



БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО ИНФОРМАЦИОННИ И
КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Кристина Иванова Динева

**ИНТЕГРИРАНЕ НА ХЕТЕРОГЕННИ ДАННИ ОТ РАЗПРЕДЕЛЕНИ
IoT УСТРОЙСТВА**

А В Т О Р Е Ф Е Р А Т

НА ДИСЕРТАЦИЯ

за присъждане на образователна и научна степен „доктор“
по докторска програма „Информатика“
професионално направление 4.6. “Информатика и компютърни науки“

Научен ръководител: доц. д-р Татяна Атанасова

София, 2020 г.

Дисертацията е обсъдена и допусната до защита на разширено заседание на
секция „Моделиране и оптимизация“ на ИИКТ-БАН, състояло се на

Дисертационният труд съдържа 166 страници, 55 фигури, 16 таблици и 175 литературни източника.

Защитата на дисертацията ще се състои на от часа в зала на блок 2 на ИИКТ-БАН на открито заседание на научно жури в състав:

- 1.
- 2.
- 3.
- 4.
- 5.

Материалите за защитата са на разположение на интересуващите се в стая на ИИКТ-БАН, ул. „Акад. Г. Бончев“, бл. 2.

Автор: **Кристина Иванова Динева**

Заглавие: **ИНТЕГРИРАНЕ НА ХЕТЕРОГЕННИ ДАННИ ОТ РАЗПРЕДЕЛЕНИ
IoT УСТРОЙСТВА**

Увод

Информационната революция е резултат от развитието на съвременните технологии и се основава на постиженията на научно-техническия прогрес. Нарастващото присъствие на WiFi и 5G безжичен достъп до интернет води до бързи темпове на развитие и увеличаващ се брой на взаимосвързани устройства.

В парадигмата на “Интернет на нещата” (IoT), обработката на данни придобити от IoT устройства, може да бъде обособена като отделна академична дисциплина, насочена към разработване и надграждане на методи и инструменти, които са важни за повишаване на производителността на труда, конкурентоспособността на производството и качеството на живот. Мрежите от взаимосвързани обекти, не само извличат информация от околната среда и взаимодействат с физическия свят, но също така използват съществуващите интернет стандарти за предоставяне на услуги за трансфер на данни в отдалечена облачна среда. Това води до генериране на огромни количества данни, които трябва да се съхраняват, обработват и представят в лесна и ефективна за интерпретиране форма. Облачните услуги могат да осигурят виртуална инфраструктура за такива процеси, които позволяват данни от устройства за наблюдение да бъдат обработени, анализирани, моделирани и визуализирани в специално разработени за тях платформи.

В същото време, развитието на “Интернет на нещата” създава значителни предизвикателства, които затрудняват пълното реализиране на неговия потенциал.

Настоящият дисертационен труд анализира методи и средства за интегриране, обработка и моделиране на хетерогенни данни, получени от разпределени IoT устройства. На тази основа са разработени и внедрени софтуерни и хардуерни решения с теоретична и практическа значимост.

Цели и задачи на дисертационния труд

От направения анализ на състоянието на изследването е формулирана целта на дисертационния труд:

Да се предложи система и инструменти за интегриране на хетерогенни данни от разпределени IoT устройства, които да позволяват тяхната обработка, моделиране и интеграция.

За тази цел се дефинират следните задачи:

1. Да се предложи методология за обработка, моделиране и интеграция на хетерогенни данни от разпределени IoT устройства.
2. Да се предложи архитектура и метод за комуникация на модулна IoT хардуерна система.
3. Да се предложи архитектура на софтуерна платформа и подход за организация на услугите за интелигентна обработка на хетерогенни данни от IoT система.
4. Да се създадат валидни модели за машинно обучение за експериментално потвърждение на разработената методология.
5. Да се покаже възможно приложение на IoT системата и инструментите за интегриране на хетерогенни данни от разпределени устройства в интелигентното земеделие.

Структура на дисертацията

Дисертационният труд е структуриран в пет глави.

В **първа глава** е направен аналитичен обзор на теоретичната база, свързана с проблемната област на дисертацията. Тя включва кратко въведение, актуалност на темата, приложения, предизвикателства и съществуващи решения на научното изследване. Мотивирана е необходимостта от създаване и прилагане на нова методология за работа с хетерогенни данни, която надгражда и подобрява съществуващите към момента.

Във **втора глава** е представена систематизирана методология за обработване, моделиране и интеграция на хетерогенни данни, извлечени от IoT устройства. Изложена е обща концептуална схема на разработката. На следващ етап подробно е описана и изяснена теоретичната методологична рамка - поредица от стъпки, групирани в четири основни етапа. В процеса на определяне на етапите е извършен обзор и сравнителен анализ на съществуващите методи и подходи при определяне на конкретни случаи, за които тяхното прилагане е правилно.

В **трета глава** е описан процесът на разработване на архитектура на IoT платформа. Тя се състои от две самостоятелни системи - хардуерна и софтуерна, които общуват помежду си. В първият раздел на главата се представя хардуерна архитектурата и нов метод за комуникация между IoT устройства. Във втори раздел се представят архитектурни решения за софтуерна реализация на сървър, базиран на микросървиси и имплементация на потребителски уеб интерфейс. За изграждането на софтуерната система е използван нов подход за организация на услуги за интелигентна обработка и обмен на данни в IoT системата.

В **четвърта глава** е направена експериментална реализация и валидиране на разработената методология. Разгледани са и са решени два типа задачи - за класификационен и регресионен анализ. За целите на тяхното решаване са преминати всички стъпки, описани в методологията. Като резултат от извършения експеримент са получени валидирани модели за машинно обучение, които са готови за интегриране в реална среда.

В **пета глава** е представено практическо приложение на създадената система. Разгледани са нуждите и ползите от нейната употреба. Направен е сравнителен анализ между съществуващите на пазара системи. На база на този анализ е съставена сравнителна характеристика, обобщаваща полезността на съществуващите системи спрямо разработената IoT система в тази дисертация.

В **Заклучението** е представено резюме на получените резултати от разработката. Определени са насоки за бъдещи изследвания и развитие. Представен е списък с научни публикации по темата и забелязани цитирания.

Глава 1 – Анализ на състоянието на изследването

1.1 Индустрия 4.0

Индустрия 4.0 представлява интензивна информационна трансформация на традиционните производствени процеси в свързана среда от данни и включва много иновативни елементи (Jazdi N., 2014), които са част от производствената технология, като: Интернет на нещата, Изкуствен интелект, Машинно обучение, Машини към машини, Големи данни и други. Също така, в Индустрия 4.0 е заложен силен акцент върху сигурността. Това не означава само сигурност на мрежите за данни и комуникации, но и защитата на данните.

1.2 Интернет на нещата

Интернет на нещата може да се разглежда като единна мрежа, свързваща физически и виртуални обекти. Може да бъде описан сценарий, при който в голям брой обекти са вградени уникално идентифицируеми изчислителни устройства. Свързаността към интернет им позволява да събират, съхраняват, споделят и анализират данни и да бъдат управлявани дистанционно посредством други устройства с интернет връзка.

1.2.1 Технологии за IoT

Технологичния стек на „Интернет на нещата“ може да бъде разделен на четири основни слоя:

- Хардуер - устройства, които представляват „нещата“ в Интернет на нещата.
- Софтуер - ключовият елемент, който прави свързаните устройства „умни“.
- Комуникации - комуникационният слой включва както решения за физическа свързаност, така и специфични протоколи, използвани в различни IoT среди.
- Платформа - мястото, където всички тези данни се събират, управляват, обработват, анализират и представят по удобен за потребителите начин.

1.3 Приложения на IoT

Интернет на нещата намира приложение в различни области като (Табл. 1.1).

Таблица 1.1 Области и приложения на IoT (Atanasova T., 2019)

Област	Приложение
Интелигентна градска среда	Умни градове, Умни сгради, Умно осветление, Мониторинг на културно поведение, Мониторинг на начина на живот, Обществена безопасност, Системи за наблюдение чрез дронаве, Преносим личен цифров асистент
Интелигентна енергийна мрежа	Интелигентна енергийна система, Умни консуматори, Умни измервателни уреди, Зелена енергия
Умно здравеопазване	Мониторинг на здравословно състояние, Прогноза за здравословно състояние, Амбиентно подпомагане на живот, Неонатологични грижи
Умен транспорт	Контрол на трафика, Умно маршрутизиране, Умно паркиране, Разпознаване на пешеходци, Умни превозни средства, Автономни превозни средства
Умен дом	Умни домашни уреди, Умни мебели, Умни термостати, Умно контакти, Умни ключалки, Умни домашни охранителни системи

Умно земеделие	Прецизно земеделие, Умен контрол на вредителите, Умно управление на добитък, Умни оранжерии и обори, Умен мониторинг на складови съоръжения.
-----------------------	--

1.4 Предизвикателства пред IoT

На база извършени анализи от Juniper Research (Sorrell S., 2018), към този момент, като най-големи предизвикателства пред успешното внедряване на IoT устройства се очертават свързаност, скорост, съвместимост, големи хетерогенни данни, интелигентен анализ и сигурност.

1.5 Хетерогенни данни

Хетерогенните данни се характеризират с голямо разнообразие по типове стойности и формати. Те често са нееднозначни и са с ниско качество поради наличието на шум и/или липсващи стойности. Интегрирането на такива разнородни данни е труден и сложен процес. Затова в този си вид те не носят голяма полза за нуждите на бизнес приложенията.

Хетерогенността е една от основните характеристики на данните придобити от IoT системи и са основна причина за проблемите при тяхното интегриране и анализиране.

1.6 Съществуващи решения за интегриране на хетерогенни данни

Интеграцията на данни поетапно развива своите подходи от исторически използваните ръчни и автоматизирани методи за интеграция до ясни, подредени научни, технически и бизнес процеси като:

ETL (Extract Transform Load) – процес за интеграция на данни, който се отнася до три основни стъпки – извличане, преобразуване и зареждане, използвани за агрегиране на данни от множество източници (Atwal H., 2020).

ELT (Extract Load Transform) – процес за интеграция на данни в три стъпки - извличане, зареждане и преобразуване (Anoshin D., 2020).

CDC (Change Data Capture) - процес за интегриране на данни, който се основава на идентифициране, улавяне и предоставяне на промените, направени в източниците на данни (Singh J., 2019).

OSEMN (Obtain Scrub Explore Model iNterpret) – стандартизиран и широко приет процес на организация на научноизследователската дейност в областта на Data Science за интеграция на данни (Guhre S., 2019).

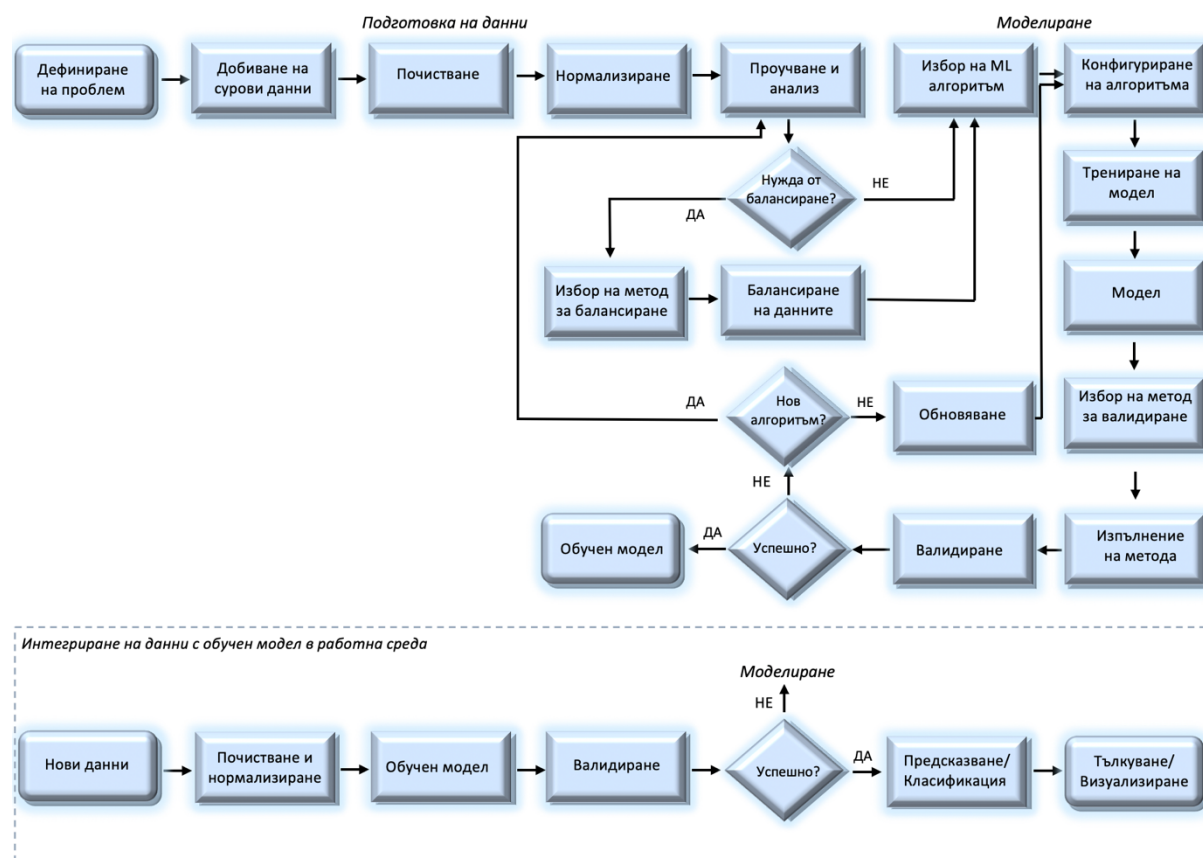
1.7 Изводи

На база направените изводи може да се обобщи, че при съществуващите проблеми за работа с хетерогенни данни от IoT системи е важно да бъдат взети навременни решения чрез създаване и използване на единен и систематизиран подход, което ще доведе до устойчиви, надеждни и възпроизводими резултати.

Глава 2 – Методология за обработка и моделиране и интеграция на хетерогенни данни

Методологията описва "общата изследователска стратегия, която очертава начина, по който трябва да се извършват научни изследвания" (Howell K., 2013) .По своята същност тя представлява система или група от методи, правила и твърдения използвани в конкретна дисциплина.

В тази глава на дисертацията е предложена методология (Задача 1), която комбинира подходи, методи, процеси, инструменти и добри практики, за да се отговори на въпроса „Как да се обработят, моделират и интегрират хетерогенни данни от разпределени IoT устройства”. Тя е приложима за всички IoT системи събиращи хетерогенни данни, които е нужно да бъдат съхранени, обработени и визуализирани. За да се добие задълбочена представа за разработената методология, на Фиг. 2.1 е представена подробна схема, в която са представени основните етапи и стъпките включени в тях.



Фигура 2.1 Подробна схема на методологията за обработка, моделиране и интеграция на хетерогенни данни

Представената методология се състои от следните етапи и придружаващи ги стъпки:

Етап 1: Дефиниране на проблем

Етап 2: Подготовка на данни:

- Добиване на хетерогенни данни от IoT устройства;
- Почистване на придобитите данни;
- Трансформация на хетерогенните данни;
- Проучване и анализ на състоянието на данните;

Етап 3: Моделиране:

- Избор на алгоритъм за машинно обучение;
- Конфигуриране на избрания алгоритъм;
- Трениране и валидиране на модела;

Етап 4: Интегриране на данни с обучен модел в работна среда:

- Добиване на нови данни;
- Почистване и трансформация на придобитите данни;
- Използване на обученния модел;
- Валидиране на резултатите;
- Тълкуване и визуализация;

Една от най-важните характеристики на методологията е, че тя описва недвусмислено не само последователността, но и отговорностите. На всеки етап от процеса е ясно какво се очаква от него, от кого той ще получи заявка за определена дейност и на кого следва да предостави резултата. Значително се улеснява идентифицирането на съществуващите добри практики. Постигането на по-голяма откритост на дейностите на всеки етап от процеса предоставя възможност за лесно откриване на допуснати грешки и бързо връщане към определен етап от процеса за отстраняването им.

2.1 Подготовка на данни

2.1.1 Придобиване на данни

Събирането на данни е първата стъпка от целия работен процес. Тя е определяща за крайния успех на зададената цел.

2.1.2 Почистване на данни

При събиране на данни от реалния свят, често се наблюдават различни причини, поради които има данни със стойност Null, NaN или NA. Често използвани подходи за решаване на този проблем са липсващите данни да бъдат заместени със средни стойности, с най-често срещани стойности или тези стойности да бъдат директно премахнати.

2.1.3 Нормализиране на данни

Поради различният вид и диапазон на данните, получени от IoT системите, е необходимо те да бъдат нормализирани. Най-популярните методи за нормализация са: Zscore, MinMax, LogNormal и TanH (Saif S., 2017) (Samariya D., 2020).

2.1.4 Проучване и анализ

Проучване, откриване, структуриране и анализ са операции, които са изключително полезни за опознаването на събраните данни. Изследването на набор от сурови данни помага избора на най-добрия подход за извършване на аналитични проучвания (Waltenburg E., 2012).

2.2 Моделиране

В парадигмата за машинното обучение, моделът се отнася до математически израз на параметрите на модела заедно с входните характеристики за всяка прогноза, клас и действие за регресионни, класификационни и подсилващи категории.

2.2.1 Анализ на алгоритми за машинно обучение

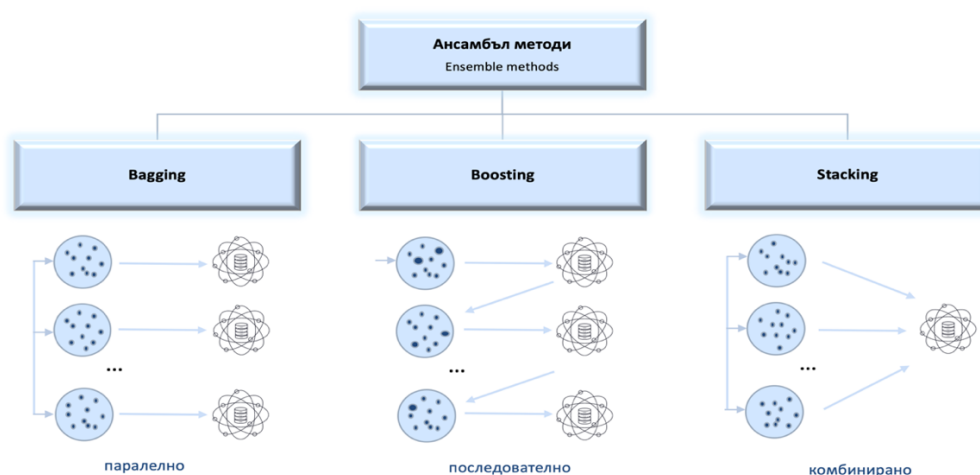
В машинното обучение съществуват различни групи алгоритми (Фигура 2.2). Избирането на подходящ алгоритъм за решаване на даден проблем зависи основно както от вида и размера на данните, така и от тяхното качество и количество (Kolchakov K., 2018).



Фигура 2.2. Типове категории на алгоритми за машинно обучение

Методи за подсилване на алгоритми

Машинното обучение и науката за данни изискват нещо повече от просто използване на съществуващи алгоритми. Един от важните моменти е да се знае кои алгоритми могат да прилагат методи за подсилване (Mease D., 2007). Тези методи се използват за създаване на така наречените ансамблови модели, които могат да подпомогнат подобряването на точността на алгоритъма и да направят модела по-стабилен. Това се постига чрез комбиниране на слаби алгоритми с ансамбъл метод (Agarwal K., 2020), (Tewari S., 2020) и (Balabanov T., 2020).



Фигура 2.3. Типове методи за подсилване на алгоритми за машинно обучение

Съществуват 3 метода за подсилване на алгоритми (фиг. 2.3):

Bagging - Предназначен да подобри стабилността и точността на алгоритмите за машинно обучение.

Boosting - Чрез този метод се редуцират пристрастията (фалшиви положителни резултати).

Stacking - Чрез този метод се подобряват предвиждания при прогнозирането.

2.2.2 Анализ на метод за валидиране

Валидирането измерва точността на модела. Тя се получава чрез разделяне на оригиналната проба в тренировъчен комплект за трениране и тест. Методите за трениране и тестване използват различни подходи за разделяне на данните (Vanwinkelen G., 2012) (Balabanov T., 2018). Тази подходи са: Train-Test Split, Cross-Validation, Early Stopping.

2.2.3 Валидиране на избрания модел

Чрез валидиране на модела се получава реална оценка за продуктивността и устойчивостта на обученения модел. То се извършва чрез различни метрики, характерни за отделните категории на обучение.

Основните метрики за класификация, чрез които се измерва ефективността на модела са базирани на матрицата на объркване. Най-често използваните са: точност, прецизност, чувствителност и средно претеглена стойност.

Основните метрики за оценка на регресионни модели са: средната абсолютна грешка, средната грешка на квадрат, относителната абсолютна грешка, относителна грешка на квадрат, средна нулева единична грешка и коефициентът на определяне.

2.2.4 Подобряване на избрания модел

Моделът, който оформя данните за обучение твърде добре, обикновено извършва прекомерно нагаждане към данните (overfitting). Подходите за справяне с прекомерно нагаждане на модела към данните са:

1. Кръстосано валидиране (Cross-validation);
2. Обучение с повече данни;
3. Минимизиране на характеристики чрез
4. Избор на променливи.

2.3 Интегриране на данни с обучен модел в работна среда

След създаване на успешен модел той бива архивиран и експортиран в структура, която съдържа обект за класификация/регресия и функция за прогнозиране. Този модел се разархивира и имплементира в работна среда, за да прави прогнози, използвайки нови данни.

Добавянето на нови данни към вече обучен модел е необходимо да се извършва при определени условия:

- Новите данни трябва да съдържат същите имена на променливите (колоните), каквито са били използвани при обучението. Функцията за прогнозиране игнорира допълнителни променливи, ако има такива.
- Форматът и видът на данните също трябва да съответстват на оригиналните данни за обучение.

2.4 Изводи

В тази глава е предложена нова систематизирана методология за обработка, моделиране и интеграция на хетерогенни данни придобити от IoT устройства. Тя описва логическа последователност от етапи и стъпки, което я прави лесна за разбиране и използване. Анализирани са предложените методи или подходи, които са най-подходящи за нейното изпълнение в зависимост от типа на разглеждания проблем и количеството и качеството на придобитите хетерогенни данни.

В резултат на това са направени следните заключения:

1. Методологията обхваща всички аспекти от работния процес – от дефиниране на проблема и поставяне на задачите, до интегриране на готово решение в работна среда.
2. Осъществяването на всеки един от етапите в методологията се извършва чрез решаване на редица по-малки задачи (стъпки), чрез което лесно може да се идентифицира наличието на проблем и да се предприемат навременни действия за неговото отстраняване.
3. Методологията е предназначена за работа с хетерогенни данни и предлага няколко метода за тяхното нормализиране и обединяване в единна хомогенна структура.
4. Методологията позволява създаване на автоматизирани процеси за работа с данни.
5. Методологията може да бъде прилагана от специалисти в различни области, работещи с хетерогенни данни.

Глава 3 - Архитектура на модулна IoT система

В тази глава на дисертацията се представя пълното разработване на система за събиране, обработка и моделиране на данни в реално или през дефинирани интервали от време. Тя е съставена от отдалечени IoT устройства (Задача 2) и разпределено облачно базирано софтуерно приложение (Задача 3).

Извличането на данни се осъществява от IoT устройства, които са интегрирани в пчелни кошери. Събраните данни се изпращат в софтуерно приложение, където данните се обработват, анализират и моделират, а резултатите могат да бъдат достъпни през потребителски интерфейси.

Мониторингът на био процесите е важен проблем и предизвикателство за изследователи и специалисти по информационни технологии. Пчеларството е един от подсекторите на селското стопанство, където могат да бъдат приложени техниките и методите за интелигентен мониторинг (Beecham, 2017). Интегрирането на информационните технологии в пчеларския процес може да подобри знанията на пчеларите за поведението на отделните пчелни колонии. Една от актуалните цели в областта на пчелния мониторинг е разработването на инструменти за непрекъснато наблюдение на пчелите в реално време, като се използват автоматични решения, избягващи излагане на пчелите на допълнителен стрес или непродуктивни дейности. Целта на тези технически средства не е да заменят, а по-скоро да подкрепят пчеларя.

3.1 Архитектура на модулна хардуерна IoT система

3.1.1 Изисквания към системата

Правилното определяне на нуждите на системата е най-важната стъпка преди планирането и разработката на хардуерната архитектура. По своята същност архитектурата се отнася до идентифицирането на физическите компоненти на системата и техните взаимовръзки.

3.1.2 Хардуерни компоненти – съвместимост и заменяемост

Съвместимостта и лесната заменяемост на хардуерните компоненти е важна характеристика на системата. Това е и едно от основните изисквания към нея, което е спазено по време на процеса на проектирането ѝ. Системата се нуждае от четири основни категории хардуерни компоненти – сензори, логически блокове, комуникационни компоненти и захранващи елементи.

3.1.3 Групиране и йерархия между хардуерните компоненти

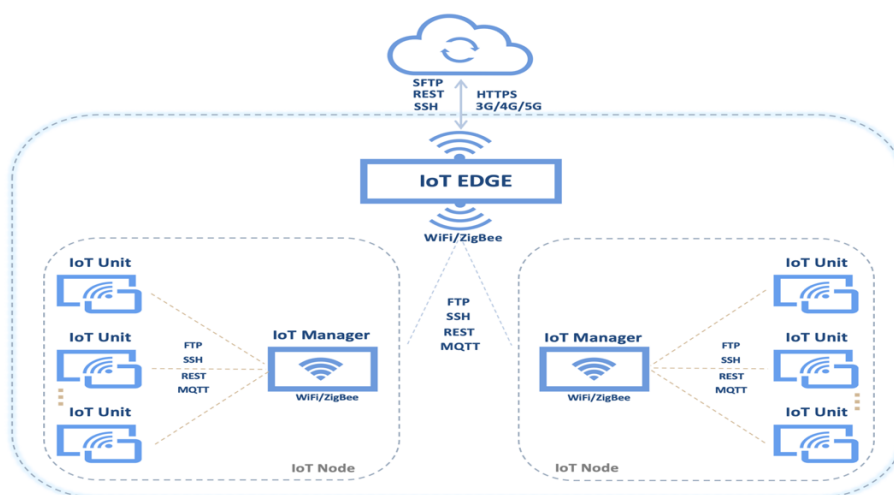
Описаните хардуерни компоненти са групирани физически и логически по специфичен за системата начин. Различните практически конфигурации на устройствата в системата позволяват да бъдат извършени разнообразни групираня, но тяхната логическа структура е постоянна.

В системата са изградени следните видове участници:

- **IoT Unit** – крайно вътрешно хардуерно устройство.
- **IoT Manager** - междинно хардуерно устройство, което притежава капацитета да служи като мениджър на няколко на брой IoT Units.
- **IoT Node** - съвкупността от всички IoT Units, за които отговаря един IoT Manager.
- **IoT Edge** – крайно външно хардуерно устройство, което е предназначено да изпълнява ролята на медиатор.

Всички IoT Nodes изпращат събраните сензорни данни и информация за състоянието си към IoT Edge. Също така IoT Edge има възможност самостоятелно да вземе данните от IoT Nodes. Това е направено с цел да се повиши стабилността при работа на системата, като се поддържа повече от един комуникационен канал между участниците в нея. От своя страна, IoT Edge на базата на дефинирани от потребителя правила, изпраща събраните данни към Облачна среда (фиг. 3.1).

IoT Edge може да получава команди от Облачната среда, тези команди могат да бъдат адресирани пряко към него, или към конкретен IoT Node. В този случай, Edge препраща информацията към конкретния IoT Node.



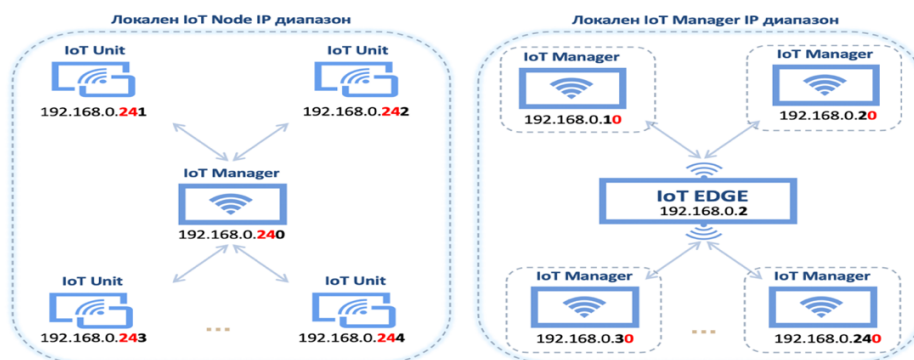
Фигура 3.1: Архитектура на модулна IoT система

3.1.4 Комуникация

Комуникацията в хардуерната система се основава на безжична свързаност (WiFi). Това е технология, която използва радио сигнали, за да предава информация между устройствата, които я поддържат.

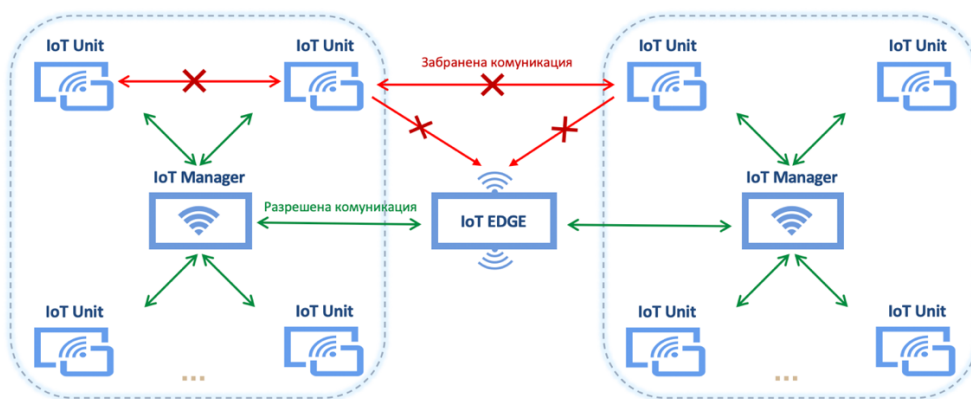
Системата се характеризира с централизирана топология в центъра на която се намира рутер, който още изпълнява ролята и на шлюз по подразбиране (Default Gateway).

За добрата организация на устройствата и семантична яснота на логическата връзка между всяко устройство с неговия IP адрес е предложен нов метод за комуникация между устройствата, основан на IP адресация сегментирана в конкретни диапазони.



Фигура 3.2: Метод за комуникация чрез IP адресация между участниците в разработената IoT системата

Така създадената системна мрежа и наложената конвенция за IP адресация, определя следните посоки комуникации, които се извършват в системата (фиг. 3.3):



Фигура 3.3: Комуникация между участниците в IoT системата

3.2 Архитектура на софтуерна система

При разработването на софтуерната система в тази дисертация е следвана утвърдената методология – Agile. Agile е гъвкава методология, при която разработката на софтуерна система се изгражда на итерации, при които изпълнението на етапите се повтаря последователно (Cohn, 2019), (Manel D., 2019).

3.2.1 Сървърна архитектура

Определяне на изискванията

Като главни изисквания към разработваната сървърната архитектура могат да бъдат определени: Услуги, Поддръжка и Гъвкавост;

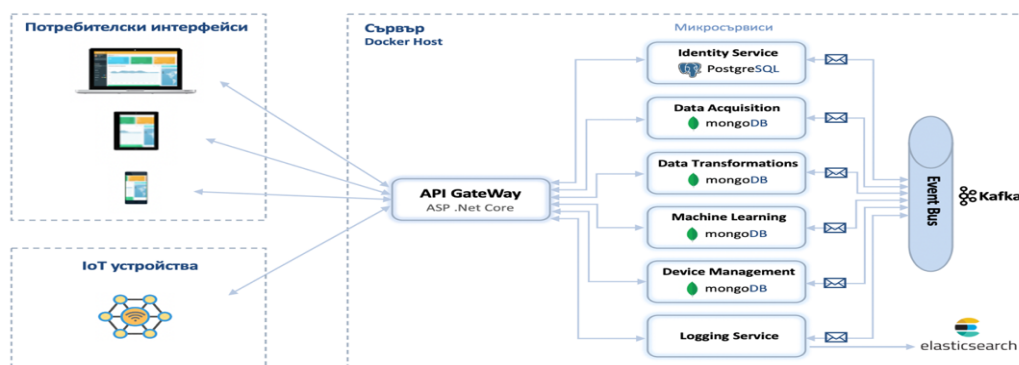
Очакваните поддържани функционалности от системата са: поддържане на потребители, работа с данни, моделиране, управление на крайни устройства и журнал;

Архитектура и Проектиране

Проектирането се отнася до създаването на софтуерна архитектура. Тя от своя страна определя цялостната структура на софтуерната система, софтуерните компоненти, свойства и отношения между тях.

След определяне на функционалните изисквания и технологичния стек за разработване на системата се определя, че най-подходяща е архитектура основана на

използването на микросървиси (Microservice architecture - MSA). Тя позволява създаване и поддържане на дистрибутирани софтуерни системи, работа с различни бази от данни и създаване на самостоятелни потребителски интерфейси, които могат да общуват със сървърната част. За изпълнение на поставените задачи е приложен нов подход за организация на услуги за интелигентна обработка на данни, състоящ се от следните реализирани микросървиси (Фиг. 3.4):



Фигура 3.4: Дизайн на предложената и реализирана в системата MSA архитектура

- **Identity Service** отговаря за всички операции, свързани с потребителя.
- **Data Acquisition** отговаря за всички data pipelines (съвкупност от всички предприети стъпки в процеса на устойчиво преместване на данни между отдалечени физически различни локации), чрез които системата събира данните от отдалечени IoT устройства.
- **Data Transformations** носи отговорност за всички трансформации върху новопостъпилите сурови данни докато те придобият нужният вид, подходящ за анализ и прилагане на модели за машинно обучение върху тях.
- **Machine Learning Service** отговаря за прилагането на готови, тренирани модели за машинно обучение за предсказване на стойности и вероятни събития.
- **Device Management Service** е отговорен за предаването на команди към IoT устройствата от потребителя.
- **Logging Service** отговаря за събирането и анализа на събитията, които са настъпили по време на работата на системата.
- **API Gateway Service** се използва за разделяне на потребителските интерфейси от системата.
- **Event Bus модул** се грижи за предаването на съобщенията до нужния получател (Alekseev I., 2015). Използваният Event Bus в системата е Apache Kafka (Hesse G., 2020) Поддържаните типове комуникации от Event Bus в системата са:

Разработване

Процесът по разработване на микросървиси започва с определяне на три основни аспекта: ограничение на контекста, определяне на типовете комуникации и консистентността на данните.

3.2.2 Потребителски интерфейс

Интерфейсът е точка на взаимодействие между потребителя и системата. Той е от важно значение за визуализацията на събраните и анализирани сензорни данни от IoT устройства и начинът, по който те биват представени на потребителя.

Определяне на изискванията

Главните изисквания по време на разработката на потребителския интерфейс са:

- Покриване на потребителски изисквания;
- Бързина при зареждане и обновяване;
- Дизайн за съвместимост с различни устройства и браузъри;
- Добре организирано и смислово подредено съдържание;
- Интуитивна и разбираема навигационна структура;
- Гъвкавост при добавяне на нови функционалности;
- Стабилност и непрекъсваемост на работа;
- Сигурност на потребителските данни;
- Лесна за поддръжка и обновяване;

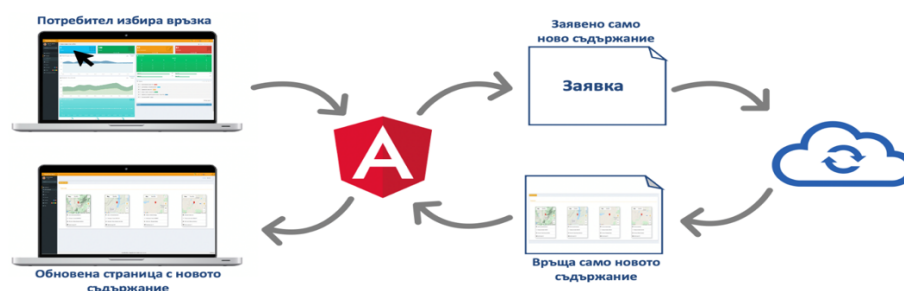
Потребителският интерфейс тематично е разделен на две части:

- Презентационен (Landing page)
- Контролен панел (Control panel)

Всички части поддържат различни функционалности, които са консистентни и се допълват взаимно.

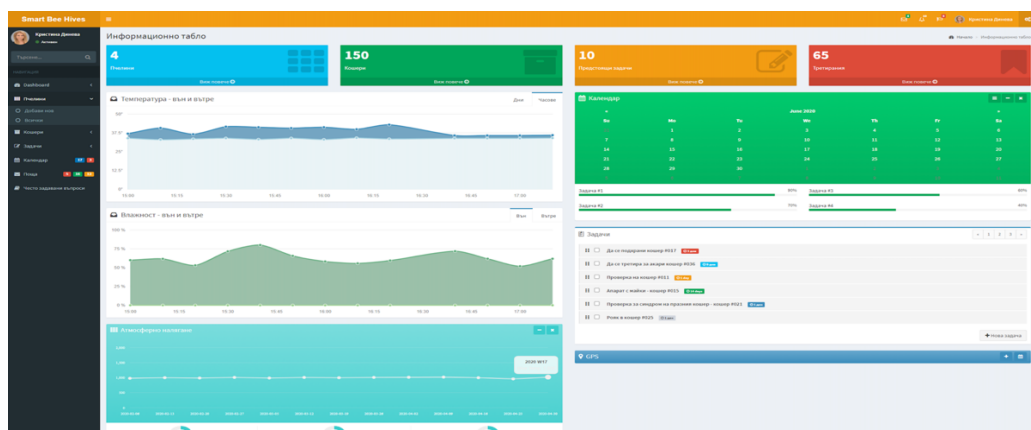
Разработка

Потребителският интерфейс в този дисертационен труд е разработен на платформата Angular, която е създадена специално за изграждане на клиентски SPA приложения (Japikse P., 2020). Наричат се SPA (Single-Page Application), защото цялото съдържание на приложението е разработено под формата на самостоятелни компоненти, които се зареждат динамично в една уебстраница, като създават впечатление, че приложението всъщност може да навигира между повече от една страница (фиг. 3.5).



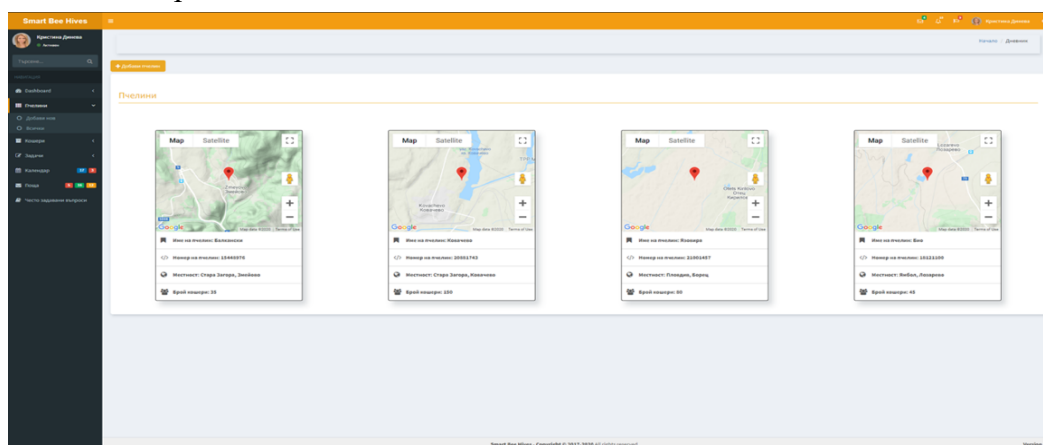
Фигура 3.5: Процес на обновяване на уеб съдържание в Angular

SPA приложенията се характеризират с редуцирано количество динамични опреснявания на страниците при получаване на данни. Това е постигнато чрез използване на AJAX комуникация със сървъра. На фиг. 3.6 „Информационно табло“ е показано уеб съдържанието, което се достъпва в приложението чрез адрес за навигация “https://smartbeehives/dashboard“. То се формира чрез извикване и зареждане на компонентите.



Фигура 3.6: Потребителски интерфейс - информационно табло

При действие от страна на потребителя да достъпи ново съдържание се изпраща заявка към сървъра за обновяване на съответните секции в страницата. Сървъра връща към клиента само новото съдържание под формата на нови компоненти и обновява страницата с тях без да бъдат повторно зареждани съществуващите елементи (фиг. 3.7), които остават непроменени.



Фигура 3.7: Потребителски интерфейс - пчелини

Този принцип на работа предоставя бързина при зареждане и обновяване на информационното съдържание и лесно добавяне на нови функционалности – чрез създаване на нови компоненти или цели модули.

3.3 Изводи

В тази глава е предложена иновативна система, изградена от IoT устройства, облачна сървърна част и потребителски уеб интерфейс, които си взаимодействат и обменят данни помежду си.

Разработената IoT хардуерна архитектура и предложеният и реализиран нов метод за комуникация между IoT устройства, основан на йерархична IP адресация, дават значителни предимства на системата:

1. Ефективна работа с различни по вид хардуерни компоненти и устройства, лесна заменяемост на вече вградените или добавяне на нови такива при нужда от промяна в мащаба или функционалностите на системата.

2. Поддръжка на различни комуникационни канали от IoT устройствата, което позволява работа с различни хардуерни компоненти в зависимост от конкретните нужди на потребителя, спецификите на терена и количеството използвани устройства.
3. Голяма част от логиката изпълнявана от устройствата може да бъде контролирана, променяна и обновявана от разстояние чрез интегрирани софтуерни подходи.
4. Справяне с по-високата консумация на енергия в системата чрез използване на хардуерни устройства за интелигентно управление на захранването и вграден часовник за реално време.
5. Хардуерната система има само една точка на достъп, което я прави с високо ниво на сигурност.

Софтуерната архитектура е от MSA тип. Тя се състои от група микросървиси, които са реализирани чрез прилагане на нов подход за организация на услуги за интелигентна обработка и обмен на данни. Това дават следните предимства на системата:

1. Съсредоточеност върху функционалностите, а не върху технологиите. Микросървисите нямат ограничения при използването на програмни езици и бази от данни за тяхното създаване и успешно функциониране. Всеки микросървис е насочен да изпълнява определена функционалност. Това прави тази архитектура адаптивна за използване в множество други процеси и в различни канали в зависимост от нуждите. Всеки микросървис е отговорен за конкретна услуга, което води до изграждане на интелигентна и многофункционална архитектура.
2. Подобрена производителност и скорост. Микросървисите могат да работят едновременно без да е необходимо да се изчакват.
3. Лесна поддръжка, гъвкавост и скалируемост на цялата система. Всеки микросървис може да се управлява независимо. При нужда и бъдещо развитие на системата може лесно да бъдат добавени нови независими едни от други микросървиси. При промяна на един микросървис или добавяне на нов не се нарушава функционирането на останалите участници в системата.
4. Автономност и мултифункционалност. Микросървисната архитектура предоставя независимост и автоматична работа на микросървисите.

Потребителският интерфейс е изграден като SPA приложение, което е построено от серия компоненти. Този вид архитектура има следните предимства:

1. Постига се висока бързина и гъвкавост чрез зареждане само на необходимото съдържание (компонент), а не на цяла страница. Непромененото съдържание се съхранява в кеша.
2. Постига се високо ниво на преизползваемост на код – създадени веднъж компоненти, могат да бъдат използвани на различни места в приложението.
3. Съществуващият интерфейс лесно може да бъде надграден и обогатен с допълнителни функционалности, като се създават нови компоненти, които ще бъдат извикани при необходимост и визуализирани в главния компонент.

Глава 4 - Експериментални резултати

В тази глава основна цел е прилагане и валидиране (Задача 4) на предложената във втора глава методология за обработка, моделиране и интеграция на хетерогенни данни получени от разпределени IoT устройства. Преминава се през всички необходими стъпки от етапите „Подготовка на данни“ и „Моделиране“ до получаване на валидни и ефективни обучени модели, които се интегрират в разработената в този дисертационен труд MSA софтуерна платформа.

4.1 Прилагане на разработената методология

Последователността на действията е важна част от изследователския процес, за да може експериментите в случай на необходимост да бъдат повторени и в бъдеще подобрили. За правилната организация на цялостния процес от зареждане на базата, през почистване, нормализиране и разделяне на данните, до прилагане, трениране и оценка на моделите, е следвана разработената методология.

След обработка на данните, от съществено значение е определянето на максимално подходящ алгоритъм за машинно обучение. Това е сложен процес, който зависи от размера, качеството и естеството на събраните данни.

Работният процес за постигане на целите на експеримента се извършва в облачна изчислителна среда - Microsoft Machine Learning Azure Studio (Abraham T., 2018), (Azure, 2020).

4.2 Избор на алгоритъм за машинно обучение

4.2.1 Класификационен анализ

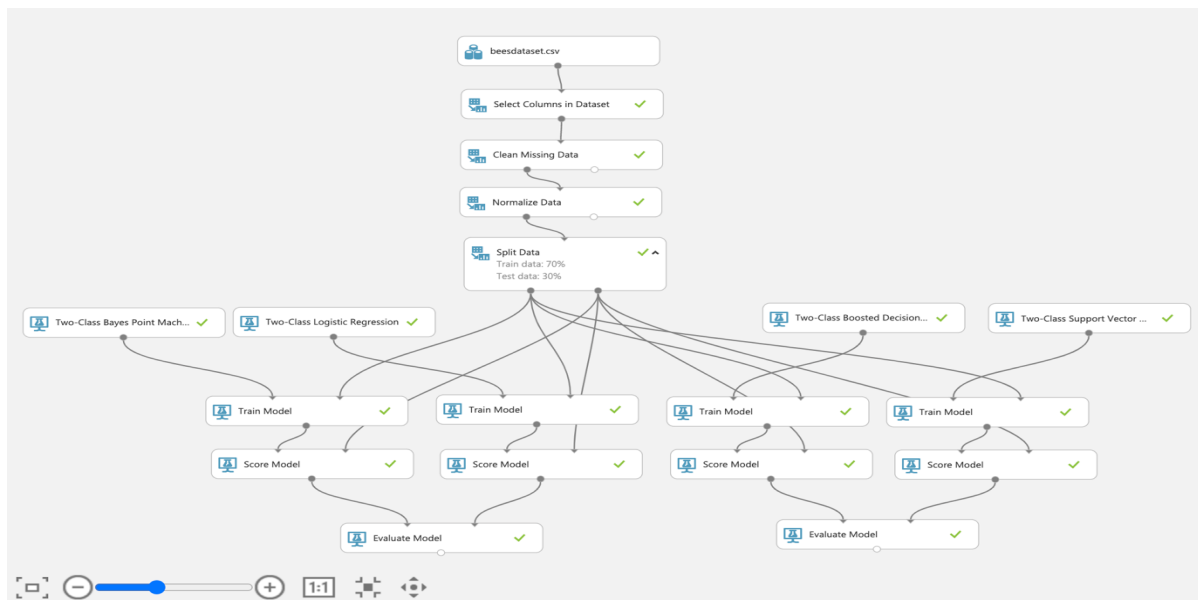
С цел получаване на валиден класификационен модел за машинно обучение се прилагат алгоритми за класификация върху хетерогенни данни от IoT устройства. Данните са придобити от център за данни за изследване на медоносни пчели “The Data Center for Honeybee Research”. Моделите биват тренирани за предсказване на един от общо два класа – 0 – пчелното семейство няма достатъчно условия да оцелее или 1 – пчелното семейство има достатъчно условия да оцелее. Тренирането се извършва на база определени параметри – вътрешна и външна температура при пчелните кошери и процент влажност т.е. ще бъде извършена биномиална класификация (binomial classification).

Трениране и тестване на модели

След зареждане на базата с данни, селектиране на необходимите колонии, почистване и нормализиране на данните, те биват разделени на две части 70:30. 70% от данните се използват за трениране на моделите, а 30% се използват за тестване. При разделянето на данните се прилага стратификация.

Прилагаме четири вида алгоритми (фиг. 4.1), избрани след извършен сравнителен анализ в Глава 2, които са:

1. Two-class Boosted Decision Tree,
2. Two-class Bayes Point Machine,
3. Two-class Logistic Regression,
4. Two-class Support Vector Machine;



Фигура 4.1 Схема на работен процес

Всеки един алгоритъм е резултат от дадена моделираща функция, която определя начина, по който работи алгоритъмът и от там и начина, по който се получават резултатите, затова е от важно значение правилната конфигурация (hyperparameter tuning) на всеки алгоритъм.

4.2.2 Оценка на моделите

За определяне успеваемостта на един модел е необходима употребата на различни метрики за оценяването му. Няма универсална формула за това – всичко зависи от конкретните данни и целите свързани с тях.

Избраните метрики за този експеримент са: точност (accuracy), прецизност (precision), чувствителност (recall) и хармонична средна стойност (F1 Score).

Изчислението на метриките за оценка на моделите е въз основа на изградените матрици за оценка на класификацията. След изчисляване на стойностите, получаваме коефициенти за всяка метрика на всеки отделен модел (таблица 4.1).

Таблица 4.1: Оценка на моделите

	Boosted Decision Tree	Bayes Point Machine	Logistic Regression	SVM
Точност	0.998	0.884	0.882	0.888
Прецизност	0.994	0.033	0.053	0.146
Чувствителност	0.993	0.004	0.008	0.017
Хармонична средна стойност	0.994	0.007	0.014	0.030

Сравнителната таблица съдържа изчислените коефициенти за всеки един алгоритъм. След извършеното обучение и тестване на моделите и разглеждането на получените от тях метрики за оценка на ефективността им, моделът Two-Class Boosted Decision Tree се откроява, като най-добър класификатор спрямо останалите. Високият резултат успеваемост на този модел, съчетан с висок резултат на чувствителност и прецизност, налага проверка на достоверността на получените резултати чрез прилагане на подходи за преодоляване наличието на недостоверни резултати от прекомерно нагаждане (overfitting).

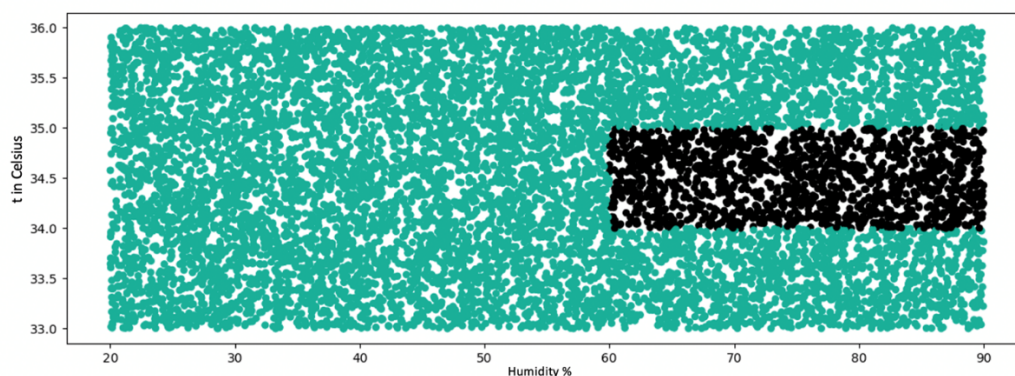
Подходи за преодоляване на прекомерното нагаждане (overfitting) на модела

Прекомерното нагаждане на модела е моделираща грешка, която възниква, когато дадена функция приляга твърде плътно към набора от данни. За да се избегне извършването на прекомерното нагаждане на модела се прилагат три основни подхода:

- Балансиране на данните;
- Избор на променливи;
- Кръстосано валидиране;

Балансиране на данните

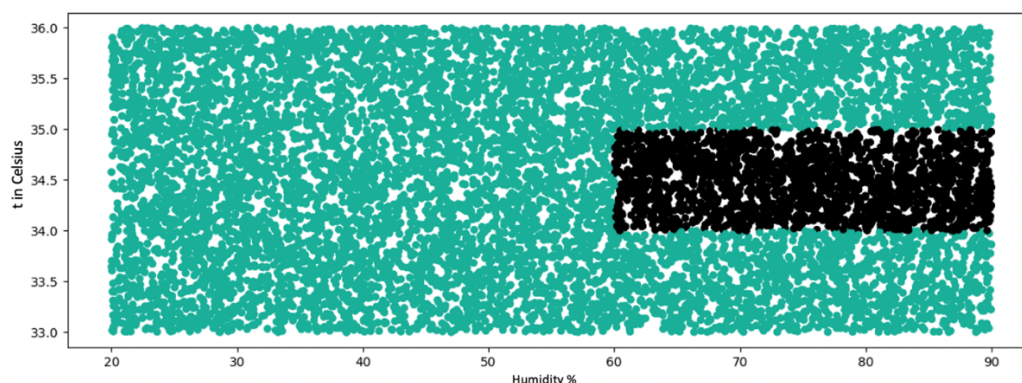
Избраният модел е Two-Class Boosted Decision Tree. С него трябва да бъде съобразен изборът на техника за балансиране на данните. Този модел е чувствителен към пропорциите на различните класове и има склонност за предпочитание към класа с най-голям дял наблюдения (наричани още мнозинство). Това често води до подвеждаща точност на модела. Затова е избрана техника за синтезиране на нови случаи на малцинство – SMOTE (Synthetic Minority Oversampling Technique).



Фигура 4.2: Разпределение на суровите данни преди прилагане на SMOTE модел

Работния процес за балансиране на данните започва с почистване на невалидни или липсващи стойности, след което се пристъпва към прилагане на алгоритъма SMOTE.

След изпълнение се образува набор от данни, който съдържа оригиналните проби плюс допълнителен брой синтетични проби от малцинството, които са създадени, така че да бъде запазено разпределението на малцинствения клас (фиг.4.3). Повишен е броят на данните без да бъдат получени нови данни, които да влияят на извличането на информация от малцинствения клас.



Фигура 4.3: Разпределение на данните след прилагане на SMOTE модел

След прилагане на алгоритъма се наблюдава увеличение на малцинствения клас от 10,39% на 37,64%, което се равнява на 27,25% повече генерирани копия на съществуващите данни. Това от една страна балансира данните, а от друга повишава размера на данните, което задължава модела да ги генерализира, за да постигне напредък. Това води до повишена точност на модела, като същевременно намалява шанса за прекомерно нагаждане.

Избор на променливи

Изборът на правилни променливи подобрява модела и предотвратява проблеми, като:

- Прекомерно нагаждане на данните.
- Данни, съдържащи излишни или неуместни променливи
- Наличие на мултиколинеарност. При мултиколинеарността наличието на две силно корелирани променливи причинява неточни изчисления на други променливи.

Поради тези причини за правилния избор на метод, определящ полезността на променливи се прилагат следните алгоритми: Корелация на Пиърсън, Корелация на Кендал и Чи-квадратна статистика.

Таблица 4.2: Получени метрики след прилагане на методи за избор на променливи

	Температура - вътре	Влажност	Температура - вън	Точност
Пиърсън	0,013381	0,569841	0,017359	0,792
Кендал	0,009175	0,458461	0,009175	0,800
Чи-квадратна статистика	0,156944	0,12615	0,000597	0,863

След прилагане на трита вида алгоритми за избор на променливи върху данните в комбинация с избрания алгоритъм *Boosted Decision Tree* за класификация на данните, може да бъде видяно от таблица 4.2, че прилагането на алгоритъма Чи-квадратна статистика извършва най-добрият избор на променливи спрямо останалите. При използването на Чи-квадратна статистика се получава 86% точност спрямо 79% точност за Пиърсън и 80% за Кендъл.

Кръстосано валидиране

Кръстосаното валидиране е важна техника в машинното обучение за оценка както на променливостта на набора от данни, така и на надеждността на всеки модел, обучен с тези данни.

Генерират се общо 10 модела по време на кръстосаната проверка, като всеки модел е обучен използвайки 8/10 от данните и е тестван върху 2/10 от тях. Когато процесът на изграждане и оценка е завършен за всички части се генерира набор от показатели за ефективност и отбелязва резултати за всяко сгъване.

Обучение и тестване на избраният модел

Важно е повторното обучение и тестване на избрания алгоритъм с цел проверка на ефективността на приложените подходи за справяне с прекомерното нагаждане на данните.

Целият работен процес се състои от осем основни стъпки.

На първа стъпка се зареждат вече балансираните данни. Новата база съдържа 27,25% новогенерирани данни. След селектиране на работните колонии, данните в тях биват почистени и нормализирани. Чрез прилагането на алгоритъма Чи-квадратна статистика се определят тежести на променливите според степента на тяхното влияние върху крайния резултат. След това алгоритъмът Boosted Decision Tree се обучава и тества, като се използва кръстосано валидиране.

Получените метрики от кръстосаното валидиране показват успеваемостта на модела да генерализира данните (таблица 4.3).

Таблица 4.3: Получени метрики след обучение и тестване на Boosted Decision Tree

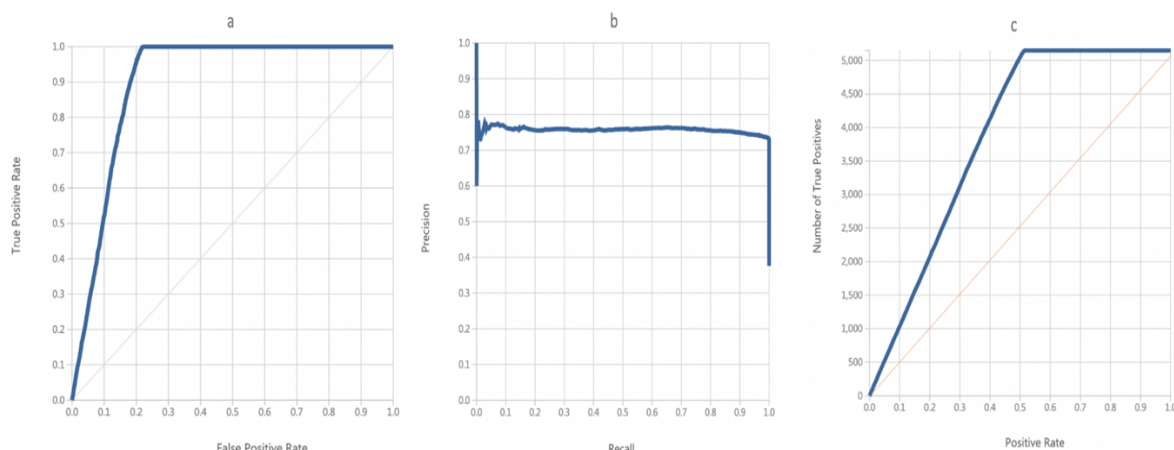
	Boosted Decision Tree
Точност	0.863
Прецизност	0.737
Чувствителност	0.989
Хармонична стойност (F1 Score)	0.845

Разпределението на резултатите биват визуализирани в три типа криви:

ROC curve – графика, показваща ефективността на модела за класификация (фиг.4.4а).

AUC (accuracy) – осигурява обобщена мярка за сигурност във всички възможни прагове за класификация (фиг.4.4b). Получените метрики показват, че прецизността на модела е в размер на 0.737, а чувствителността е в размер на 0.989. След тяхното изчисление получения коефициент е 0.902, което показва, че избраният алгоритъм е балансиран и има висок процент на успеваемост.

Lift curve – мярка за ефективността на класификационен модел, изчислен като съотношението между резултатите, получени с и без прогнозния модел (фиг.4.4с).



Фигура 4.4 Визуално разпределение на резултатите (ROC-AUC-LIFT)

Благодарение на визуалното представяне на резултатите чрез различни по тип диаграми може лесно да бъде установена ефективността на използвания модел преди да бъде приложен в реална среда.

4.2.3 Регресионен анализ

Регресионният анализ е съвкупност от статистически методи за оценка на характера на взаимоотношенията между променливите.

Прилагане на методи за избор на променливи

С цел постигане на висока точност, намаляване на шума и времето за обучение прилагаме 5 различни метода. След прилагане на различните методи (таблица 4.4) за избор на променливи се установява, че най-подходящият метод за целите на изследването е корелация на Пиърсън, която описва най-добре степента на влияние на променливите. Тя показва, че променливите величини са количествени, имат нормално разпределение и зависимостта между тях е линейна.

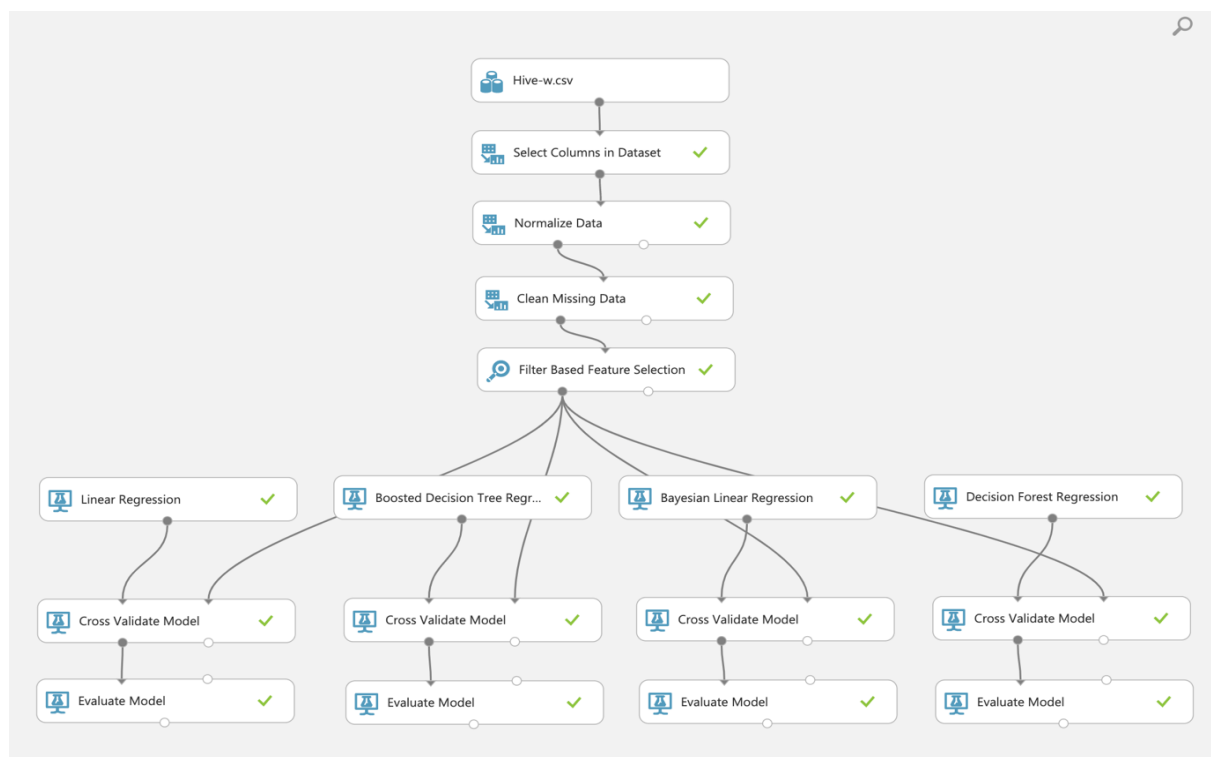
Таблица 4.4: Сравнение на получените резултати при прилагане на различни методи за избор на променливи.

	Температура	Влажност	Атмосферно налягане	Вятър - скорост	Часови диапазон
Пиърсън	0,58	0,28	0,44	0,09	0,01
Кендал	0,33	0,18	0,45	0,029	0,001
Чи-квадратна статистика	146,86	97,07	235,20	40,84	2,29
Фишър	2,45	1,84	1,06	0,92	0,91
Спирмън	0,47	0,26	0,61	0,038	0,002

Работен процес

След извършен избор на колони за работа от базата, нормализиране на данните, изчистване на липсващите данни и направен избор на променливи за работа се извършва

обучение на четири модела - Linear Regression, Boosted Decision Tree Regression, Bayesian Linear Regression and Decision Forest Regression (фиг. 4.5).



Фигура 4.5: Работен процес

Резултати от извършен работен процес

Получените резултати след извършено обучение и тестване на избраните модели са определящият фактор за правилен избор на модел. В таблица 4.5 са представени получените метрики след трениране и тестване на четири регресионни модела.

Таблица 4.5: Сравнение на получени метрики от приложените четири алгоритъма

	Linear Regression	Boosted Decision Tree Regression	Bayesian Linear Regression	Decision Forest Regression
Mean Absolute Error	0.031631	0.100482	0.066806	0.087543
Root Mean Squared Error	0.092386	0.145703	0.098254	0.138833
Relative Absolute Error	0.229573	0.729287	0.484872	0.635377
Relative Squared Error	0.209455	0.520975	0.236910	0.473003
Coefficient of Determination	0.790545	0.479025	0.763090	0.526997

Таблично представеното сравнението на получените метрики за всеки един от приложените регресионни модели показва, че линейната регресия е моделът, който най-добре описва статистическата връзка между разглежданите променливи.

4.3 Изводи

В тази глава се прилага и валидира предложената в Глава 2 методология за обработка, моделиране и интеграция на хетерогенни данни от разпределени IoT устройства. Експериментът е извършен в облачната изчислителна среда на Microsoft - Azure Machine Learning Studio. Изградени са работни схеми, съдържащи всички необходими стъпки от представената методология, които са необходими за решаването на дефинираните проблеми.

Разгледани и са решени задачи за класификация и регресия. При решението на класификационния проблем са разгледани два варианта с цел да бъдат обхванати и успешно валидирани всички стъпки от етапите „Обработка на данни“ и „Моделиране“ от методологията.

1. В резултат на извършените експерименти са направени следните заключения: Методологията е гъвкав и систематичен процес, който е необходимо да бъде следван, за постигане на правилен резултат. Пропускането на стъпка в процеса на работа води до overfit or underfit.
2. Изборът на един измежду няколко алгоритъма, решаващи една и съща задача, зависи от неговите качествени показатели. Като едни от най-важните качествени показатели на моделите се приема времето, необходимо за тяхното изпълнение, точността на получените резултати, прецизността и чувствителността на данните към промени.
3. Правилното конфигуриране на всеки алгоритъм значително увеличава точността на получения краен резултат.
4. Балансирането на данни при решаване на класификационни задачи играе ключова роля за успешното генерализиране на данните от модела.
5. Определянето на най-пряко зависещи променливи към предмета на разглежданата задача повишават точността и прецизността на модела, което води до най-добри крайни резултати.

С цел независимост на IoT системата от Azure облачни услуги за обработка и моделиране на данни, е разработено решение за валидиране на предложената методология с използване на програмен език Python. Решението следва всички стъпки представени в методологията. Моделите за машинно обучение (за класификация и регресия) и тяхната правилна конфигурация се съхраняват във файлов формат Pickle.

Глава 5 – Практическо приложение на разработената IoT система

В тази глава на дисертацията се отразява едно от многото практически приложения на предложената модулна IoT система. Разгледана е значимостта от интегриране на мониторингова система за сектора, ползите и положителните ефекти, които тя оказва върху отрасъла. Изготвено е кратко описание на съществуващите аналогични решения на пазара на база, на което е направен сравнителен анализ на тези решения спрямо разработената в този дисертационен труд система.

С цел по-лесна разпознаваемост на представената IoT система, тя е наречена SmartBeeHives.

IoT системата SmartBeeHives съчетава основните седем компонента, които ICT обхваща – софтуер, хардуер, транзакции, комуникация, данни, интернет достъп и облачни изчисления. Чрез съчетанието на тези компоненти системата извършва цялостен процес (фиг. 5.1) - събиране, предаване, обработка, анализ, моделиране, визуализиране на данни и предсказване на събития.



Фигура 5.1. SmartBeeHives работен процес

Процесът започва от имплементиране на хардуерни устройства в пчелни кошери, които събират хетерогенни данни за микроклимата в кошера и атмосферните условия извън него. Тези данни се предават към облачна среда посредством интернет. В облачната среда се извършва трансформация на данните с цел извършване на точни изчисления и получаване на полезна информация. Получената информация се визуализира в потребителски интерфейс, където потребителят може да следи развитието на пчелните семейства в кошерите и да бъде уведомен за бъдещо настъпване или моментно наличие на заплахата.

Технологиите и пчелите

Днес пчеларството е в челните редици на селскостопанските дейности, които се извършват широко в света. Поради развитието на технологиите и намаляването на

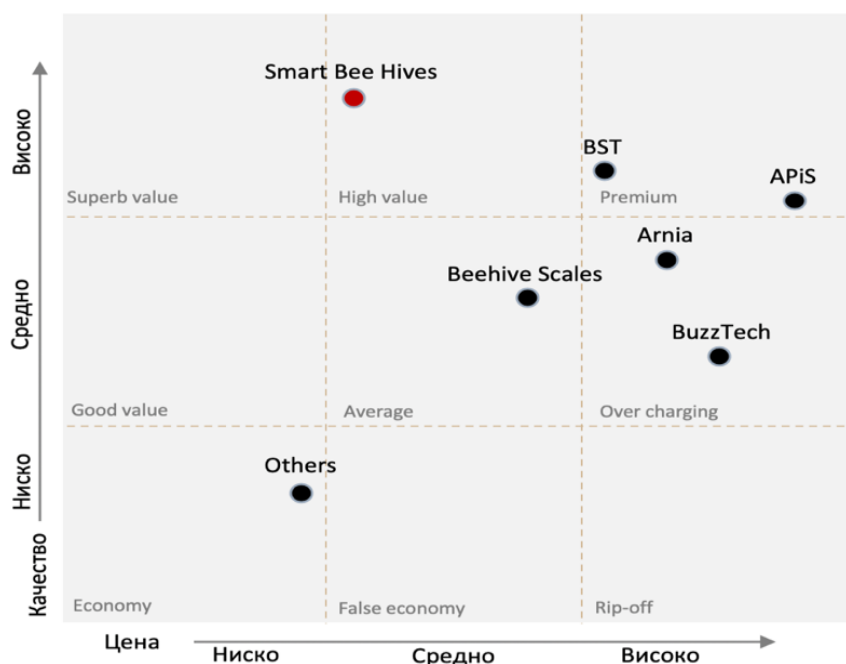
дейностите по пчеларството поради екологичните и климатични характеристики навсякъде по света, тя се превърна във важна област на научноизследователска дейност (Braga A., 2020), (Beecham, 2017).

Преглед на съществуващи системи и разработки

Все по-често последните технологични изложения обръщат внимание на „Интернет на нещата“ (IoT) като технология, която спомага за устойчивото развитие и опазване на околната среда. Мнозина проекти свързани с IoT и опазване на околната среда получават финансиране и започват своята развойна дейност. Тези проекти са свързани със земеделието и опазването на пчелните колонии. Част от тях са: Arnia, ApiS, Bee Smart Technology, BeeHive Scales и BuzzTech.

Матрица цена-качество

Позиционирането на продуктите в един от деветте квадранта на матрицата на Котлър (Kotler, 2011) показва съотношението цена-качество и как то се възприема от потребителите. Възприеманата стойност от потребителите е съществен фактор при изграждането на мнение за продукта (фиг. 5.2).



Фигура 5.2 Матрица цена-качество

Златната среда за въвеждане на нов продукт е позиционирането му в квадрант “high value” – средна цена и високо качество спрямо предлаганите на общия пазар подобни продукти. SmartBeeHives попада именно в тази златна среда, където качеството надхвърля цената. Това дава възможност на потребителите достъпно да се запознаят с продукта и да направят сравнение с продуктите позициониращи се в квадрант “Premium”, където високата цена отговаря на високо качество и от потребителска гледна точка това е полезна инвестиция.

Изводи

В тази глава е представено практическо приложение на разработената система за моделиране и интегриране на хетерогенни данни придобити от IoT устройства.

Като сфера на приложение е разгледан сектор пчеларство. Анализирана е нуждата от интегриране на нови технологии в този отрасъл и ползите, до които те ще доведат. Изготвен е сравнителен анализ между съществуващите системи на пазара, на база на който е направена икономическа обосновка на практическото приложение на системата в този сектор.

В резултат на това могат да бъдат направени следните изводи:

1. Новите технологии дават възможност на пчеларите да направят дейностите си по-ефективни и оптимизирани, като предоставят възможност за отдалечен мониторинг и ранно известяване при настъпване на отклонения в нормалните параметри в пчелния кошер.
2. Интегрираните IoT устройства позволяват събиране на данни за всички важни вътрешните параметри на пчелните семейства.
3. Съществуващите системи са насочени към големи пчелни ферми и не покриват нуждите на малки производители и любители на пчеларството. Предложената модулна система в този дисертационен труд позволява еднакво добре да бъдат задоволени функционалните изисквания както на големи производители, така и на любители.
4. Системата SmartBeeHives единствена съчетава основните седем компонента на ICT.
5. Системата позволява интегриране на вече съществуващи устройства, които да изпращат данни към софтуерната платформа с цел анализ и визуализация на данните.
6. Системата е много гъвкава поради разнообразието си от хардуерни компоненти от които тя може да бъде изградена.

Заклучение - резюме на получените резултати

В дисертационния труд е подробно анализирана значимостта на данните, придобити от разпределени IoT устройства, и тяхната роля за непрекъснато подобряване на бизнес процеси свързани с тях. Посочени са проблемни области и са дефинирани конкретни проблеми.

Предложена е нова систематизирана методология, която се състои от последователни етапи за обработка, моделиране и интегриране на хетерогенни данни. Методологията е валидирана в Облачна изчислителна среда, където са обучени и потвърдени регресионни и класификационни модели за машинно обучение, които се използват за получаване на полезни знания от събраните данни и прогнозиране на бъдещи събития.

Разработена и внедрена е обща хардуерна и софтуерна скалируема система от модулен тип за събиране и обработка на данни от разпределени IoT устройства. Тя

поддържа ефективна работа с голям брой различни по вид интегрирани хардуерни компоненти и устройства, които общуват чрез използване на различни комуникационни протоколи.

Практическите резултати от работата на системата се визуализират в специално изграден потребителски интерфейс, който представлява точка на взаимодействие между потребителя и системата.

Разгледана е сфера на практическа приложимост на разработената система и е анализирано пазарното ѝ позициониране спрямо други съществуващи системи.

С оглед на работата извършена в този дисертационен труд и резултатите, получени в хода на изследванията и изложени по-горе, могат да бъдат формулирани следните **научно-приложни приноси**:

1. Разработена е методология за интегриране на хетерогенни данни придобити от разпределени IoT устройства, която позволява обработка, моделиране и интеграция на тези данни, като са селектирани и избрани:

- методи за работа с хетерогенни данни;
- класификационни и регресионни алгоритми за машинно обучение;
- метрики за оценка и валидиране на получени прогнозни резултати.

2. Разработена е архитектура на модулна хардуерна IoT система, състояща се от сензорни, управляващи и комуникационни модули. Разработен е нов метод за комуникация между IoT устройства, основан на йерархична IP адресация с предложената логическа схема на групиране „Снежинка“.

3. Проектирана и реализирана е софтуерна MSA архитектура за съхранение, обработка и анализ на хетерогенни данни. Разработен е нов подход за организация на услуги за интелигентна обработка и обмен на данни в IoT системата, който повишава надеждността на функционирането на системата.

4. Създадени са модели за машинно обучение, които експериментално потвърждават разработената методология и са интегрирани в реализираната MSA софтуерна архитектура.

5. Показано е възможно приложение на разработената IoT система за интегриране на хетерогенни данни в интелигентно земеделие. Направен е сравнителен анализ на функционалните характеристики и пазарното позициониране на съществуващи аналогични системи, чрез който е доказана икономическата ефективност и целесъобразност на разработената IoT системата.

Насоки за бъдещи изследвания

Основните насоки за бъдещи изследвания върху тематиката на дисертацията включват:

- Интегриране на нови платформи за работа с големи данни в реално време като Hadoop и Spark.
- Изследване на процеси за подобряване на UX при използване на потребителския интерфейс през мобилни устройства.
- Изследване на сигурността на системата и анализ на потенциални нови атакуващи вектори. Интегриране на автоматизирани системи за сигурност.
- Изследване на нови хардуерни компоненти за съвместимост към съществуващите такива. Анализ и подобряване на жизнения цикъл на внедрените хардуерни компоненти.

Публикации по темата на дисертационния труд

1. **Dineva K.**, Atanasova T., Methodology for Data Processing in Modular IoT System. Distributed Computer and Communication Networks, 22-st International Conference, DCCN 2019, Springer Nature Switzerland AG 2019. V. M. Vishnevskiy et al. (Eds.): DCCN 2019, LNCS 11965, **2019**, https://doi.org/10.1007/978-3-030-36614-8_35, pp. 457 – 468, **Q2, SJR:0.283**
2. **Dineva K.**, Atanasova T., Integrated Systems With Embedded Sensors for Digital Agriculture. 19th International Multidisciplinary Scientific GeoConference SGEM 2019, 19, 6.1, SGEM, **2019**, ISBN:978-619-7408-88-1, ISSN:1314-2704, DOI:10.5593/sgem2019/6.1/S25.098, pp. 761-768, **Q4, SJR:0.232**.
3. **Dineva K.**, Atanasova T., Regression Analysis on Data Received from Modular IoT System. ESM'2019, EUROSIS-ETI, **2019**, pp. 114-120, **Scopus**.
4. **Dineva K.**, Atanasova T., Security in IoT Systems, *Proceedings 19th International Multidisciplinary Scientific Geoconference SGEM*, **2019**, Vol. 19, Informatics, Geoinformatics and Remote Sensing, Issue 2.1, ISBN 978-619-7408-79-9, ISSN 1314-2704, DOI:10.5593/sgem2019/2.1, pp. 576-577, **Q4, SJR:0.232**.
5. **Dineva K.**, Atanasova T., ICT-based Beekeeping using IoT and Machine Learning, *Distributed Computer and Communication Networks*, 21-st International Conference, DCCN 2018, Revised Selected Papers, Vladimir Vishnevskiy, Dmitry Kozyrev (Eds.), Springer, Communications in Computer and Information Science (CCIS). 919, Springer, **2018**, ISBN:978-3-319-99446-8, ISSN:1865-0929, DOI: <https://doi.org/10.1007/978-3-319-99447-5>, pp. 132-143, **Q3, SJR:0.188**
6. **Динева К., Атанасова Т.**, Подходи и методи за анализ и обработка на данни в мониторингова система за пчелни кошери, Годишник на департамент „Телекомуникации“, НБУ, **2018**, ISSN: 2534-854 X (online) No: 5, pp. 37-46.
7. **Dineva K.**, Atanasova T., Applying Machine learning against beehives Dataset. *18-th International Multidisciplinary Scientific Geoconference - SGEM 2018*, 18, 6.2, **2018**, ISBN:978-619-7408-51-5, ISSN:1314-2704, DOI:10.5593/sgem2018/6.2, pp. 35-42, **Q4, SJR:0.232**.
8. **Dineva K.**, Atanasova T., OSEMN process for working over data acquired by IoT devices mounted in beehives. Current Trends in Natural Sciences, 7, 13, University of Pitesti, **2018**, ISSN:2284-953X, pp. 47-53.
9. **Dineva K.**, Atanasova T., Computer System Using Internet of Things for Monitoring of Bee Hives, SGEM GeoConference, 27 - 29 November 2017, ISSN:1314-2704, DOI:10.5593/SGEM_GeoConference, **2017**, Vol. 17, Issue 63, pp. 169-176, **Q4, SJR:0.232**.
10. **Dineva K.**, Atanasova T., Model of Modular IoT-based Bee-Keeping System. European Simulation and Modelling Conference ESM'2017, EUROSIS-ETI, **2017**, ISBN:978-492859-00-6, 404-406, **Scopus**.
11. **Динева К.**, Аналитичен обзор на съществуващите компютърни системи за мониторинг на пчелни кошери, Международна научна конференция UNITECH'2017, 17-18 Ноември **2017**, Габрово, България, ISSN: 1313-230X, стр. II-148-II-152.
12. **Динева К.**, Интернет на нещата и устойчиво развитие на земеделието, Международна научна конференция AUTOMATICS AND INFORMATICS'2017, 4-6 Октомври **2017**, София, България, , Sofia, Bulgaria, 2017, ISSN:1313-1850, стр. 309-312.

Забелязани цитирания

I. Dineva, K., Atanasova, T.: Model of Modular IoT-based Bee-Keeping System, ESM'2017, Lisbon, EUROSIS-ETI, 404-406 (2017).

Цитиращи трудове:

1. Todor Balabanov, Ivan I. Blagoev, Zornitsa Atanassova, Greedy Genetic Algorithm Hybrid Solution of 1D Stock Cutting Problem, International Scientific Conference UNITECH 2018, November 2018, Gabrovo, Bulgaria, ISSN 1313-230X, pp.307-312.
2. S. Šakanović, N. Dogru, D. Kečo and J. Kevrić, "Short-Term Prediction of Honey Production in Bosnia and Herzegovina using IoT," *2019 8th Mediterranean Conf. on Embedded Computing (MECO)*, Budva, Montenegro, 2019, pp. 1-4. doi: 10.1109/MECO.2019.8760012.
3. Johannes Pirhonen. "A data transmitter using the GSM network", Bachelor's thesis, University of Tampere, Finland, Bachelor's Degree Program in Information and Electrical Engineering, Electrical Engineering, December 2019.
4. Riste Poposki, Dejan Gjorgjevikj, Precision Apiculture - IoT System for remote monitoring of honeybee colonies, 17th International Conference on Informatics and Information Technology - CIIT2020, Ss. Cyril and Methodius University in Skopje, 2020.

II. Dineva, K., Atanasova, T.: Computer System Using Internet of Things for Monitoring of Beehives, Vienna, SGEM GeoConference, vol. 17, Issue 63, 169-176 (2017).

Цитиращи трудове:

5. Balabanov T., I. I. Blagoev, Z. Atanassova, Greedy Genetic Algorithm Hybrid Solution of 1D Stock Cutting Problem, International Scientific Conference UNITECH 2018, November 2018, Gabrovo, Bulgaria, ISSN 1313-230X, pp.307-312.
6. Braga, A.R., Rabelo, J.C., Callado, A.C., da Rocha, A.R., Freitas, B.M., Gomes, D.G. Beenotified! a notification system of physical quantities for beehives remote monitoring | [Beenotified! um sistema de notificação de grandezas físicas para monitoramento remoto de Colmeias de abelhas], *Revista de Informatica Teorica e Aplicada*, 27(3), pp. 50-61, 2020

III. Dineva, K., Internet of Things in Help of Sustainable Agricultural Development, International Conference AUTOMATICS AND INFORMATICS'2017, 4-6 October 2017, Sofia, Bulgaria, John Atanasoff Society Of Automatics And Informatics, ISSN:1313-1850, pp.309-312, (2017).

Цитиращ труд:

7. Atanasova T., N. Bakanova, I. Blagoev, Analysis of Data from OIS To Discover and Model Process-Oriented Information, NVU "Vasil Levski", 14-15 June 2018, Tom 9, pp. 106-111.

IV. Dineva, K., Atanasova, T., OSEMN Process For Working Over Data Acquired By IoT Devices Mounted In Beehives, Current Trends in Natural Sciences, 7, 13, University of Pitesti, 2018, ISSN:2284-953X, 47-53, <http://www.natsci.upit.ro>.

Цитиращи трудове:

8. Jae Deok Son, Sooho Lim, Dong-In Kim, Giyoung Han, Rustem Ilyasov, Ural Yunusbaev and Hyung Wook Kwon, Automatic Bee-Counting System with Dual Infrared Sensor based on ICT, *Journal of Apiculture* 34(1): 47-55 (2019) DOI: 10.17519/apiculture.2019.04.34.1.47
9. Blagoev I., Application of Time Series Techniques for Random Number Generator Analysis, Proceedings of XXII Int. Conference DCCN 2019, September 23-27, 2019, Moscow, Russia, pp. 437-446. ISBN 978-5-209-09683-2, 2019.
10. Yadhunath R. , S. Srikanth, A. Sudheer and S. Palaniswamy, "Identification of Criminal Activity Hotspots using Machine Learning to Aid in Effective Utilization of Police Patrolling in Cities with High Crime Rates, " 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), IEEE, Bengaluru, India, 2019, pp. 1-6.

11. Braga, A. R., Gomes, D. G., Rogers, R., Hassler, E. E., Freitas, B. M., Cazier, J. A. "A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honeybee colonies", *Computers and Electronics in Agriculture*, Volume 169, February **2020**.
12. Braga A.R., Gomesa D. G., Freitas B.M., Cazier J.A. "A cluster-classification method for accurate mining of seasonal honey bee patterns", *Ecological Informatics*, Elsevier, **2020**
13. Braga, A. R., Modelos de Classificação Para Predição do bem Estar de Colônias da Abelha Apis Mellifera, Doutorado em Engenharia de Teleinformática, Fortaleza **2020**
14. Da Silva, Daniel; Rodrigues, Ícaro; Braga, Antonio; Nobre, Juvêncio; Freitas, Breno; Gomes, Danielo. An Autonomic, Adaptive and High-Precision Statistical Model to Determine Bee Colonies Well-Being Scenarios. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA), 11, 2020, Evento Online. Anais do XI Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais. Porto Alegre: Sociedade Brasileira de Computação, June **2020**, p. 31-40. ISSN 2595-6124. DOI: <https://doi.org/10.5753/wcama.2020.11017>.

V. Dineva, K., Atanasova, T. Applying machine learning against beehives dataset. SGEM 2018, 18, 6.2, SGEM 2018, (**2018**), ISBN:978-619-7408-51-5, ISSN:1314-2704, DOI:10.5593/sgem2018/6.2, 35-42.

Цитиращи трудове:

15. Braga, A. R., Gomes, D. G., Rogers, R., Hassler, E. E., Freitas, B. M., Cazier, J. A. "A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honeybee colonies", *Computers and Electronics in Agriculture*, Volume 169, February **2020**.
16. Braga, A. R., Modelos de Classificação Para Predição do bem Estar de Colônias da Abelha Apis Mellifera, Doutorado em Engenharia de Teleinformática, Fortaleza **2020**

VI. Dineva K., Atanasova T., Security in IoT Systems, SGEM 2019, Vol. 19, Informatics, Geoinformatics and Remote Sensing, Issue 2.1, ISSN 1314-2704, pp. 576-577, (**2019**)

Цитиращ труд:

17. P. Panev, S. Dimitrov, Innovative Technology for Increasing the Efficiency in Tubular Furniture Production Machine, 8th International Conference on Advanced Technologies (ICAT'19), August 26-30, 2019, Sarajevo, Bosnia and Herzegovina, E-ISBN: 978-605-68537-4-6 pp. 338-341, **2019**.

VII. Dineva, K., Atanasova, T. Regression analysis on data received from modular IOT system. ESM'2019, EUROSIS-ETI, (**2019**), ISBN: 978-9492859-09-9, pp. 114-120.

Цитиращ труд:

18. Tomov, P., Zankinski, I., Balabanov, T. "Training of Artificial Neural Networks for Financial Time Series Forecasting in Android Service and Widgets", Scientific journal "*Problems of Engineering Cybernetics and Robotics*", vol. 71, pp. 50-56, **2019**.

VIII. Dineva, K., Atanasova, T. ICT-based Beekeeping using IoT and Machine Learning. Distributed Computer and Communication Networks, 21-st International Conference, DCCN 2018, 919, Springer, (**2018**), ISBN:978-3-319-99446-8, ISSN:1865-0929, <https://doi.org/10.1007/978-3-319-99447-5>, pp. 132-143.

Цитиращ труд:

19. Pešović, U., D. Marković, S.Đ. Đurašević. "Remote monitoring of beehive activity". *Acta Agriculturae Serbica*, Vol. XXIV, 48 (**2019**); 157-165

Награди

1. Награда на ИИКТ-БАН за отлични научни постижения през 2019 г. в категория „Докторанти“.

Библиография

1. **Cohn J.** Scrum Mastery + Agile Leadership: The Essential and Definitive Guide to Scrum and Agile Project Management [Book]: Jeff Cohn, 2019.
2. **Manel D. Fraser S.** Agile Processes in Software Engineering and Extreme Programming – Workshops [Conference]: 20th International Conference on Agile Software Development. - Montréal : Springer, 2019.
3. **R. Hertzog R. Mas** The Debian administrator's handbook: Debian Jessie from Discovery to Mastery [Book]: Freexian, 2015.
4. **Alekseev I. Nikitinskiy M.** Eventbus module for distributed openflow controllers [Conference]: 17th Conference of Open Innovations Association (FRUCT), IEEE, 2015.
5. **Hesse G. Matthies Ch., Rabl T., Uflacker M.** How Fast Can We Insert? A Performance Study of Apache Kafka [Journal]: Cornell University, 2020.
6. **Japikse P. Grossnicklaus K., Dewey B.** Building the Spy Store Web Application with Angular [Journal]: Berkeley, CA : Apress, 2020.
7. **Brada A. Gomes D., Rogers R., Hassler E., Freits B., Cazier J.** A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies [Journal]: Elsevier, 2020. - Vol. 169.
8. **Beecham Research** Smart Farming: The sustainable way to food [Report]: Beecham Research, 2017.
9. **Kotler Ph.** Philip Kotler's Contributions to Marketing Theory and Price [Journal]: Emerald Group Publishing limited. - 2011. - pp. 87-120.
10. **Jazdi N.** Cyber physical systems in the context of Industry 4.0 [Journal]. - Romania : IEEE International Conference on Automation, Quality and Testing, Robotics, 2014.
11. **Atanasova T.** Methods for Processing of Heterogeneous Data in IoT Based Systems [Conference]: DCCN. - Moscow : Springer, 2019. - pp. 524-535.
12. **Sorrell S.** The Internet Of Things: Consumer, Industrial & Public Service 2018-2023 [Report]: Juniper Research, 2018.
13. **Atwal H.** DataOps Technology. In: Practical DataOps [Book]: Apress, 2020.
14. **Anoshin D. Shirokov D., Strok D.** Getting Started with Cloud Analytics, Jumpstart Snowflake [Book]: Apress, 2020.
15. **Singh J. Saravanan B., Prased V.** [Patent]: 16/147, 976. - US, 2019.
16. **Guhr S. Martenson J., Laser H.** Data Science as a Service - Prototyping for an Enterprise Self-Service Platform for Reproducible Research [Conference]: IARIA - The Fifth International Conference on Fundamentals and Advances in Software Systems Integration: FASSI, 2019.
17. **Howell K.** An introduction to the philosophy of methodology [Book] : SAGE, 2013.
18. **Saif S. Garba A., Awwalu J., Arshad H., Zakaria L.** Performance Comparison of Min-Max Normalisation on Frontal Face Detection Using Haar Classifiers [Journal]: PERTANIKA, 2017.
19. **Samariya D. Aryal S., Ting K.** A new effective and efficient measure for outlying aspect mining [Journal]: Cornell University, 2020.
20. **Waltenburg E. McLauchlan W.** An Introduction to Exploratory Data Analysis [Book Section]: Exploratory Data Analysis: A primer for undergraduates. Purdue e-Pubs, 2012.
21. **Kolchakov K. Monov V.** Query Conflicts In An Algoorithm For NoN Conflict Scheduling For Crossbar Commutator [Conference] : International Conference Automatics and Informatics`2018. - Sofia : SAI, 2018.
22. **Mease D. Wyner A. J. and Buja A.** Boosted Classification Trees and Class Probability Estimation [Journal]: JMLR, 2007.

23. **Agarwal K. Uniyal P., Virendrasingh S., Duff V.** Spam Mail Classification using Ensemble and Non- ensemble Machine Learning Algorithms [Book Section]: Springer SIST, Springer, 2020.
24. **Tewari S.** Ensemble Methods: An intelligent Modeling Approach [Conference]: EAGE Annual 2020, EAGE, 2020.
25. **Balabanov T. Ivanov S., Ketipov R.,** Solving Combinatorial Puzzles with Parallel Evolutionary Algorithms [Conference]: Large-Scale Scientific Computing: 12th International Conference, LSSC 2019, Springer, 2020.
26. **Vanwinkelen G. Blockeel H.** On estimating Model Accuracy with repeated Cross-Validation [Conference]: Belgian-Dutch Conference on Machine Learning, 2012.
27. **Balabanov T. Atanasova T., Blagoev I.,** Activation Function Permutation for Multilayer Perceptron Training [Conference]: International Conference on Big Data, Knowledge and Control Systems Engineering, Sofia : IEEE, 2018.
28. **Abraham T. Parasher Sh.** Hands-on Machine Learning with Azure [Book]: Packt, 2018.
29. **Azure** Microsoft Azure Machine Learning Studio [Online], 2020. <https://studio.azureml.net/>.
30. **Braga A. Gomes D., Rogers R., Hassler E., Freits B., Cazier J.,** A method for mining combined data from in-hive sensors, weather and apiary iinspections to forecast the health status of honey bee colonies [Journal]: Elsevier, 2020. - Vol. 169.

Оформление: GOST стандарт