РЕЦЕНЗИЯ

на

на дисертационен труд, представен за получаване на образователната и научна степен „доктор”

Автор на дисертационния труд: ас. Антоанета Петрова Йорданова

Тема на дисертационния труд: „Приложение на дейта майнинг методи за статистическо моделиране”

Заявител за откриване на процедурата: катедра “Математически анализ”, ПУ “Паисий Хилендарски”

Докторска програма – Математическо моделиране и приложение на математиката

Рецензент: проф. д-р Михаил Д. Тодоров, кат. Математическо моделиране и числени методи, ФПМИ, ТУ – София, зап.Р33-1451/27.04.2021 г. на Ректора на ПУ „Паисий Хилендарски”

Кратки биографични данни за дисертантката

Антоанета Петрова Йорданова е родена през 1969 г. Завършва Математическа гимназия във Враца през 1987 г. В периода 1987-92 г. следва математика в СУ „Св.Климент Охридски”, където завършва семестриално. По-късно през 2014-о и социално дело“. В 2015 г. се дипломира като магистър по приложна математика и статистика в ПУ „Паисий Хилендарски“. Междувременно в периода 1994-1999 г. следва в СА „Димитър Ценов“ – Свищов, където се дипломира като магистър по застрахователнзавършва В периода 2014-17 г. е хоноруван асистент по информатика и информационни технологии в Тракийския университет – Стара Загора. В периода 2017-20 г. е редовен докторант в ПУ „Паисий Хилендарски“. От октомври 2020 г. е асистент в Медицинския колеж към ТрУ – Стара Загора.

Представената дисертация има обем от 119 стр., формат А4, в.т.ч. 43 фигури, 26 таблици и библиография от 105 работи.

1. Актуалност на дисертационния труд

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.

От друга страна 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. Дисертацията е посветена на търсенето на връзки и симбиоза между аграрната наука и статистичското математическо моделиране. Подобен вид дейност изисква еднакво добро познаване на реални процеси и явления и начините за тяхното регулиране от една страна и съответен математически инструментариум за обработка на информация, оптимизация и прогнози от друга. В дисертацията 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).

Тематиката е с ясен интердисциплинарен фундамент и с многобройни приложения, което е достатъчна обосновка и мотивация за провеждане на изследванията. Всичко това предполага нужната математическа квалификация и практически знания, които дисертантката очевидно притежава.

1. Анализ на състоянието на проблема

Дейта майнинг методите са ново направление, което датира от 90-те години на 20-ти век. При тях целта е получаването или извличането на полезни знания от данните, откриването на модели на данните, подпомагането на вземане на експертни решения. Към тях принадлежат *а)* метод на класификационни и регресионни дървета CART, който може да доведе до редукция на размерността, осигуряване на надежни оценки при малки извадки и пр.; *б)* Метод на Случайните гори (Random Forest) е един от най-силните и мощни алгоритми. Такъв 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.

2. Методика на изследванията

Изследването е свързано със събирането, съхранението и обработката на значителни обеми информация. Това налага използването на специализиран и високо-производителен статистически софтуер за анализ на емпиричните данни. В дисертационния труд се използва алгоритъмът на CART-EBag метода, включен в софтуерния пакет на Salford Predictive Modeler на компанията Minitab. Методът CART-Ensembles and Bagging (CART-EBag) се прилага за пръв път за статистистическо моделиране на данни от областта на аграрните науки. Оценка на качеството на моделите се извършва с помощта на коефициента на детерминация R2, средната абсолютна грешка (MAD), средната абсолютна процентна грешка (MAPE), средната абсолютна грешка (MSE) и средната квадратична грешка (RMSE). Методиката на изследване е подчинена на основната цел, а именно 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.

3. Характеристика и оценка на получените резултати

В Глава 1, която има и уводен характер е направен сравнителен и критичен анализ на използваните модели. По този начин са ясно очертани насоките на изследванията, проведени от дисертантката и представени в настоящата дисертация. 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). Построените математически модели за зависимостта на изследвания продуктивен признак от линейните признаци показват кои линейни признаци са важни предиктори за млечността. Включването на фермата в CART Модел 2 променя подредбата на предикторите по важност, получени от CART Модел 1. Анализът на грешките показва нормалност на тяхното разпределение и за двата обсъждани модела.

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. Разгледани са 3 основни задачи, резултатите и изводите от чието решаване ще бъдат проследени по-долу.

В Глава 2 са представени резултати от проведено изследване на основните връзки, класификация и прогнозиране при многомерни данни, използвайки метода CART. За разглежданата извадка от 97 наблюдения на говеда от 4 ферми в България са изградени две дървета за вземане на решения за изследване на зависимостта на 305-дневния млечен добив за кравите от Холщайн по отношение на промените в 13 независими променливи. Описани са

основни линейни характеристики на животните от породата Холщайн. Изходните данни са класифицирани в пет основни групи.

Вторият CART Модел 2 допълнително включва фермата, където кравите се отглеждат, като 13-та независима променлива. Разширеният модел обяснява 70% от данните.

С получените правила за двата модела може да се направят прогнози за млечността преди края на лактациите. Прилагането на метода CART изяснява насоките за селекция по отделни линейни признаци и може да помогне възможно развитие на индексите за селекция в зависимост от тяхното значение.

В Глава 3 са построени и изследвани два RF модела, на база на наборите от променливи от предходната глава. Направено е сравнение на съответстващите модели по следните показатели R2, RMSE, MAE, MSE, MAPE. CART техниката е използвана за регресионен анализ на набор данни със зависима променлива 305 дневна млечност *MilkY305*. В единия от моделите, CART Модел 1 са използвани 12 предиктора и е получено дърво с 5 крайни възела и дълбочина 3. Този модел обяснява около 48% от варирането на зависимата променлива. В другия модел, CART Модел 2, като предиктор е включена като фактор и фермата *FarmN*. Получено е дърво с 5 крайни възела и дълбочина 3. CART Модел 2 описва 70% от зависимата променлива. Получените модели чрез CART техниката подлежат на подобряване чрез подходящи стойности на задаваните контролни параметри. Получените два модела в рамките на настоящото изследване са сравнени с построените по-рано CART модели за същия набор данни и е установено значителното повишаване на коефициента на детерминация с метода RF. 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 2). В RF моделите недвусмислено се отчита и състоянието на скакателните стави при животните, докато тази променлива не участва в CART моделите. Получените CART модели са сравнително по-груби и могат да се подобрят чрез рафиниране на задаваните контролни параметри. Признаците от линеен тип, включени в двата RF модела, са сравнително лесни за измерване. Те наистина биха могли да улеснят прогнозирането на добива на мляко в процеса на оценка на животните. Забележителното увеличение на коефициента на детерминация R2 при RF моделите - 95% / 95% срещу 48% / 70%, както и същия ред на предикторите по влиянието им в сравнение с CART моделите показва по-голямата ефективност на RF метода.

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.

Изследвана е зависимостта на усреднената 305-дневна млечност на крави от породата Холщайн от 12 екстериорни признака. Установена е мултиколинеарност на показателите. С метода на главните компоненти и Варимакс ротация са получени 11 компонента. Приложена е регресия с главните компоненти и е получено регресионно уравнение. Установено е, че усреднената 305-дневна млечност зависи от пет от изследваните показатели. Моделът описва 53% от данните. Направена е интерпретация на резултатите. Може да се направи изводът, че полученият регресионен модел е статистически валиден и може да се използва за изводи и заключения.

Построени са 2 групи модели – с началните променливи и с главните компоненти и е приложен методът CART- EBag за 10, 15, 20 и 25 дървета в ансамбъла за всяка от групите. Получени са 8 модела, които са сравнени по групи и е избран такъв, който приближава най-добре предсказаните стойности до наблюдаваните – с минимална стойност за RMSE и максимална за R2.
Аналогично са получени още 8 модела с прилагане на метода CART-EBag като към първоначалните 12 предиктора е добавена и фермата като средови фактор. Моделите са сравнени и са избрани оптималните по групи по отношение на коефициента на детерминация. Тези стойности на коефициента са значително над постигнатия R2=0,533 при прилагането на регресията с главните компоненти. При моделите с главни компоненти съответстващите фактори по относителен принос за добива на мляко са: ширина на вимето, ширина на гърдите и изглед на задни крака отзад. Първите два фактора имат устойчиво влияние и присъствие и за двете групи модели.

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 R2, RMSE, MAE, MSE, MAPE и е избран е най-добър модел. Направена е интерпретация на модела.

CART- EBag моделирането има редица предимства: дава по-добри резултати при комбиниране с метода на главните компоненти; може да се приложи и ако зависимата променлива не е с разпределение близко до нормалното; може да се приложи при предиктори изцяло от ординален тип. Недостатък е, че за модели с над 25 дървета статистиките започват да се влошават и затова не е целесъобразно броят на дърветата да се увеличава.

За обработка на данните са използвани Метод на главните компоненти и Регресия с главните компоненти. Регресията е приложена за генерираните главни компоненти с варианта стъпкова регресия. Този метод е подходящ, тъй като е установено, че е налице мултиколинеарност между изследваните променливи и директното прилагане на линейна регресия е затруднено. С помощта на SPSS е приложен факторен анализ с МГК и са извлечени 12 главни компонента.

4. Преценка на авторската справка

Авторската справка отразява приносите и акцентите в дисертацията като

цяло. Приносите имат научен, но в много по-голяма степен научно-приложен характер. Представените анализи и формулираните критерии могат да се използват успешно в животновъдството в реални условия. Според мен те могат да бъдат интерес за млеконадоя и хранително-вкусовата промишленост.

Оценявам приносите като колективни, но с водеща роля на дисертантката и под ръководството на научния ръководител. Всички те могат да бъдат причислени към направлението „Обогатяване на съществуващи знания с цел приложения в практиката".

5. Критични бележки по трудовете и литературна осведоменост на дисертантката

Дисертацията прави много добро впечатление. Получените резултати и изводи изискват огромен обем от работа, симулации и обработка на информация. Според мен може да послужи и като наръчник за практическа реализация в селското стопанство и в частност в животновъдството. Написана е на правилен български език, изложението е стегнато и логически последователно. Нямам критични бележки по същество. Налице е и задълбочено познаване на литературата по разглежданите в дисертацията въпроси, видно от въведението, което прави получените резултати още по-убедителни. Литературната осведоменост на дисертанта се основава на най-нови източници.

6. Публикации по дисертацията

Резултатите са докладвани на няколко конференции и семинари. Публикувани са в сборници доклади и научни съобщения и научни статии в Пловдив, *Agricultural Science and Technology* и в *AIP CP*. Статиите са в съвторство с научния ръководител на дисертантката. *AIP CP* има SJR=0.182 (2020).

Други данни за публикациите могат да се видят в представената таблица.

Таблица: Справка за трудовете

Статии – 3 бр.	У нас - 2 бр. *Agricultural Science and Technology*, конференция в Пловдив В чужбина - 1 бр. *AIP CP*
Доклади на международни научни прояви – 4 бр.	У нас - 3 бр. (*Тракийски университет, AMiTaNS'20*)
Участие в проекти и курсове за квалификация – 8 (5+3)	У нас – 5 бр. (*ФНИ – МОН, Еразъм – Италия и др.*)

7. Приложение на резултатите в практиката

Получените в дисертацията резултати имат приносен характер към

приложната статистика и модерното животновъдство. Наред с теоретичните от съществена важност са обаче научно-приложните приноси, тъй като тук се работи с масиви от реална входяща информация, която след съответна статистическа обработка и интелигентен анализ извежда информация, която има също тъй реалистичен и адекватен характер. Според мен на базата на проведените изследвания може да се говори за създадена технология, която може да се използва ефективно в реални условия. Нейната успешна реализация и надграждане е и пожеланието ми към дисертантката и нейния научен ръководител.

8. Преценка на автореферата

Авторефератът отразява правилно и пълно съдържанието на дисертационния труд.

9. Лични впечатления

Познавам дисертантката дистанционно. Имах възможност да слушам чрез телемост нейна презентация по тематиката на дисертацията на конференцията AMiTaNS'20, която се проведе on-line миналата година. Впечатленията ми са, че е деен и инициативен човек с перспектива за бизнес развитие и научно израстване, което е гаранция, че ще приложи успешно натрупаните знания и резултати в по-нататъшното си кариерно развитие.

Заключение

Отчитайки значимостта на проведените изследвания и след справка с ППЗРАСРБ в ПУ, както със специфичните изисквания на ФМИ мога да твърдя, че представената дисертация отговаря напълно и количествено, и качествено на препоръчителните наукометрични критерии на ПУ за присъждане на научни степени. Въз основа на гореизложеното си позволявам да препоръчам убедено на членовете на уважаемото НЖ да гласуват даване на ОНС „доктор" на Антоанета Петрова Йорданова, Професионално направление 4.5. Математика, докторска програма „Математическо моделиране и приложение на математиката".

СЪСТАВИЛ:

Проф. д-р Михаил Тодоров
кат."Матем. моделиране и числени методи",
ФПМИ при ТУ - София

30 май 2021 г.
София