

Пловдивски университет „Паисий Хилендарски“
Факултет по математика и информатика
Катедра „Математически анализ“

Антоанета Петрова Йорданова

**ПРИЛОЖЕНИЕ НА ДЕЙТА МАЙНИНГ МЕТОДИ
ЗА СТАТИСТИЧЕСКО МОДЕЛИРАНЕ**

АВТОРЕФЕРАТ

на дисертационен труд
за присъждане на образователна и научна степен “Доктор”
в област на висше образование 4. Природни науки, математика и информатика
Професионално направление 4.5 Математика
Докторска програма: Математическо моделиране и приложение на математиката

Научен ръководител:

проф. дмн Снежана Георгиева Гочева-Илиева

Пловдив 2021 г.

Дисертационният труд е обсъден и насрочен за защита на разширен катедрен съвет на катедра „Математически анализ” при Факултет по математика и информатика на Пловдивски университет „Паисий Хилендарски”, град Пловдив, проведен на 16.04.2021 г.

Дисертационният труд е с общ обем от 118 страници, в които 26 таблици и 43 фигури, с увод, четири глави и библиография, състояща се от 105 източника. Списъкът на авторските публикации включва 3 заглавия.

Защитата на дисертационния труд ще се състои на 22.06.2021 от ... часа в Заседателна зала на Нова сграда на Пловдивски университет „Паисий Хилендарски” на открито заседание на научно жури в състав:

1. Проф. д-р Михаил Димов Тодоров
2. Проф. д.с.н. Живка Илиева Герговска
3. Доц. д.м.н. Иванка Миткова Желева
4. Доц. д-р Христина Николова Кулина
5. Проф. д.м.н. Снежана Георгиева Гочева-Илиева

Номерацията на формулите, цитиранията, примерите, таблиците и фигурите съвпада с тяхната номерация в дисертационния труд.

Материалите по защитата са на разположение на интересувашите се в секретариата на ФМИ, нова сграда на ПУ, каб. 330, всеки работен ден от 8:30 до 17:00 часа.

Автор на дисертационния труд: Антоанета Петрова Йорданова

Заглавие: Приложение на дейта майнинг методи за статистическо моделиране

Съдържание

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД	4
Актуалност на проблемите.....	4
Цел и задачи на дисертационния труд.....	4
Структура на дисертационния труд.....	5
КРАТКО СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД	6
ГЛАВА 1. Анализ на състоянието на проблема	6
1.1. Дейта майнинг методи за моделиране и изследване на данни за 305 дневната млечност при крави порода Холщайн.....	6
ГЛАВА 2. Моделиране на зависимости в многомерни данни за крави порода Холщайн чрез метода CART	7
2.1. Формулиране на задачата.....	7
2.2. Описание на данните.....	7
2.3. Описателна статистика на многомерните данни.....	8
2.4. Построяване на CART модели на 305-дневната млечност и анализ.....	9
ГЛАВА 3. Моделиране на зависимости в многомерни данни за крави порода Холщайн чрез Random forest	13
3.1. Постановка на задачата.....	13
3.2. Описание на данните.....	13
3.3. Построяване на Random forest модели и анализ.....	14
3.4. Обсъждане на резултатите.....	15
ГЛАВА 4. Моделиране на зависимости в многомерни данни за крави порода Холщайн чрез регресия с главните компоненти и CART- Ensembles and Bagging (CART-Ebag)	16
4.1. Постановка на задачата.....	16
4.2. Описание на използваните данни и променливи.....	16
4.3. Описателна статистика на многомерните данни.....	17
4.4. Резултати от прилагането на регресия с главните компоненти.....	17
4.5. Резултати от прилагане на метода CART-Ebag.....	20
Заклучение	22
Претенции за научни и научно-приложни приноси.....	22
Перспективи за бъдеща работа.....	23
Списък на публикации по дисертационния труд.....	23
Апробация на научните и научно-приложни резултати.....	23
А) Участия в проекти.....	23
Б) Обучение в докторантски програми, курсове и семинари.....	24
В) Доклади , изнесени на научни форуми.....	24
Д Е К Л А Р А Ц И Я за оригиналност.....	25
Благодарности	26
Библиография	26

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблемите

Извличането на знания от данните (data mining) е една от динамично равиващите се области на познанието поради подобряването и внедряването на нови техники и алгоритми от областта на машинното обучение, разпознаването на образи, статистиката, невронните мрежи и визуализацията на данни във все повече области от науката и практиката.

В настоящия дисертационен труд се развиват и прилагат най-съвременни дейта майнинг методи за статистическо моделиране и анализ на многомерни данни от животновъдството, основна област от аграрните науки: класификационни и регресионни дървета (CART), ансамблови методи: CART Ensemble and Bagging (CART EBag) и Случайни гори (RF).

Интензивната практика в съвременната аграрна област се нуждае от по-прецизни и точни методи на изследване, които да осъществяват многостранен и задълбочен анализ на връзките между изследваните признаци и показатели, наложени от години в областта. Общоприето е математическите методи да се схващат като най-разпространен подход за описание и моделиране на обектите и явленията, за обяснение и предсказване на наблюдаваните феномени.

Биологичните обекти, за които най-често се събират данни при експерименти в животновъдството, съществуват чрез съответен жизнен цикъл. Те се характеризират с редица **фенотипни и генотипни показатели**, които представляват интерес за повишаване на продуктивността и биха имали икономическия ефект. **Средата на обитание** и технологията на отглеждане (фермата като фактор) също е предмет на изучаване от аграрните специалисти поради тяхното пряко въздействие върху биологичните единици. Данните, използвани в настоящия дисертационен труд илюстрират горните три аспекта фенотип, генотип и среда.

Цел и задачи на дисертационния труд

Основната цел на настоящия дисертационен труд е да се приложат дейта майнинг с машинно обучение методи за статистическо моделиране и изследване на зависимости в емпирични данни от животновъдството.

Обект на изследването е 305-дневната млечна продуктивност на крави от породата Холщайн-Фрезийска, отглеждани във ферми на територията на България.

Хипотеза на изследването е разкриването на зависимостта на млечността от екстериорни признаци и средовия фактор и определяне на тези от тях, които оказват най-голямо влияние за увеличаване на млечността.

За постигане на поставената цел на дисертационния труд се формулираха следните **задачи**:

1. Да се построят и изследват модели на 305-дневната млечност в зависимост от линейни екстериорни признаци и фермата с метода на класификационните и регресионни дървета CART. Да се установи степента на влияние на факторите в CART моделите, от които зависи млечността и се идентифицират основните от тях.
2. Да се построят и изследват модели на млечността в зависимост от факторите с ансамбловия метод на случайните гори Random Forest (RF). Да се установи влиянието на факторите в моделите.
3. Да се проведе моделиране на зависимостта в многомерни данни чрез факторен анализ и метода на регресия с главните компоненти.
4. Да се построят и изследват модели за 305-дневната млечност за разширена извадка от данни с метода CART Ensemble and Bagging (CART-EBag) с началните

променливи и с факторни променливи. Да се определи приносът на предикторите в CART-EBag моделите.

5. Да се проведе сравнение на резултатите от приложените методи и получените модели и качествата им за предсказване на стойностите на зависимата променлива спрямо реалните данни.

Структура на дисертационния труд

Дисертационния труд е структуриран в увод, 4 глави, заключение, библиография и е с общ обем от 119 печатни страници.

Глава 1 представя основните понятия и методи по избраната тема. Разглеждат се някои проблеми и задачи при изследванията в животновъдната наука и практика по отношение на породата Холщайн. Описани са следните статистически методи за изследване на зависимости и класификации като: линеен регресионен анализ, метод на класификационните и регресионни дървета (CART), факторен анализ и регресия с главните компоненти (PCA), случайни гори (RF), CART Ensemble and Bagging (CART EBag). Тук се прави и кратък преглед на литературата и на методи, използвани за оценка на екстериора на животните.

Глава 2 представя резултатите от приложението на метода CART за обработка и анализ на реални емпирични данни от областта на животновъдството. Построени са два модела за зависимостта на продуктивния признак 305 - дневна млечност от 13 признака: 12 от линеен тип и един средови фактор - фермата на отглеждане на животните.

В **Глава 3** са построени и изследвани два RF модела, на база на наборите от променливи от **Глава 2**. Направено е сравнение на съответстващите модели по следните показатели R^2 , RMSE, MAE, MSE, MAPE. Направени са изводи относно ефективността на двата метода при малки ($n=97$) извадки.

Глава 4 представя в първата си част приложението на метода на регресия с главните компоненти (РГК) за анализ на разширена извадка на данните от Глава 2 за 160 крави от породата Холщайн, като данните за млечността са трансформирани до усреднена млечност по ферми. Получени са 11 главни компонента и е построен линеен регресионен модел, в който извлечените главни компоненти участват като независими променливи. Направена е интерпретация на резултатите от модела.

Във втората част от Глава 4 с извлечените главни компоненти и началните променливи са построени 16 модела с фактора Ферма и без него чрез прилагане на метода CART Ensembles and Bagging. Направен е сравнителен анализ на моделите по следните показатели R^2 , RMSE, MAE, MSE, MAPE и е избран е най-добър модел. Направен е интерпретация на модела.

Заклучението е представено кратко резюме, списък на публикациите по дисертационния труд, апробации и са систематизирани претенциите за научни приноси на докторанта в областта на използването на описаните статистически методи в животновъдната наука.

Библиографията включва 105 източника.

Дисертационният труд е разработен към катедра “Математически анализ” на Факултета по математика и информатика на Пловдивския университет “Паисий Хилендарски” в периода 2017-2021 година.

КРАТКО СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. Анализ на състоянието на проблема

1.1. Дейта майнинг методи за моделиране и изследване на данни за 305 дневната млечност при крави порода Холщайн

При дейта майнинг методите целта е получаването или извличането на полезни знания от данните, откриването на модели на данните, подпомагането на вземане на решения, както се разглежда от едни от първите публикации по темата [37, 26].

Метод на класификационни и регресионни дървета CART

Някои задачи, които могат да намерят решение чрез приложение на CART методите:

- Намаляване на размерността на задачата – брой променливи и/или случаи;
- Избор на съществени предиктори за модела;
- Осигуряване на стабилни модели при малки извадки;
- Класификация в степенувани групи.

Метод на Случайните гори (Random forest)

Сред съществуващите методи за класификация Класификаторът на случайните гори (RF) се оценява като един от най-силните и мощни алгоритми [14, 67].

Има случаи, в които регресионните CART модели може да покажат неустойчивост в предсказването под влияние на много отдалечени от средната стойност измервания, неправилен подбор и несъществени предиктори, предиктори с малки вариации в стойностите, и други. Също така се счита, че е възможно получаването на преопределени модели, когато моделът започва да отчита и промени в грешките, съотнесими с грешката на измерванията [3]. В такива случаи се счита за по-подходящо използването на ансамблови методи като Случайните гори и др. Такъв метод е и методът CART Ensemble and Bagging (CART-EBag), в който моделирането се осъществява с множество дървета (ансамбъл от дървета), в комбинация с Bagging алгоритъм (известен още като bootstrap aggregation, или пакетиране). При него се прилага техника, подобна на Random Forest (RF). Съществуват множество ML (Machine Learning) методи, включващи тези техники. В дисертационния труд се използва алгоритъмът на CART-EBag метода, включен в софтуерния пакет на Salford Predictive Modeler на компанията Minitab [70].

Методът CART-Ensembles and Bagging (CART-EBag)

Методът CART-EBag се прилага за пръв път за статистическо моделиране на данни от областта на аграрните науки.

Оценка на моделите [77]

Оценката за качеството на моделите ще бъде извършено с помощта на коефициента на детерминация R^2 , средна абсолютна грешка (MAD), средната абсолютна процентна грешка (MAPE), средната абсолютна грешка (MSE) и средната квадратична грешка (RMSE) по формулите:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1.5)$$

Тук и по-надолу y_i са значенията на променлива y , $\hat{y}_i$ са предсказаните ѝ стойности от модела, n е брой на наблюденията в извадката.

$$MAD = \frac{\sum_{i=1}^n |y_i - \bar{y}|}{n} \quad (1.6)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (1.7)$$

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (1.8)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (1.9)$$

ГЛАВА 2. Моделиране на зависимости в многомерни данни за крави порода Холщайн чрез метода CART

В тази глава се прилага методът Класификационни и регресионни дървета (CART) за статистическо моделиране на изследваните данни за крави от порода Холщайн в България. Представя се обработката на данни от ординален и номинален тип, които представляват екстериорни и продуктивни признаци за извадка от породата.

2.1. Формулиране на задачата

Основен приоритет при изследване на зависимостите между характеристиките на млечните породи крави е установяването на онези признаци, които влияят на тяхната продуктивност с най-голяма сила. В този смисъл обект на изследването е зависимостта между продуктивния признак 305 дневна млечност от линейните признаци, като се наблегне на задачата да се определят онези, които имат най-голямо влияние.

По този начин се цели намаляване на броя необходими линейни признаци за предсказването на стойността на млякото при кравите и намаляването на разходите за измервания по екстериора на кравите.

Формулирани са следните конкретни изследователски задачи:

- 1) Статистически анализ с приложение на метода CART за избор на адекватни модели.
- 2) Диагностика на грешките, анализ и оценка на построените модели.
- 3) Приложение на построените модели за класификация и прогноза.

Изследването е проведено със статистическия софтуер SPSS [71].

2.2. Описание на данните

В изследването, представено в тази глава, се използват данни от лактации на 97 Холщайн крави, събрани в периода 2013-2014 от четири ферми, разположени в различни региони в България. Измерени и оценени са 12 признака от линеен тип според инструкциите на Международния Комитет по Регистрация в Животновъдството (ICAR) за преценка по тип и екстериор от 2014 г. [40], както и продуктивен признак – 305- дневна млечност, който се разглежда като зависим от екстериорните признаци. В извадката са обхванати животни от първа до трета лактация.

Описанието на използваните променливи в изследването са дадени в таблица 2.1. В четвърта колона на таблица 2.1. са представени характеристиките, които отговарят на първата и последната стойност от скалата за всеки признак. Обикновено първата отговаря на нежелана характеристика на признака, а последната на добро или отлично състояние.

Таблица 2. 1. Описание на променливите, използвани в анализа

Променлива	Име на променливата	Тип	Мярка
305-дневна млечност	MilkY305	Scale	Kg
Височина	Stature	Ordinal	1-9 Ниска - Висока
Ширина на гърдите	ChestW	Ordinal	1-9 Тясна - Широка
Ширина на крупата	RumpW	Ordinal	1-9 Тясна - Широка
Задни крака (отзад)	RLRV	Ordinal	1-9 X образни -Прави
Задни крака (отстрани)	RLSW	Ordinal	1-5 трансформирана Слонова - Съблевидни
Скакателни стави	HockD	Ordinal	1-9 С отоци -Плоски

Развитие на костите	Bone	Ordinal	1-9 Дебели- Тънки
Ъгъл на копито	FootA	Ordinal	1-5, трансформирана Малък - Голям
Дълбочина на копито	FootD	Ordinal	1-9 Плитко - Дълбоко
Ширина на вимето	UdderW	Ordinal	1-9 Тясно - Широко
Походка	Locomotion	Ordinal	1-9 Лоша - Нормална
Куцота	Lameness	Ordinal	1-3 Ходи неравномерно - Много куца
Ферма	FarmN	Nominal	Номер на ферма: 1, 2, 3, 4

Трансформация е приложена за линейните признаци Ъгъл на копитото и Задни крака (отстрани) тъй като при тях желаната стойност е средната, 5 или 6. Маринов установява, че кравите с идеална постановка на задните крака – поглед отстрани с описание 5 или 6, и Ъгъл на копитото 45°, съответващо на ниво 5, имат отчетена най-висока млечност за нормална лактация, около 8000 кг [103]. При кравите с описания 1 или 9 млечността е по-ниска.

2.3. Описателна статистика на многомерните данни

Описателна статистика на използваните променливи:

305 Day Milk yield – 305 дневна млечност, измерва се в килограми, означава се в млечното животновъдство още като млечна продуктивност, 305 дневен млеконадой или мляко. При измерването на признака 305 дневна млечност има особености, тъй като той е приравнен млеконадой към 305 дни. Разпределението е близко до нормалното.

Коефициент на вариация

$$V\% = \frac{SD}{\bar{x}} * 100 = 33.5\% > 33\% ,$$

което води до извода за нееднородност на извадката. В изследванията на животновъдната наука коефициентът на вариация се тълкува и като възможност за промяна в изследвания признак. Или признакът 305 дневна млечност за тази извадка показва, че една добра селекция, може да доведе до промяна от 33 %.

Стойности на описателните статистика за линейните признаци - независимите променливи са в таблица 2.3. Направените тестове за нормалност за разпределенията на независимите променливи Колмогоров-Смирнов дават като резултат, че не всички от тях са нормално разпределени.

Таблица 2. 2. Описателна статистика за линейните признаци - независимите променливи за Глава 2, N=97

	Mean	Std. Error of Mean	Median	Std. Deviation	Variance	Skewness	Kurtosis	Range
RLRV	4,96	0,113	5	1,117	1,248	0,586	1,341	6
RLSW	4,04	0,104	4	1,02	1,04	-0,746	-0,356	4
FootA	4,47	0,064	5	0,631	0,398	-0,792	-0,364	2
FootD	6,49	0,113	7	1,11	1,232	-0,384	-1,342	3
UdderW	6,32	0,162	6	1,598	2,553	-0,415	0,203	7
Locomotion	5,55	0,122	6	1,199	1,438	-0,167	-0,359	5
Lameness	1,51	0,061	1	0,597	0,357	0,729	-0,416	2
Stature	5,27	0,155	5	1,524	2,323	-0,088	-0,317	8
ChestW	7	0,136	7	1,339	1,792	-0,293	-0,554	5
RumpW	6,24	0,119	6	1,171	1,37	0,24	-0,769	5
HockD	5,33	0,154	5	1,512	2,286	-0,008	0,021	8
Bone	6,67	0,122	7	1,205	1,453	-0,064	-0,469	5

2.4. Построяване на CART модели на 305-дневната млечност и анализ

2.4.1. CART Модел 1

Методът CART се приложи за зависимата променлива MilkY305 и 12 екстериорни признака от ординален тип. Поставени са контролни параметри за дървото: минимален брой случаи в родителски възел на дървото -20, минимален брой случаи в краен възел на дървото – 10. Зависимата променлива е MilkY305. Получено е дърво с дълбочина 3, 9 възела от които 5 крайни.

Полученото дърво на решенията е показано на Фигура 2.5. В началото всички 97 случая (наблюдения) се поставят в коренния възел 0 на дървото. Средната им стойност е 7583 кг за 305 - дневна млечност. Тези случаи са разделени по отношение на първата променлива UdderW (ширина на вимето) на две групи: тези, които имат стойности на UdderW по-малка или равна на 7,5, и останалите със стойности за UdderW по-големи от 7,5. В левия възел 1 са класифицирани 75 случая с прогнозна (т.е. средна) стойност за MilkY305 - 6968 кг.

В десния възел 2 са класирани 22 случая със стойности на UdderW по-високи от 7,5 и очакваната им стойност за MilkY305 е 9680 кг. Възелът 2 е терминален (краен). Той не се разделя повече, няма наследници поради избраните ограничения на модела. За всяка разделяща променлива в дървото на решенията се изчислява степента на важност по отношение на нейния принос за подобряване на точността на модела.

Следващата процедура на разделяне продължава от възел 1. Сега е установено, че по-важна променлива е ChestW (ширина на гърдите) със съответния праг от 6.5. Информацията за двата наследника, възли 3 и 4, е изчислена и представена, както в първия етап на алгоритъма. Чрез повтаряне на тези стъпки се получава дървото на решенията, показано на фигура 2.5.

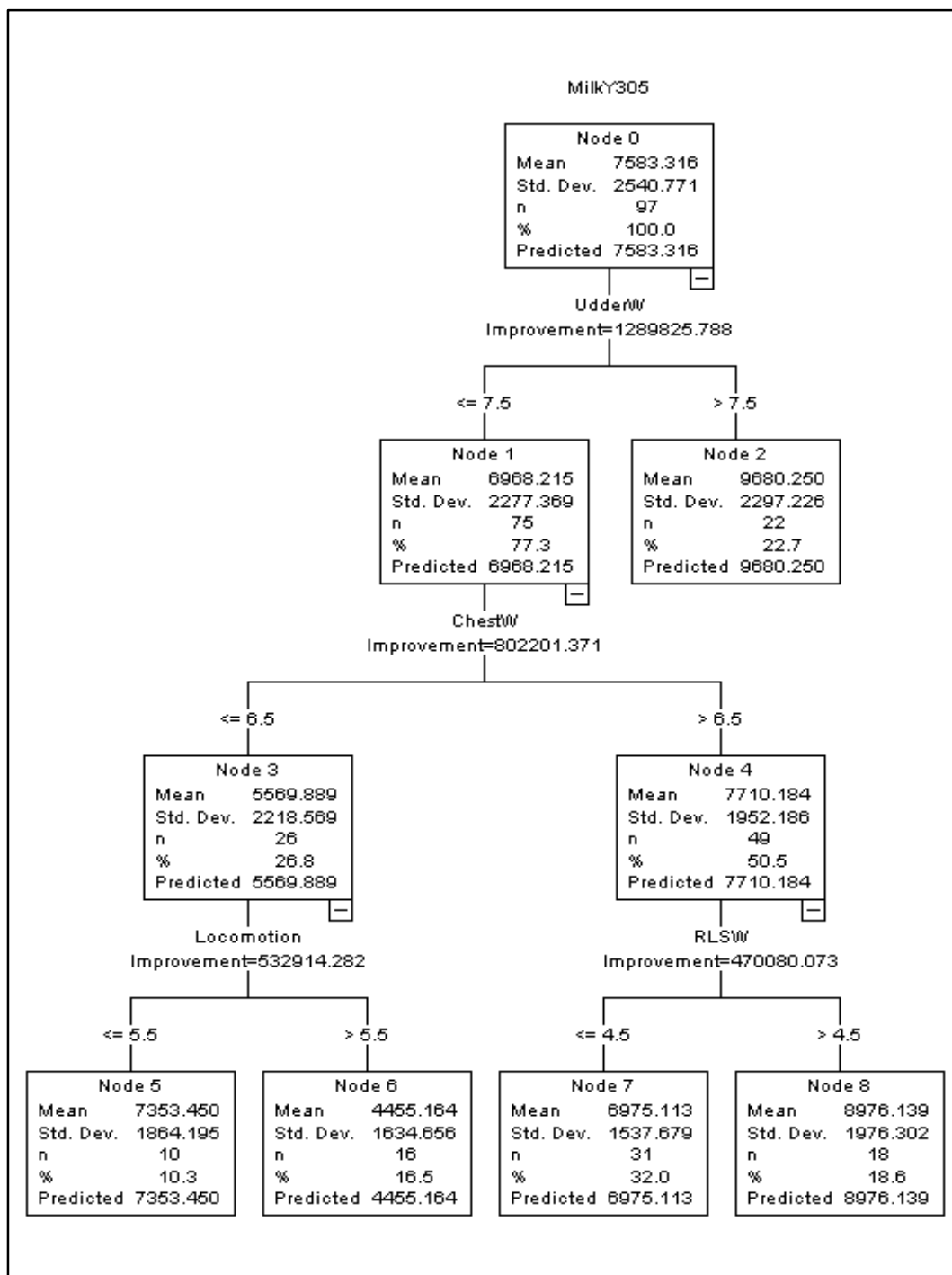
Дървото на решенията може да се използва за класификация и прогнозиране на нови данни, като се следват правилата докато се класифицират новите случаи в някой от терминалните възли на дървото.

Таблица 2. 5. Описание на правилата и предсказани стойности за пет крайни възела от дървото на решения за CART Модел 1.

Крайни възли	Брой на случаите	Правила	Предсказана стойност за MilkY305-d, кг
Възел 2	22	UdderW>7.5	9680
Възел 5	10	UdderW ≤ 7.5, ChestW ≤ 6.5, Locomotion ≤ 5.5	7353
Възел 6	16	UdderW ≤ 7.5, ChestW ≤ 6.5, Locomotion > 5.5	4455
Възел 7	31	UdderW>7.5, ChestW>6.5, RLSW ≤ 4.5	6975
Възел 8	18	UdderW>7.5, ChestW>6.5, RLSW>4.5	8976

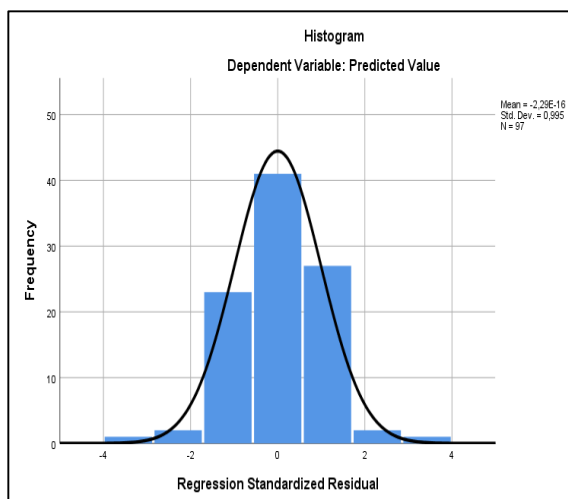
Подробно приносът на линейните фактори върху млечността за 305 дни е следния: с най-голямо значение има признакът ширина на вимето (считан за 100%), а останалите са нормализирани спрямо него, следват походка 48,1%, височина 41,2%, ширина на гърди 39%, куцота 27,6%, задни крака изглед отстрани 22,9% и т.н.

Дървото на решенията може да се разглежда като поредица от правила, определени от неравенства за разделяне. Пълното описание на правилата за получените крайни възли на дървото е дадено в таблица 2.5.

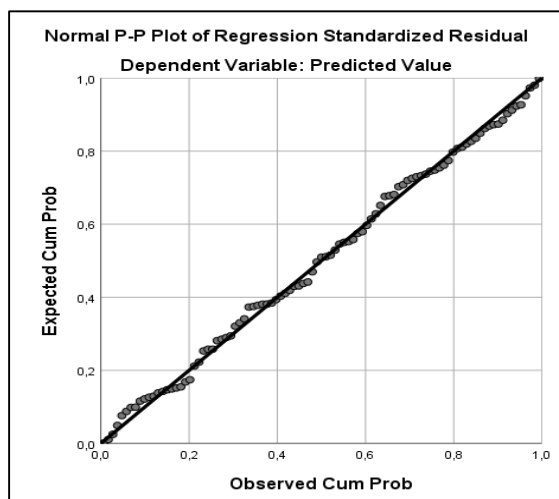


Фигура 2. 5. Дърво на решенията, получено за променливата MilkY305 (305 - дневна млечност) и 12 екстериорни признака – CART Модел 1.

Диагностика на остатъците



Фигура 2. 7. Хистограма на остатъците за CART Модел 1



Фигура 2. 8. P-P графика на остатъците CART Модел 1

Анализът на разпределението на грешките показва, че то е близко до нормалното, което се потвърждава нагледно и от хистограмата на фигура 2.7 (средната стойност на остатъците е 0, а стандартното отклонение $s=0.995$, много близко до 1) и от P-P графиката на остатъците от фигура 2.8 (разположението на стойностите е S-образно около 45° права).

2.4.2. CART Модел 2

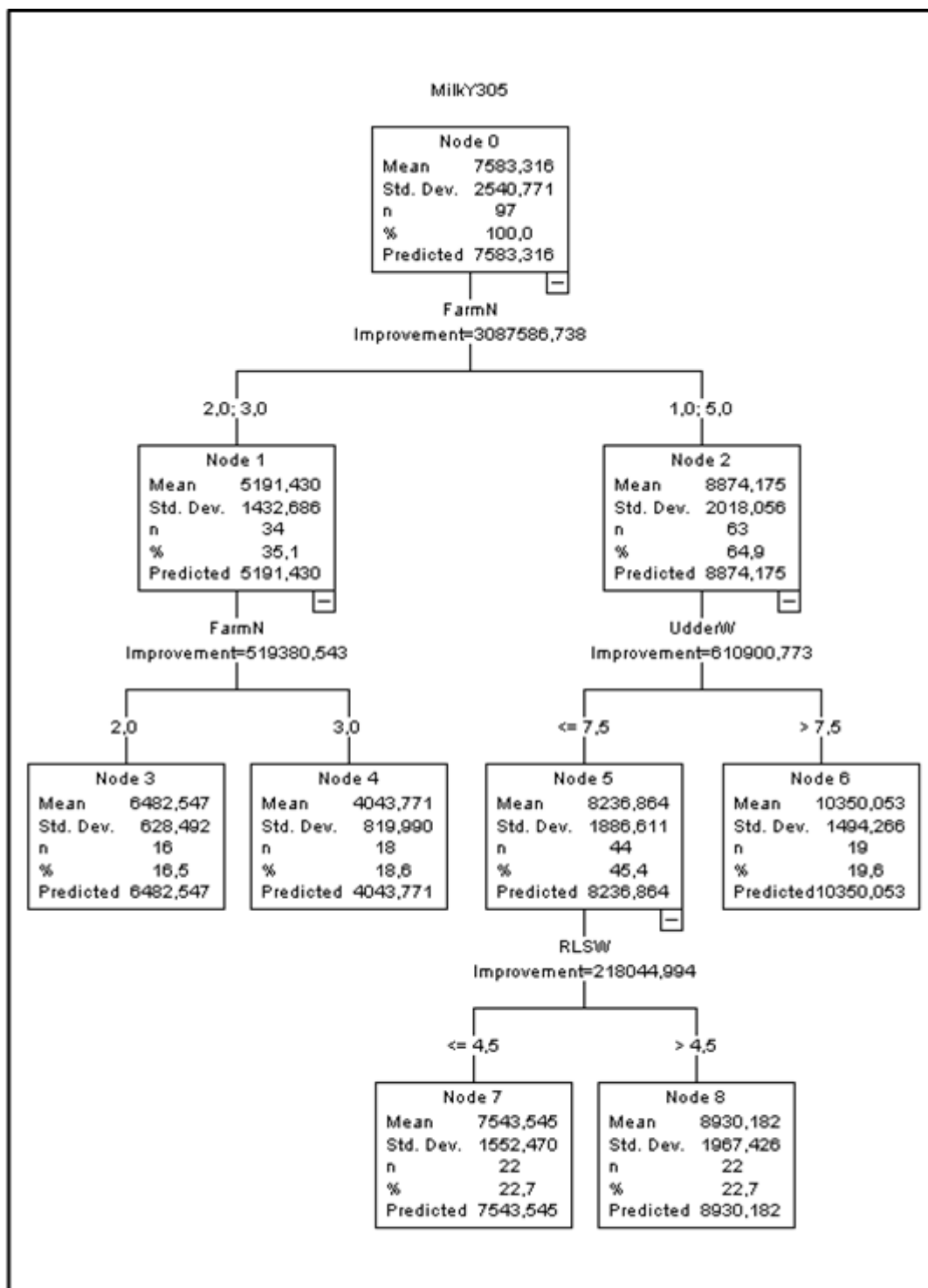
За да се анализира значението и влиянието на средовите фактори, методът CART се прилага отново при същите фактори, като включва допълнителна променливата Ферма (FarmN). Поставени са контролни параметри за дървото: Минимален брой случаи в родителски възел-20, минимален брой случаи в краен възел – 10 и 10 степенна кросвалидация. Полученото дървото на решенията е с 9 възела, 5 от които са крайни (фигура 2.9).

При спазване на резултатите по отношение на разпределението на данните в петте крайни възела се вижда, че наблюденията от фермите 1 и 5, класифицирани във възлите 2 и 8, имат най-високи стойности за 305-дневната млечност. Крайните възли 6 и 8 класифицират случаите с най-големите средни стойности на променливата MilkY305, съответно 10350 и 8930 кг. Всички случаи в терминален възел 6 и половината от възел 8 са от фермата 1. Другите две ферми показват по-ниски резултати

Проведеният анализ разкрива значението на фермата като най-съществен фактор, който е свързан и с условията за хранене и отглеждане на добитъка.

Изграденият CART Модел 2 има коефициент на детерминация $R^2 = 0,701$ и описва 70% от данните.

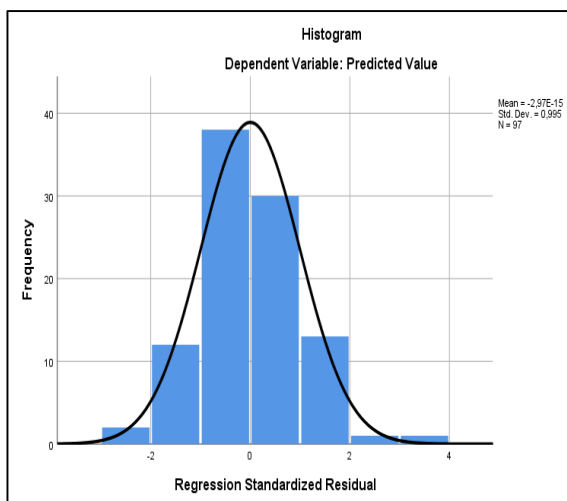
За CART Модел 2 относителният принос на факторите, влияещи на 305-дневната млечност на изследваната извадка, са: фермата Farm 100%, ширина на виме 49,8%, ширина на гърди 35,7%, скакателни стави 15,4%, развитие на кости 13,9% и т.н. Подредбата на факторите в CART Модел 2 е различна от тази на CART Модел 1.



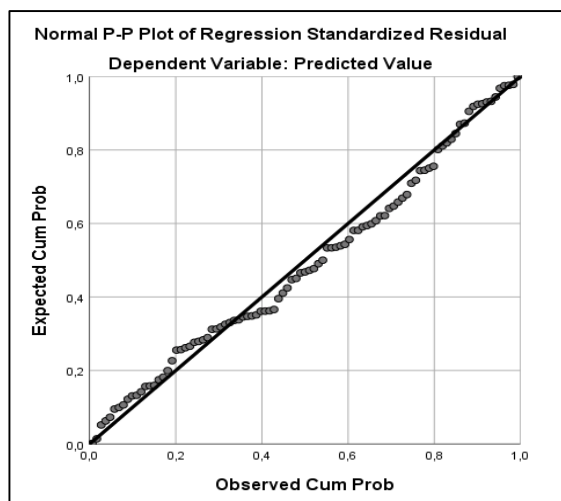
Фигура 2. 1. Дърво на решенията за CART Модел 2

Анализ на разпределението на грешките от CART Модел 2.

Разпределението на грешките в CART Модел 2 е нормално с ниво на значимост 0.2 (тест на Kolmogorov-Smirnov). Заключение се потвърждава и от представена на фигура 2.10 хистограма и P-P графиката на остатъците е изобразена на фигура 2.11.



Фигура 2.10. Хистограма на остатъците на CART Модел 2



Фигура 2.11. P-P графика на остатъците за CART Модел 2

Получените модели, описващи важните линейни признаци, позволяват създаване на прогнози за млечната продуктивност на базата на получените правила за крайните възли в моделите. Направено е тълкуване и обяснение на получените резултати.

Прилагането на метода CART изяснява насоките за селекция по отделни линейни признаци и може да помогне за възможно развитие на индексите за селекция в зависимост от тяхното значение.

Резултатите от Глава 2 са публикувани в [P1].

ГЛАВА 3. Моделиране на зависимости в многомерни данни за крави порода Холщайн чрез Random forest

3.1. Постановка на задачата

Целите на настоящата глава са:

1. Прилагане на RF модел за линейни характеристики на крави Холщайн - фризийска порода и
2. Сравнение на RF и CART модели за един и същ набор от данни.

Приложението на RF модели не е особено популярно в областта на животновъдните науки, където изследванията са посветени на фактори с голямо влияние върху продуктивността. Важно е да се построи модел за предсказване на 305 дневната млечност на базата на лесни за поддръжка признаци като линейните, които са част от външния вид на животното. В предишната глава са представени и обсъдени CART модели за изследване на връзката на същите предиктори с 305 дневната млечност. Затова и втората цел на формулираната задача е да се сравнят тези модели с новите RF модели, получени за същото множество данни.

3.2. Описание на данните

В главата се представят модели по алгоритъма RF за набора от данни, който се използва и в Глава 2.

За данните е приложена същата система от раздел 2.2 за кодиране в скала от 1 до 9 за групата признаци в развитие и трансформация в скала от 1 до 5 за признаците с два противоположни недостатъка (ъгъл на копито и задни крайници, изглед отстриани).

Използваните в настоящото изследване линейни екстериорни признаци и означението на съответните им променливи са представени в Таблица 2.1.

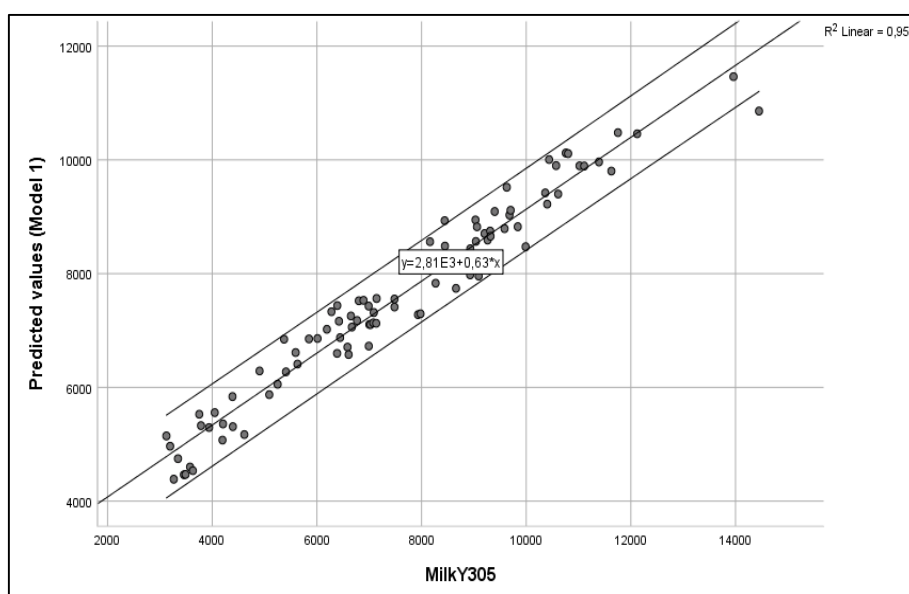
3.3. Построяване на Random forest модели и анализ

С алгоритъма RF са построени два нелинейни регресионни модела със зависима променлива – 305 дневна млечност *MilkY305*. Моделите са генерирани с помощта на Salford Predictive Modeler [70].

Както се посочва от [20] проблемът с колинеарността на предикторите може да доведе да отклонения в тълкуваните резултати. Корелационната матрица на разглежданите тук 13 предиктора показва силна обратно пропорционална корелационна зависимост между походката *Locomotion* и куцотата *Lameness*. Корелационният коефициент на Spearman's rho = - 0.908 с ниво на значимост 0.01. По тази причина изключваме променливата *Lameness* от множеството на предикторите.

3.3.1. RF Модел 1

В първия модел (RF Модел 1) като предиктори са включени само линейни характеристики на животните, а именно: *Stature*, *ChestW*, *RumpW*, *UdderW*, *Bone*, *HockD*, *RLRV*, *RLSW*, *FootA*, *FootD*, *Locomotion*. Построен е модел с 500 случайни дървета, при случаен избор на 3 от 11 предиктора за всяко дърво и минимум 2 случая в родителски възел.



Фигура 3.2. Сравнение на стойностите на зависимата променлива с предсказанияте от RF Модел 1 стойности с 5% доверителен интервал

Моделът описва 95% от вариацията на зависимата променлива с RMSE = 995.01. В построения RF Модел 1 най-съществена роля за млекодобива играе ширината на вимето (считано за 100%). Нормализираното влияние на останалите предиктори спрямо него са както следва: ширина на гърди (50.86 %), скакателни стави (25.49 %), походка (17.54 %), следват задни крака, поглед от страни и степента на развитието на костите.

Разпределението на грешките в RF Модел 1 е нормално с ниво на значимост K-S

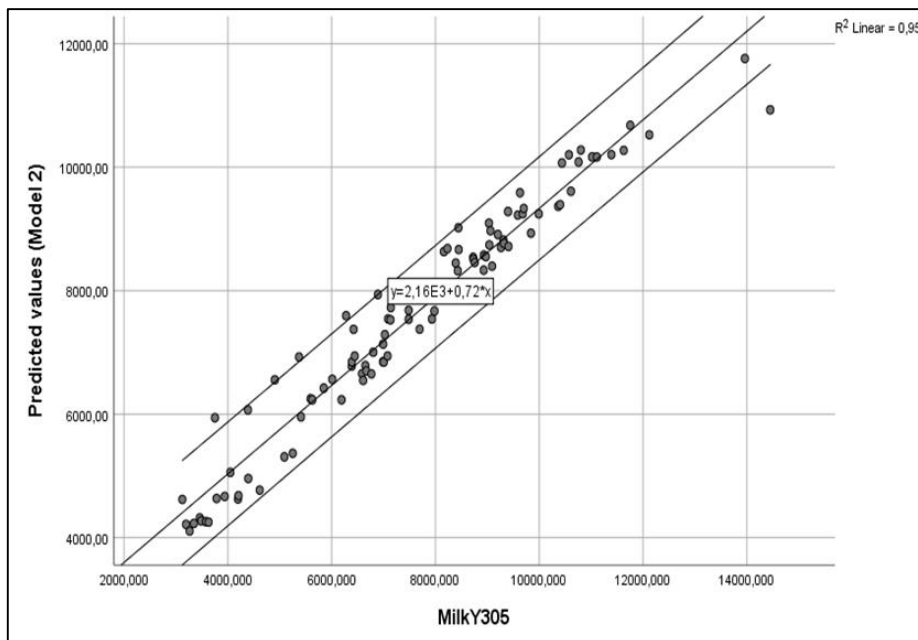
Sig.=0,848 (тест на Kolmogorov-Smirnov). Моделът е статистически значим при $p < 0,0001$.

На Фигура 3.2. са сравнени стойностите на зависимата променлива с предсказанияте от модела стойности с 5% доверителен интервал.

3.3.2. RF Модел 2

Вторият RF модел (RF Модел 2) е построен, като към предикторите е добавена и фермата, в която се отглеждат животните. Това е свързано с местоположението, честота на доене, технология на отглеждане и др. За модела са зададени същите контролни параметри

- 500 случайни дървета, при случаен избор на 3 от 12 предиктора и минимум 2 случая в родителски възел. RF Модел 2 описва 95% от вариацията на зависимата променлива с $RMSE = 825.97$. В този модел фермата е с най-голямо влияние върху млекодобива (считано за 100%), следвана отново от ширина на виме (16.44%), ширина на гърди (9.26%), скакателни стави (4.42%) и походка (3.33%).



Фигура 3. 7. Сравнение на стойностите на зависимата променлива с предсказанияте от RF Модел 2 стойности с 5% доверителен интервал

Моделът е статистически значим при $p < 0.0001$.

Разпределението на грешките в RF Модел 2 е нормално с ниво на значимост 0.2 (тест на Kolmogorov-Smirnov). На Фигура 3.7 са сравнени стойностите на зависимата променлива с предсказанияте от Модел 2 стойности с 5% доверителен интервал.

3.4.Обсъждане на резултатите

В Таблица 3.1 са представени стойностите на предсказващите статистики на двата модела, които се изчисляват по формулите, зададени в раздел 1.2 по-горе с номера от (1.5) до (1.9). Коефициентите на детерминация за двата модела са много близки, но RF Модел 2 има по-ниска стойност на $RMSE=825.97$ кг.

Таблица 3. 1. Прогнозни статистики на получените RF Модел 1 и RF Модел 2

Модел	RMSE	MSE	MAD	MAPE	R ²
RF Модел 1	995.013	990050.881	808.458	0.130	0.953
RF Модел 2	825.97	682237.999	617.663	0.09718	0.951

$RMSE$ =Root mean squared error; R^2 =Coefficient of determination;

Получените два модела в тази глава са сравнени с CART моделите от Глава 2 и е установено значителното повишаване на коефициента на детерминация с метода RF, а именно – от 48 % на 95 % за моделите без участие на фермата като предиктор, и от 70% на 95% за моделите с участието на фермата като предиктор. Може да се отбележи, че сред линейните признаци на животните, ширината на вимето се нарежда на първо място във всички модели, ширината на гърдите участва на второ място (с изключение на CART Модел

1). В RF моделите се отчита и състоянието на скакателните стави при животните, докато тази променлива не участва в CART моделите. Наблюдава се, че във всички модели се отчита влиянието на походката, а височината на животните е отбелязана само в моделите с CART. Посочените сравнения са обобщени в Таблица 3.3. Получените CART модели са сравнително по-груби и могат да се подобрят чрез рафиниране на задаваните контролни параметри.

Таблица 3. 3. Сравнение на RF и CART моделите

	RF Модел 1	CART Модел 1	RF Модел 2	CART Модел 2
R²	0.95	0.48	0.95	0.70
Относителен принос на предикторите	UdderW ChestW HockD Locomotion	UdderW Locomotion Stature ChestW	Farm UdderW ChestW HockD Locomotion	Farm UdderW ChestW Locomotion Stature

Наблюдава се еднаква подредба по важност на предикторите от RF Модел 1 и след включването на новия предиктор Фермата в RF Модел 2. Резултатите показват стабилност на получените модели.

Дори за малки извадки (в случая 97 записа) както и за качество на данните (ординални за предикторите) са приложими както методите на CART, така и RF. Силата на двата метода за малки извадки е потвърдена. Използваният софтуер Salford Predictive Modeler работи достатъчно добре и с малки извадки от данни.

Увеличението на коефициента на детерминация R^2 при RF моделите, както и същия ред на предикторите по влиянието им в сравнение с CART моделите показва по-голямата ефективност на RF метода и стабилност на моделите.

Резултатите от Глава 3 са публикувани в [P3].

ГЛАВА 4. Моделиране на зависимости в многомерни данни за крави порода Холщайн чрез регресия с главните компоненти и CART- Ensembles and Bagging (CART-Ebag)

4.1. Постановка на задачата

Целите на настоящата глава са:

1. Прилагане на РГК за моделиране на зависимостта на 305-дневната млечност на крави от породата Холщайн в зависимост от линейните екстериорните признаци на база на извадка от 160 наблюдения;
2. Прилагане на ансамблов метод CART Ensembles and Bagging за моделиране на зависимостта с вариране на броя на дърветата с началните променливи и с факторни променливи;
3. Сравняване на получените модели и избор на най-добър CART-Ebag модел.

4.2. Описание на използваните данни и променливи

В изследването са използвани данни от проведени измервания на $N=160$ Холщайн-Фризийска порода крави в периода 2013-2014 г. за продуктивния признак – млечност за 305 дни и 12 линейни признака на екстериора, които са описани в Таблица 2.1, Глава 2.

По данните за всяка от четирите ферми е пресметната средната стойност за известните данни на продуктивния признак 305 - дневна млечност. Въз основа на възходящо сортиране е съставена зависимата рангова променлива ordFerma със стойности от 1 до 4.

4.3. Описателна статистика на многомерните данни

В Таблица 4.1. са дадени резултатите от описателната статистика на изходните данни с добавената рангова променлива ordFerma. Тъй като стойностите на асиметрията (Sk) и ексцеса (Ku) за всички променливи са близки до 0, то разпределението е близко до нормалното.

Таблица 4. 1. Описателна статистика за независимите променливи за извадката в Глава 4, N=160.

	Mean	Std. Error of Mean	Median	Mode	Std. Deviation	Variance	Skewness	Kurtosis	Range
Stature	4,67	0,144	5	5	1,818	3,305	-0,215	-0,54	8
ChestW	6,51	0,13	7	6	1,644	2,704	-0,359	-0,543	6
RumpW	6,09	0,093	6	5	1,173	1,376	0,254	-0,672	5
RLRW	4,93	0,114	5	5	1,447	2,095	0,044	0,272	7
RLSW	3,96	0,08	4	5	1,009	1,017	-0,594	-0,62	4
HockD	5,29	0,116	5	5	1,469	2,158	0,033	-0,143	8
Bone	6,14	0,12	6	7	1,519	2,308	-0,268	-0,451	7
FootA	4,51	0,047	5	5	0,593	0,352	-0,756	-0,392	2
FootD	6,41	0,087	7	7	1,101	1,212	-0,304	-1,441	3
UdderW	5,72	0,152	6	6	1,92	3,688	-0,364	-0,489	8
Locom	5,33	0,105	5	5	1,33	1,77	-0,11	-0,511	6
Lameness	1,64	0,053	2	1	0,667	0,445	0,554	-0,701	2

4.4. Резултати от прилагането на регресия с главните компоненти

За обработка на данните са използвани Метод на главните компоненти и Регресия с главните компоненти, описани в Глава 1, раздел 1.2. Регресията е приложена за генерираните главни компоненти с варианта стъпкова регресия. Този метод е подходящ, тъй като бе установено, че е налице мултиколинеарност между изследваните променливи и директното прилагане на линейна регресия е затруднено [20].

От изчислената корелационната матрица за всички 12 независими променливи. се установява мултиколинеарност и условия за прилагане на МГК както следва:

- детерминанта 0,020;
- тест на Kaiser–Meyer–Olkin: $0,658 > 0,5$ и тест на Bartlett за сферичност със значимост $< 0,0001$.

Това означава, че провеждането на МКГ е адекватно [41].

Също така бяха установени следните значими корелационни коефициенти между зависимата променлива ordFerma и признаците, съответно: с променливата UdderW (0,665); със Stature (0,498); с ChestW (0,407); с Bone (0,345).

С помощта на SPSS е приложен факторен анализ с МГК и са извлечени 12 главни компонента. Получените собствени стойности на корелационната матрица с разпределение на общата дисперсия са показани на **Таблица 4.2.**

Получената матрица L с факторните тегла (loadings), след завъртане с метода Варимакс е показана в Таблица 4.3. За прегледност теглата под 0,3 не са показани. Забелязва се групиране на куцота и походка в един компонент, като всичките главни компоненти са много добре отделени. За 12-тия компонент няма отделна променлива. След пренебрегване на 12-тия компонент, описващ едва 0,731% от общата дисперсия на данните анализът се повтаря с извличане на 11 фактора (компонента).

Таблица 4. 1. Разпределение на общата дисперсия след прилагане на МГК. 4.2.

Total Variance Explained			
Component Компоненти	Initial Eigenvalues Начални собствени стойности		
	Total	% of Variance	Cumulative %
1	3,072	25,600	25,600
2	1,938	16,146	41,746
3	1,298	10,814	52,560
4	1,114	9,285	61,845
5	1,040	8,666	70,511
6	0,854	7,121	77,632
7	0,712	5,936	83,568
8	0,675	5,624	89,191
9	0,515	4,290	93,482
10	0,386	3,213	96,694
11	0,309	2,575	99,269
12	0,088	0,731	100,000

След процедурата на завъртане на компонентите по метода Варимакс се получават 11 ортогонални помежду им компоненти, означени с $pc1$, ..., $pc11$. Те обясняват 99,269% от общата дисперсия на извадката. Запомняме извлечените латентни променливи за провеждане на регресионен анализ.

Таблица 4. 3. Завъртяна матрица по метода Варимакс с 12 главни компонента

Rotated Component Matrix												
	Component											
	$pc1$	$pc2$	$pc3$	$pc4$	$pc5$	$pc6$	$pc7$	$pc8$	$pc9$	$pc10$	$pc11$	$pc12$
Походка	0,975											
Куцота	-0,974											
ШКк		0,94										
ШГк			0,947									
ВШк				0,919								
Задни крака отзад					0,983							
ДКк						0,965						
З.крака отстрани							0,992					
Ъгъл на копито								0,989				
Скак.става									0,993			
РКк										0,88		
Вк											0,85	

При извличане на 12 главни компоненти в завъртяната матрица (Таблица 4.3) се забелязва групиране на куцота и походка в един компонент, като всичките главни компоненти са много добре отделени. За 12-тия компонент няма отделна променлива.

Като се пренебрегне този 12-ти компонент, описващ съгласно Таблица 4.2 много малка част от общата дисперсия на данните (едва 0,721%), анализът се повтаря с извличане на 11 фактора (компонента). След процедурата на завъртане на компонентите по метода

Варимакс се получават 11 ортогонални помежду им компоненти, чиито тегла са показани на Таблица 4.4.

Означаваме компонентите с $pc1, \dots, pc11$. Те обясняват 99,269% от общата дисперсия на извадката, по Таблица 4.4.

Таблица 4. 2. Матрица на 11 завъртени компоненти, получени по метода Варимакс.

Променлива	Компоненти										
	$pc1$	$pc2$	$pc3$	$pc4$	$pc5$	$pc6$	$pc7$	$pc8$	$pc9$	$pc10$	$pc11$
Locom	0,98	-0,03	-0,01	0,01	0,01	0,04	0,05	0,01	-0,01	0,01	-0,01
Lam	-0,97	0,00	0,00	-0,01	-0,02	-0,01	0,00	0,02	0,06	-0,03	-0,03
RumpW	-0,04	0,94	0,12	0,12	0,03	0,14	-0,03	-0,02	-0,02	0,16	0,19
ChestW	-0,01	0,11	0,95	0,16	0,03	0,10	0,01	0,00	-0,03	0,16	0,16
UdderW	0,01	0,13	0,18	0,92	0,13	0,01	0,09	0,00	-0,04	0,18	0,23
RLRW	0,03	0,03	0,03	0,10	0,98	-0,03	0,06	0,01	-0,03	0,11	0,06
FootD	0,04	0,13	0,10	0,01	-0,03	0,96	-0,03	0,13	0,01	0,14	0,05
RLSW	0,04	-0,02	0,01	0,07	0,06	-0,03	0,99	0,02	0,03	-0,04	0,05
FootA	-0,01	-0,02	0,00	0,00	0,00	0,12	0,02	0,99	-0,06	-0,04	0,05
HockD	-0,06	-0,01	-0,03	-0,04	-0,02	0,01	0,03	-0,06	0,99	-0,05	0,00
Bone	0,05	0,20	0,19	0,20	0,15	0,18	-0,05	-0,05	-0,08	0,88	0,19
Stature	0,03	0,27	0,22	0,29	0,09	0,06	0,08	0,08	0,01	0,21	0,85

Запомняме извлечените латентни променливи (11-те главни компонента) за провеждане на регресионен анализ.

Построяване на регресионния модел по метода РГК

За установяване на зависимост между $ordFerma$ и получените латентни променливи $pc1, \dots, pc11$ провеждаме многомерен регресионен анализ чрез стъпкова регресия. Полученият модел има нестандартизиран вид:

$$\hat{ordFerma} = 0,657pc4 + 0,350pc11 + 0,328pc3 + 0,224pc10 - 0,190pc6 + 2,369 \quad (4.1)$$

Стандартизираният вид е:

$$\hat{zordFerma} = 0,555zpc4 + 0,296zpc11 + 0,277zpc3 + 0,189zpc10 - 0,160zpc6 \quad (4.2)$$

Моделът и коефициентите му са статистически значими при ниво 0,01.

Полученият модел има корелационен коефициент $R = 0,730$ и коефициент на детерминация $R^2 = 0,533$, т.е. моделът обяснява 53% от извадката. Това е удовлетворително като се има предвид, че усреднената 305-дневна млечност за всяка ферма зависи и от други фактори, като възраст на кравата, генетичен произход, здравни характеристики, условия на хранене, отглеждане и други.

От уравнение (4.1) се вижда, че най-голямо влияние върху усреднената млечност ($ordFerma$) имат 5 латентни променливи: $pc4$, $pc11$, $pc3$, $pc10$ и $pc6$. Те съответстват на изходните екстериорни характеристики: ширина на вимето, височина на кравата, ширина на гърди, развити кости и дълбочина на копитото. Уравнение (4.2) показва относителното им влияние в модела, най-същественият фактор е ширината на вимето (0,555), след това височината на кравата (0,296), ширината на гърдите (0,277) и т.н.

Остатъците са нормално разпределени, т.е. допуснатите от модела грешки са случайни и нямат систематичен характер. Моделът е статистически значим при ниво 0,01.

4.5. Резултати от прилагане на метода CART-Ebag

За данните, описани в точка 4.3, чрез софтуера Salford Predictive Modeler 7.8 е приложен методът CART-Ebag, описан в Глава 1, раздел 1.2. със следните базови настройки: ограничения за броя случаи в родителския и наследствения възел 10/5, Кросвалидация - 10, брой дървета 10, 15, 20 и 25. Получени са 16 модела, от които 8 са с включен фактор Ферма и 8 без Ферма. Всяка от тези групи е разделена по отношение на използването на Метода главните компоненти, извлечени в раздел 4.5.2.

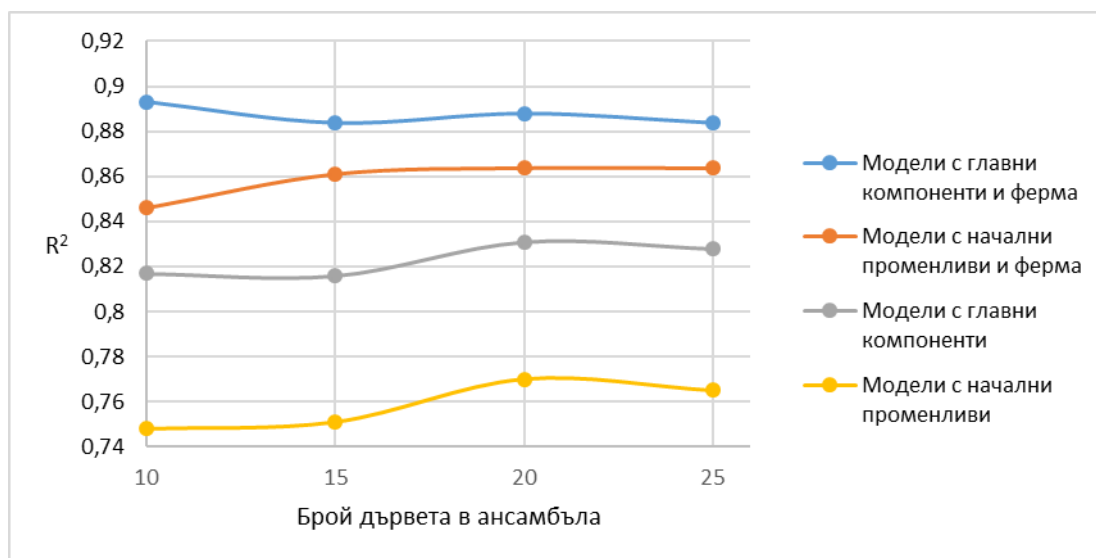
Моделите са оценени и сравнени на база получените от SPM резултати за следните показатели за оценка на грешките: RMSE, MSE, MAD, описани в раздел 1.2.

Построени са точкови графики чрез SPSS 26 и линейни графики за сравняване на R^2 на всички модели заедно, както и по групи чрез MS Excel.

Получени са и е направен анализ на остатъците по следната методика:

- Хистограма на стандартизираните чрез регресия остатъци;
- Р-Р графика на стандартизираните остатъци на модела
- Графика за оценка на хетероскедастичността

Графиката на фигура 4.8 показва значително над 10 пункта подобрене на моделите с главни компоненти по отношение на моделите с начални променливи. При моделите с главни компоненти с ферма и без ферма най-добрите резултати са при ансамбъл от 10 дървета и след това се влошават. При моделите без главни компоненти най-добрият резултат е при 20 дървета без ферма и при ансамбъл от 25 с ферма. Най- висок коефициент на детерминация $R^2=0,894$ има при модела с главни компоненти и с ферма CART-Ebag10PCF, който е получен при ансамбъл с 10 дървета (таблица 4.13)



Фигура 4. 8. Сравнение на R^2 за четирите групи модели за различните ансамбли от дървета – 10, 15, 20 и 25

Таблица 4.13. Сравнение на най-добрите модели от четирите групи

Модел	CART-Ebag20	CART-EbagE10PC	CART-Ebag25F	CART-Ebag10PCF
R^2	0,770	0,818	0,864	0,894
RMSE	1240.32	1157,41	915.98	816.01

При моделите CART-Ebag20 и CART-Ebag25F, които са без главни компоненти се запазва подредбата дори при добавянето на фермата като фактор (таблица 4.14). Аналогично и при моделите с главни компоненти се запазва подредбата на извлечените фактори с влияние. Това показва стабилност на моделите получени с този метод.

Таблица 4. 14. Принос на предикторите за избраните модели

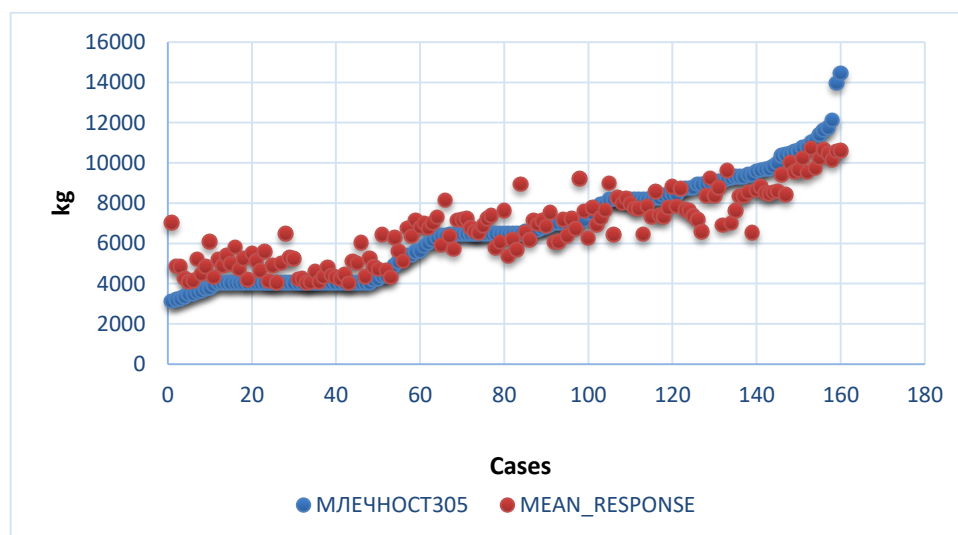
CART-Ebag20	CART-EbagE10PC	CART-Ebag25F	CART-Ebag10PCF
	Ферма	Ферма	
UDDERW	FAC4_2- (UdderW)	UDDERW	FAC4_2- (UdderW)
CHESTW	FAC3_2 (ChestW)	CHESTW	FAC3_2 (ChestW)
STATURE	FAC5_2 (RLRW)	STATURE	FAC5_2 (RLRW)
BONE	FAC1_2	BONE	FAC1_2

Резултатите за модел CART-Ebag10PCF

В избрания модел CART-Ebag10PCF най-съществена роля за млекодобива има ФЕРМАТА (считано за 100%), ширината на вимето, ширина на гърдите, задни крака отзад. (Таблица 4.14). Прогнозните статистики за модела са описани в таблица 4.15.

Таблица 4. 15. Прогнозна статистика за модел CART-Ebag10PCF.

	Name	Score Result
RMSE		816,01045
MSE		665 873,05732
MAD		496,37122
MAPE		0,07397
R-Sq		0,89368



Фигура 4. 12. Наблюдаваните и предсказани стойности за модел CART-Ebag10PCF с

$$R^2=0,894$$

Изводи за CART- EBag моделирането:

1. Дава по-добри резултати при комбиниране с метода на главните компоненти.
2. Може да се приложи и ако зависимата променлива не е с разпределение близко до нормалното.
3. Може да се приложи при предиктори изцяло от ординален тип.
4. Не се предсказват добре отдалечените случаи, в частност най-високите стойности на зависимата променлива, което е общ недостатък на ансамбловите усредняващи предсказванията методи.
5. За модели с над 25 дървета статистиките започват да се влошават и затова не е целесъобразно броят на дърветата да се увеличава.
6. Прилагането му се осъществява по-лесно чрез SPM, отколкото РГК в SPSS.

Част от резултатите от Глава 4 са публикувани в [P2].

Заклучение

Главният извод, който можем да направим е, че целта на настоящия дисертационен труд и поставените задачи са постигнати. Основната част от получените резултати са публикувана в 3 научни публикации. Изнесени са 4 доклада на научни форуми и семинари.

Претенции за научни и научно-приложни приноси

В дисертационния труд са получени и анализирани голям брой статистически модели за изследване зависимостта на 305-дневната млечност на крави от породата Фрезийска-Холщайн от екстериорните признаци и фермата. Двата основни приноса от гледна точка на математическото моделиране и приложение на математиката са:

- За пръв път за реални данни от животновъдството в България за моделирането им са приложени дейта майнинг методи с машинно обучение от типа на CART, Random Forest и CART Ensembles and Bagging.
- Методът CART Ensembles and Bagging се прилага за първи път в литературата в областта на животновъдните науки.

Постигнатите конкретни научни и научно-приложни резултати могат да се систематизират както следва:

1. С помощта на метода CART са построени регресионни дървета на зависимостта на 305 дневната млечност от 12 линейни признаци и фермата като средови фактор при крави тип Фрезийска-Холщайн от 4 ферми на територията на България ($n=97$). Като основни фактори за по-голяма млечност са идентифицирани фермата (условия на средата на отглеждане, здравни грижи и други) и някои линейни признаци. Моделите достигат съвпадение с измерените данни до $R^2=70\%$.

2. Построени и анализирани са RF модели за същата извадка данни, изследвани с CART моделите. Потвърдени са резултатите, като е достигнато обяснение на реалните данни с коефициент на детерминация R^2 до 95% и средноквадратична грешка $RMSE=825.97$ кг.

3. За друга извадка данни ($n=160$) е въведена рангова променлива за 305-дневната млечност. Построени са модели с помощта на регресия с главните компоненти с използване на 12 екстериорни признака. Достигнат е $R^2=77\%$.

4. С приложение на ансамбловия дейта майнинг метод CART Ensembles and Bagging е проведено моделиране на ранговата променлива за 305-дневната млечност в зависимост от 12 екстериорни признаци и фермата. Оценено е влиянието на всички изследвани фактори за обяснение на млечността. Най-добрите модели от тази група обясняват реалните наблюдения с $R^2=0,894$ и точност $RMSE=816,01$ кг.

5. Проведено е сравнение на получените резултати от моделите. Установено е, че най-добри предсказващи статистики има методът RF, следван от CART Ensemble and Bagging и CART. Всички модели са кросвалидирани и е проведен анализ на остатъците им. Използваните методи работят добре и с малки извадки и ординален тип данни.

В следващата таблица е показана връзката между основните получени резултати, поставената цел и задачи, приносите и публикациите на дисертационния труд.

Таблица. Връзка на основните резултати с целите и публикациите.

Принос	Цел	Задача	Параграф	Публикации
1	1	1	2.1. - 2.4.	[P1]
2	1	2	3.1. – 3.3	[P3]
3	1	3	6.1. – 4.5	[P2]
4	1	4	4.1 – 4.4, 4.6	подадена
5	1	5	2.5, 3.4, 4.7	[P1], [P3]

Перспективи за бъдеща работа

Разработените методи и подходи могат да се прилагат в следващи изследвания, които могат да бъдат насочени към:

1. Моделиране на 305-дневната млечност с по-голяма извадка, за всички 24 екстериорни признака.
2. Моделиране за откриване на зависимости между 305 - дневната млечност и други признаци (репродуктивни, комфорт, поведение и т.н.).
3. Разширяване обхвата на методите за изследване на продуктивните и репродуктивни характеристики за Холщайн-Фризийската порода с невключени в настоящата работа методи (напр. SEM, нелинейно моделиране).
4. Прилагане на дейта майнинг методите за моделиране на данни от други области на животновъдните науки, в селскостопанските науки и други.

Списък на публикации по дисертационния труд

- [P1] Yordanova A., Gocheva-Ilieva S., Kulina H., Yordanova L., Marinov I., Classification and regression tree analysis in modeling the milk yield and conformation traits for Holstein cows in Bulgaria, Agricultural science and technology, volume 7(2), pp. 208-213, 2015. ISSN 1313-8820. <http://agriscitech.eu/classification-and-regression-tree-analysis-in-modeling-the-milk-yield-and-conformation-traits-for-holstein-cows-in-bulgaria/>
- [P2] Кулина Х., Йорданова А., Моделиране на зависимости в многомерни данни за Холщайн-Фризийска порода крави чрез регресия с главните компоненти. Сб. доклади на Осма национална конференция „Образованието и изследванията в информационното общество“, Пловдив, 28-29 май, 2015, с. 180-186. https://www.adis.org/ERIS_conference/2015/sbornik-ERIS.2015.pdf
- [P3] Yordanova A., Hristina Kulina H., Random Forest Models of 305-Days Milk Yield for Holstein Cows in Bulgaria AIP Conference Proceedings 2302, 060020 (2020); <https://doi.org/10.1063/5.0034778>

Апробация на научните и научно-приложни резултати

А) Участия в проекти

Част от резултатите са апробирани и използвани в следните университетски проекти:

1. Научен проект МУ-17-ФМИ-003/2018/, Млади учени и докторанти, тема: „Дейта майнинг моделиране и приложни софтуерни системи“ с ръководител проф. д.м.н. Снежана Гочева-Илиева, Фонд „Научни изследвания“ при ПУ „Паисий Хилендарски“, (за периода 25.04.2017/30.11.2018 г.)
2. Научен проект МУ19-ФМИ-010, Млади учени и докторанти, тема: “Иновативни ИКТ модели, прогнози и симулации в областта на физиката, екологията, икономиката, здравеопазването и образованието” с ръководител доц. д-р Христина Кулина, Фонд „Научни изследвания“ при ПУ „Паисий Хилендарски“.
3. Научен проект МУ21-ФМИ-015, Млади учени и докторанти, тема: “Високоэффективни предсказващи алгоритми с машинно обучение за моделиране на данни от екологията, индустрията и образованието” с ръководител доц. д-р Христина Кулина, Фонд „Научни изследвания“ при ПУ „Паисий Хилендарски“.
4. 5ОУП/2015 г. – „Технологична и информационна инфраструктура за електронно и дистанционно обучение в Тракийски университет“, 2015-2016, Тракийски университет, Стара Загора
5. 2018 г.- „Изграждане на Център за Иновативни Образователни Технологии“, 2018-2019 г., Тракийски университет, Стара Загора

Б) Обучение в докторантски програми, курсове и семинари

1. Европейска програма „Учене през целия живот“, Програма Еразъм+ - студентска мобилност (докторантска), University of L'Aquila, Италия, 2019 с продължителност 5 месеца, 26.02.2019 – 29.07.2019 г.
2. Курс „Introduction into Biostatistics with R“, Stara Zagora, Trakia University, University of Southampton, June 05-08, 2018.
3. Обучение за „Европейски сертификат за компютърни умения (ECDL Start Certificate)/Тракийски университет;
4. Посетени са 7 курса към Докторантско училище – Тракийски университет

В) Доклади , изнесени на научни форуми

Част от резултатите, получени в дисертационния труд са докладвани на следните национални и международни конференции, научни сесии и семинари:

- [D1] Участие в „Юбилейна научна конференция 20 години Тракийски университет“, 19-20 май 2015 г., Стара Загора,
 - [D2] Осма национална конференция „Образованието и изследванията в информационното общество“, Пловдив, 28-29 май, 2015
 - [D3] Семинар по Математическо моделиране и приложения, 22 юни 2020 г., <http://www.fmi-plovdiv.org/index.jsp?ln=1&id=3567>
 - [D4] Twelfth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences, Albena, Bulgaria, June 24-29, 2020 <http://2020.eac4amitans.eu/2.html>
-

Д Е К Л А Р А Ц И Я за оригиналност
по чл. 36, ал. 1(8) от ПРАСПУ

от Антоанета Петрова Йорданова

редовен докторант към катедра „Математически анализ”
на Факултет по математика и информатика
при Пловдивския университет „Паисий Хилендарски“

Във връзка с провеждането на процедура за придобиване на образователната и научна степен „доктор” в Пловдивския университет „Паисий Хилендарски” и защита на представения от мен дисертационен труд,
декларирам:

Резултатите и приносите на проведеното дисертационно изследване, представени в дисертационния ми труд на тема „Приложение на дейта майнинг методи за статистическо моделиране“ са оригинални.

ДЕКЛАРАТОР :.....
/Антоанета Йорданова/

09.04.2021
гр. Пловдив

Благодарности

Изказвам благодарност на научния си ръководител проф. дмн Снежана Гочева-Илиева за въвеждане в интересната област на изследване на зависимости в емпирични данни със съвременен софтуер и оказаната помощ при разработката на дисертационния труд. Също така съм благодарна и на доц. д-р Христина Кулина за съвместната работа при апробирането на резултатите.

На проф. дсн Живка Гергowska от катедра “Животновъдство - преживни животни и млекарство”, д-р Ивайло Маринов, доц. д-р Тончо Пенев. за предоставените данни и консултациите по предметната област на дисертационния труд. На проф. д-р Лина Йорданова от катедра „Информатика и математика“ на Тракийски университет - Стара Загора за постоянна ангажираност към моята работа и за всеотдайната подкрепа.

Благодаря на ръководството на Факултета по математика и информатика, в лицето на Декана проф. д-р Антон Илиев, за съдействието, съпричастността и административната подкрепа при обучението ми във ФМИ.

Благодаря на колегите от ФМИ за ценните съвети и взаимната подкрепа в процеса на обучение за периода.

Благодаря на семейството ми за разбирането и помощта.

Библиография

- [1] Almeida T., E. Kern, D. Daltro, J. Braccini Neto, C. McManus, A. Thaler Neto and J. Cobuci. 2017. Genetic associations between reproductive and linear-type traits of Holstein cows in Brazil. *Revista Brasileira de Zootecnia*. 2017 r., Vol. 46, 2, pp. 91–98.
- [2] Amit, Y. and Geman, D. 1997. Shape quantization and recognition with randomized trees. *Neural Computation*. 1997 r., Vol. 9, pp. 1545-1588.
- [3] Apté, C. and Weiss, S. 1997. Data mining with decision trees and decision rules. *Future Generation Computer Systems*. 1997 r., Vol. 13 (2-3), pp. 197–210.
- [4] Baker, Z. K., V. K. Prasanna. 2005. Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs. 2005 r.
- [5] Bauer, E. and Kohavi, R. 1999. An Empirical Comparison of Voting Classification Algorithms. *Machine Learning*. 36, No. 1/2, 1999 r., pp. 105-139.
- [6] Bermingham M., Berry D. and Cromie A. 2006. Change in the conformation of Irish HolsteinFriesian dairy cows over the past decade. *Moorepark Production Research Centre*. Fermoy, Co. Cork, Ireland: The Irish Cattle Breeding Federation, 2006.
- [7] Berry D. P., Buckley F., Dillon P., Evans R. D. and Veerkamp R. F. 2004. Genetic relationships among linear type traits, milk yield, bodyweight, fertility and somatic cell count in primiparous dairy cows. 2004, Vol. 43 (2), pp. 161-176.
- [8] Berry DP, Buckley F, Dillon P, Evans RD, Rath M and Veerkamp RF. 2003. Genetic Relationships among body condition score, body weight, milk yield, and fertility in dairy cows. *Journal of Dairy Sciences*. 2003, Vol. 86, pp. 2193-2204.
- [9] Biggs D., B. de Ville, E. Suen. 1991. A method of choosing multiway partitions for classification and decision trees. 1991, Vol. 18, pp. 49-62.
- [10] Bouška J, Vacek M, Stipkova M, Nemcova E and Pytloun P. 2006. The relationship between conformations of dams and daughters in Czech Holsteins. *Czech Journal of Animal Sciences*. 2006, pp. 236-240.
- [11] Breiman L., Friedman J. H., Olshen R. A. and Stone C.J. 1984. *Classification and regression trees*. Monterey, CA, US: Wadsworth and Brooks, 1984.
- [12] Breiman, L. 1996a. Bagging Predictors. *Machine Learning*. 1996a, Vol. 26, 2, pp. 123-140.
- [13] Breiman, L. 1996b. Out-of-bag estimation. [Онлайн] 1996b

- [14] Breiman, L. 2001. Random forest. Machine Learning. 2001, Vol. 45, pp. 5-32.
- [15] Breiman, L. 1998. Randomizing Outputs To Increase Prediction Accuracy. неизв. : Technical Report 518, Statistics Department, UCB, 1998.
- [16] Breiman, L. 2000b. Some infinity theory for predictor ensembles. Technical Report 579, Statistics Dept. UCB, 2000b.
- [17] Breiman, L. 1999. Using adaptive bagging to debias regressions, Technical Report 547, Statistics Dept. UCB. 1999 r.
- [18] Breiman, L., Kutler, A. Random Forest (description and programing codes in Fortran) . [Онлайн] https://web.archive.org/web/20080622230434/http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm .
- [19] Cak B., Keskin S. and Yilmaz O. Asian Journal of Animal and Veterinary Advances. 2013. Regression tree analysis for determining of affecting factors to lactation milk yield in brown Swiss cattle. 2013 r., Vol. 8, 6, pp. 677-682.
- [20] Carsten T., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Marquéz, B. Gruber, B. Lafourcade, P. Leitão, T. Münkemüller, C. McClean, P. Osborne, B. Reineking, B. Schröder, A. Skidmore, D. Zurell and S. Lautenbach. 2012. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2012 , Vol. 36, pp. 27–46.
- [21] Celik S. 2019. Comparing predictive performances of tree-based data mining algorithms and MARS algorithm in the prediction of live body weight from body traits in Pakistan goats. Pakistan Journal of Zoology. 2019, Vol. 51, 4, pp. 1447–1456.
- [22] Dallago G., D. M. de Figueiredo, P. C. de Resende Andrade, R. A. dos Santos, R. Lacroix, D. Santschic and D. Lefebvre. 2019. Predicting first test day milk yield of dairy heifers. Computers and Electronics in Agriculture. 2019, Vol. 166, pp. 105032.
- [23] Dietterich, T. 1998. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization. Machine Learning. 1998, Vol. 1-22.
- [24] Eydurán E., Yilmaz I., Tariq M. M. and Kaygisiz A. 2013. Estimation of 305-D milk yield using regression tree method in brown Swiss cattle. The Journal of Animal and Plant Sciences. 2013, Vol. 23, 3, pp. 731-735.
- [25] Eydurán, E., A. E. Karahan. C. Tirink, O. Mayis, M. Türkoğlu, M. Masood. 2017. Comparison of predictive performances of MARS and CART algorithms through R software. 2017 <https://www.researchgate.net/publication/321493039> .
- [26] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. From Data Mining To Knowledge Discovery in Databases. AAAI Press/ The MIT press, Massachusetts Institute of Technology, 1996. ISBN 0-262 56097-6 .
- [27] Field, A. 2005. Factor analysis using SPSS, Chapter 15. Discovering statistics using SPSS. Research Method II. London: Sage, 2005.
- [28] France, J., & Thornley, J. M. 1984. Mathematical Models in Agriculture. неизв. : Butterworth&Co Ltd., 1984.
- [29] Gergovska Z., T. Miteva, T. Angelova, D. Yordanova and J. Mitev. 2012a. Relation of milking temperament and milk yield in Holstein and Brown Swiss cows. Bulgarian Journal of Agricultural Science. 2012a , Vol. 18, pp. 771-777.
- [30] Gergovska, Zh., Y. Mitev, T. Angelova, D. Yordanova, Zh. Krastanov and K. Uzunova. 2012. Effect of body condition score at calving on 305-day and test-day milk yield in Holstein-Friesian and Brown Swiss cows. J. Fac. Vet. Med. 2012 , Vol. 38.
- [31] Gocheva-Ilieva S, H Kulina, A Ivanov. 2021. Assessment of Students' Achievements and Competencies in Mathematics Using CART and CART Ensembles and Bagging with Combined Model Improvement by MARS. 2021, 9 (1), 62.

- [32] Grove, A. and Schuurmans, D. 1998. Boosting in the limit: Maximizing the margin of learned ensembles. 1998
- [33] Grzesiak, W. , D. Zaborski. 2012. Examples of the Use of Data Mining Methods in Animal Breeding. intechopen.com. 2012 <https://www.intechopen.com/books/data-mining-applications-in-engineering-and-medicine/examples-of-the-use-of-data-mining-method>
- [34] Grzesiak, W., J. Wojcik And B. Binerowska. 2003. Prediction of 305-day first lactation milk yield in cows with selected regression models. Arch. Tierz. 2003, Vol. 46, 3, pp. 215-226.
- [35] Grzesiak, W., Lacroix, R., Wójcik, J. and Blaszczyk, P. 2003. A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records. Canadian Journal of Animal Science. 2003, Vol. 83(2), pp. 307-310.
- [36] Harvey, W.R. 1987. Mixed model least-squares and maximum likelihood computer program. User`s guide for LSMMW PC-1 version. Columbus: Ohio State Univ., 1987
- [37] Heikki, Mannila. 1996. Data mining: machine learning, statistics, and databases. IEEE, 1996.
- [38] Ho, T.K. 14–16 August 1995. Random decision forests. 14–16 August 1995, pp. 278–282.
- [39] Ho, T.K. 1998. The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence. 1998, Vol. 20, 8, pp. 832-844.
- [40] ICAR. 2014. International Agreement of Recording Practices. Conformation recording of dairy cattle. 2014 <http://aberdeenangus.ro/wp-content/uploads/2014/03/ICAR.pdf> .
- [41] Izenman, A. 2008. Modern multivariate statistical techniques. Springer, 2008.
- [42] Jing, He. 2009. Advances in Data Mining: History and Future. 2009
- [43] Jolliffe I. T., , 2nd edn. 2002. Principal component analysis. 2nd. New York: Springer, 2002.
- [44] Karacaören, B., Kadarmideen, H. N. 2008. Principal component and clustering analysis of functional traits in Swiss dairy cattle. 2008, Vol. 32, 3, pp. 163-171.
- [45] Karamizadeh, S. , Abdullah, Sh. M., Manaf, A. A., Zamani, M., Hooman, A. 2013. An Overview of Principal Component Analysis. Journal of Signal and Information Processing. 2013, Vol. 4, pp. 173-175.
- [46] Kass, G. V. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data, 1980, Vol. 29, 2, pp. 119-127.
- [47] Kern, E. L., Cobuci J. A., Costa, C.N. and Pimentel, C. M. 2014. Factor analysis of linear type traits and their relation with longevity in Brazilian Holstein cattle. Asian Australasian Journal of Animal Sciences. 2014, Vol. 27, pp. 784-790.
- [48] Krupa E, Wolfová M, Peškovičová D, Huba J and Krupová Z. 2005. Economic values of traits for Slovakian Pied cattle under different marketing strategies. Czech Journal of Animal Science. 2005, Vol. 60, pp. 483-492.
- [49] Liaw, A. and M. Wiener. 2002. Classification and Regression by randomForest. 2002, Vol. 2, 3, pp. 18-22.
- [50] MacHugh, D. E., Loftus, R. T., Cunningham, P., Bradley, D. G. 1998. Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. 1998, Vol. 29, pp. 333-340.
- [51] Marinov I., Zh. Gergovska. 2015a. Effect of age at first calving on conformation traits in Black-and-White cows. 2015a, Vol. 7, 3, pp. 303 - 308.
- [52] Marinov, I. 2014. Selection for linear traits for legs and feet and its significance for dairy cattle breeding. 2014, Vol. 6, 2, pp. 115 - 123.
- [53] Marinov, I., T. Penev and Zh. Gergovska. 2015b. Factors affecting linear type traits in black-and-white cows. International Journal of Current Microbiology and Applied Sciences. 2015b, Vol. 4, pp. 374–383.

- [54] Mark, T. 2004. Applied genetic evaluations for production and functional traits in dairy cattle pp. 2004, Vol. 87, pp. 2641-2652.
- [55] Miglior, F. , B.L. Muir, B.J. Van Doormaal. 2005. Selection indices in Holstein cattle of various countries. 2005 , Vol. 88, pp. 1255-1263.
- [56] Murphy, M. D., M. J. O'Mahony, L. Shalloo , P. French, and J. Upton. 2014. Comparison of modeling techniques for milk-production forecasting. J. Dairy Sci. 2014, Vol. 97, pp. 3352–3363.
- [57] Nemcova E., Stipkova M and Zavadilova L. 2011. Genetic parameters for linear type traits. Czech Journal of Animal Science. 2011, pp. 157–162.
- [58] Nguyen Q., R. Fouchereau, E. Frénod, C. Gerard and V. Sincholle. 2020. Comparison of forecast models of production of dairy cows combining animal and diet parameters. Computers and Electronics in Agriculture. 2020, Vol. 170, pp. 105258.
- [59] Ostling, A. 2016. Can methods of machine learning be used to better predict lactation curves for bovines? . Uppsala University: Department of Statistics, 2016.
- [60] Pant, S. D., Schenkel, F. S., Verschoor, C. P., You, Q., Kelton, D. F., Moore, S. S., Karrow, N. A. 2010. A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in Holstein cattle. 2010, Vol. 95, 3, pp. 176-182.
- [61] Pantelic V., Samolovac L., Aleksic S., Trivunovic S., Petrovic M. M, Ostojic-Andric D. and Novakovic Z. 2010. Heritability of type traits in first calving Black and White cows. ArchivTierzucht. 2010, Vol. 53, pp. 545–554.
- [62] Powell, R.L., H.D. Norman. 2006. Major advances in genetic evaluation techniques. 2006, Vol. 89, pp. 1337-1348.
- [63] Pradeepkumar, D. и Ravi, V. 2017. Forex rate prediction: A hybrid approach using chaos theory and multivariate adaptive regression splines. 2017, Vol. 515, pp. 219–227.
- [64] Pryce JE, Esslemont RJ, Thompson R, Veerkamp RF, Kossabati MA and Simm G. 1998. Estimation of genetic parameters using health, fertility and production data from a management recording system for dairy cattle. Animal Science. 1998, Vol. 66, pp. 577-584.
- [65] Rougoor, C. W., Sundaram, R., van Arendonkd, J. A. M. 2000. The relation between breeding management and 305-day milk production, determined via principal components regression and partial least squares. 2000, Vol. 66, pp. 71–83.
- [66] Roxström A, Strandberg E, Berglund B, Emanuelson U and Philipsson J., 2001. Genetic and environmental correlations among the female fertility traits, and between the ability to show oestrus and milk production in dairy cattle. Acta Agriculturae Scandinavica, Section A - Animal Science. 2001, Vol. 51, pp. 192-199.
- [67] Schapire, R. ,Freund Y. Bartlett P. and Lee W. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics. 1998, Vol. 26, 5, pp. 1651-1686.
- [68] Shook, G.E. 2006. Major advances in determining appropriate selection goals. 2006, Vol. 89, pp. 1349-1361.
- [69] Sinha, R., B. Sinha, R. Kumari, M.R. Vineeth, A. Verma, and I. D. Gupta. 2020. Principal component analysis of linear udder type traits and their relation with milk yield and composition in indigenous Sahiwal cattle. Asian Australasian Journal of Animal Science. 2020
- [70] SPM. 2010. Salford Predictive Modeler. 2010 <https://www.minitab.com/en-us/products/spm/> .
- [71] SPSS. 2019. SPSS IBM Statistics 25. 2019 <http://www-01.ibm.com/software/analytics/spss/> .

- [72] Statsoft. 2013. Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: (Electronic Version). 2013 <http://www.statsoft.com/textbook/> .
- [73] Steinberg D., M. Golovnya,. 2007. CART 6.0 User's Guide. 2007
- [74] Steinberg, D. L. and Colla P. 1995. CART: tree-structured non parametric data analysis. Salford Systems. 1995
- [75] Steinberg, D., Cardell N. S. 1998b. The hybrid CART-Logit model in classification and data mining. Keystone, CO : American Marketing Association, 1998b. Eighth Annual Advanced Research Techniques Forum.
- [76] Steinberg, D., Cardell, N. S. 1998a. Improving data mining with new hybrid methods. Boston, MA: Presented at DCI Database and Client Server World, 1998a.
- [77] Tibshirani, R. 1996. Bias, Variance, and Prediction Error for Classification Rules, Technical Report.: Statistics Department, University of Toronto, 1996.
- [78] Topal M, Aksakal V, Bayram B and Yaganoglu AM. 2010. An analysis of the factors affecting birth weight and actual milk yield in Swedish red cattle using regression tree analysis. The Journal of Animal and Plant Sciences. 2010, Vol. 20, pp. 63-69.
- [79] Venkatadri, M. 2011. A review on Data mining from Past to the Future. 2011, Vol. 15, 7, pp. 19-22.
- [80] Vukasinovic, N., J. Moll, N. Kiinzi. 1997. Factor analysis for evaluating relationships between herd life and type traits in Swiss Brown cattle. Livestock Production Science. 1997, Vol. 49, pp. 227-234.
- [81] Wen S., Buyukada, M., Evrendilek, F., Liu, J. 2019. Uncertainty and sensitivity analyses of co-combustion/pyrolysis of textile dyeing sludge and incense sticks: Regression and machine-learning models. 2019, 151 (May 2020), pp. 463-474.
- [82] WHFF. 2015. World Holstein Friesian Federation. Април 2015 <http://www.whff.info/index.php>
- [83] White, J.M. 1974. Genetic parameters of conformational and managemental traits. 1974, Vol. 57, pp. 1267-1278.
- [84] Wilhelm A. 2004. Data and Knowledge Mining . [автор на книга] W. Härdle, Y. Mori Eds. J. E. Gentle. Handbook of Computational Statistics. Heidelberg: Springer, 2004, pp. 789-820.
- [85] Yordanova A., S. Gocheva-Ilieva, H. Kulina, L. Yordanova and I. Marinov. 2015. Classificatoin and regression tree analysis in modeling the milk yield and conformation traits for Holstein cows in Bulgaria. Agricultural science and Technology. 2015, Vol. 7, 6, pp. 208–213.
- [86] Yordanova, A., Yordanov, M. 2018. Web based system for choosing a statistical method for data processing. 2018, Vol. 6, 2, pp. 138-147.
- [87] Zaborski D., W. Proskura, W. Grzesiak, J. Różańska-Zawieja and Z. Sobek. 2019. The comparison between random forest and boosted trees for dystocia detection in dairy cows. Computers and Electronics in Agriculture. 2019, Vol. 163, pp. 104856.
- [88] Войникова, Д. 2013. Линеен модел на изходната лазерна мощност на CuBr лазер с помощта на CART. 2013, Vol. 2, pp. 15-20.
- [89] Въндев, Д. 2003а. Записки по приложна статистика 1. Софийски университет “Св. Климент Охридски”. 2003а <https://www.matematika.bg/visha-matematika/uchebnici/applstat1.pdf>
- [90] Въндев, Д. 2003б. Записки по приложна статистика 2. Софийски университет “Св. Климент Охридски”, С. 2003б
- [91] Герговска Ж., Цветанова, Я., Л. Йорданова. 2004. Сравняване на смесени линейни модели за оценка на развъдната стойност на говеда за мляко на база продуктивност

- за нормална лактация и за контролен ден. Животновъдни науки. 2004, Vol. 4, pp. 62-66.
- [92] Герговска, Ж. et al. 2010. Наръчник по животновъдство. Тракийски университет, 2010
- [93] Герговска, Ж. 2000. Влияние на някои екстериорни линейно описателни признаци върху млечната продуктивност при черно-шарени крави Национална конференция с международно участие "Постижения в областта на аграрните и обществени науки", СУБ – Стара Загора 1-2. 2000, Vol. 1, pp. 84-88.
- [94] Герговска, Ж. 2008. Проучване на различни показатели за устойчивост на лактацията при крави за мляко. Животновъдни науки. 2008, pp. 145-152.
- [95] Гочева, С. 2015. Бизнес статистика с SPSS. Пловдив, 2015.
- [96] Гочева-Илиева, С. 2013. Приложна математика. Пловдив Университетско издателство „Паисий Хилендарски“ (електронно издание), 2013. ISBN 978-954-423-847-6.
- [97] Желязков, Е., Я. Цветанова. 2002. Ръководство за упражнения по Генетика. Стара Загора: СД Контраст, 2002. ISBN -954-9887-32-4.
- [98] Животновъдство. 2014. Отборът на най-добрите животни е решаващ в селекционния процес. 2014
- [99] Захариев З., А. Алексиев, М. Ангелов, И. Карабалиев. 1988. Черношарено говедо. София: Земиздат, 1988.
- [100] Йорданов, М. 2017. WebStat. 2017 <http://webstat.uni-sz.bg>
- [101] Йорданова Л., Г. Кирякова, П. Велева, Н. Ангелова, А. Йорданова. 2019. Статистически софтуер за обработка на експериментални данни. 1. Стара Загора: Академично издателство "Тракийски университет", 2019. ISBN: 978-954-338-152-4.
- [102] Кръстанов, Ж., Герговска Ж., Ангелова Т., Йорданова Д., Карабашев В., Калайджиев Г. 2013. Първа оценка на развъдната стойност на биците и кравите на Асоциацията за развъждане на Черношареното говедо. 2013
- [103] Маринов, И. 2016. Линейни екстериорни признаци и връзката им с продуктивни, репродуктивни и здравословни показатели при черно-шарени крави. Дисертация за ОКС "Доктор". Стара Загора, България, 2016
- [104] Николов, В. et al. 2012. Технология на говедовъдството. Академично издателство на Аграрния университет. Пловдив: Академично издателство на Аграрния университет, 2012.
- [105] Стоянова, Р. 2007. Влияние на някои линейно описателни признаци на тялото и морфологични показатели на вимето върху млечната продуктивност и скорост на доене при черно-шарени крави. Дисертация за ОКС „Доктор“. Стара Загора, България: 2007

Plovdiv University “Paisii Hilendarski”
Faculty of Mathematics and Informatics
Department of Applied Mathematics and Modeling

Antoaneta Petrova Yordanova

APPLICATION OF DATA MINING METHODS
FOR STATISTICAL MODELING

Author’s summary of a dissertation

for the acquisition of the educational and scientific degree “Doctor”

Area of Higher Education: 4. Natural Sciences, Mathematics and Informatics

Professional Stream: 4.5. Mathematics

Doctoral Program: Mathematical modeling and application of Mathematics

Dissertation Tutor: Prof. PhD, DSc Snezhana Georgieva Gocheva-Ilieva

Plovdiv, 2021