широки граници от 45.20% до 99.00%.

Оценка на лечебните заведения за определена група МКБ съгласно създадената методика е представена в табличен вид.

Таблица 4.23. Класификация на лечебни заведения за МКБ група С15

МКБ група	Лечебно заведение	Оценка
С15 /хранопровод/	СБАЛ по белодробни болести Св.София ЕАД	**висока**
	МБАЛ ТОКУДА Болница София АД	**висока**
	МБАЛ Царица Йоанна ЕАД, гр. София	**висока**
	СУБАЛ Еврохоспитал - Пловдив	**висока**
	МБАЛ Св. Марина ЕАД, гр. Варна	**ниска**
	МБАЛ Плевен ЕАД	**ниска**
	УМБАЛ Св. Георги ЕАД, гр. Пловдив	**ниска**
	КОЦ - Бургас ЕООД	**ниска**
	МБАЛ Каспела ЕООД, гр. Пловдив	**ниска**
	КОЦ - Русе ЕООД	**ниска**
	ВМА - Министерство на отбраната	**ниска**
	СБАЛОЗ д-р М. Марков ЕООД, гр. Варна	**ниска**
	МБАЛ Проф. Ст. Киркович АД, гр. Стара Загора	**ниска**

Таблица 4.22. Класификация на лечебни заведения за МКБ група С34

МКБ група	Лечебно заведение	Оценка
С34 /бронх и бял дроб/	СБАЛ по белодробни болести Св.София ЕАД	висока
	МБАЛ Св. Марина ЕАД, гр. Варна	висока
	УСБАЛ по онкология ЕАД, гр. София	висока
	ВМА - Министерство на отбраната	висока
	МБАЛ ТОКУДА Болница София АД	висока
	СБПФЗ - Хасково ЕООД	висока
	МБАЛСМ Н.И.Пирогов ЕАД	висока
	МБАЛ Св. Анна АД, гр. Варна	висока
	V МБАЛ - София ЕАД	висока
	УМБАЛ Св. Георги ЕАД, гр. Пловдив	ниска
	МБАЛ Плевен ЕАД	ниска
	МБАЛ Проф. Ст. Киркович АД, гр. Стара Загора	ниска
	КОЦ - Бургас ЕООД	ниска
	КОЦ - В. Търново ЕООД	ниска
	МБАЛ - Пловдив АД	ниска
	ОДПФЗС - Русе ЕООД	ниска
	МБАЛ - Търговище АД	ниска
	УМБАЛ-Стара Загора ЕАД	ниска
	КОЦ - Пловдив ЕООД	ниска

Таблица 4.24. Класификация на лечебни заведения за МКБ група С71

МКБ група	Лечебно заведение	Оценка
С71 /главен мозък/	УМБАЛ Св. Ив. Рилски ЕАД, гр. София	висока
	МБАЛСМ Н.И.Пирогов ЕАД	висока
	МБАЛ ТОКУДА Болница София АД	висока
	ВМА - Министерство на отбраната	висока
	МБАЛ Св. Марина ЕАД, гр. Варна	висока
	МБАЛ Царица Йоанна ЕАД, гр. София	висока
	МБАЛ Плевен ЕАД	висока
	МБАЛ Св. Анна АД, гр. София	висока
	УСБАЛ по онкология ЕАД, гр. София	висока
	УМБАЛ Св. Георги ЕАД, гр. Пловдив	ниска
	МБАЛ Св. Анна АД, гр. Варна	ниска
	МБАЛ - Бургас АД	ниска
	МБАЛ Проф. Ст. Киркович АД, гр. Стара Загора	ниска
	МБАЛ - Хасково АД	ниска

ИЗВОДИ КЪМ ГЛАВА ЧЕТИРИ

- Създадена е методика за класифициране на онкологични лечебни заведения на база преживяемост на пациент и топографски код на заболяване.
- Анализирани и обработени са данните за онкологично болни пациенти и лечебни заведения приложили лечение на регистрирани пациенти съгласно новата методика.
- Създадени и обучени са сто петдесет и шест класификатора за класифициране на онкологични лечебни заведения, за дванадесет съвкупности от данни чрез методи:
  - Дърво на решения;
  - К – най-близък съсед;
  - Ансамблов;
  - Наивен Бейсов;
  - Опорни вектори.
- Оценени са създадените модели чрез метрики за оценка на качество и сравнение на обучени класификатори – точност, прецизност, пълнота, F-мярка, класификационна грешка, време за обучение и точност на предсказване. На база получени резултати са формулирани следните изводи:
  - Класификационните модели създадени с алгоритъм CART, обучени с данни от големи тестови извадки, постигат най-голяма точност в три от пет случая. Точността на класификатора е между 99.53% и 99.97%. Най-висока точност при предсказване за тестови набор данни имат класификаторите TreeCART, GaussianSVM. Точността на предсказване на тези модели е в граници от 99.38% до 100.00%.
  - Класификаторите създадени с метода на опорните вектори, обучени с данни от малки извадки, регистрират най-голяма точност при три от пет обучителни множества (93.02%, 98.44%, 99.41%, 99.67%, 100.00%).
- Създадените седем модела, които класифицират шестдесет и три лечебни заведения обхващащи сто и девет МКБ-та са съпоставени по точност, допусната класификационна грешка, ROC криви, необходимо време за обучение и точност на предсказване. На база получените резултати са формирани следните заключения:
  - Три класификатора - TreeCART, GaussianSVM и BaggedTrees - реализират висока точност над 99.5% и малка класификационна грешка под 0.5%;
  - Моделите TreeCART, TreeC4.5 и NaiveBayes изисква най-малко време за обучение в сравнение с моделите SVM и Bagged Trees.

- Точността на предсказване за малка по-обем тестови извадка е най-висока при класификатори TreeCART. Разлика от 0.01% точност спрямо тези модели се реализира от класификатор KNN (Таблица 4.19). При средни и големи по обем извадки моделът SVM, създаден с метода на опорните вектори (Таблица 4.20 и Таблица 4.21) постига най-голяма точност. Близки стойности за точност при предсказване регистрира и моделът TreeCART.

Изследвайки поведението на класификаторите върху двадесет и девет различни извадки, част от които са случайно генерирани от тестовото множество, а други съдържащи данни само от един и същи клас се формират следните изводи:

- Моделът SVM достига 100%, в случаи, когато данните в извадката са от един и същи клас.
- Моделът TreeCART е устойчив и неговата точност не зависи от обема и съдържанието на извадката.
- Моделът NaiveBayes има най-слабо представяне за всички оценъчни параметри.
- На база на проведени експерименти, получени резултати и направения сравнителен анализ, отчитайки комплексното представяне на всеки един от моделите, за всеки от приложените методи на машинно обучение, за решаване на поставената задача, се отличава модела TreeCART създадени с метода дърво на решения.

ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

НАУЧНО – ПРИЛОЖНИ ПРИНОСИ

1. Създадени са регресионни модели за предсказване стадия на онкологичното заболяване.
2. На база корелационни зависимости са определени най-значимите характеристични описатели за предсказване стадия на онкологично заболяване.
3. Разработена е методика за класифициране на онкологични лечебни заведения.

ПРИЛОЖНИ ПРИНОСИ

1. Реализирани са пет метода за създаване на класификационни и регресионни модели в машинното обучение: Дърво на решения, Наивен Бейсов класификатор, К-най-близък съсед, Опорни вектори и Ансамблов.

2. Проведени са експерименти с реални данни за изследвания на два основни алгоритъма за създаване на дърво на решения - CART и C4.5.

3. Проведени са експериментални изследвания с помощта на софтуер върху данни за регистрирани пациенти с онкологични заболявания и лечебните заведения, които представляват извадка от Националния Раков Регистър.

4. Класификацията на лечебните заведения дава възможност да се направи информиран избор за провеждане на лечение.

ПУБЛИКАЦИИ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

1. **Тодорова** М., Н. Николов. Машинно обучение. Методи за извличане на знания от данни чрез обучение на класификатор. VIII International Scientific Conference "Engineering. Technologies. Education. Safety", 08-11.06.2020 , Borovets, Bulgaria, ISSN 2535-0315, Volume 2, стр. 73-76
2. **Тодорова** М., Н. Калчева, Г. Маринова, Н. Николов. Обзор и класификация на методи и задачи в Data Mining. VIII International Scientific Conference "Engineering. Technologies. Education. Safety", 08-11.06.2020 , Borovets, Bulgaria, ISSN 2535-0315, Volume 2, стр. 77-80
3. **Тодорова** М. Сравнителен анализ на CART, ID 3, C4.5 И C5.0 алгоритми. Предимства и недостатъци. Списание „Компютърни науки и технологии", ISSN 1312-3335, Година XVIII, Брой 1/2020, стр. 111-117
4. **Todorova** M. Application of Machine Learning Methods for Determining the Stage of Cancer. International Conference "Automatics and Informatics 2020" (ICAI 20), ISBN 978-1-7281-9308-3, pp.1-4, DOI: 10.1109/ICAI50593.2020.9311355, Scopus
5. **Todorova** M., G. Marinova, Methods of machine learning in oncology. XXIX International Scientific Conference Electronics - ET2020, September 16 - 18, 2020, Sozopol, Bulgaria, **DOI:** 10.1109/ET50336.2020.9238263, Scopus
6. Маринова Г., М. **Тодорова**. Обзор на техники за извличане на данни в онкологията и психоонкологията", списание „Компютърни науки и технологии", ISSN 1312-3335, Година XVIII, Брой 1/2020, стр. 141-146

Специални благодарности на:

доц. д-р инж. Недялко Николов, доц. д-р инж. Виолета Божикова, доц. д-р инж. Мариана Стоева, гл. ас. д-р инж. Нели Калчева и на всички, които са били съпричастни към моята дисертационна работа.

Annotation

The dissertation focuses on the research of different methods and algorithms for classification and regression and their application in medical oncology. Based on the set goal, two tasks have been formulated and implemented to assist medical specialists in determining the stage of cancer and patients in choosing a medical institution where to receive treatment.

The dissertation research is related to four main aspects such as:

1. Analysis and data processing for malignant disease patients and medical institutions.
2. Creating regression models to predict the stage of cancer.
3. Creating methodology for determining the rating of a medical institution. On the base of developed methodology are trained and tested classifiers that determine the rating of hospitals.
4. Evaluation and comparison of models using selected metrics. Objective assessment of experimental results

Ninety regression models are developed with three machine learning methods, such as Decision Tree, Support Vector Machines, and Ensemble. The models are trained to determine the stage of cancer-based on one predictor variable, three, four, or six using four independent data set. Base on conducted experimental researches and obtained results on the evaluation parameters from the training and test set is concluded that the model Fine Tree is best satisfactory for predicting the stage of malignancy neoplasm. This regression model was created using six predictor variables - Sex, Age Group, ICD, Tumor Size, Lymph Node Status and Metastases.

One hundred and fifty-six classifiers were created according to the methodology developed for determining, the rating of the medical institution by using thirteen data sets. Experimental studies were conducted with five machine learning methods, such as Decision Tree, K - Nearest Neighbor, Ensemble, Naive Bayes Classifier, and Support Vector Machin. Metrics used to quality evaluate and compare the different classifiers are accuracy, precision, recall, F-measure, classification error, training time, Receiver Operating Characteristic curves and prediction accuracy. The influence of data volume is studied over the accuracy of the classifier. The choice of the TreeCART classifier for predicting the rating of the medical institution is justified.