С Т А Н О В И Щ Е

на дисертационен труд за присъждане на образователна и научна степен „доктор“ по област на висше образование 4. Природни науки 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**

Член на научното жури: проф. д.м.н. Снежана Георгиева Гочева-Илиева, Пловдивски университет „Паисий Хилендарски“, Факултет по математика и информатика

1. Общо представяне на процедурата и докторанта

Съгласно заповед Р33-1451/27.04.2021 г. на Ректора на Пловдивския университет „Паисий Хилендарски“ (ПУ) съм определена за член на научното жури по процедурата за защита на настоящия дисертационен труд. На първото заседание на НЖ бях избрана да подготвя становище. Получила съм всички официални документи за изготвяне на становището ми.

Докторантката Антоанета Йорданова най-напред завършва семестриално специалност Математика в СУ „Климент Охридски“. След това, следвайки повика на времето се ориентира към по-атрактивно образование и завършва СА ”Димитър А. Ценов”, гр. Свищов, специалност Застраховане и социално дело (бакалавър и магистър). През 2015 г. завършва магистратура по приложна математика във ФМИ на ПУ с дипломна работа по тема, близка до темата на дисертационния й труд. От 1 август 2017 г. е зачислена като моя редовна докторанта към бившата катедра „Приложна математика и моделиране“ на ФМИ на ПУ. Обучението й в докторантура включваше разучаване на нови методи и софтуер, за да навлезе в новата наука за данните, и преодоляване на големи трудности с реални данни от областта на животновъдните науки. По време на докторантурата тя завърши и множество полезни курсове за докторанти към Тракийския университет – Стара Загора. Също така беше на 5-месечно обучение в рамките на Европейската програма „Учене през целия живот“, Програма Еразъм+ (докторантска мобилност) в университета в Акила, Италия. За нуждите на катедрата по времето на докторантурата подготви и проведе лабораторни упражнения по дисциплината Маркетингови изследвания на студентите от 1ви курс, специалност Бизнес информационни технологии, редовно обучение във ФМИ на ПУ. Взе редовно участие в научните семинари по приложение на математиката към катедрата, участва с доклади на конференции и др., беше част от екипите на три проекти към НПД на ПУ. В

момента Антоанета Йорданова работи като асистент в Медицинския колеж към Тракийски университет - гр. Стара Загора.

2. Актуалност на тематиката и познаване на проблема

Темата на дисертационния труд на Антоанета Йорданова е изцяло в направлението на математическото моделиране и приложение на математиката. Проблемната област е част от селскостопанските науки и по-конкретно - животновъдните науки. Тази област има огромно значение за стопанството на всяка държава и тук стоят множество научни предизвикателства. В България в тази област няма други изследвания с използване на модерните и мощни дейта майнинг методи. Актуалността на темата е обоснована от възможността за изучаване на факторите, които влияят на 305-дневната млечност на крави от породата Холщайн-Фрезийска. Следва да обърна внимание, че в сравнение със световните стандарти, добивът за българските ферми е доста по-нисък от средните показатели в много страни. Основната цел бе да се проучат и демонстрират възможностите на дейта майнинг методите за моделиране на този тип данни и то при ограничен обем на извадката. Докторантката разучи водещи софтуерни продукти за моделиране с предсказващи статистически техники – CART, ансамбловите методи CART Ensemble and Bagging и Random Forest, и голям брой литературни източници, както се вижда от библиографията от 105 източника. Считам, че нивото й на познаване на проблема е достатъчно високо.

3. Обща характеристика и оценка на резултатите на дисертационния труд

Дисертационният труд е разработен на 119 печатни страници и съдържа увод, четири глави, заключение с декларация за оригиналност и библиография. Онагледен е с 43 фигури и 26 таблици. Резултатите са публикувани в 3 публикации, едната от тях - е реферирана в SCOPUS. Основната цел на изследванията е *да се приложат дейта майнинг с машинно обучение методи за статистическо моделиране и изследване на зависимости в емпирични данни от животновъдството.*

Глава 1 съдържа анализ на състоянието на проблема с обект на изследванията, използваните дейта майнинг методи, литературен обзор и актуалност на проблемите в двете области.

Глава 2 е посветена на моделиране на многомерните данни с метода на класификационните и регресионни дървета – CART. Търсят се регресионни модели за предсказване и определяне на факторите, които определят 305-дневната млечност на база на емпирична извадка с измервания от четири ферми в България. Изходни независими променливи са 12 екстериорни признаци на крави от порода Холщайн-Фрезийска – височина на животното, ширина на вимето, ширина на крупа, ширина на гърдите, развити кости и други, и номер на фермата. Признаците са от ординален, а фермата е от номинален тип. Построени са предсказващи регресионни дървета, от

които най-добрият CART модел обяснява данни със 70%, което се дължи основно на фермата, а следващите фактори се класират в реда: ширина на вимето, ширина на гърдите и скакателни стави.

Глава трета описва моделиране на данните от Глава 2 с ансамбловия дейта майнинг с машинно опбучение метод на случайните гори (Random Forest, RF). Подробно са описани и изследвани два RF модела за изследване на зависимостта на 305-дневния млеконадой с 13-те независими независими променливи. Най-напред, с 12-те екстериорни признака като предиктори се построява модел RF1. Той описва 95% от данните и определя като основни фактори (по изходящ ред на принос в модела): ширина на вимето, ширина на гърдите, скакателни стави и походка. С отчитане и на фактора ферма, следващият модел RF2 предсказва данните със същата степен на съвпадение от 95%, като в този случай определя фермата за основен фактор, а останалите признаци са в същия ред, както в модел RF1. В качествен и количествен аспект се установява, че за изследваната извадка методът RF е значително по-ефикасен от CART. Разбира се, тук значение има и настройката на хиперпараметрите, но най-вече това се дължи на по-добрите предсказващи качества на RF. Направено е подробно статистическо изследване на грешките на моделите.

Последната Глава 4 съдържа резултати от прилагане на новия ансамблов метод CART Ensemble and Bagging, с който са построени голям брой модели, означени като EBag, за разширена извадка емпирични данни със същия тип данни. В тази глава е изследвана зависимостта на усреднена по ферми 305-дневна млечност на крави от породата Холщайн от 12 екстериорни признака, със и без добавяне на фермата. Установена е мултиколинеарност на екстериорните показатели. С метода на факторния анализ от тях са генерирани 11 взаимноортогонални компонента. Чрез метода на регресия с главните компоненти с тях е построен модел на съвпадение със зависимата променлива от 53%. С началните променливи или с 11те копонента, със или без фермата са построени и изследвани подробно 16 EBag модели, като се задават различен брой дървета в ансамбъла и други хиперпараметри. Сред тях като модел с най-високи статистически показатели е моделът с предиктори 11 факторни променливи и фермата, достигащ коефициент на съвпадение със зависимата променлива от 0.894 или 89.4%. Факторите с най-голям принос относно зависимата променлива усреднена 305-дневна млечност моделите са фермата, ширина на вимето, ширина на гърдите и изглед на задни крака отзад. Направено е общо сравнение на всички получени модели. Установи се, че изборът на EBag моделите е улеснен, тъй като те се саморегулират по статистики. По-конкретно, в дисертационния труд в случая при задаване на над 25 дървета статистиките започват да се влошават и затова не е целесъобразно броят на дърветата да се увеличава. В същото време, не се предсказват добре отдалечените случаи, в частност най-високите стойности на зависимата променлива, което е общ недостатък на ансамбловите методи, усредняващи предсказванията. *Основен принос в*

тази глава е именно приложението на този метод, който не е използван досега в литературата за данни от областта на аграрните науки.

Всички модели в дисертацията, построени с дейта майнинг методи, са крос-валидирани за избягване на преопределяне. Направен е детайлен анализ на резидиумите за проверка на тяхната статистическа пригодност за интерпретиране на резултатите.

5. Оценка на приносите на дисертационния труд

Приемам напълно така формулираните приноси. Считам, че участието на докторантката в публикациите е водещо. По същество приносите могат да се класифицират като научно-приложни. В частност, важен резултат е, че методът CART Ensemble and Bagging се прилага за първи път за данни от разглежданата проблемна област.

6. Критични бележки и препоръки

Нямам критични бележки. Препоръчвам в бъдеще докторантката да разшири обхвата на проблемната област и да прилага овладяните знания и умения за статистическо моделиране с мощните интелигентни методи. Пожелавам й успех!

ЗАКЛЮЧЕНИЕ

На база на гореизложеното заключавам, че е постигнато високо ниво на получените резултати и приноси от Антоанета Йорданова. За цялостната й подготовка, обучение и дисертационен труд моята оценка е ПОЛОЖИТЕЛНА. На база на това убедено предлагам на уважаемото научно жури по процедурата за защита да присъди образователната и научна степен „доктор“ на Антоанета Петрова Йорданова по професионалното направление 4.5 Математика, докторска програма Математическо моделиране и приложение на математиката.

Дата: 25.05.2021 г. **ЧЛЕН НА ЖУРИТО:**

/проф. д.м.н. С. Гочева-Илиева/