

РЕЦЕНЗИЯ

от доц. дн Иванка Миткова Желева

Русенски университет „Ангел Кънчев“

на дисертационен труд за присъждане на образователната 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*

Автор: *Антоанета Петрова Йорданова*

Тема: „**ПРИЛОЖЕНИЕ НА ДЕЙТА МАЙНИНГ МЕТОДИ ЗА СТАТИСТИЧЕСКО МОДЕЛИРАНЕ**“

Научен ръководител: *проф. д.м.н. Снежана Георгиева Гочева-Илиева от Пловдивски университет „Паисий Хилендарски“*

1. Общо описание на представените материали

Със заповед № Р33-1451 от 27.04.2021 г. на Ректора на Пловдивския университет „Паисий Хилендарски“ (ПУ) съм определена за член на научното жури за осигуряване на процедура за защита на дисертационен труд на тема „**ПРИЛОЖЕНИЕ НА ДЕЙТА МАЙНИНГ МЕТОДИ ЗА СТАТИСТИЧЕСКО МОДЕЛИРАНЕ**“ за придобиване на образователната 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*. Автор на дисертационния труд е Антоанета Петрова Йорданова – докторантка в редовна форма на обучение към катедра Математически анализ на факултета по математика и информатика на Пловдивски университет “Паисий Хилендарски“ с научен ръководител: *проф. д.м.н. Снежана Георгиева Гочева-Илиева от Пловдивски университет „Паисий Хилендарски“*.

В периода 10.05.2021 г. – 11.05.2021 г. се проведе първото **неприсъствено заседание** на научното жури, на което бях определена за рецензент на дисертационния труд на Антоанета Петрова Йорданова.

Представеният от докторантката комплект материали е в съответствие с Чл.36 (1) от Правилника за развитие на академичния състав на ПУ и включва следните документи:

- ✚ молба до Ректора на ПУ за разкриване на процедурата за защита на дисертационен труд;
- ✚ автобиография в европейски формат;
- ✚ протокол от катедрения съвет, свързан с докладване на готовността за откриване на процедурата и с предварително обсъждане на дисертационния труд;
- ✚ дисертационен труд;
- ✚ автореферат;

- ✚ списък на научните публикации по темата на дисертацията;
- ✚ копия на научните публикации;
- ✚ декларация за оригиналност и достоверност на приложените документи;
- ✚ справка за спазване на специфичните изисквания на съответния факултет (само за докторантите зачислени до 04.05.2018 г.);

Докторантката е приложила 3 публикации.

Представеният от Антоанета Йорданова комплект материали е в съответствие с изискванията на Закона и Правилника за развитие на академичния състав на Пловдивски университет. Документите са акуратно подготвени и правят много добро впечатление.

2. Кратки биографични данни за докторанта

Антоанета Петрова Йорданова е родена на 29.04.1969 г. Средното си образование е завършила в Математическата гимназия в гр. Враца през 1987 г. След това, в периода 1987 – 1992 г., е учила математика в Софийския университет, като е завършила семестриално. В периода 1994 – 1999 завършва специалността Застраховане и социално дело в Стопанската академия в Свищов и получава образователна и квалификационна степен Магистър, а в периода 2014 – 2015 завършва още една магистратура по Приложна математика и статистика в Пловдивски университет. Зачислена е в редовна докторантура към същия университет през 2017 г. По-голямата част от трудовата ѝ заетост е в периода 1994 – 2015 и е в София. Работила е 5 години като специалист по имуществено застраховане, 6 години като търговски директор на фирма за производство на хляб и хлебни изделия, и 10 години като управител на фирма за книгоиздаване. В периода 2014 – 2017 е хонорован асистент по информатика и информационни технологии в Аграрния факултет на Тракийския университет, а след това става асистент в Медицински колеж на същия университет.

Антоанета Йорданова има публикувани 7 научни статии – от тях 6 са на английски език и една е на български, както и е издадено едно учебно пособие на български език.

Участвала е в 5 научни проекта, финансирани от ПУ, в 3 проекта, финансирани от Тракийски университет, участвала е с доклади в няколко международни и национални научни конференции. Взела е участие и в 9 курса за повишаване на квалификацията, организирани от Тракийски университет, била е на специализация в Италия по програмата за студентска мобилност Еразмус за 5 месеца през 2019 г.

Антоанета Йорданова използва английски и руски език на добро ниво. Притежава много добри комуникационни, организационни, мотивационни умения, както и умения за управление на проекти, на хора, на кризисни ситуации.

3. Актуалност на тематиката и целесъобразност на поставените цели и задачи

Една от най-бързо развиващите се области на науката сега е извличането на знания от данни (data mining). Това е процес на откриване на смислени корелации, зависимости, повтарящи се образци, тенденции и аномалии в големи масиви от данни чрез използване на техники и алгоритми от областта на машинното обучение, разпознаването на образи, статистиката, невронните мрежи и визуализацията на данни. За разлика от извличането на данни посредством заявки към база данни, извличането на знания от данните генерира не очевидна и предварително известна, потенциално полезна информация. Data mining представлява процес на анализ на съхранявани бази данни за извличане на нова полезна информация чрез разкриване на дълбоките и скрити взаимоотношения между на пръв поглед неиз-

вестни или несвързани една с друга величини. Важна негова особеност е, че той осигурява възможност за обработка на многомерни масиви и извличане на многомерни зависимости, като същевременно разкрива изключителните ситуации – данни и случаи, не включващи се в общите закономерности. Появата на Data mining анализа е свързана също с необходимостта от усъвършенстване на техниките за запис и съхранение на данни, получавани обикновено от хиляди източници в колосални потоци от информация в най-различни области. Развитието на съвременните технологии за такава (дейта майнинг) преработка на данните се обуславя от следното: Неограничен обем на данните; Голямо разнообразие и разнородност на данните (количествени, качествени и текстови); Необходимост от конкретни и разбираеми резултати; Разработване на инструменти за обработка на данните, предоставящи възможност за лесното им използване. Следователно извличането на знания от данни (data mining) е една от най-динамично развиващите се области на познанието 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 една от основните области на аграрните науки – животновъдството.

Интензивната практика в животновъдството в съвременните условия 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 тази област. Вече е общоприето мнението, че математическите методи и модели са най-разпространен, сигурен и адекватен подход за описание и изследване на обекти и явления, за обяснение и предсказване на наблюдавани феномени в различни предметни области, в това число и в животновъдството..

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 икономически 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 тези три аспекта фенотип, генотип и среда. За тяхното изследване са използвани следните статистически методи: метод на класификационни и регресионни дървета (CART), ансамблови методи: CART Ensemble and Bagging (CART EBag) и метод Случайни гори (RF).

В дисертационния труд ясно са дефинирани неговите цел и задачи.

Основната цел на рецензирания дисертационен труд е *да се приложат дейта майнинг методи с машинно обучение 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 са формулирани следните **задачи**:

1. Да се построят и изследват модели на 305-дневната млечност в зависимост от линейни екстериорни признаци и характеристики на 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 ансамбовия 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.

Всичко това ме убеждава, че тематиката на дисертационния труд е актуална и поставените в него цел и задачи са актуални и значими в научно и научно-приложно отношение.

4. Познаване на проблема

Представените документи по защитата на дисертационния труд показват, че докторантката познава много добре състоянието на разглежданите в дисертацията проблеми, както по отношение на характеристиките и изследването на млечността на кравите от породата Холщайн - Фрезийска, така и по отношение на математическите методи за изследване и анализ на данни. Тя разбира добре и оценява творчески литературния материал по темата на дисертацията.

5. Методика на изследването

В дисертационния труд се използват следните статистически методи: линеен регресионен анализ, корелационен анализ, метод на класификационните дървета (CART), факторен анализ и регресия с главните компоненти (PCA), случайни гори (RF), ансамблови методи CART Ensemble and Bagging (CART EBag).

Избраната методика на статистическо изследване и моделиране **позволява** да се постигане поставената цел на дисертационния труд и да се получи адекватен отговор на задачите, решавани в него.

6. Характеристика и оценка 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 за моделиране и изследване 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), ансамблови методи 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. Направена 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 Пловдивски университет “Паисий Хилендарски“. Написан е ясно и подробно и е оформен много добре. Обсъждан е в катедрата и на различни научни форуми и по-голямата част от резултатите му са публикувани в реферирани научни списания. Това за мен означава, че материалът, върху който се градят приносите на дисертационния труд, е достоверен.

7. Приноси и значимост на разработката за науката и практиката

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 моделирането на реални данни от животновъдството в България 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.

Всичко казано до тук ми дава основание да приема изцяло и оценя високо научните и научно-приложните приноси на дисертационния труд на Антоанета Йорданова.

7. Преценка на публикациите по дисертационния труд

По темата на дисертацията са представени 3 публикувани научни публикации. Те могат да бъдат класифицирани по следния начин: статия, публикувана в списание – 1 брой; доклади на международни или национални научни конференции – 2 броя; една от тях е в материалите на международна научна конференция и има SJR ранг, на английски език са 2 броя, а на български език – 1 брой, с един съавтор – 2 броя; с трима и повече съавтори – 1 брой.

Броят и качеството на публикуваните статии отговарят на изискванията на ЗПАЦРБ, на Правилника за приложенията му, както и на Правилниците на Пловдивски университет за присъждане на образователната и научна степен “доктор“.

9. Лично участие на докторантката

От представените материали по защитата на дисертационния труд на Антоанета Петрова Йорданова става ясно, че формулираните приноси и получени резултати са нейно лично дело и са получени под общото ръководство на научния ръководител.

10. Автореферат

Дисертационният труд е представен в автореферат, който е с обем от 32 страници. В началото на автореферата е направена обща характеристика на дисертационния труд, на неговата актуалност и съдържание, посочени са приносите му, последвани от списък с публикациите на автора по темата на дисертацията и накрая е представена използваната литература – общо 105 заглавия.

Съдържанието на автореферата съответства на съдържанието на дисертационния труд.

Оценката ми за автореферата е, че той отговаря на общоприетите изисквания и отразява вярно съдържанието и приносите на дисертационния труд.

11. Критични забележки и препоръки

Нямам съществени критични бележки.

Считам, че представеният дисертационен труд е отличен атестат, както за автора му, магистър Антоанета Петрова Йорданова и научния ръководител проф. Дн Снежана Гочева - Илиева, така и за катедрата, факултета и университета, в които те работят. Представеният

дисертационен труд може да послужи за стабилна основа на бъдещо сътрудничество и съвместна дейност между Тракийски университет и Пловдивски университет в областта на прилагане на математическото моделиране в аграрните науки. Бих приветствала такова сътрудничество и съвместна работа по общи проекти, свързани с математическо моделиране и приложение на математиката в аграрните и ветеринарни науки.

12. Лични впечатления

Не познавам лично докторантката Антоанета Петрова Йорданова. Но по представените материали за защита на дисертационния труд на тема **„ПРИЛОЖЕНИЕ НА ДЕЙТА МАЙНИНГ МЕТОДИ ЗА СТАТИСТИЧЕСКО МОДЕЛИРАНЕ“**, мога да съдя, че тя е задълбочен научен работник с критично отношение към литературните данни и проявява задълбоченост и акуратност при получаване на резултатите от изследванията по темата на дисертацията и при оформяне на публикациите и документите по защитата.

13. Препоръки за бъдещо използване на дисертационните приноси и резултати

Приемам и одобрявам посочените от Антоанета Йорданова перспективи за бъдеща работа по темата на дисертацията. По-конкретно, разработените методи и подходи в дисертацията 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.

ЗАКЛЮЧЕНИЕ

Дисертационният труд **съдържа научни и научно-приложни резултати, които представляват оригинален принос в науката** и **отговарят на всички** изисквания на Закона за развитие на академичния състав в Република България (ЗРАСРБ), Правилника за прилагане на ЗРАСРБ и съответния Правилник на ПУ „Паисий Хилендарски“. Представените материали и дисертационни резултати **напълно** съответстват на специфичните изисквания на Факултета по математика и информатика, приети във връзка с Правилника на ПУ за приложение на ЗРАСРБ.

Дисертационният труд показва, че докторантката Антоанета Петрова Йорданова **при- тежава** задълбочени теоретични знания и професионални умения по научна специалност Математическо моделиране и приложение на математиката като **демонстрира** качества и умения за самостоятелно провеждане на научно изследване.

Поради гореизложеното, убедено давам своята **положителна оценка** за проведеното изследване, представено от рецензираните дисертационен труд, автореферат, публикувани статии, постигнати резултати и приноси и **предлагам на почитаемото научно жури да присъди на Антоанета Петрова Йорданова образователната 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.**

28.05.2021 г.

Рецензент:

Доц. дмн Иванка Желева