Advances in Science, Technology & Innovation
IEREK Interdisciplinary Series for Sustainable Development

Fadi Al-Turjman *Editor*

The Smart IoT Blueprint: Engineering a Connected Future

Guiding Principles and Practical Strategies for Seamless Integration

IEREK RESEARCH & KNOWLEDGE ENRICHMENT

Springer

Advances in Science, Technology & Innovation

IEREK Interdisciplinary Series for Sustainable Development

Advances in Science, Technology & Innovation (ASTI) is a series of peer-reviewed books based on important emerging research that redefines the current disciplinary boundaries in science, technology and innovation (STI) in order to develop integrated concepts for sustainable development. It not only discusses the progress made towards securing more resources, allocating smarter solutions, and rebalancing the relationship between nature and people, but also provides in-depth insights from comprehensive research that addresses the **17 sustainable development goals (SDGs)** as set out by the UN for 2030.

The series draws on the best research papers from various IEREK and other international conferences to promote the creation and development of viable solutions for a **sustainable future and a positive societal** transformation with the help of integrated and innovative science-based approaches. Including interdisciplinary contributions, it presents innovative approaches and highlights how they can best support both economic and sustainable development, through better use of data, more effective institutions, and global, local and individual action, for the welfare of all societies.

The series particularly features conceptual and empirical contributions from various interrelated fields of science, technology and innovation, with an emphasis on digital transformation, that focus on providing practical solutions to **ensure food, water and energy security to achieve the SDGs.** It also presents new case studies offering concrete examples of how to resolve sustainable urbanization and environmental issues in different regions of the world.

The series is intended for professionals in research and teaching, consultancies and industry, and government and international organizations. Published in collaboration with IEREK, the Springer ASTI series will acquaint readers with essential new studies in STI for sustainable development.

ASTI series has now been accepted for Scopus (September 2020). All content published in this series will start appearing on the Scopus site in early 2021.

Fadi Al-Turjman
Editor

The Smart IoT Blueprint: Engineering a Connected Future

Guiding Principles and Practical Strategies for Seamless Integration

Springer

Editor
Fadi Al-Turjman
Artificial Intelligence, Software, and Information
Systems Engineering Departments
Research Center for AI and IoT
AI and Robotics Institute
Near East University
Nicosia, Mersin 10, Turkey

ISSN 2522-8714 ISSN 2522-8722 (electronic)
Advances in Science, Technology & Innovation
ISBN 978-3-031-63102-3 ISBN 978-3-031-63103-0 (eBook)
https://doi.org/10.1007/978-3-031-63103-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

HR Tracking System in the AI and IoT Era

Abdulmunaim Saleh, Ibrahim Ame, and Fadi Al-Turjman

Abstract

The goal of this project is to create a thorough human resources (HR) tracking system that will give CEOs access to up-to-date information on the performance and advancement of their employees by the use of artificial intelligence and internet of things. Using modern technology, the system will collect and evaluate information about individual contributions, project milestones, and key performance metrics. Excellent monitoring systems will provide the CEO with a comprehensive picture of the productivity of the staff, allowing for data-driven decision-making to raise total organizational productivity. CEOs will have an easy-to-use interface to monitor and evaluate employee performance using the platform's customized reports, interactive visualizations, and user-friendly dashboards. This HR monitoring system, which prioritizes efficiency and openness, aims to improve communication, encourage responsibility, and support the company's strategic use of human resources.

Keywords

Artificial intelligence · Human resource · Chief executive officer

A. Saleh (✉) · I. Ame
Department of Artificial Intelligence Engineering, AI and Robotics Institute, Near East University, Nicosia, Mersin 10, Turkey
e-mail: abdulmunim.swaleh20@gmail.com

F. Al-Turjman
Artificial Intelligence, Software, and Information Systems Engineering Departments, AI and Robotics Institute, Near East University, Nicosia, Mersin 10, Turkey
e-mail: Fadi.alturjman@neu.edu.tr

F. Al-Turjman
Research Centre for AI and IoT, Faculty of Engineering, University of Kyrenia, Kyrenia, Mersin 10, Turkey

1 Introduction

Personnel management and employee activity monitoring are important tasks for any firm in the modern world, where technology is advancing quickly. Not only can affect staff monitoring speed operations, but it also guards against potential violations and wasteful use of work time. An intelligent monitoring and tracking service is a comprehensive software solution designed to observe and analyze staff activity. It may include various functions, such as constant and systematic tracking of working hours and the completion of production tasks. Such a service's main objective is to analyze an employee's activities at work and assess their efficiency (Nazarov and Nazarov 2023).

The track HR is a system that helps you gain a real-time insight into the performance of your employees. Because as a CEO or a managing figure in a particular institution you can be able to track employees' productivity especially tracking the employees time spend on a particular task (https://clickup.com/blog/employee-productivity-tracking-tools/). This kind of system can be used in various environments, such as hospitals, where the hierarchy will be different because the doctor will have full authority over all hospital staff (https://clickup.com/blog/employee-productivity-tracking-tools/) It can also be used in schools or other educational institutions where the principal or head teacher will have full access to this system that tracks all teachers also can be used in publishing institutions (https://trackobit.com/glossary/employee-tracking-software). This demonstrates the significance of this initiative in our daily lives.

This project will be able to track employee's progress and also predict how a particular company will do in the future if it will be successful or not according to the way the employees are set to complete their task. This will be done by the use of AI and IOT. This system is really helpful for organization to monitor their employees' productivity and can help lots of companies to thrive in future.

F. Al-Turjman (ed.), *The Smart IoT Blueprint: Engineering a Connected Future*, Advances in Science, Technology & Innovation, https://doi.org/10.1007/978-3-031-63103-0_1

2 Related Works

In this particular paper (Nazarov and Nazarov 2023) overseeing working hours of employees entails handling security issues as well as monitoring and controlling work. Measures for staff control include efficiency classification, process enhancement, and process adherence verification. An intelligent service that gathers and analyzes data on computer activities with an emphasis on screen monitoring is one suggested solution to the problem of tracking working hours based on specific tasks and goals specified by management. They concluded that thus, the creation of an intelligent service to track workers' working hours will automate activities related to managing of people, enhance the precision and dependability of work time data, and guarantee a more effective use of the company's time and resources.

In this research paper (Tamim et al. 2021) a thorough explanation of Class Insight can be found in the student monitoring system. It provides an explanation of the system's functions and goals. It also explains how the system operates, including its interfaces, how it operates under restrictions, and how it responds to outside stimuli. With the help of this machine learning-based student tracking and monitoring system, teachers can electronically submit assessments to pupils. It offers resources to help teachers and students manage their reading lists, assignments, and other duties. As kids read, the app will monitor their face and eyes and update their progress automatically. Instructors are able to monitor task updates in real time as a result. They will also be informed of how long the pupil spent on each reading material page and whether or not the picture shows the student's face. A report will be produced from this.

In this research (Cotton et al. 1981) the PROCTOR system, an online interactive tool for monitoring and evaluating students in PSI-format courses, was explained. With this method, a tester module takes over a lot of the regular managerial responsibilities that are often given to the student proctor, including administration and scoring of the quiz, performance evaluation, and thorough record keeping. Additionally, PROCTOR comes with a multifunctional editor that lets a teacher input, review, or update any data found in the disk files the tester accesses. Aspects about the utilization and execution of this system are examined, along with its benefits for staffing, quality assurance, and management in PSI courses.

This study (Kis and Alexandru 2015) aims to describe an application based on RFID technology. RFID technology is used in this application to provide an easy-to-use and practical online item tracking solution. It offers a user-friendly web interface and adheres to the same procedures and guidelines as a track and trace software in the real world. It allows the user to search, purchase, and follow the item's whereabouts in real time. It is also easily customizable for a variety of uses, including personal, business, and military.

Using Android smartphones, in study (Nirmal et al. 2016) created an employee surveillance system that included GPS tracking and monitoring with 3G connectivity in this study. The system monitors calls, data consumption, web activity, and document revisions made by employees. Its objectives are to limit the improper use of resources during business hours and to stop the unlawful distribution of confidential company information. With the aid of technology, managers may evaluate staff commitment and advance the organization.

In this research survey (Thompson and Molnar 2023) on Employee Monitoring Applications (EMAs) in Canada, they were able to find a surge in adoption, driven by the need to manage remote workforces during the pandemic. Despite companies justifying EMAs for security, their analysis in this paper reveals that a disconnect between these justifications and their actual use. While security concerns were cited, current application focuses more on coercion and control for productivity and efficiency, rather than care. Many believe that EMAs violate privacy, which undermines employee-management trust according to this survey research.

In study (Kalyani and Bangar 2015) developed an Android-based employee monitoring system with admin and employee applications connected to a centralized server. The system records every action taken by employees, including sending and receiving SMS messages, calling records, whereabouts, data consumption, and web surfing. 3G network technology facilitates user-friendly communication. By saving time, lowering effort, and stopping illegal usage of business phones intended for office use, the system seeks to increase management efficiency.

According to this study (Tomczak et al. 2018) its objective is that the real-time data that Electronic Performance Monitoring (EPM) systems collect can be used for performance appraisal, training and development, logistical tracking, wellness programs, employee safety, and more. Despite the organizational benefits of EPM, these systems can have adverse effects on employee satisfaction, organizational commitment, fairness perceptions, and employee behavior. This research provides evidence, however, that these downfalls can be mitigated by implementing these systems with employee attitudes and privacy perceptions in mind. So using theory and empirical research evidence, they offer five recommendations for maximizing the positive effects and minimizing the negative effects of EPM.

This study (Peshave 2014) aimed at analyzing the productivity management system adopted by the hotels and the

challenges faced by them is implementation of the same. Although it looks simple from the face of it, productivity management is a big challenge to organizations especially when the product is in the form of a service. The characteristics of service industry make productivity management in such industries more difficult and challenging. Hospitality industry being a part of such a service industry faces a similar problem. Problems pertaining to measurement of productivity, lack of awareness about the importance of tracking employee productivity, lack of knowledge and initiative of management in implementation of the Employee Productivity Management System (PMS) are a few challenges that are faced by the hospitality industry in implementing the PMS. This study aimed at analyzing the challenges faced by the hospitality industry in implementation of employee productivity management system and to identify the methods to overcome them for effective implementation of the PMS in hotels.

This paper (Enefiok and Uzochukwu 2016) describes the development of an employee tracking system for Android that allows for thorough activity monitoring of employees. The system includes time-off requests and schedule information, which are essential parts of personnel tracking that let managers know when staff members are due at the office or other designated areas. For managers looking for effective staff monitoring via mobile phones, the method is quite helpful.

Kasliwal et al. (2016) developed an Android-based employee tracking and monitoring system. They offered various security profiles on the same smartphone for their study. They made use of a dynamic database tool, which pulls information or data from a centralized database. When an employee enters the company's grounds, they give him a different mode. All employee phone information, including SMS history, incoming and outgoing calls, employee locations, data consumption, web browser history, and details of unauthorized call history, is tracked via smart phones. It is required that employees possess an Android phone, and manager activities are also observed.

Priti et al. (2015) worked on utilizing an Android application to monitor the employee's smartphone. The software on this system is operated by Android-based mobile phones. The employee's mobile device should be Android-based, but managers are free to use any sort of device as long as they have an SMS format for receiving alerts from employees. Conveniently, the alerts are also kept on a centralized server, together with information about incoming calls, texts, and multimedia communications, as well as timely updates on the employee's whereabouts and attendance. A manager may thereafter get the information on their employee's cell activity by logging into the central server. The managers can use this technology to keep an eye on their employees using their smartphones.

In this study (Sultana et al. 2015) a smart location-based time and attendance tracking system using an Android application was developed. They suggested a time- and location-based smart system. Attendance tracking system that runs on a smartphone application for Android lowers the requirement for a second biometric scanner device. An organization's exact location can be found using GPS technology. The GPS on each employee's smartphone can pinpoint their whereabouts.

3 Methodology

The system is built using an array of languages, tools, and techniques to achieve its aim. We have the front end built with HTML and CSS for the UI views, and JavaScript for the events and AJAX calls to the backend. On the backend we are using PHP to communicate with our MySQL database where we house all the tasks, details about users and notification data.

The front end consists of a dashboard that conforms to the user who is logged in. When a CEO or an administrator logs in, they are presented with a full view of links that will let them view all the details from top to bottom. This includes managing their settings as well as seeing supervisors, managers, and workers underneath them, and being able to create new projects and tasks for them at will.

3.1 Hierarchy and Access

The side panel of the dashboard morphs for different users. We created a number-based role system, where we can trickle down the actions for each user. The lower the number, the more important you are in the system, and the more you can do. The number zero is reserved for admins who can oversee major settings in the system. The number one would be for the CEO and it trickles down to the number 4 being the workers at the bottom who only get to see the tasks they are supposed to do.

3.2 Projects, Tasks and Collaborators

The UI consists of a top navigation bar with quick links to "Add Project", "Add Subtask", "View Projects and Tasks" as well as "View Gantt Chart". These links enable quick access to popups or views to be able to add new details quickly or view them on command. Projects, tasks, and their corresponding subtasks share the same structure in our database. They are inherently the same thing, using abstraction to render them differently. The goal of our system was to create an infinitely based task system, where you could

subtask a task over and over as many times as you liked. We didn't want to limit the managers of a project. Some tasks could require only 3 levels of sub-tasking, whilst others required maybe 5–10 layers. Thus, we developed a linked-list approach with our fields in the database by linking a task to its parent task and so on. If a task has a parent task ID of zero, then that means it is the project itself and is at the top. When the task has a specific ID that isn't zero, then that means it's underneath another task. This enables us to link those tasks together and be able to traverse a long list of hierarchies very efficiently by just calling the children which contain certain parent task Identifiers.

When adding a new task, you will be presented with input boxes for adding the task name, the parent task name for subtasks, the date range for how long the project/task should last, as well as the collaborators who will work on it. The collaborators are bubbled through a filter list, that enables the user to search for them quickly. Once submitted, the collaborators are saved into their table, with a reference to the project they are supposed to be working on. A notification is sent to their dashboard asking them to accept or deny the collaboration request to a project. Once agreed the status is changed for that request. If it is denied, a subsequent notification is sent back to the creator of the project that the collaboration request was denied.

There are multiple ways to view the projects and tasks currently being done. The simplest way is to view them as a table view in the "View Projects and Tasks", which will show you a list of the total number of projects and tasks, as well as their details. This view lets you go to a specific page of the project or task so you can view more of its details.

3.3 Gantt View

Another way to view projects and tasks is by the Gantt View. This gives you a more visual outlook on the projects as a timeline. The Gantt View expresses the project itself at the top, taking the maximum width of the view and then displays its children below according to when they start and when they end as well as their status in 3 strong colors. Black means that everything is going smoothly. Green to indicate that the task is done, and red to warn that something has gone wrong. The 2 latter colors will bubble notification to the person in charge of the project, notifying him that something has gone right, or wrong.

3.4 System Architecture

Below is the system architecture of the system showing who has authority of the system and how the system operates. So here the CEO and the admin have full control over the system, they can access everything. They can add task, add subtask, view task and view the Gantt chart. The Managers and Supervisors have minimal access of the system where they can only add subtask, view task and view Gantt chart. For the workers they only have access to view the task and view the Gantt chart only (Fig. 1).

4 Results

Here in this particular section we will see the outcome of our project and discuss about most important features in our research like the Gantt chart feature and adding of projects/subprojects feature.

Below is how the project Dashboard will look like.

As you can see above from Fig. 2 there are many features in the page like: adding new project, adding subtasks, ongoing projects, the staffs (managers, supervisors, workers) with the CEO and the Gantt chart itself where the monitoring and tracking of employees' progress will happen.

So when you click on add project on the dashboard in Fig. 2 this section pops out (Fig. 3) and the user is required to key in the project name, start date and the deadline of the projects, also to add in employees' names that will work on this project and describe the project itself. After filling up and pressing add new task button the project will be added to the sets of projects at the bottom as shown in Fig. 2. Upon adding a project, the CEO will receive recommendations from the AI-powered system. The algorithm will present him/her with the most accurate and suitable employment possibilities to choose from as you can see in Fig. 3 the system suggested software engineer since the project is about building a website. But this part of the system has not yet been built due to some little bit of challenges of finding the correct and most accurate algorithm model (Fig. 4).

Also when the user clicks on the subtask on the dashboard in Fig. 1, this subtask section pops up, where the user will key in the subtask Information of a particular project. He will choose the project that he/she wants to add a subtask in that particular project.

So when the user clicks on Gantt chart on the dashboard in Fig. 1, this section of the website comes up, it's the section where the CEO and managers will see the progress of the projects, but keep in mind that the CEO will have more clearance on viewing the Gantt chart, he/she will be able to view the progress of all projects. The manager will only be able to view the projects progress that they are in charge of.

In the Fig. 5, we see the projects that are ongoing and we can see the start date and time of the BUILD WEBSITE project which is on 01/12/2023 and we can see the deadline of it is on 31/12/2023 at 00:00.

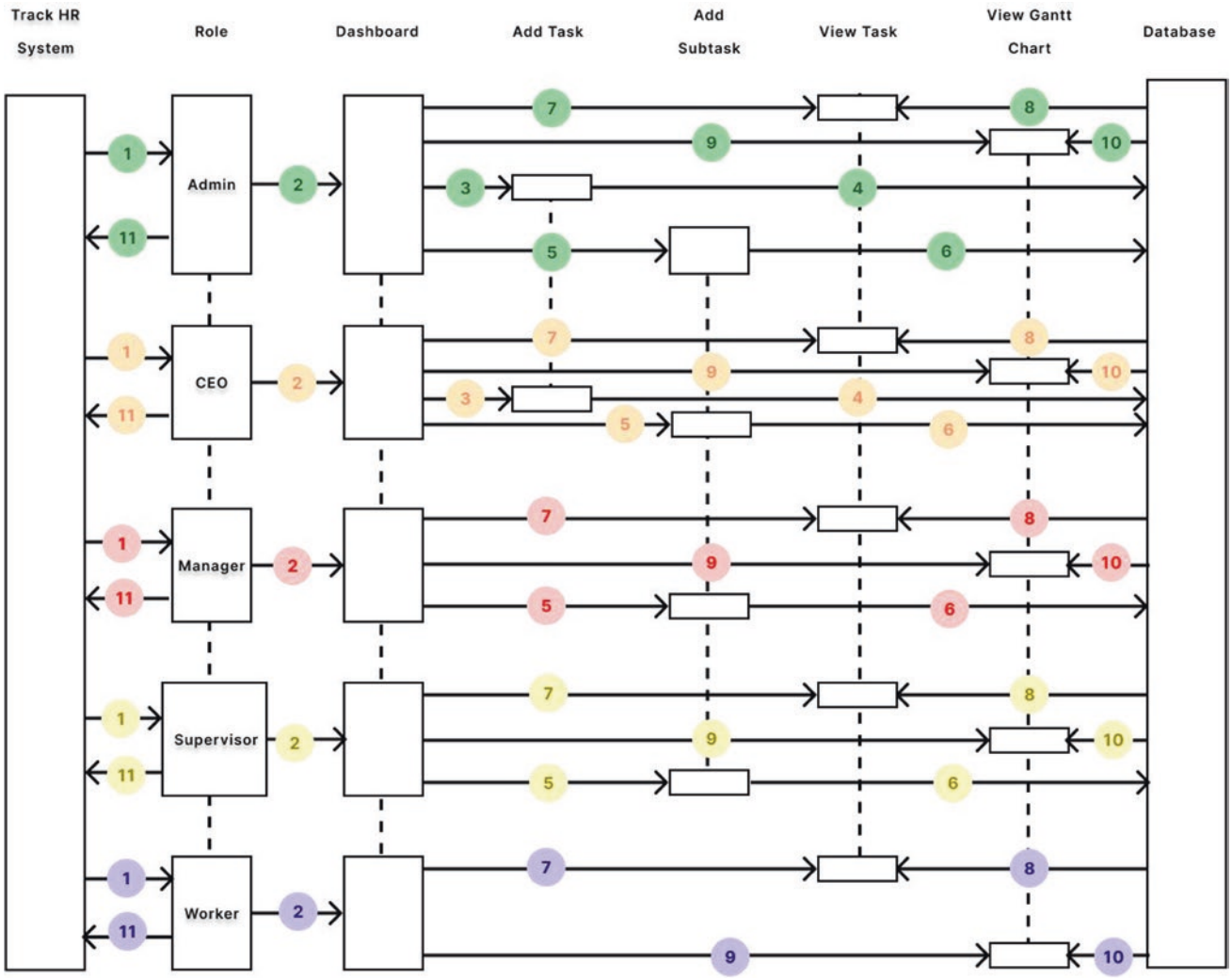

Fig. 1 The system architecture of track HR

So when you click on the build website project the above Fig. 6 displays the subtask. We can see the subtask of the project of building a website. Each subtask has a starting date and time with a deadline too. As you can see there are several subtasks in this project which include the following: building the backend, build frontend, call to action, responsive design, add a login page etc.…

The above Fig. 7 shows if the task/project is still going on and that the deadline has not passed yet. The black color indicates and the small black dot indicates that the project is still on going as observed it's still in the middle.

So when the task is completed the above Fig. 8 displays all the subtask indicated green color with a small ticked circle to the end showing the subtask of building the backend and add a login page was done and completed successfully within the deadline given.

So when the deadline has passed, the above Fig. 9 displays all the subtask indicated red color with a small x circle to the end of the project or subtask showing the subtask has not been completed and the final date given to submit the subtask has passed meaning the subtask of building the backend and setting up the database was incomplete.

Below is a table of comparison of previous research highest accuracy on models (Table 1).

From all the above mentioned researches, our research is unique in its own way. Our project is a web based application that tracks the work progress of the employee's making sure the work given to them is completed before the deadline unlike the other researches from GururamaSenthilvel (2023), Rista et al. (2020), Chowanda (2022), Gopal et al. (2022) and Jabbar et al. (2020). In our project we will try to get the best and accurate algorithm model of 99.99%, where the CEO will be able to choose a suitable employee for a specific project based on the employee's good specifications in a particular field as I have mentioned in Fig. 3.

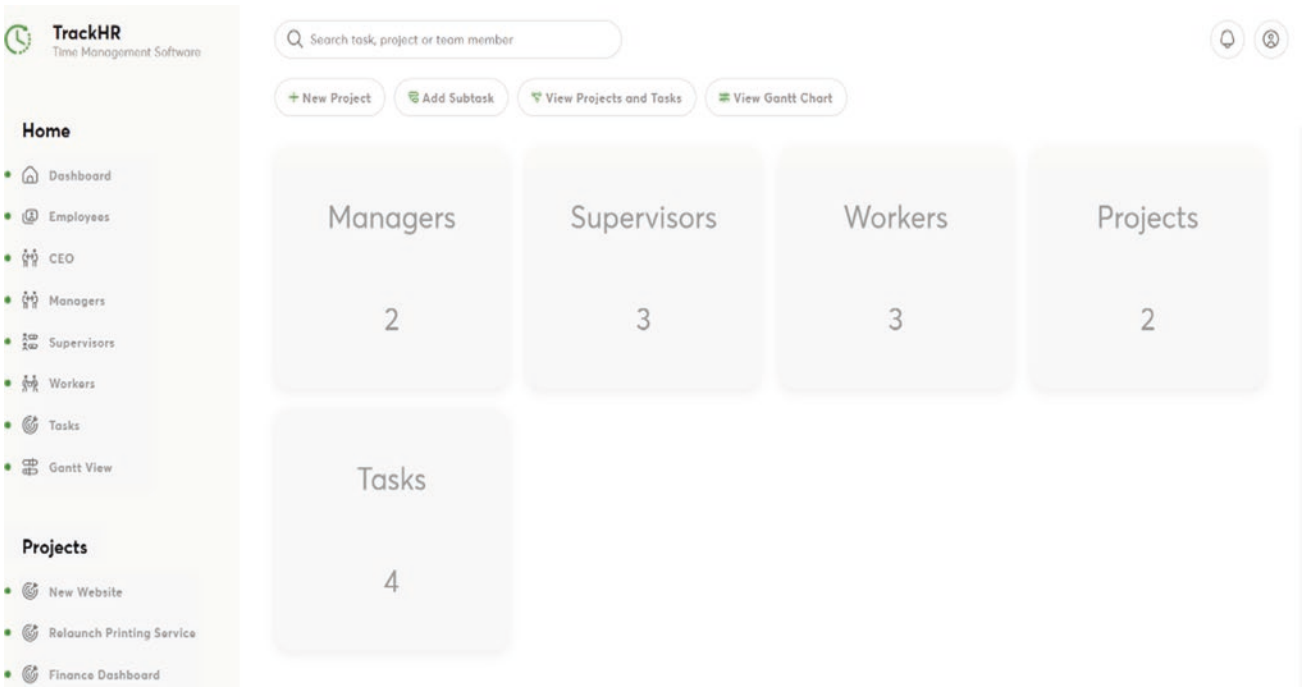

Fig. 2 The project page which displays the dashboard

Fig. 3 Adding a project

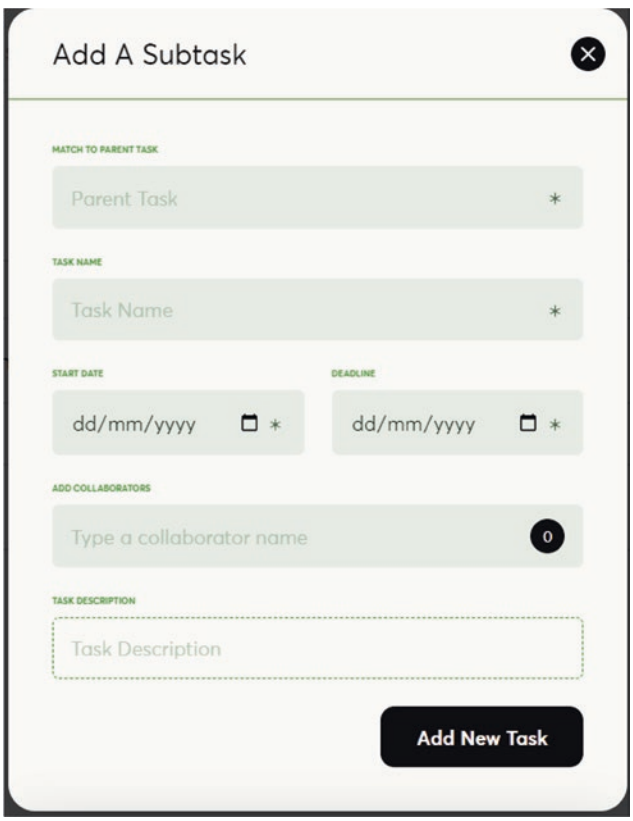

Fig. 4 Subtask section pops out a form

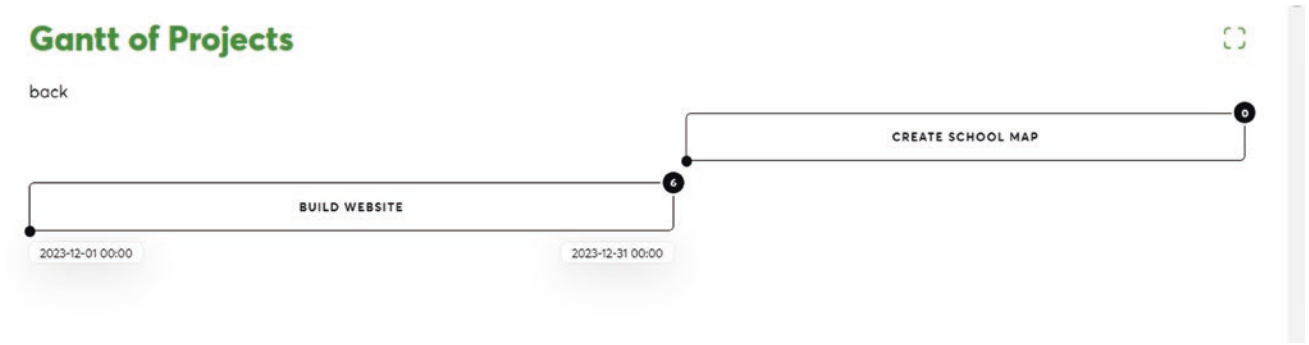

Fig. 5 The Gantt projects is displayed

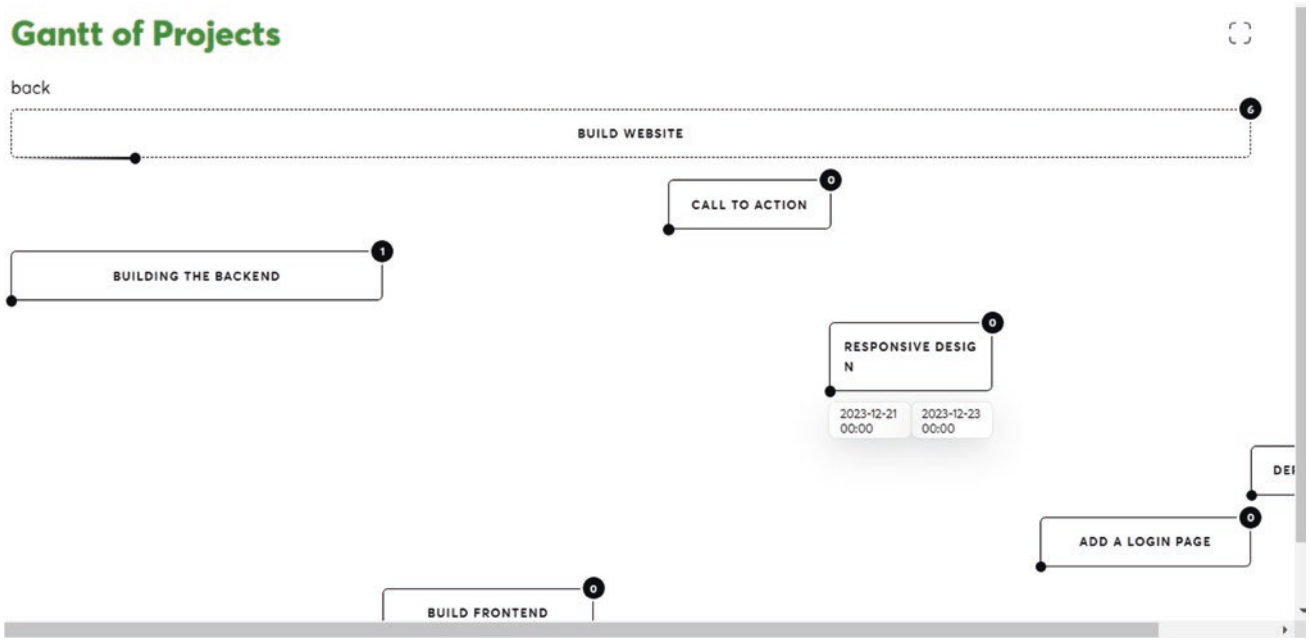

Fig. 6 Gantt of subtask is displayed

Fig. 7 Displays the on-going project

Fig. 8 Displays the projects that are completed

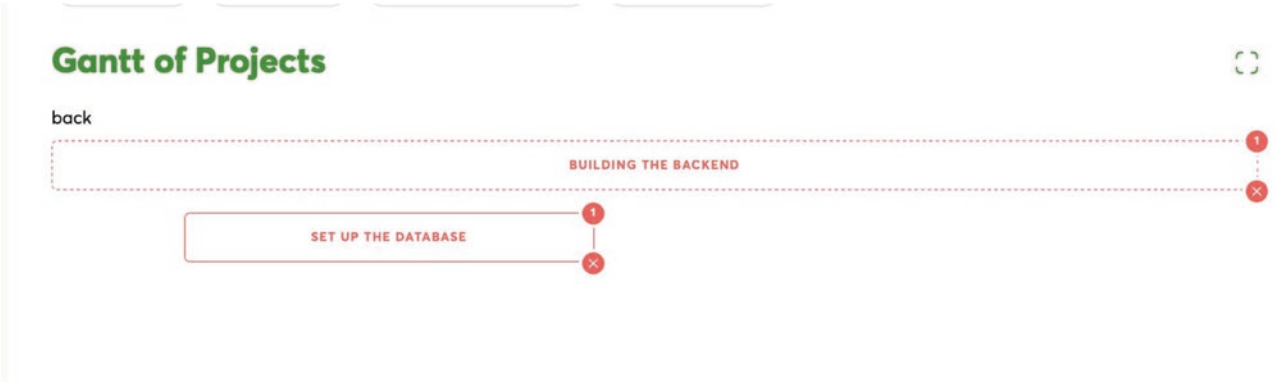

Fig. 9 It shows the deadline has passed

Table 1 Results comparison with other previous researches

Topic	Models	Highest accuracy
AI enabled employee Performance Prediction using Comprehensive Learning Metrics (GururamaSenthilvel 2023)	Naïve Bayes, Random Forest, Support Vector Machine, Logistic Regression	RF—95.833%
Predicting and Analyzing Absenteeism at Workplace Using Machine Learning Algorithms (Rista et al. 2020)	Naïve Bayes, Random Forest, Support Vector Machine, Logistic Regression	Naïve Bayes—99.81%
Machine Learning Face Recognition Model for Employee Tracking and Attendance System (Chowanda 2022)	SVM, ResNet-50, ResNet-152 and Mobile-Net	SVM:—Training: 100% Validation: 99.90%
Analysis of Employee Surveillance System using Deep Learning Models (Gopal et al. 2022)	Mobile-Net, Sequential, VGG-16 and ResNet50	Mobile-Net:—Training: 98.3% Validation: 94.27%
Driver Drowsiness Detection Model Using Convolutional Neural Networks Techniques for Android Application (Jabbar 2020)	D2MLP-FLD, D2CNN-FLD, VGG-16, and AlexN	VGG-16—90.5%

5 Conclusion

In conclusion, the TRACK HR project's use of modern technology is a major step in revolutionizing human resources management. CEOs are given a powerful tool to obtain real-time insights into staff performance through the system's deployment of artificial intelligence and the Internet of Things, which encourages a data-driven approach to decision-making.

The project's ability to change how organizations function by encouraging openness, communication, and the tactical use of human resources is what will ultimately determine its success. Through the provision of a comprehensive solution for tracking and monitoring employee development, the TRACK HR system enables businesses to maximize workflow, stick to deadlines, and speed up project completion.

To summarize, the TRACK HR project has succeeded in developing an advanced HR monitoring system and has the potential to provide substantial benefits for organizations through improved productivity, accountability, and decision-making within the HR domain.

References

D. Nazarov and A. Nazarov, "Intelligent service for monitoring the activities of an employee of an organization," in 2023 IEEE 25th Conference on Business Informatics (CBI), IEEE, Jun. 2023, pp. 1–5. https://doi.org/10.1109/CBI58679.2023.10187419.

Devin Swan, "https://clickup.com/blog/employee-productivity-tracking-tools/."

TrackObit, "https://trackobit.com/glossary/employee-tracking-software."

H. Tamim, F. Sultana, N. Tasneem, Y. Marzan, and M. M. Khan, "Class Insight: A Student Monitoring System with Real-time Updates using Face Detection and Eye Tracking," in 2021 IEEE World AI IoT Congress (AIIoT), IEEE, May 2021, pp. 0213–0220. https://doi.org/10.1109/AIIoT52608.2021.9454176.

J. W. Cotton, C. R. Crowell, L. R. Quintanar, and K. L. Grant, "SESSION V COMPUTER-AIDED INSTRUCTION PROCTOR: An on-line student evaluation and monitoring system for use with PSI format courses," Behavior Research Methods & Instrumentation, vol. 13, no. 2, pp. 121–127, 1981.

P. KIS and M. ALEXANDRU, "REAL TIME MONITORING AND TRACKING SYSTEM FOR AN ITEM USING THE RFID TECHNOLOGY," Review of the Air Force Academy, vol. 13, no. 3, pp. 71–76, Dec. 2015, https://doi.org/10.19062/1842-9238.2015.13.3.12.

M.D. Nirmal, Rohit Koul, Halne Atul, Gagare Tejaswita, and Kharde Mayura, "Employee Surveillance System Using Android Smart Phone," IJARIIE-ISSN(O)- 2395–4396, vol. Vol-2, no. Issue-2, 2016.

D. E. Thompson and A. Molnar, "Workplace Surveillance in Canada: A survey on the adoption and use of employee monitoring applications," Canadian Review of Sociology, vol. 60, no. 4, 2023, https://doi.org/10.1111/cars.12448.

B. P. S. Kalyani and Shamal Bangar, "Employee Monitoring System Using Android Smart Phone, International Journal on Recent and Innovation Trends in Computing and Communication," ISSN: 2321–8169 , vol. Volume: 3, no. Issue: 2, 2015.

D. L. Tomczak, L. A. Lanzo, and H. Aguinis, "Evidence-based recommendations for employee performance monitoring," Bus Horiz, vol. 61, no. 2, 2018, https://doi.org/10.1016/j.bushor.2017.11.006.

M. A. Peshave, "An Analysis of Challenges Faced by the Hospitality Industry in Implementation of Employee Productivity Management System," The International Journal Of Business & Management, vol. 2, no. 5, 2014.

E. Enefiok and O. Uzochukwu, "An Android based Employee Tracking System," Int J Comput Appl, vol. 153, no. 3, 2016, https://doi.org/10.5120/ijca2016911985.

Sonal Kasliwal, Sushma Kotkar, and H.D.Gadade, "Employee Tracking and Monitoring System Using Android," Jan. 2016.

PRITI P. DAFALE, NILIMA N. MANDAL, and DIVYAMALA B. THAKARE, "MONITORING EMPLOYEE'S SMARTPHONE USING ANDROID APPLICATION," May 2015.

S. Sultana, A. Enayet, and I. J. Mouri, "A Smart, Location Based Time and Attendance Tracking System using Android Application," International Journal of Computer Science, Engineering and Information Technology, vol. 5, no. 1, 2015, https://doi.org/10.5121/ijcseit.2015.5101.

P. GururamaSenthilvel, N. Duraimutharasan, H. R. Esther T, P. N. Periyasamy and R. K. K, "Artificial Intelligence enabled Employee Performance Prediction using Comprehensive Learning Metrics," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1–7, https://doi.org/10.1109/ACCAI58221.2023.10201125.

A. Rista, J. Ajdari and X. Zenuni, "Predicting and Analyzing Absenteeism at Workplace Using Machine Learning Algorithms," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2020, pp. 485-490, https://doi.org/10.23919/MIPRO48935.2020.9245118.

A. Chowanda, J. Moniaga, J. C. Bahagiono and J. Sentosa Chandra, "Machine Learning Face Recognition Model for Employee Tracking and Attendance System," 2022 International Conference on Information Management and Technology (ICIMTech), Semarang, Indonesia, 2022, pp. 297–301, https://doi.org/10.1109/ICIMTech55957.2022.9915078.

P. G. Gopal, N. S. Subhash and R. A. Anil, "Analysis of Employee Surveillance System using Deep Learning Models," 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA, Pune, India, 2022, pp. 1-6, https://doi.org/10.1109/ICCUBEA54992.2022.10010874.

R. Jabbar, M. Shinoy, M. Kharbeche, K. Al-Khalifa, M. Krichen and K. Barkaoui, "Driver Drowsiness Detection Model Using Convolutional Neural Networks Techniques for Android Application," 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2020, pp. 237–242, https://doi.org/10.1109/ICIoT48696.2020.9089484.

Customer Relationships Management (CRM) Application for Customer Segmentation via RFM Analysis and K-Means Clustering

Fadi Al-Turjman and Arda Karaderi

Abstract

This paper introduces a web-based solution utilizing PostgreSQL, Python, and PHP (Laravel Framework) to automate Recency, Frequency, and Monetary (RFM) analysis and K-Means clustering for customer segmentation. The study addresses the inefficiencies in the manual method and positions the developed system as a transformative tool for businesses. The research aimed to automate and enhance customer segmentation through a web application, exploring the effectiveness of the solution in automating RFM analysis and clustering via experiments. The developed software is a user-friendly, integrated platform that not only saves time but also minimizes errors inherent in manual analyses. The findings highlight the system's effectiveness in automating RFM analysis and the added value of comparative insights derived from clustering outcomes. Differentiating the system from the solutions on the market, clustering via Artificial Intelligence (AI) was added to the system. This enables users to compare what the clustering algorithm thinks the groups of users should be against what the RFM analysis produces. Finally, by providing full control of their data to users, the system aims to align with GDPR standards, ensuring both efficiency and legal compliance.

Keywords

Customer relationship management · Customer segmentation · K-Means algorithm · Machine learning · Clustering

F. Al-Turjman · A. Karaderi (✉)
Artificial Intelligence, Software, and Information Systems Engineering Departments, Research Center for AI and IoT, AI and Robotics Institute, Near East University, 99138 Nicosia/TRNC Mersin 10, Turkey
e-mail: 20213703@std.neu.edu.tr

1 Introduction

Relationships with customers can make or break the success of a business. For a business, there is nothing more vital than its customers. The whole objective of running a for-profit business is to sell products or services to profit, scale and expand. However, not every customer means the same for a business. Some customers are more profitable. They might be spending more money, spending more time on the services, having a longer cash-positive relationship with the company or making bigger purchases overall. This leads to the need to understand the potential of each customer to maximize profit (Kim et al. 2006).

According to Martin (2011), by understanding the customers and their needs, companies can create more specialized products and/or services to create and promote value for their customers. This will eventually lead to higher profitability. But which customers should be targeted for the next email campaign?

Rather than investing in understanding each customer, customers can be categorized into groups based on their attributes, and those can be used as the key decision-making criteria for the earlier questions. For example, a business can segment its customers according to things like their demographics or spending behaviours (Tarver et al. 2023). In a hypothetical example, that means if a customer is classified as a top spender in a software agency, the business can offer in-person onboarding service to them while not offering the same service to regular clients. This means a top-paying customer feels more exclusive and is more likely to stay loyal to the brand while other top-paying customers are receiving the benefit of having faster service deliveries. This is a result of offering an exclusive service to a segment of customers and managing other valuable resources like employees on more profitable tasks.

British Airways (2023) applies a similar strategy. Everyone can join their flight point collection programme (Avios). However, as customers spend more money on their

F. Al-Turjman (ed.), *The Smart IoT Blueprint: Engineering a Connected Future*, Advances in Science, Technology & Innovation, https://doi.org/10.1007/978-3-031-63103-0_2

services, they increase their tier (segment) and give them exclusive benefits to positively impact customers' brand loyalty and boost customers' spending behaviours. At the time of writing the paper, regular customers can use Avios to collect points (to tier up) and save their preferences on the system and higher-tiered customers are eligible for first-class check-ins and boardings.

According to Murphy and Kvilhaug (2022) is a scoring model used in the analysis phase of marketing processes to segment customers of a company by their behaviour. RFM is an acronym used for describing recency (R, how recent a customer's latest purchase was), frequency (F, in what intervals does customer places orders) and monetary (M, the financial outlay associated with their activities).

If a business has records regarding its customers, it can utilize unsupervised machine learning via clustering to group (segment) its unlabeled customers (Google for Developers 2022). The project developed for this paper is a web application that enables businesses to upload their customer order data into the system to perform RFM analysis and clustering via the K-Means algorithm. The application has a statistics page allowing a user to quickly read and understand their company performance, customers and employees pages to list customers and employees, and pages to view details about individual customers and employees and list their orders. (Historical analysis scores for customers are also shown on their profiles.) Moreover, the application includes a page where users can view the history of analysis run on the system. Finally, the application has a preferences section that enables users to delete their data, delete their account, change their password and set thresholds for colouring their performance indicators.

Before the development of the application, several experiments were run with DBSCAN and K-Means clustering algorithms to find out which algorithm is more suitable for the development of a generalised web solution that is accessible by an easy-to-use user interface.

Two different datasets were used for performing the experiments. The first dataset contained 2747 rows of order details belonging to 298 orders from an automobile company (Dosad 2023). The second dataset contained 9800 rows of order details belonging to 4922 unique orders (Sahoo 2021). It has been identified that the K-Means algorithm is a better suit for the development of such an application.

The rest of the paper is organized as the following: Sect. 2 discusses the existing research in the literature and its results. Details of experiments performed to decide which clustering algorithm to use were discussed in Sect. 3. Moreover, the details of the developed application were also shared in Sect. 4. Finally, the conclusion and suggestions for future work were discussed in Sect. 5.

2 Literature Review

According to Sheshasaayee and Logeshwari (2018), CRM segmentation models are outdated. The reason behind this statement is the noise and excessiveness in the datasets. In their research, Sheshasaayee and Logeshwari (2018) propose a new method that utilises recital segmentation and RFM analysis. RFM is an acronym used for Recency-Last, Frequency and Monetary. Recency-last refers to the consumption interval of the customer. Frequency is a label used for the periodic number of transactions and Monetary is the amount they spend in a particular period. For RFM-based segmentation, the K-Means clustering algorithm was used. The output of the research suggests that the K-Means algorithm can be used for this sort of segmentation while they only used the frequency of the RFM model to form the clusters. Further studies can be conducted on this research and dataset via other parameters.

In their research, Pramono et al. (2019) conducted research on the beauty industry in Indonesia. They claim that to survive the competitiveness of the sector, a brand needs to understand and utilise Customer Relationship Management (CRM). Thus a brand can utilise different strategies to grow its customer base. The goal of their study is to create customer clusters that have a similar lifetime value. They used the K-Means algorithm to form the clusters.

Essayem et al. (2022) claim that offering targeted products and services is a must for a successful business. Moreover, businesses are working on their marketing to create and use new strategies that allow them to retain their existing customers while attracting new ones. To offer services, discounts and perform targeted advertising companies frequently use RFM-based clustering via K-Means clustering. In their research, Essayem et al. (2022) used point-of-sale (POS) data of a store. They conducted further analysis via Silhouette analysis. The output of the research concluded with two clusters, one of them containing customers who are likely to stop their relationship with the store while the other contained loyal and continuous customers of the store.

In their research, Arul et al. (2022), used a dataset that contained data from nearly 300 customers. In their paper, they discussed how they used K-Means clustering to segment customers. The customers originated from an e-commerce website. The reason that they have chosen this topic to research is to show how K-Means clustering can resolve the issue of categorising customers into segments (groups, clusters in machine learning terms) according to their behaviour. They believe that segmentation has a crucial role in business, whether it's the launch of a new product, increasing sales to retaining existing customers.

Maryani and Riana (2017) support that companies face the issue of identifying potential customers and applying the right strategy according to the customer specifics to maximize the benefit to the company. In their research, the objective was set to perform clustering on the RFM model's monetary parameter to categorize customers. For forming the clusters, the K-Means clustering algorithm was used. Moreover, their research was not limited to forming clusters. Therefore, they adopted a decision tree (J48) to classify customers further. To collect the data, they manually analysed the sale transaction bills created for customers. Overall, the study used 326 records. To validate the clusters that the K-Means algorithm identified, the Davies Bouldien index was performed. To profile the customers, Maryani and Riana (2017) used Grid Hill's economic theory. The paper concludes with recommendations to the company about their Customer Relationship Management.

According to Chen et al. (2017), customer transaction data needs to be clustered and has a vital role in understanding customer behaviours. Their study focuses on retail and e-commerce companies. The research's objective is to provide a new clustering algorithm called PurTreeClust. As their objective is to compare their proposed algorithm to the field-tested existing algorithms, 10 different datasets containing real-life transactions of customers were used. They claim that companies frequently organize their products into product trees. Their goal was to prove that a personalised product tree could be produced. Many different approaches exist for choosing the right number of clusters to be formed. During the research, the team proceeded with Gap statistics to identify the correct number of clusters to be formed by the algorithm. The proposed algorithm was evaluated against six other popular clustering algorithms on all of the datasets (10). It was observed that it performed better than others on 8 of the 10 datasets. The six algorithms compared were, DBSCAN, HAC-SINGLE, HAC-MEAN, HAC-COMPLETE, NCut and RCut. The research concludes that for massive customer transaction data, the PurTreeClust algorithm showed significant effectiveness and scalability against the compared algorithms.

According to Tavakoli et al. (2018)'s research, the RFM model is an effective analysis for a company to target their customers and apply marketing strategies accordingly. In their research, the RFM model's simplicity has been criticized as it doesn't account for the customer's relationship with the business and the customer's environment as well as behavioural changes. The research takes Digikala to build their research around. All of the purchase data of the Digikala and customer data between the dates January 1st, 2014 to December 30th, 2017 were extracted and processed to form the dataset to be used in the research. The research utilised the K-Means algorithm and used the Frequency and Monetary aspects of the RFM model to form the clusters. The research resulted in a campaign that boosted the number of purchases and the value of properties customers placed into their baskets.

Qiu et al. (2020) claim that a part of customers are called silent customers. Silent customers are considered to be very easy to lose. As discussed earlier, it costs more to attract a new customer to retain an existing one. Therefore, silent customers can be considered a high-risk segment. To segment those customers and take a marketing action to prevent churn K-Means++ clustering was performed on the dataset. To validate the number of clusters to be formed, the Calinski-Harabasz index was used. The dataset used in the research contained 125,296 silent customers. The data was sourced from an anonymous telecom company in Fujian. The dataset contained 26 features for each record. To conclude the research a set of marketing suggestions as well as the research output was handed over to the company.

3 Experiments

K-Means and Density-based spatial clustering of applications with noise (DBSCAN) were algorithm candidates for the application's clustering functionality. According to Google Developers (2022), K-Means is an algorithm that is used to put raw data into k clusters, where k is a user-specified number of clusters.

On the other hand, DBSCAN is a clustering algorithm where clusters need to have a certain data density to be clustered together. According to both Spoor (2022) and Jain (2023), the Silhouette Coefficient is considered the key metric to evaluate the quality of clusters created.

To decide which algorithm to use, experiments were run on two datasets. The first dataset contained 2747 rows of data belonging to 298 orders from an automobile company (Dosad 2023). The second dataset contained 9800 rows of data belonging to 4922 orders (Sahoo 2021). In experiments on both datasets, the K-Means algorithm performed better thus, it is used as the clustering algorithm of the application. This section justifies the decision via a discussion of the results of the experiments performed.

3.1 Methodology

This section explains the environments of the experiments and the metrics used for evaluating the results of the experiments. Davies-Bouldin Index (DBI) and Silhouette Score were used for evaluating the results of the experiments. The definitions of both metrics are the following:

Davies-Bouldin Index (DBI). The Davies-Bouldin Index (DBI) is a measure of the similarity between each cluster

and its most similar cluster, which is used to evaluate the quality of a clustering algorithm. A lower DBI score indicates a better clustering solution (Davies and Bouldin 1979).

Silhouette Score. The Silhouette Score measures the similarity of each point to its cluster compared to other clusters. Similar to the Davies-Bouldin Index, The Silhouette Score also has a simple evaluation. Unlike the Davies-Bouldin Index, as the Silhouette Score score gets higher, better clustering is achieved (scikit-learn 2022).

As the application's objective is to provide a generic solution as a service that can be accessed from the internet, at this stage there shouldn't be too much fine-tuning to be done by the user on the algorithms. Therefore, except for the core parameters, none of the parameters were changed from the defaults during the experimentation. For the K-Means algorithm, the number of clusters was left to be optimised by the system while for the DBSCAN, EPS (maximum distance between two samples) and the minimum number of points required in a group of points to be considered as a core point were attempted to be optimised by the system.

Both datasets went through the same pre-processing. First, as both datasets contained multiple products for each order, to eliminate the effect of duplicate values on frequency calculations, orders were condensed into single lines containing attributes; order number, customer name, order total (total cost of products within that order) and order date. Afterwards, the R, F and M scores are calculated with the following formulas:

Recency. The recency was calculated by finding the difference in days between a defined date (differs for each dataset) and the most recent order date for each customer.

Frequency. In the experiments, the frequency of customers was calculated via grouping and counting the number of unique order dates.

Monetary. The order totals for each customer is summed to achieve the total monetary value.

To define scores via these calculations, scoring over 5 via the quantile that the calculation falls into for each calculation was used. This process is repeated for all calculations resulting in new columns called Recency Score, Frequency Score and Monetary Score. To calculate the RFM score, scores for each customer in each category are concatenated (e.g. R = 1. F = 3, M = 5 results in an RFM score of 135). Afterwards, the newly formed data is passed forward to algorithms for experimenting.

3.2 Experiment 1: Working with Bare Minimum Customisation

To cluster with the K-Means algorithm, the number of clusters has to be defined. In this experiment, the script used for finding optimal values was put to test to validate the accuracy of the function (Fig. 1).

From Table 1, it can be observed that 2 clusters do produce the best results according to the Silhouette Score & Davies-Bouldin Index (DBI). When the clusters formed are used for labelling data on the RFM plots the results displayed in Fig. 2 are produced.

In order to perform clustering with the DBSCAN algorithm, optimal values for minimum samples and EPS parameters have to be identified. As two parameters were used, many combinations were produced. Thus, the five top-performing combinations were exported. The outputs of the script are the following (Tables 2, 3 and Fig. 3).

When the top-performing minimum samples and EPS combination for DBI are used for labelling the RFM plot the results displayed in Fig. 4 are produced. (2 minimum variables and 0.5 EPS.)

When the top performing minimum samples and EPS combination for Silhouette Score is used for labelling RFM plot the results displayed in Fig. 5 are produced. (3 minimum variables and 68 EPS.)

Table 1 Silhouette Score and the Davies–Bouldin Index for different K values on dataset 1

Clusters (n)	Silhouette Score	Davies-Bouldin Index
2	0.429526	0.887353
4	0.389357	0.956442
7	0.358468	0.986180
3	0.377234	1.027592
6	0.333131	1.028629
8	0.353869	1.054525
10	0.336390	1.066779
9	0.319979	1.079013
5	0.357059	1.099822

```
2 clusters is the new best and has the Silhouette Score of 0.42952609685340365 and has the Davies-Bouldin Index of 0.8873529231126174
The highest Silhouette Score of 0.42952609685340365 is achieved with 2 clusters.
```

Fig. 1 Output of the return values of the script on dataset 1 for K-Means algorithm

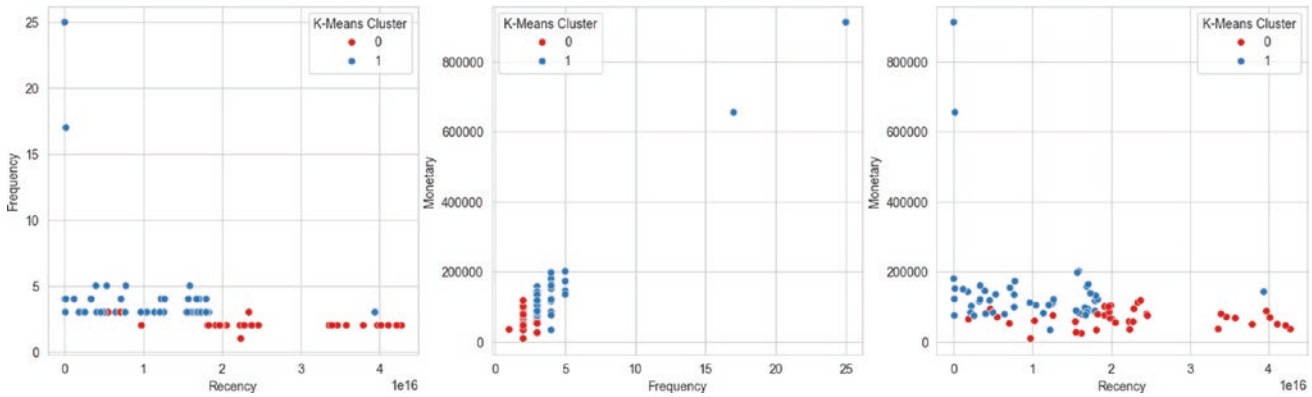

Fig. 2 Clusters formed via K-Means algorithm for dataset 1 displayed on RFM plots

Table 2 Silhouette score and the Davies–Bouldin Index for different minimum sample and EPS value combinations for dataset 1

Minimum samples	EPS	Silhouette Score	Davies-Bouldin Index
68	3.0	0.219271	0.590172
7	0.5	0.219271	0.590172
5	0.5	0.219271	0.590172
6	0.5	0.219271	0.590172
4	0.5	− 0.033464	0.746104

The table is ordered by the DBI index

Table 3 Silhouette Score and the Davies–Bouldin Index for different minimum sample and EPS value combinations for dataset 1

Minimum samples	EPS	Silhouette Score	Davies-Bouldin Index
2	0.5	0.499693	1.947125
23	1.5	0.375485	0.883537
37	2.0	0.348628	1.030629
35	2.0	0.348628	1.030629
36	2.0	0.348628	1.030629

The table is ordered by Silhouette Score

Experiment Conclusion. Even though the DBSCAN algorithm produced a lower DBI score than the lowest score of the K-Means algorithm and produced a slightly higher best value for the Silhouette Score when the RFM graphs were plotted, it's been identified that the DBSCAN formed too many clusters for the top performing Silhouette Score combination and did not cluster properly for best performing DBI combination. For the DBI combination, the algorithm clustered the majority of points together and put the others into the outliers group (− 1) which is not acceptable.

Even though K-Means performed slightly poorer in the evaluation metrics, the results produced were acceptable. Thus, for this experiment, it's been concluded that as the objective of the software is to provide a generic solution that any user can use without too much tweaking, the K-Means algorithm is a better choice.

3.3 Experiment 2: Performance of Algorithms on a Different Dataset

The same script was adapted to the second dataset to benchmark the performance of the algorithm on different datasets. The outputs for the K-Means algorithm are shown in Table 4, and Fig. 6. The plots created for Dataset 2 with K-Means labels are shown in Fig. 7.

For the K-Means algorithm, it's been observed that on both datasets the best Silhouette Score and Davies-Bouldin Index (DBI) metrics were close to each other. This suggests that the algorithm shows a similar performance on different datasets of different sizes. For both datasets, the script identified that two clusters would provide the optimal results. By checking the results in Tables 1 and 4, the results produced by the application were accepted as accurate. (Optimal parameter identifying script prioritizes the Silhouette Score over the Davies-Bouldin Index.)

For the DBSCAN algorithm, the steps performed in Experiment 1 were repeated as well. The top combinations for the Silhouette Score and Davies-Bouldin Index (DBI) identified by the script are shown in Table 5 and Fig. 8.

```
2 minimum samples with 0.5 EPS value is the new best and has the Silhouette Score of 0.49969285115728646 and has the Davies-Bouldin Index of 1.9471249173606162
The highest Silhouette Score of 0.49969285115728646 is achieved with 2 minimum samples and  0.5 EPS value.
```

Fig. 3 Output of the return values of the script

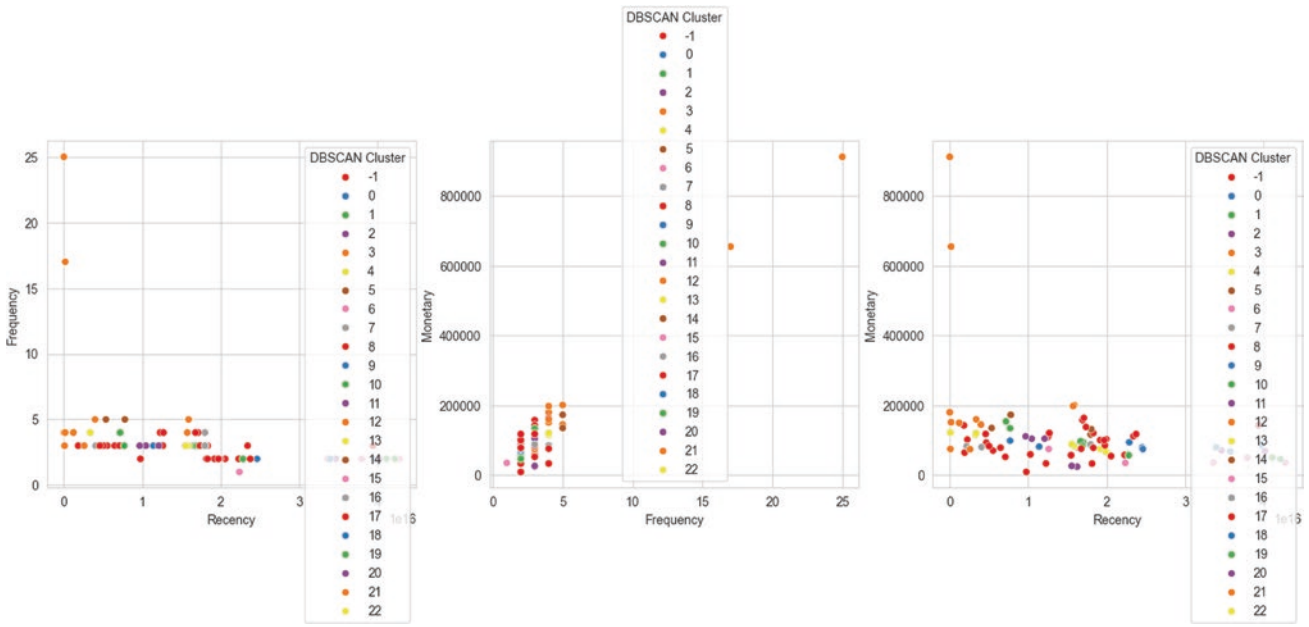

Fig. 4 Clusters formed via DBSCAN for dataset 1 displayed on RFM plots for top Silhouette Score value pair

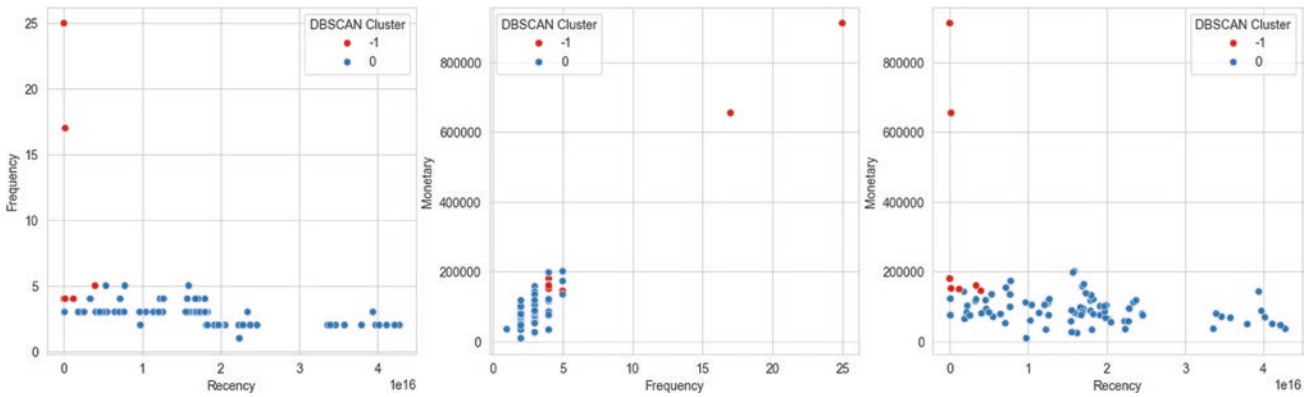

Fig. 5 Clusters formed via DBSCAN for dataset 1 displayed on RFM plots for top DBI index pair

Table 4 Silhouette Score and the Davies–Bouldin Index for different K values on dataset 2

Clusters (n)	Silhouette Score	Davies-Bouldin Index
2	0.371098	1.098131
7	0.333255	1.018275
6	0.323334	1.006696
5	0.321925	1.086013
4	0.319824	1.113885
3	0.316369	1.235760
10	0.303290	1.108314
9	0.291309	1.083346
8	0.275016	1.143418

Comparing the results on dataset 2 with the previous results on dataset 1, it's observed that a noticeably higher Silhouette Score has been achieved. This suggests that with the default values provided by the algorithm, consistency of results cannot be achieved. Therefore, even though the literature suggests that it can produce better results than K-Means, it requires more fine-tuning. Moreover, as the targetted end users of this application may not be capable of fine-tuning clustering algorithms, the K-Means algorithm seemed like a better fit. Revising the results displayed in Figs. 4, 5 and 9, it is concluded that the DBSCAN algorithm needs advanced controls from the user to produce meaningful results. Thus, this experiment is also concluded with K-Means being a better pick for the development of an application this project aims for.

```
2 clusters is the new best and has the Silhouette Score of 0.37109761257067175 and has the Davies-Bouldin Index of 1.098130872530381
The highest Silhouette Score of 0.37109761257067175 is achieved with 2 clusters.
```

Fig. 6 Output of the return values of the script on dataset 2 for K-Means algorithm

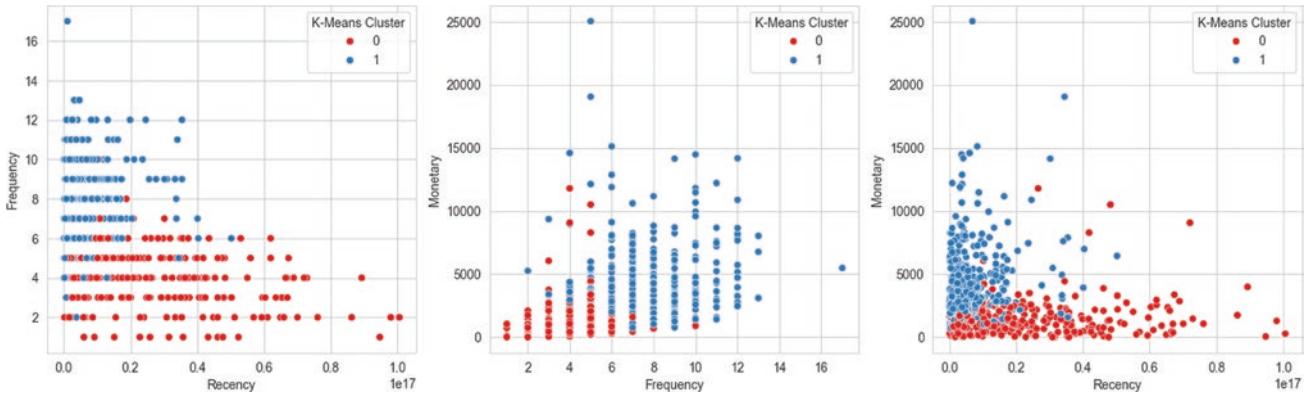

Fig. 7 Clusters formed via K-Means for dataset 2 displayed on RFM plots

Table 5 Silhouette Score and the Davies–Bouldin Index for the different minimum sample and EPS value combinations for dataset 2

Minimum samples	EPS	Silhouette Score	Davies-Bouldin Index
2	0.5	0.984661	1.297059
3	0.5	0.916528	1.499931
4	0.5	0.840187	1.517622
5	0.5	0.750432	1.220569
6	0.5	0.710648	1.206200

The table is ordered by Silhouette Score

4 Implementation of the Software Solution

This project aimed to create a complete software product that performs RFM analysis and clustering on data uploaded to the system. After the analysis, the user is provided with a dashboard to explore the results and receive insights.

The system contains the following components to achieve the results:

```
2 minimum samples with 0.5 EPS value is the new best and has the Silhouette Score of 0.9846610499354685 and has the Davies-Bouldin Index of 1.2970593104147525
The highest Silhouette Score of 0.9846610499354685 is achieved with 2 minimum samples and  0.5 EPS value.
```

Fig. 8 Output of the return values of the script on dataset 2 for DBSCAN algorithm

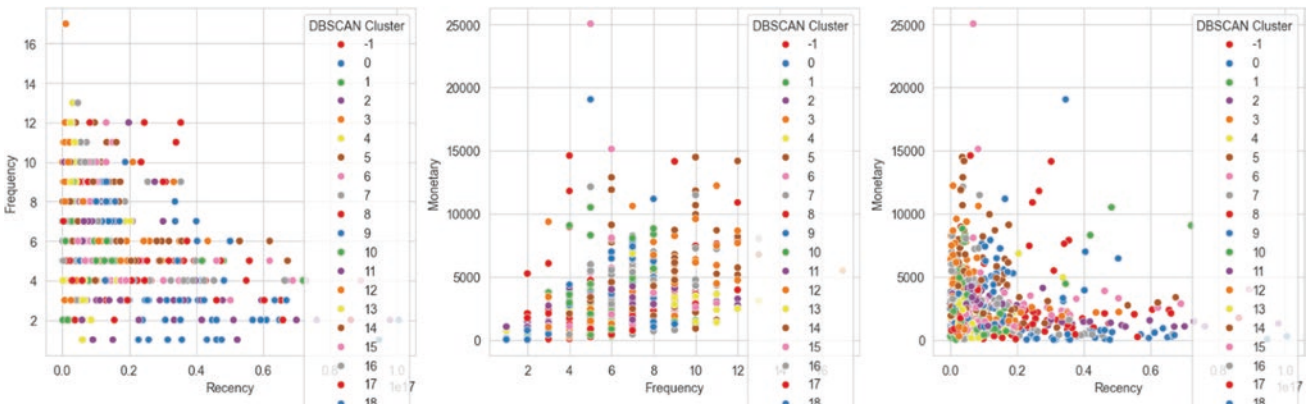

Fig. 9 Clusters formed via DBSCAN algorithm with the top performing Silhouette Score pair displayed on RFM plots

- A database management system (PostgreSQL), and a database,
- Python application that performs the clustering on the data to produce results,
- A PHP-Laravel application, which is used to create the dashboard that the end user uses, performs RFM calculations and analysis.

As described in Fig. 10, the user is able to sign in to his/her account to access the dashboard. Once the user accesses the dashboard, the user is able to explore the previous analysis that has been done, view details about individual customers or employees, explore segments, receive insights on the segments identified by the system, compare clustering results to old-school RFM analysis results or upload new data to the system to initiate the new analysis.

As stated by Khajvand et al. (2011) for a business to succeed it needs to understand the market and its customers. Therefore, to understand its customers and receive insights on customer behaviors companies widely adopt RFM analysis methodology. There are less than 10 software products in the first 60 search results of the term "RFM analysis software" on Google. This shows that not many software are capable of producing such analysis.

However, when the other results are analysed, the results are focused on the importance of RFM analysis and display guidance on how to perform the analysis using tools like Excel etc. This states that RFM analysis is an important tool to utilize by businesses.

As an employee needs to do the analysis, it creates a workload on the employee and a cost on the business. The core benefit of the application is the time saved on performing RFM analysis. As the system performs the analysis, employee time will be saved. Therefore, the employee can use that time to do other tasks that can bring more profits to the business. Moreover, this will result in cost savings for the company. Other major benefits of the application include the minimization of the errors in the analysis process that might be caused by humans and the use of artificial intelligence (AI) to provide comparative results to the old-school RFM analysis results with clusters. People from the same cluster can be in different RFM customer segments, this provides a great lead on what customers from different RFM segments share and how a customer from a lower-profiting segment can be upgraded to a higher segment.

The application's core values are to provide abstraction, and control over data and deliver value to the end user. Unlike applications on the market, the system does not require any integrations to work. Therefore, as an abstract platform that can operate on its own, the data must be uploaded to the system by the user. Moreover, the user is able to delete or update all the information linked to their

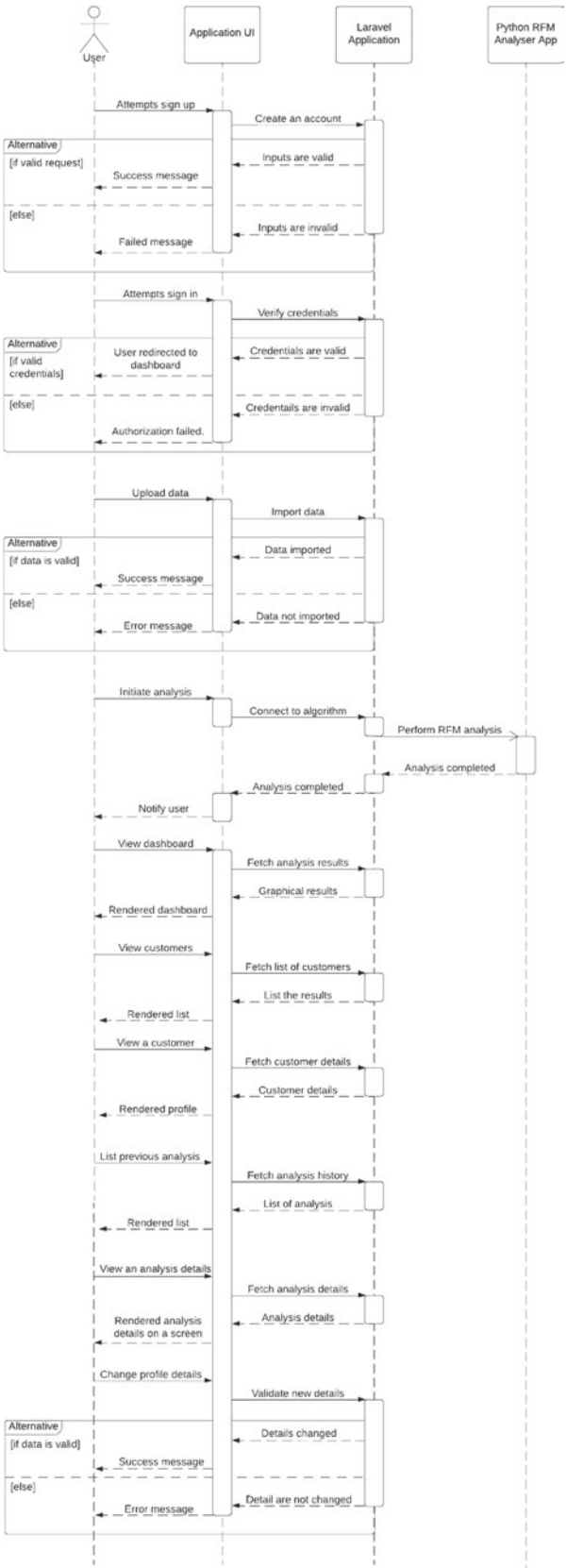

Fig. 10 Sequence diagram

account if they wish. The processes will occur in compliance with the General Data Protection Regulation (GDPR) which is a law in the European Union (European Union 2023).

4.1 Validation of Results Produced by the Developed Software

To validate the implementation of the web application, a random sampling method was used. To avoid bias, Python's Pandas library was used for sampling 15 random results from the script's output.

After the sampling, the same dataset used for producing results in the script was run through the application. Then the RFM scores and the clusters assigned by the application were cross-checked with the results of the script. The results indicated that all samples were clustered the same way on both runs. However, a slight difference in F score calculation on sample 2 was observed. It was concluded that the difference was due to performing similar calculations on two different programming languages. Python 3 was used for experimentation and PHP was used for development of the final application. All of the results are shown in Table 6.

5 Conclusion and Future Work

According to the literature, RFM analysis is a powerful tool for businesses. The inefficiencies in performing RFM analysis manually prompted the exploration of a comprehensive web application that has RFM analysis and clustering capabilities.

With the proposed application, on top of data exploration, the users can perform RFM analysis and K-Means clustering to group customers into categories. Thus, the developed system fulfills the objective of the research which was to automate the customer segmentation process by providing a user-friendly and efficient application.

This paper has navigated through the experiments on different clustering algorithms to decide which one of the algorithms can be used for developing a solution that performs better on multiple datasets. Based on the results of the experiments, the development of a web application that utilises PostgreSQL, Python, and PHP (Laravel Framework) to create an application that enables users to explore data of orders of their business and their customers was discussed.

Moreover, the paper discussed the benefits of the developed system in automating RFM analysis and offering comparative insights through K-Means clustering. Looking forward, future research can be done on the refinement and expansion of clustering techniques, adding further functionality to the interactiveness of the system. Moreover, the current stage of development provides an opportunity to investigate the long-term impact of automated customer segmentation on businesses, analyzing how it influences decision-making, profitability, and customer retention over extended periods.

Table 6 RFM scores and clusters produced by the script compared to RFM scores and clusters produced by the web application

Randomly sampled experiment results for K-Means algorithm—dataset 2						Results on application on dataset 2 via K-Means	
Sample #	Recency Score	Frequency Score	Monetary Score	RFM Score	K-Means cluster	RFM Score	K-Means cluster
1	3	4	5	345	0	345	0
2	2	4	1	*241*	1	*251*	1
3	3	2	4	324	1	324	0
4	4	4	2	442	1	442	0
5	4	4	3	443	1	443	0
6	3	1	1	311	0	311	1
7	3	2	3	323	0	323	1
8	5	4	4	544	1	544	0
9	2	2	3	223	0	223	1
10	5	4	4	544	1	554	0
11	5	5	4	554	1	554	0
12	2	4	4	244	1	244	0
13	3	4	2	342	1	342	0
14	4	1	2	412	0	412	1
15	2	3	5	235	1	235	0

References

Arul, V., Kumar, A., & Agarwal, A. (2022, March 09). Segmenting Mall Customers Data to Improve Business into Higher Target using K-Means Clustering. *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. https://doi.org/10.1109/ICAC3N53548.2021.9725630

British Airways. (2023, December 7). *Tiers and benefits | Executive Club*. British Airways. Retrieved December 7, 2023, from https://www.britishairways.com/en-us/executive-club/tiers-and-benefits

Chen, X., Fang, Y., Yang, M., Nie, F., Zhao, Z., & Huang, J. Z. (2017, October 26). PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data. *IEEE Transactions on Knowledge and Data Engineering*, *3*(30), 559 - 572. IEEE. https://doi.org/10.1109/TKDE.2017.2763620

Davies, D. L., & Bouldin, D. W. (1979, April). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224 - 227. IEEE. https://doi.org/10.1109/TPAMI.1979.4766909

Dosad, D. (2023). *Automobile Sales Dataset*. Kaggle. Retrieved December 18, 2023, from https://www.kaggle.com/datasets/ddosad/auto-sales-data

Essayem, W., Bachtiar, F. A., & Priharsari, D. (2022, June 18). Customer Clustering Based on RFM Features Using K-Means Algorithm. *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*. IEEE. https://doi.org/10.1109/CyberneticsCom55287.2022.9865572

European Union. (2023). *General Data Protection Regulation (GDPR) Compliance Guidelines*. General Data Protection Regulation (GDPR) Compliance Guidelines. Retrieved December 1, 2023, from https://gdpr.eu

Google Developers. (2022, July 18). *Clustering Algorithms | Machine Learning*. Google for Developers. https://developers.google.com/machine-learning/clustering/clustering-algorithms

Google for Developers. (2022, July 18). *Clustering Algorithms | Machine Learning*. Google for Developers. Retrieved December 7, 2023, from https://developers.google.com/machine-learning/clustering/clustering-algorithms

Jain, S. (2023, May 23). *DBSCAN Clustering in ML | Density based clustering*. GeeksforGeeks. Retrieved January, 2024, from https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011, February 22). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57-63. https://doi.org/https://doi.org/10.1016/j.procs.2010.12.011

Kim, S.-Y., Jung, T.-S., Suh, E.-H., & Hwang, H.-S. (2006, July 1). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, *31*(1), 101–107. https://doi.org/10.1016/j.eswa.2005.09.004

Martin, G. (2011, May 01). The Importance Of Marketing Segmentation. *American Journal of Business Education (AJBE)*, *4*(6), 15. https://doi.org/10.19030/ajbe.v4i6.4359

Maryani, I., & Riana, D. (2017, October 30). Clustering and profiling of customers using RFM for customer relationship management recommendations. *2017 5th International Conference on Cyber and IT Service Management (CITSM)*. https://doi.org/10.1109/CITSM.2017.8089258

Murphy, C., & Kvilhaug, S. (2022, November 19). *What Is Recency, Frequency, Monetary Value (RFM) in Marketing?* Investopedia. https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp

Pramono, P. P., Surjandari, I., & Laoh, E. (2019, July 13). Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method. *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*. https://doi.org/10.1109/ICSSSM.2019.8887704

Qiu, Y., Chen, P., Lin, Z., Yang, Y., Zeng, L., & Fan, Y. (2020, May 04). Clustering Analysis for Silent Telecom Customers Based on K-means++. *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE. https://doi.org/10.1109/ITNEC48623.2020.9084976

Sahoo, R. (2021). *Superstore Sales Dataset*. Kaggle. Retrieved January 15, 2024, from https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting/data

scikit-learn. (2022). *Selecting the number of clusters with silhouette analysis on KMeans clustering*. Scikit-learn. Retrieved December 13, 2023, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Sheshasaayee, A., & Logeshwari, L. (2018, May 12). Implementation of Clustering Technique Based RFM Analysis for Customer Behaviour in Online Transactions. *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. https://doi.org/10.1109/ICOEI.2018.8553873

Spoor, J. M. (2022, September 30). Improving customer segmentation via classification of key accounts as outliers. *Journal of Marketing Analytics*, *11*, 747–760. Springer. https://doi.org/10.1057/s41270-022-00185-4

Tarver, E., Khartit, K., & Perez, Y. (2023, October 31). *Market Segmentation: Definition, Example, Types, Benefits*. Investopedia. https://www.investopedia.com/terms/m/marketsegmentation.asp

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018, December 30). Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study. *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*. IEEE. https://doi.org/10.1109/ICEBE.2018.00027

Densely Connected CNN-Based XAuNet1.0 for Brain Tumor Classification via MRI Images in the IoT Era

Hadjer Benyamina and Fadi Al-Turjman

Abstract

Brain tumor classification stands as a pivotal challenge within the realm of Computer-Aided Diagnosis. This research delves into the binary classification of brain images derived from varied angles in Magnetic Resonance Imaging scans of the human brain. The proposed classification model is built on deep transfer learning paradigm, implementing pre-trained Densely Connected Convolutional Neural Networks and "Image Net" weights to autonomously extract features from the inputs of MRI brain images based on DenseNets. Our model was trained on a large dataset of two classes "Yes" and "No" taking into consideration the importance of hyper parameters tuning role and the usage of regularization techniques effectively. Therefore, our results suggest that our proposed model XAuNet1.0 performs exceptionally well in accurately identifying instances of both classes, making it a robust solution for real-world tests and live usage through deploying the model in an accessible platform or IoT health care systems.

Keywords

Brain tumor · DenseNets · Classification

H. Benyamina (✉)
Artificial Intelligence, Software, and Information Systems Engineering Departments, AI and Robotics Institute, Near East University, Mersin 10, Turkey
e-mail: 20205682@std.neu.edu.tr; hadjer.benyamina2024@gmail.com

F. Al-Turjman
Faculty of Engineering, Research Center for AI and IoT, University of Kyrenia, Kyrenia, Cyprus
e-mail: fadi.alturjman@neu.edu.tr

F. Al-Turjman
Artificial Intelligence Engineering Department, AI and Robotics Institute, Near East University, Mersin 10, Turkey

1 Introduction

A tumor is caused by uncontrolled growth and abnormal spread of body cells that can result in death if not treated. Normally, cells multiply and die in an orderly way to be replaced by new cells. However, abnormal growth can keep happening to form a mass or a lump called a tumor. According to the Centers for Disease Control and Prevention, cancer is the second most frequent cause of mortality after heart disease. The American Cancer Society reported 2022 an estimated number of 18.280 deaths versus 25.050 new cases of brain and other nervous system cancer. In both adults and children, brain tumors are one of the most fatal and threatening cancers. Moreover, they have the lowest survival rate. Consequently, early diagnosis has an important role in preventing more spread of the tumor and identification of the right treatment plan. Examining MRI images of the brain with tumors is an effective technique to distinguish brain tumors which is a difficult process even for a professional oncologist or a neurosurgeon which means that any tiny human error in the diagnosis will lead to undesirable consequences. The number of brain tumor cases is increasing gradually which makes manual techniques tedious, time-consuming, and erroneous. MRI is well-known for effectively finding some cancers like breast, brain, and spinal cord tumors. Using MRI medical professionals can even tell sometimes if a tumor is or is not cancerous. Furthermore, it helps plan treatment such as surgeries or radiotherapy (American Cancer Society 2021). Brain tumor diagnosis begins usually with MRI scans. When the MRI shows that the brain has a tumor then professionals head towards further investigations for example performing a biopsy procedure or surgery depending on the tumor type and severity of the patient's condition (Cancer. Net 2019). And here comes computer-aided systems that are always a handful for medical professionals to avoid errors, and time-consuming procedures and reduce the need for manual-based investigations. In other words, developing

F. Al-Turjman (ed.), *The Smart IoT Blueprint: Engineering a Connected Future*, Advances in Science, Technology & Innovation, https://doi.org/10.1007/978-3-031-63103-0_3

systems that can classify brain tumors is an important key for more accurate results, early diagnosis and even making treatment plans which is possible through Convolutional Neural Networks (CNNs) that have proven interesting results in treating classification problems and the computer vision field globally. Teaching a machine to recognize images was a huge challenge for decades until research could finally simulate the visual recognition system of the human brain.

In 1963, Lawrence Roberts (Roberts 1963) suggested the possibility of extracting 3D geometrical information from a 2D view. This was the first breakthrough in computer vision. Several algorithms and models were suggested along this journey such as the generalized cylinder model, pictorial structure model, and convolutional neural networks known as CNNs or ConvNets. They are not much different from classical neural networks. ConvNets are both structured on neurons based on weights learned from data. There are 3 main layers in a CNN: the convolution layer, the pooling layer, and the fully connected layer which is the output layer. ConvoNets can arrange their neurons in three dimensions: width, height, and depth. Each layer transforms the 3D input into a 3D output of neurons using an activation function. To summarize, CNNs are deep neural networks that share their parameters across space. Moreover, CNNs accept matrices as an input differently from multi-layer perceptrons that only accept vectors which guarantee the preservation of the spatial structure of the image during the process. These characteristics and more are what allowed CNNs to be a top candidate when addressing computer vision problems under the supervised Machine Learning (ML) paradigm. The three major learning paradigms are supervised learning, unsupervised learning, and reinforcement learning. Each paradigm corresponds to a specific learning task.

1.1 Supervised Learning

Supervised learning allows the discovery of relationships between input data which can be dependent or independent parameters. This relationship is represented in the form of a "Model". Building a model is the objective of supervised ML. It generates predictions based on proofs of uncertainty. The learning functions are built from labeled training data consisting of a set of training samples, or Models. The machine learns via observation. The more observations a machine gets, the more it learns and improves its performance to level up the accuracy of its predictions when dealing with new data that was not part of the learning phase (Maimon and Rokach 2010).

1.2 Unsupervised Learning

Labels of classes in this type of learning are unknown and the number of trained classes is unknown. It is referred to as "Clustering" because there is no trained model in unsupervised learning. Clustering is used to discover classes within the data (Han et al. 2012; Aggarwal 2015). Real-life applications of unsupervised learning are numerous and diverse. Companies specializing in transportation and logistics use anomaly detection to identify obstacles or expose defective mechanical parts and condition monitoring (Del Campo Barraza 2017). In the medical context, anomaly detection can also be used to detect anomalies, lesions, or abnormalities such as brain tumors, without explicitly training the model for that specific pathology (Chatterjee et al. 2023), unsupervised anomaly detection in medical images enables models to learn visual representations from unlabeled data (Iqbal et al. 2023). Unsupervised learning techniques are also applied in customer and market segmentation to create customer personas for better marketing and targeting campaigns (Paranavithana et al. 2021; Joppe 2016), data preparations and visualization, recommender systems and engines such as Amazon's frequently bought together recommendations, Natural Language Processing models for complex syntax learning, generating genuine grammatical sentences (Solan et al. 2005), and computer vision (Chen et al. 2022).

1.3 Reinforcement Learning

Another subcategory of ML is reinforcement learning. It is based on agent intelligence in taking actions depending on the environment to maximize cumulative rewards, in other words: Sequential decision-making (François-Lavet 2018; Sutton and Barto 2014). Various papers have proposed reinforcement learning-based trajectory optimization, motion planning, dynamic pathing, and scenario-based learning policies in highways for self-driving cars such as AWS DeepRacer (AWS DeepRacer—Le Moyen Le plus Rapide de Commencer à Apprendre Le Machine Learning 2024). Google Data Centers also uses reinforcement learning agents to cool the system helping reduce energy spending by up to 40% (Gamble and Gao 2018). NLP reinforcement learning is used for text summarization (Paulus et al. 2017), question answering (Choi et al. n.d.), and even machine translation (Grissom II et al. 2014). Reinforcement learning has a say in healthcare as well. RL can optimize treatment policies for chronicle disease and critical care based on previous experiences, it supports automated medical diagnosis, and other domains (Yu et al. 2020).

Related Works. Several machine-learning techniques are used to classify brain tumor datasets using Support Vector Machines (Ozsoz et al. 2021), K-Nearest Neibours Algorithms (Altman 1992), and Convolutional Neural Networks (CNNs) (Akella et al. 2023). Our research focuses on using CNNs as the state of art proving the high efficacy of using CNNs in the field of computer vision generally (Mubarak et al. 2022) and medical imaging specifically (Salehi et al. 2023).

The winners of ImageNet Challenge 2014 (Simonyan and Zisserman 2014) submitted a CNN model known as VGGNet16 and VGGNet19 that secured them first place in the localization track and second place in the classification one. Moreover, the model generalizes well on other datasets and even a wide range of tasks. Such achievement marked a significant milestone in computer vision with the proposed VGGNet16 and VGGNet19 CNN models. The VGGNet architecture showcased its effectiveness in handling diverse challenges within the competition. This versatility highlighted the robust nature of the VGGNet models, making them valuable tools not only for image classification and localization but also for various applications beyond the initial challenge.

A year later, He et al. (2015) introduced an innovative approach to Deep Learning (DL) with their CNN architecture known as Deep Residual Learning. The proposed model, particularly the ResNet50 and ResNet101 variants, achieved remarkable success by securing first place in both the image classification and localization tracks. Notably, the introduction of residual learning, which involves the use of residual blocks, allowed for the creation of significantly deeper networks without suffering from vanishing gradient problems. This breakthrough design facilitated the training of extremely deep neural networks, leading to improved performance and generalization across various datasets and tasks.

In 2016, a groundbreaking contribution in the field of DL was presented by Huang et al. (2016) that addresses the challenges and limitations of conventional convolutional networks by introducing the Dense Convolutional Network (DenseNet). The key innovation of DenseNet lies in its dense connectivity pattern, effectively addressing challenges such as the vanishing-gradient problem, strengthens the flow of features through the network, encourages feature reuse, and significantly reduces the number of parameters compared to traditional architectures which demonstrated the superior performance of DenseNet across various tasks, solidifying its significance in advancing the state-of-the-art in DL. The focus of this paper goes to the depth of the CNN architecture that leads to better accuracy compared to previously published papers in the same field.

Brain Tumors Detection Using AI. Pathak et al. (2019) proposed a CNN classifier for brain MRI images to indicate any presence of tumor in addition to tumor area calculation using the watershed algorithm for segmentation operation. The proposed classifier achieved a very good accuracy with low complexity. Another paper using DL to segment tumor areas from brain MRI images was proposed by Sajjad et al. (2019) after applying data augmentation to treat the lack of data issue while training the model based on a novel CNN system for multi-grad brain tumor classification task. The proposed system was evaluated on both raw data and augmented data achieving convincing performance compared to existing methods. Ayadi et al. (2021) proposed a multilayer CNN to perform brain tumor MRI image classification. The proposed model was evaluated on three datasets and attained satisfying performance. The proposed method in Pashaei et al. (2018) consists of extracting hidden features from brain tumor MRI images using CNNs then a Kernel Extreme Learning Machine algorithm performs a feature-based classification task which has shown promising results compared to other classifiers such as support vector machines and radial base function. Abiwinanda et al. (2018) implemented a simple CNN architecture that recognizes the three most common brain tumor types trained on a publicly available dataset. The architecture has no prior region-based segmentation, yet it achieved higher accuracies compared to the one that applied segmentation algorithms. Naseer et al. (2021) focus on early diagnosis possibility through a CNN trained on a benchmark dataset containing brain tumors MRI images. The proposed system outperforms other suggested systems by achieving a very satisfying average accuracy and almost perfect specificity while hitting 100% correct diagnosis for two datasets that were evaluated. Alanazi et al. (2022) proposed a novel transfer deep-learning model to classify a multi-class dataset of brain tumors by re-using a 22-layer binary classifier to adapt the neuron weights to their model through a transfer learning technique. To approach the computational time issue lying within CNNs approaches, Bacanin et al. (2021) proposed a metaheuristic method based on an improved firefly algorithm (Jati et al. 2013) developing an automatic system for brain tumor multi-class classification that has outperformed similar approaches with considerable classification accuracy.

2 Methodology: XAuNet1.0 Implementation Process

In this section, we provide details about our research design and explain the practical side of the research including dataset source, image preprocessing techniques and the Convolutional Neural Network architecture used to train the most stable and reliable generated model XAuNet1.0 based on pre-trained DesneNets and evaluated by comparison to