DATA ANALYTICS 2024

The Thirteenth International Conference on Data Analytics

ISBN: 978-1-68558-187-9

September 29 - October 03, 2024

Venice, Italy

DATA ANALYTICS 2024 Editors

Timothy Haas, University of Wisconsin-Milwaukee, USA

Clement Leung (SSE), The Chinese University of Hong Kong, Shenzhen, China

DATA ANALYTICS 2024

Forward

The Thirteenth International Conference on Data Analytics (DATA ANALYTICS 2024), held between September 29th, 2024, to October 3rd, 2024, in Venice, Italy, continued a series of international events on fundamentals in supporting data analytics, special mechanisms, and features of applying principles of data analytics, application-oriented analytics, and target-area analytics.

Processing terabytes to petabytes of data or incorporating non-structural data and multi- structured data sources and types require advanced analytics and data science mechanisms for both raw and partially processed information. Despite considerable advancements in high performance, large storage, and high computation power, there are challenges in identifying, clustering, classifying, and interpreting a large spectrum of information.

We take here the opportunity to warmly thank all the members of the DATA ANALYTICS 2024 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to DATA ANALYTICS 2024. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the DATA ANALYTICS 2024 organizing committee for their help in handling the logistics of this event.

We hope that DATA ANALYTICS 2024 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in the field of data analytics.

DATA ANALYTICS 2024 Chairs

DATA ANALYTICS 2024 Steering Committee
Wolfram Wöß, Institute for Application Oriented Knowledge Processing | Johannes Kepler University, Linz, Austria
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
George Tambouratzis, Institute for Language and Speech Processing, Athena R.C., Greece
Les Sztandera, Thomas Jefferson University, USA
Ivana Semanjski, Ghent University, Belgium
Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands

DATA ANALYTICS 2024 Publicity Chairs
Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

DATA ANALYTICS 2024
Committee

DATA ANALYTICS 2024 Steering Committee

Wolfram Wöß, Institute for Application Oriented Knowledge Processing | Johannes Kepler University, Linz, Austria
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
George Tambouratzis, Institute for Language and Speech Processing, Athena R.C., Greece
Les Sztandera, Thomas Jefferson University, USA
Ivana Semanjski, Ghent University, Belgium
Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands

DATA ANALYTICS 2024 Publicity Chairs

Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

DATA ANALYTICS 2024 Technical Program Committee

Arianna Agosto, University of Pavia, Italy
Irfan Ahmed, Virginia Commonwealth University, USA
Raed Ibrahim Alharbi, University of Florida, USA
Madyan Alsenwi, Kyung Hee University, Global Campus, South Korea
Katie Antypas, Lawrence Berkeley National Laboratory, USA
Vincenzo Arceri, University of Parma, Italy
Najet Arous, University of Tunis Manar, Tunisia
Abderazek Ben Abdallah, The University of Aizu, Japan
Sadok Ben Yahia, Tallinn University of Technology, Estonia
Soumia Benkrid, Ecole Nationale Supérieure d'Informatique, Algeria
Flavio Bertini, University of Parma, Italy
Nik Bessis, Edge Hill University, UK
Sandjai Bhulai, Vrije Universiteit Amsterdam, Netherlands
Jean-Yves Blaise, UMR CNRS/MC 3495 MAP, Marseille, France
Jan Bohacik, University of Zilina, Slovakia
Vincenzo Bonnici, University of Parma, Italy
Marco Calderisi, Kode srl, Pisa, Italy
Ozgu Can, Ege University, Turkey
Wandercleiton Cardoso, University of Genoa, Italy
Julio Cesar Duarte, Instituto Militar de Engenharia, Rio de Janeiro, Brazil
Junghoon Chae, Oak Ridge National Laboratory, USA
Richard Chbeir, Université de Pau et des Pays de l'Adour (UPPA), France
Daniel B.-W. Chen, Monash University, Australia
Yujing Chen, VMware, USA
Stefano Cirillo, University of Salerno, Italy
Giovanni Costa, ICAR-CNR, Italy
Bi-Ru Dai, National Taiwan University of Science and Technology, Taiwan

Antonio Jiménez Martín, Universidad Politécnica de Madrid, Spain
Dimitrios Karapiperis, International Hellenic University, Greece
Ashutosh Karna, HP Inc. / Universitat Politecnica de Catalunya, Barcelona, Spain
Srinivas Karthik V., Huawei Technologies, India
Christine Kirkpatrick, San Diego Supercomputer Center - UC San Diego / CODATA, USA
Weikun Kong, Tsinghua University, China
Alina Lazar, Youngstown State University, USA
Kyung Il Lee, Reinhardt University, USA
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
Clement Leung, Chinese University of Hong Kong, Shenzhen, China
Yuening Li, Texas A&M University, USA
Zhao Liang, University of São Paulo, Brazil
Ninghao Liu, Texas A&M University, USA
Weimo Liu, Google, USA
Fenglong Ma, Pennsylvania State University, USA
Ruizhe Ma, University of Massachusetts Lowell, USA
Massimo Marchiori, University of Padua, Italy / European Institute for Science, Media and Democracy, Belgium
Mamoun Mardini, College of Medicine | University of Florida, USA
Miguel A. Martínez-Prieto, University of Valladolid, Spain
Alfonso Mateos Caballero, Universidad Politécnica de Madrid, Spain
Archil Maysuradze, Lomonosov Moscow State University, Russia
Abbas Mazloumi, University of California, Riverside, USA
Gideon Mbiydzenyuy, Borås University, Sweden
Ryan McGinnis, Thomas Jefferson University, USA
Letizia Milli, University of Pisa, Italy
Yasser Mohammad, NEC | AIST | RIKEN, Japan / Assiut University, Egypt
Thomas Morgenstern, University of Applied Sciences in Karlsruhe (H-KA), Germany
Lorenzo Musarella,University Mediterranea of Reggio Calabria, Italy
Azad Naik, Microsoft, USA
Roberto Nardone, University Mediterranea of Reggio Calabria, Italy
Alberto Nogales, Universidad Francisco de Victoria | CEIEC research center, Spain
Ameni Youssfi Nouira, RIADI Laboratory - ENSI, Tunisia
Panagiotis Oikonomou, University of Thessaly, Greece
Ana Oliveira Alves, Polytechnic Institute of Coimbra & Centre of Informatics and Systems of the University of Coimbra, Portugal
Riccardo Ortale, Institute for High Performance Computing and Networking (ICAR) - National Research Council of Italy (CNR), Italy
Moein Owhadi-Kareshk, University of Alberta, Canada
Yu Pan, University of Nebraska-Lincoln, USA
Massimiliano Petri, University of Pisa, Italy
Hai Phan, New Jersey Institute of Technology, USA
Antonio Pratelli, University of Pisa, Italy
Yiming Qiu, Rice University, USA
Víctor Rampérez, Universidad Politécnica de Madrid (UPM), Spain
Andrew Rau-Chaplin, Dalhousie University, Canada
Ivan Rodero, Rutgers University, USA
Sebastian Rojas Gonzalez, Hasselt University / Ghent University, Belgium

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Evaluating 21st Century Skills Development through Makerspace Workshops in Computer Science Education

Petros Papagiannis and Georgios Pallaris
Department of Computer Science
Cyprus College
Limassol, Cyprus
e-mail: p.papagiannis@cycollege.ac.cy, g.pallaris@cycollege.ac.cy

***Abstract*—This study evaluates the effectiveness of incorporating makerspace workshops into computer science education by assessing 21st-century skills—critical thinking, collaboration, communication, and creativity—before and after the intervention. Using a pre-test and post-test approach with the "21st Century Skills Survey Instrument," the study quantifies the impact of makerspace activities on student skill development. Participants included students enrolled in two computer science courses at Cyprus College. Statistical analysis, conducted using Python, revealed significant improvements across all assessed skills, indicating that makerspace workshops enhance essential competencies needed for the modern workforce. These findings provide valuable insights into how experiential learning environments can transform traditional computer science education, promoting a more interactive and engaging learning experience. Future research should focus on larger, more diverse samples and explore specific components of makerspace activities that most effectively contribute to skill development.**

Keywords-makerspace; computer science education; 21st-century skills; pre-post study; educational impact.

I. INTRODUCTION

Makerspaces provide innovative learning environments that empower students to develop essential skills necessary for success in an increasingly technological world while also enhancing their engagement levels [1]. This paper examines the impact of makerspace workshops integrated into two computer science courses: Introduction to Programming and Computer Architecture. In these workshops, students are encouraged to think critically, collaborate, communicate effectively, and engage creatively. As technology continues to advance and permeate educational settings, this research seeks to quantify the effects of such modern teaching methods on students' learning outcomes through an experimental design. By focusing on the development of 21st-century skills, this study aims to provide valuable insights into how makerspaces can transform traditional computer science education, fostering a more interactive and engaging learning experience. This paper will discuss the methodology, results, and implications of integrating makerspaces into the computer science curriculum. The makerspace workshops included activities such as collaborative coding projects, 3D printing tasks, and interactive problem-solving sessions. These activities were designed to specifically target and enhance critical thinking, collaboration, communication, and creativity. Future studies could analyze which specific activities have the most significant impact on each skill area.

The courses were designed to provide hands-on experience in both "Introduction to Programming" and "Computer Architecture" courses. In the "Introduction to Programming" workshop, students engaged in building simple physical computing projects using microcontrollers and sensors. These projects were designed to reinforce programming concepts like loops, conditionals, and functions by applying them in a tangible context. In the "Computer Architecture" workshop, students constructed basic digital circuits and simulated CPU operations using breadboards and logic gates. These activities were intended to deepen their understanding of hardware components and the underlying principles of computer architecture. Each workshop spanned four weeks, with weekly sessions lasting two hours.

The structure of the paper is as follows: In Section 2, we discuss the theoretical framework and related work. Section 3 presents the methodology, including the design of the makerspace workshops. Section 4 details the data collection and analysis process. Section 5 discusses the findings, and Section 6 concludes with implications for practice and future research directions.

II. LITERATURE REVIEW

A. Introduction to Makerspaces in Education

Makerspaces have become integral to modern educational environments, promoting active, hands-on learning that aligns with constructivist and constructionist theories [2], [3]. These spaces enable students to engage in creative problem-solving and collaboration, crucial for developing skills needed in today's technological world [1].

B. Impact of Makerspaces on 21st-Century Skills

Research indicates that makerspaces significantly enhance critical thinking, collaboration, communication, and creativity. Blikstein [4] demonstrated that makerspace activities improve students' problem-solving abilities and critical thinking. Halverson and Sheridan [5] found that these environments foster collaboration and enhance communication skills through peer interactions and feedback mechanisms.

C. Makerspaces in Computer Science Education

The integration of makerspaces into computer science education has shown to improve student engagement and learning outcomes. Martinez and Stager [6] argue that the project-based

nature of makerspaces is well-suited to computer science, which often involves designing and building technological solutions. Litts [7] supports this by highlighting that makerspaces help bridge the gap between theoretical knowledge and practical application, enhancing students' understanding of complex concepts.

D. Methodological Approaches to Assessing Makerspaces

Assessing the impact of makerspaces involves various methodological approaches. Kelley et al. [8] utilized pre-test and post-test designs to measure changes in skills, which is a method also adopted in this study. Qualitative methods, such as interviews and observational studies, provide additional insights into the student experience and the effectiveness of specific activities [9].

E. Challenges and Considerations

Despite the benefits, implementing makerspaces poses challenges, including cost and resource availability [10]. Ensuring equitable access to these resources is crucial for maximizing their educational impact [11].

F. Future Directions

Future research should explore the long-term impact of makerspace participation on student skills and career outcomes. Longitudinal studies tracking students over multiple years could provide insights into the sustained benefits of makerspace activities [12]. Additionally, investigating which specific components of makerspace activities are most effective could help educators design more impactful interventions [13].

III. METHODOLOGY

A. Participants

Participants for this study were students attending makerspace workshops embedded in two computer science courses: Introduction to Programming and Computer Architecture, delivered at Cyprus College, Limassol. Participants were selected based on their enrolment in the specified courses and their willingness to participate in the study. The courses were conducted over 12 weeks, followed by 2 weeks of examinations, with each week including three 50-minute sessions. The integration was led by the course instructor/researchers and an educational technologist, following the Learning Experience Design (LXD) framework [14].

The teaching approach combined traditional methods with autonomous problem-solving. Sessions began with theoretical concepts using slides, videos, and lectures, followed by tasks of increasing difficulty to encourage independent problem-solving. The instructor facilitated learning through guidance and feedback without providing direct solutions. Outside class, students actively engaged in collaborative projects using the makerspace and digital tools, fostering peer learning and the practical application of theoretical knowledge. This iterative problem-solving process promoted teamwork, resilience, and lifelong learning. We conducted the questionnaire on 23 respondents over one semester within an academic year—in about four months.

While the sample size for this study was limited to 23 students, future research should aim to include a larger and more diverse sample to improve the generalizability of the findings. Efforts to replicate this study across multiple institutions will provide more comprehensive insights.

B. Instruments

The survey instrument utilized in this study was developed using LimeSurvey, an open-source online survey tool, which was adapted to align with the "21st Century Skills Survey Instrument" methodology. This tool allowed for the customization of survey questions to ensure they were tailored specifically to measure the development of skills such as critical thinking, collaboration, and creativity within the context of the makerspace activities. We employed a pre-test and post-test approach using the "21st Century Skills Survey Instrument" [8]. This tool assesses students' abilities in four key areas: Collaboration, Communication, Creativity, and Critical Thinking. Students first completed an initial test, participated in course content and makerspace activities, and then took a post-test to determine any improvements in these areas. We composed two sets of survey instruments for the collection of data on the students' 21st-century skills:

- Pre-Assessment Survey: This was administered before the workshop to determine the baseline capacities of the students in terms of critical thinking, collaboration, communication, and creativity.
  - "I am confident in my ability to revise drafts and justify revisions with evidence." (Critical Thinking)
  - "I am confident in my ability to follow the rules for team decision-making." (Collaboration)
  - "I am confident in my ability to organize information well." (Communication)
  - "I am confident in my ability to understand how knowledge or insights might transfer to other situations or contexts." (Creativity)
- Post-Assessment Survey: This survey re-evaluates the same skills after the complication of the workshop, testing for changes and developments that could have been made post-intervention. The questions were the same as those in the pre-assessment, so we can easily match the responses.

C. Data Collection

Data was collected through web-based survey applications administered before the commencement of the makerspace workshop and after the seminar concluded. This study received approval from the Institutional Review Board (IRB) of Cyprus College, ensuring ethical standards were met. All participants provided informed consent, and measures were taken to ensure data privacy and confidentiality. Future research should continue to prioritize these ethical considerations, particularly when expanding to larger and more diverse samples. This method was intended to directly correlate the changes in the activities performed throughout the workshop with a change

in skills. We created the survey online, and the link was forwarded electronically. Once a survey link is active, that survey remains available for one week. Since this was an online survey, respondents' answers would be anonymous to preserve the information's secrecy and provide more unbiased information.

D. Data Preparation

Data preparation involved cleaning and standardization to compare the two datasets. We used Python for statistical analysis. Missing values were handled by either dropping respondents with significant gaps in their data or imputing where appropriate based on the distribution of other responses.

Identification of Comparable Columns: In this study, "similar constructs" refers to the alignment of survey questions that target the same cognitive or skill-based dimension across both pre- and post-workshop surveys. For example, questions that assess critical thinking in the pre-survey were mapped directly to corresponding questions in the post-survey to ensure consistent measurement of this construct. "Comparable columns" thus refer to the specific data columns that contain responses to these aligned questions in both datasets.

Given that the pre and post questionnaires were identical, the data preparation process involved ensuring that the responses were directly comparable. This involved standardizing the data formats and verifying that the response scales remained consistent across both surveys. Missing values were addressed by either omitting respondents with significant data gaps or using imputation techniques where appropriate, based on the distribution of responses within the dataset. For example, if a respondent missed one or two questions, their missing responses were imputed using the median or mode of the other responses to that question:

- Identification of Comparable Columns: Data columns for the pre-and post-datasets were observed to evaluate similar constructs. For example, questions measuring critical thinking were aligned in both surveys on this dimension. Column mapping follows we performed for this matching:

```
column_mapping = {
    "pre": {
        "CARPROSQ[SQ001]": "CARPROSQ[SQ001]_post",
        "CT[001]": "CT[001]_post",
        "COL[SQ001]": "COL[SQ001]_post",
        "COM[SQ001]": "COM[SQ001]_post",
        "CRE[SQ001]": "CRE[SQ001]_post",
    }
}
```

Figure 1. Column Mapping.

- Cleaning and Standardization: Critical cleaning steps included handling missing values—such as ensuring that response scales were consistent and that data formats across the two surveys were standardized. Missing values were dealt with by either dropping a respondent if there were significant gaps in their data or imputing where it seemed appropriate based on the distribution of other responses.
- Data Coding: Ensured all the response scales were standardized so any Likert scale responses, say 1-5, were placed on both the pre and post-data sets. For example, it was meant to have a "4" in the pre-assessment, a direct equivalent of a "4" in the post-assessment.
- Merging Data: Paired pre and post-survey data based on unique participant identifiers to conduct a comparison analysis. This merging allowed a side-by-side comparison of each student's responses before and after the workshop.

E. Data Analysis and Visualization

In this study, we used Python to clean, process, and visualize the pre- and post-assessment data. The following code was employed to read the CSV files containing the assessment data, clean the column names, and generate histograms to compare the distribution of responses before and after the makerspace workshop. The Python code used for data cleaning and visualization is available upon request or can be accessed at https://github.com/petranpap/21st-Century-Skills-Data.

IV. RESULTS

A. Descriptive Statistics

Descriptive statistics were compiled for pre- and post-datasets to provide an initial understanding of respondents' answers' distribution and central tendencies across all the skills surveyed. Figure 2 shows the distribution of pre-workshop survey results for critical thinking, collaboration, communication, and creativity, providing a baseline for comparison against post-workshop data. Figure 3 illustrates the post-workshop survey results, highlighting the shifts in responses that occurred following the intervention.

TABLE I. PRE-ASSESSMENT DESCRIPTIVE STATISTICS.

Skill	Mean	Median	Std Dev	Min	Max
Critical Thinking	3.8	4.0	0.9	2	5
Collaboration	4.1	4.0	0.8	3	5
Communication	3.9	4.0	0.7	2	5
Creativity	4.0	4.0	0.8	3	5

TABLE II. POST-ASSESSMENT DESCRIPTIVE STATISTICS.

Skill	Mean	Median	Std Dev	Min	Max
Critical Thinking	4.2	4.0	0.7	3	5
Collaboration	4.4	4.0	0.6	3	5
Communication	4.3	4.0	0.7	3	5
Creativity	4.5	4.5	0.6	4	5

B. Comparative Analysis

Paired t-tests were conducted that compared the pre and post-responses for each skill area to test if there were statistically significant changes. As shown in Figure 4, the comparative histograms provide a visual representation of the changes between pre- and post-workshop survey results, reinforcing the statistical findings from the paired t-tests.

TABLE III. PAIRED T-TEST RESULTS.

Skill	t-value	p-value
Critical Thinking	3.5	0.001
Collaboration	4.2	0.0005
Communication	3.8	0.0008
Creativity	4.5	0.0002

These results indicate significant improvements in all skill areas assessed, with p-values well below the standard threshold 0.05. The statistical analysis, conducted using Python, included a detailed examination of paired t-tests for each skill area. The results showed statistically significant improvements with p-values well below the 0.05 threshold, confirming the positive impact of makerspace workshops. Including confidence intervals for these improvements can provide additional statistical robustness.

C. Visualizations

Box plots were created to visualize how the responses' distribution looks in both pre- and post-datasets. These support the ability to explore changes in responses and determine if any trends appear to be significant. Figure 5 highlights the improvements in critical thinking, collaboration, communication, and creativity skills post-workshop, showing the distribution shifts and indicating which skills had the most significant enhancement.

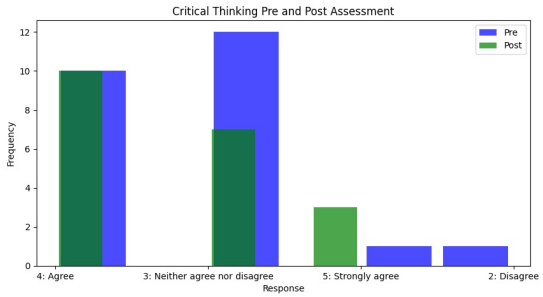

Figure 2. Histogram of Pre-workshop Survey Results.

V. DISCUSSION

Results of this study indicate that participation in the makerspace workshop significantly enhanced students' critical thinking, collaboration, communication, creativity skills. The differences in mean scores of pre and post-assessment, combined with low p-values from paired t-tests, demonstrate that the workshop positively influenced these essential skills.

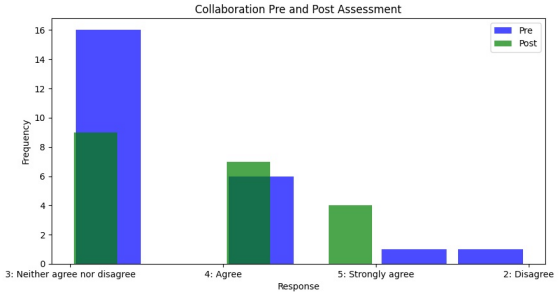

Figure 3. Histogram of Post-workshop Survey Results.

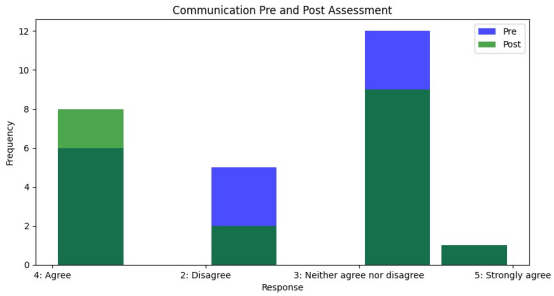

Figure 4. Comparative Histogram of Pre- and Post-workshop Survey Results.

A. Critical Thinking

Enhanced essential thinking skills suggest that the hands-on, project-based nature of these makerspace activities encouraged students to interact with the material in more profound ways, thoughtfully analyze problems, and develop more powerful reasoning skills. Critical thinking is foundational for students studying computer science, as they must solve problems and develop algorithms. This improvement suggests that the workshop successfully cultivated an environment for students to exercise and enhance their analytical skills.

B. Collaboration

Improvement in collaboration skills indicates that teamwork and peer interaction are integral to makerspace activities. This is essential in computer science, which often involves working in teams to develop software, solve problems, and innovate. The makerspace workshop provided many opportunities for students to collaborate, share ideas, and develop collaborative strategies.

C. Communication

Better communication skills can be attributed to the frequent presentations, project discussions, and feedback from peers and faculty. Effective communication is crucial for explaining complex technical concepts, documenting code, and collaborating with team members. The iterative process of sharing and refining ideas in the makerspace environment likely enhanced these skills.

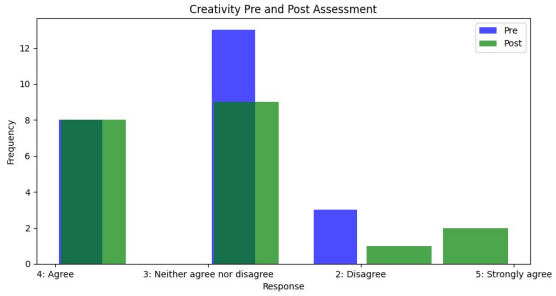

Figure 5. Histogram of Skill Improvement Across Different Dimensions.

D. Creativity

The notable rise in creativity scores mirrors the role of the makerspace in providing an open, flexible environment for experimentation, risk-taking, and exploring new ideas without fear of failure. Creativity is vital in computer science, where innovative solutions and technologies are continuously developed. The makerspace workshop encouraged students to think outside the box and explore new paths, fostering creative problem-solving abilities.

E. Alignment with Previous Research

These findings align with earlier research showing that experiential, hands-on learning environments like makerspaces can significantly enhance 21st-century skills [4], [5], [11]. The results highlight the utility of makerspaces in computer science education, which requires practical, project-based learning.

F. Limitations

This study has several limitations. The sample size was small and conducted at a single institution, potentially limiting its generalizability. Additionally, self-reported information may contain biases. Future studies should replicate these findings with larger, more representative samples and investigate the long-term impact of makerspace participation on students' skills. To mitigate potential biases associated with self-reported data, future studies should consider incorporating objective measures of skill improvement, such as performance-based assessments and peer evaluations.

G. Recommendations for Future Research

While this study focused on immediate skill improvements, future research should investigate the long-term impact of makerspace workshops on student skills. Longitudinal studies tracking students over several semesters could provide valuable insights into the sustained benefits of makerspace integration. Future research should focus on the generalizability of these findings across diverse educational settings and explore best practices for implementing makerspaces. Longitudinal studies following students over time could provide insights into the lasting effects of makerspace experiences. Additionally, investigating how makerspaces can enhance diversity and inclusion in computer science learning could help mitigate current disparities, ensuring that all students benefit from these innovations.

VI. CONCLUSION

This study provides compelling evidence that makerspace workshops can significantly enhance critical 21st-century skills among computer science students. The improvements in critical thinking, collaboration, communication, creativity highlight the value of integrating hands-on, experiential learning environments into the curriculum. These findings suggest that maker spaces play a crucial role in preparing students for the demands of the modern workforce by fostering essential skills relevant to their academic success and future professional endeavors. Educators and institutions should consider the benefits of incorporating maker spaces into their programs and explore ways to maximize the impact of these environments on student learning outcomes. For instance, embedding maker space activities within the core curriculum, providing faculty training on facilitating maker space projects, and ensuring access to various tools and resources can enhance the effectiveness of these spaces.

ACKNOWLEDGMENT

We would like to thank our colleagues at the Department of Computer Science at Cyprus College for their continuous support and encouragement throughout the study. Special thanks go to the students who participated in the makerspace workshops and provided valuable feedback through their survey responses. Your insights have been crucial to the success of this study. We also acknowledge the assistance provided by the educational technologists who helped integrate the makerspace activities into the curriculum. Your expertise and dedication have been invaluable.

REFERENCES

[1] K. D. Julian and D. J. Parrott, "Makerspaces in the library: Science in a student's hands," *Journal of Learning Spaces*, vol. 6, no. 2, pp. 13-21, 2017. [Online]. Available: https://files.eric.ed.gov/fulltext/EJ1152687.pdf.

[2] S. Papert, *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books, Inc., 1980.

[3] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.

[4] P. Blikstein, "Digital fabrication and 'making' in education: The democratization of invention," in *FabLabs: Of Machines, Makers and Inventors*, Transcript Publishers, 2013, pp. 1-21.

[5] E. R. Halverson and K. Sheridan, "The maker movement in education," *Harvard Educational Review*, vol. 84, no. 4, pp. 495-504, 2014.

[6] S. L. Martinez and G. Stager, *Invent to Learn: Making, Tinkering, and Engineering in the Classroom*. Constructing Modern Knowledge Press, 2013.

[7] B. K. Litts, "Making learning: Makerspaces as learning environments," Ph.D. dissertation, University of Wisconsin-Madison, 2015.

[8] T. Kelley, J. Knowles, J. Han, and E. Sung, "Creating a 21st Century Skills Survey Instrument for High School Students," *American Journal of Educational Research*, vol. 7, no. 8, pp. 583-590, 2019.

[9] K. M. Sheridan, E. R. Halverson, B. K. Litts, L. Brahms, L. Jacobs-Priebe, and T. Owens, "Learning in the making: A comparative case study of three makerspaces," *Harvard Educational Review*, vol. 84, no. 4, pp. 505-531, 2014.

[10] L. Martin, "The promise of the maker movement for education," *Journal of Pre-College Engineering Education Research (J-PEER)*, vol. 5, no. 1, pp. 4, 2015.

[11] B. Bevan, J. P. Gutwill, M. Petrich, and K. Wilkinson, "Learning through STEM-rich tinkering: Findings from a jointly negotiated research project

taken up in practice," *Science Education*, vol. 99, no. 1, pp. 98-120, 2015.

[12] K. A. Peppler and S. Bender, "Maker movement spreads innovation one project at a time," *Phi Delta Kappan*, vol. 95, no. 3, pp. 22-27, 2013.

[13] A. Hira, C. H. Joslyn, and M. M. Hynes, "Classroom makerspaces: Identifying the opportunities and challenges," in *IEEE Frontiers in Education Conference (FIE) Proceedings*, IEEE, 2014, pp. 1-5.

[14] M. Schmidt and R. Huang, "Defining learning experience design: Voices from the field of learning design & technology," *TechTrends*, vol. 66, pp. 1-10, 2021. [Online]. Available: https://doi.org/10.1007/s11528-021-00656-y.

Ontology-Based Integration of Occupational Health Data: Method and Case Studies

Cassandra Barbey
Department of Pollutant Metrology, INRS
Univ Rennes, Inserm, EHESP, Irset- UMR_S 1085
Vandœuvre-lès-Nancy, France
Email: cassandra.barbey@inrs.fr

Malika Smaïl-Tabbone
LORIA
Université de Lorraine, CNRS, LORIA, UMR 7503,
Vandœuvre-lès-Nancy, France
Email: malika.smail@loria.fr

Nathalie Bonvallot
Univ Rennes, Inserm, EHESP, Irset - UMR_S 1085
Rennes, France
Email: nathalie.bonvallot@ehesp.fr

Frédéric Clerc
Department of Pollutant Metrology, INRS
Vandœuvre-lès-Nancy, France
Email: frederic.clerc@inrs.fr

***Abstract*— The data related to occupational health exhibit diverse characteristics and are not inherently designed to interoperate; however, they contain complementary information. The integration of such data has the potential to enhance the current understanding of occupational health risks. Therefore, the objective of this study is to analyse heterogeneous data derived from 10 French occupational databases provided by 6 French institutes. An Ontology-Based Data Integration (OBDI) approach was employed, involving the mapping of data sources to a domain-specific ontology, namely the Occupational Exposure Ontology (OExO). In addition to OExO, four other ontologies were utilised: the Occupational Exposure Thesaurus (TEP) for occupational nuisances or hazards, the International Classification of Diseases (ICD-10) for medical conditions, the French Nomenclature of Activities (NAF) for industry sectors, and the Professions and Socio-professional Categories (PCS) for occupational classifications. The integration of these data is primarily achieved through the concept of the "occupational group", defined as a cohort of individuals of the same gender, engaged in the same occupation, and employed within the same industry sector. The study presents two case studies derived from the integrated knowledge base: a quantitative analysis identifying occupational groups with the highest exposure to nuisances and disease prevalence, and a qualitative analysis evaluating the consistency of information associated with each nuisance and disease.**

Keywords*— *ontologies; integration data; heterogeneous data; occupational health.

I. INTRODUCTION

Workers are exposed to several occupational nuisances that can have an effect on their safety or health and lead to occupational accidents or diseases. Moreover, interactions between these nuisances can affect health differently, reducing the effectiveness of risk mitigation measures often designed for single nuisances. The implementation of relevant preventive actions requires knowledge about these interactions, which is still limited.

In France, several national organisations collect occupational nuisance data and health data from surveys, declarations, or surveillance systems. These databases have different characteristics (objectives, collection method, target population, etc.) and provide much information but they were not designed to be used jointly. Some databases have been created to be representative of the population of French workers, while others exhibit more restricted scope. The information they contain may also be different, in line with their initial objective (to monitor, reference, describe, encourage, analyse, group together, etc.).

Analytical methodologies whose aim is to use occupational health data from several databases related to occupational risks prevention together have been identified. Some studies focus on a specific subject, such as that of L. Rollin et al. [1] about the occupational diseases faced by women in the homecare sector. Others are part of a larger project, such as Datamining project [2], in which both administrative recorded data and data from surveys are used. Following the same path of integrating health-related data related to the surveillance of elderly people, Dandan et al. [3] attempted to formalise the knowledge using ontologies for integrating data from sensors, surveys and personal health records. The DataPOST project is part of this trend, with the aim of developing a methodology for extracting knowledge about occupational nuisances and health outcomes, while relying on an ontological approach in order to bring together information from ten databases.

For this purpose, the data are integrated using an Ontology-Based Data Integration (OBDI) approach [4], and used to qualify and quantify multiple nuisances and health effects. The statistical unit used for analysis is named “occupational group”, which is a set of individuals of the same sex sharing the same occupation and working in the same sector of activity. The variables used for the definition of an occupational group are defined in all databases. These integrated data are then used in various analyses. We selected two examples that will be presented in Sections 5 and 6:

- A quantitative analysis to measure the degree of exposure of occupational groups to several nuisances and diseases by creating relevant indicators.
- A qualitative analysis to check the consistency of the data on each nuisance and disease provided by the databases and the validity of the relevant indicators constructed.

The rest of the article is structured into 5 sections: the introduction is followed by Sections 2 and 3, which define OBDI and OExO; Section 4 details the general representation of the data using Sections 2 and 3. Sections 5 and 6 present the case studies, detailing the methodology and results. Section 7 concludes and presents the outlook for the future.

II. ONTOLOGY-BASED DATA INTEGRATION (OBDI) APPROACH

The ODBI approach aims at integrating heterogeneous data by leveraging on an ontology that contains a semantic description of the concepts and their relationships in a domain of interest. This approach is built around three elements: the data sources, a heterogeneous repository where data are stored; the domain ontology, a formal description of that domain made by the organisations involved; and the mapping between them acting as the reconciliation structure (Figure 1).

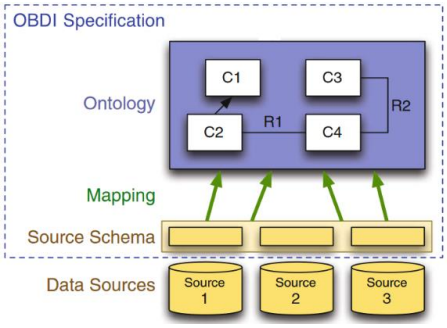

Figure 1. Ontology-based data integration adapted from Calvanese et al. (2017) [4].

This approach will be our methodological starting point. The context of our study is more complex, and several ontologies are required to integrate the available data and their relationships. It is necessary to have an integrated representation of data and ontologies similarly to the Occupational Exposure Ontology (OExO) [5].

III. OCCUPATIONAL EXPOSURE ONTOLOGY (OEXO)

The central knowledge representation used for this study relies on OExO, which is itself an extension of the Exposure science Ontology [6]. OExO consists of four central nuisance concepts: receptor, stressor, event and outcome; each of which is described by several child terms and attributes. The receptor is an individual worker or a population of workers that may be exposed to a stressor. The stressor represents an agent, activity or event that can affect the nuisance receptor, a chemical substance for example. The interaction between the two is called an exposure event that can lead to a health outcome, a disease for example (Figure 2).

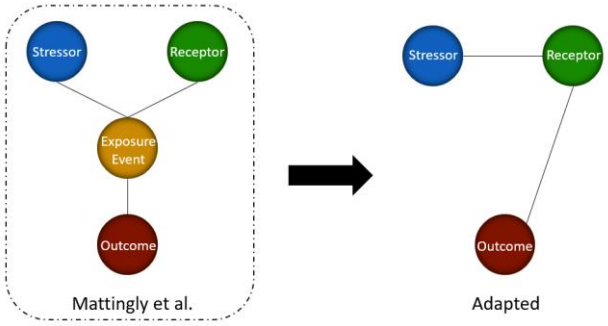

Figure 2. Main concepts of the exposure science ontology from Mattingly et al. (2012) [6] and their adaptation to our context

In this work, the receptor corresponds to the occupational group and the stressor to the occupational nuisance. In this work, unlike to OExO, the exposure event concept is not defined and is therefore not used. However, the available information in our databases allows us to link the occupational group the outcome.

IV. GENERAL REPRESENTATION OF THE HEALTH OCCUPATIONAL DATA USING ODBI AND ADAPTED OEXO

In order to integrate health occupational data using ODBI and adapted OExO main concepts, we grouped together several ontologies and related them to the available data, in the form of a knowledge base formalised as a conceptual data model (Figure 3). The following paragraphs will describe the ontologies we used for the Stressor/Receptor/Outcome concepts, the data sources, and the mappings we had to define between the latter and the former.

A. Ontologies

Four ontologies are used:

- The Occupational Exposure Thesaurus (TEP) [7], used to characterise and group the nuisances. This is a reference system designed in 2014 by the French agency for health safety to collect uniformly data on occupational nuisances. TEP is organised in 8 hierarchical levels representing around 8,300 nuisance concepts.
- ICD-10 is the tenth revision of the international classification of diseases [8], an international compilation on the causes and consequences of human disease designed and maintained by the World Health Organisation. It provides a common health language by using around 150 000 codify clinical terms [9]. It consists of 22 chapters subdivided into several blocks of three-character categories which can also be subdivided into four-character subcategories.
- The statistical classification of economic activities in the European Community is used to organise the information about economic and social activities. In this case, its French version named NAF [10] is used for the

definition of the occupational group. It is divided into 5 nested levels.

- The Professions and Socio-professional Categories (PCS) [11] results from a statistical classification conducted by the National Institute of Statistics and Economic Studies that brings together occupations from the same social background. This ontology is used for the definition of the occupational groups. It is divided into 4 nested levels of job designations.

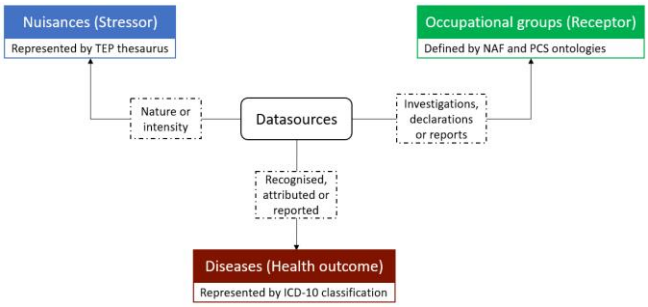

Figure 3. General representation of used data and ontologies, linked to the adapted OExO ontology.

B. *Data sources and their schema*

Ten data sources are used. Six of them provide information on occupational nuisances: SUMER [1], C2P [12], COLCHIC and SCOLA [13], COLPHY [14] and MatGene [15]; one data source concerns occupational diseases: AT-MP [16]; and three on both: Evrest [1], MCP [1] and RNV3P [1]. These data sources are described in Table I.

TABLE I. TABLE DESCRIBING THE DIFFERENT DATA SOURCES AND THE INFORMATION THEY CONTAIN

Data source	Collection method	Original statistical unit	Content	Example
Sumer	National surveys	Worker	340 columns representing nuisances to which workers are exposed to.	Nuisance to lead [yes ; no]
C2P	Regulatory Declarations	Worker	10 columns representing nuisances to which employers declared the worker are exposed to.	Nuisance to repetitive movements [yes ; no]
Colchic / Scola	Sampling and analysis of workplace air by specialised chemistry laboratories	Measurement	460 columns representing the measurement of the Intensity of the concentration of 230 substances in the air with regards to the regulatory limit value.	Lead Intensity [moderate ; high ; very high]
Colphy	Historical measurements and sampling	Measurement	4 columns representing the measurement of the intensity of the emissivity of 2 physical nuisances with regards to the regulatory limit value.	Whole body vibration Intensity [moderate ; high ; very high]
MatGene	Historical and census information	Occupational group	4 columns representing the nature and intensity of nuisance according to occupation.	Night work [yes ; no]
AT-MP	Medical consultation	Worker	86 columns representing occupational recognised diseases among workers.	Spondylopathies [yes ; no]
Evrest	Systematic occupational health interviews	Worker	45 columns representing the percentage of workers concerns by nuisance. 15 columns representing clinical signs. 15 columns representing first treatment.	Noise [yes ; no] Treatment for hearing problems [yes ; no]
MCP	Compulsory professional medical consultation	Worker	720 columns representing nuisances associated with reported occupational diseases and 56 columns representing these work-related diseases.	[Allergic contact dermatitis] [Chemical agents]
RNV3P	Medical consultation with a specialist of CCPPE	Health problem	129 columns representing the occupational diseases identified and 908 columns representing the nuisances probably linked to these diseases.	[Scoliosis] [Heavy loads ; Awkward postures]

C. *Mapping between data schemas and ontologies*

The mapping between data and ontologies is a 3-stage process:

- Mapping to define “occupational groups”: All combinations of variables relating to sector of activity (NAF), occupation (PCS) and sex are created.
- Mapping to standardise “health outcomes”: each variable in the AT-MP, MCP and RNV3P data are associated to the disease codes present in the ICD-10 classification.
- Mapping to define occupational nuisances: each nuisance variable in the data sources is associated to the nuisance it represents in the TEP. For example, the “extreme temperatures” variable in C2P, which refers to all temperatures below or equal to 5°C or at least equal to 30°C, will be linked to the “extreme thermal environment” nuisance in the TEP. This part was carried out on data from multiple sources (SUMER, MatGene, C2P, Evrest, COLCHIC/SCOLA, COLPHY, MCP, RNV3P).

The integrated data are then used in two case studies, the results of which are presented for the construction sector. 12,835 occupational groups were created, including 816 for the construction sector. Information is available for 308 nuisances and 174 diseases. The analyses were carried out using RStudio software, an integrated development environment (v.4.3.2) [17].

V. CASE STUDY ONE: QUANTITATIVE ANALYSIS

An indicator is created for each nuisance and disease to represent its importance for each occupational group.

A. *Indicator construction method for simple nuisance or disease*

The nature of information contained in databases can be:

• "Quantification": number of workers exposed to the nuisance (SUMER and MatGene).

• "Declarative": number of workers who declared (or have been declared by their employer) to be exposed to the nuisance (C2P and Evrest).

• "Intensity": number of intensity assessments to the nuisance recorded and maximum intensity of the nuisance (COLCHIC/SCOLA and COLPHY).

• "Plausibility": number of occupational nuisances which are the cause of worker's diseases as assessed by occupational physicians (MCP and RNV3P).

• "Disease": number of occupational diseases (AT-MP, MCP and RNV3P).

As the data are heterogeneous, the values stored for each type of information do not have the same scale (see Table 1 for an example). Therefore, the raw values were converted into non-parametric values. To achieve this, the original data source values were discretised into a scale of 1 to 10, with 1 representing 'very low' and 10 representing 'very high.' For each type of information, a score was computed: the arithmetic average of the discretised values divided by 10, ensuring the score value is between 0 and 1. The four nuisance scores were summed up to define a nuisance indicator. The disease score was used as is.

Figure 4 shows the main stages in the creation and construction of indicators.

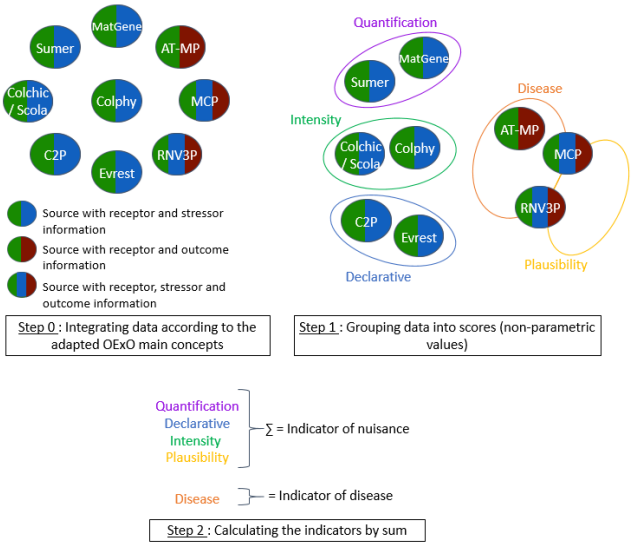

Figure 4. Main stages in the indicator construction method.

B. *Example: focus on the occupational group "43_632_1"*

The occupational group "43_632_1" represents skilled male workers in the "special construction" activity sector. This group is exposed to manual handling of heavy loads and suffering from arthrosis. The data confirms nuisance in the SUMER, C2P, MCP and RNV3P sources, as well as the presence of diseases in the AT-MP, MCP and RNV3P sources (Table II).

Indicator of nuisance "manual handling of heavy loads": The "quantification" score corresponds to the value present in the SUMER data source, MatGene having no information concerning this nuisance. The "declarative" score corresponds to the value present in the C2P data source, Evrest having no information concerning this nuisance. The "intensity" score is 0, COLCHIC/SCOLA and COLPHY having no information concerning this nuisance. The MCP and RNV3P data sources provide information; the "plausibility" score is calculated using both sources.

Indicator of disease "arthrosis": The AT-MP, MCP and RNV3P data sources provide information; the "disease" score is calculated using all three sources.

TABLE II. SUMMARY TABLE OF DATA ON MANUAL HANDLING OF HEAVY LOADS AND ARTHROSIS FOR THE OCCUPATIONAL GROUP "43_632_1"

	Data source	***Original data source value***	***Discretised value***	***Scores***	***Indicator***
Manual handling of heavy loads	Sumer	301,078.6	10	Quantification : 0.7	Nuisance indicator : 2.7
	MatGene	/	/		
	C2P	764.14	10	Declarative : 1	
	Evrest	/	/		
	Colchic /Scola	/	/	Intensity : 0	
	Colphy	/	/		
	MCP	321	10	Plausibility : 1	
	RNV3P	362	10		
Arthrosis	MCP	45	10	Disease : 10	Disease indicator : 10
	RNV3P	27	10		
	AT-MP	204	10		

C. *Heatmap: visualisation of nuisances and diseases indicators in the construction sector*

In the construction sector, 308 indicators for nuisances and 174 indicators for diseases were created. All the indicators are then represented in the form of a heatmap showing the occupational groups most at risk and most affected by diseases (Figure 5). The x-axis represents the various occupational groups, arranged in ascending order based on the number of exposures or diseases. Occupational groups with fewer exposures or diseases are positioned on the left, while those with a greater number of exposures or diseases are placed on the right. The y-axis displays the nuisances or diseases, ordered by the cumulative sum of their respective indicators. The upper portion of the y-axis corresponds to nuisances or diseases for which a large

number of occupational groups are exposed or affected, whereas the lower portion indicates nuisances or diseases associated with a smaller number of exposed or affected occupational groups.

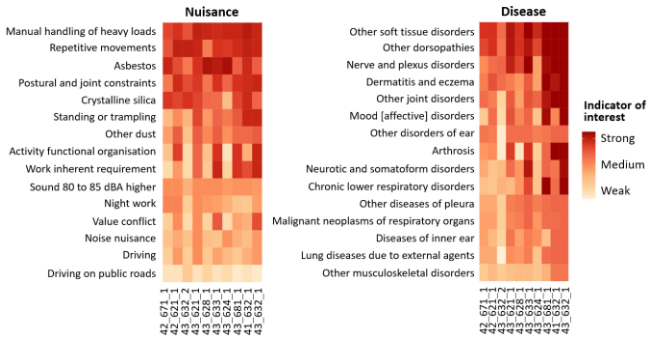

Figure 5. Heatmap of indicators by occupational group and by nuisance and disease: focus on 10 sickest occupational group and 15 nuisances and diseases.

For example, the occupational group "43_632_1" mentioned above is the most exposed and the most affected by occupational diseases because there is a lot of high valued indicators for nuisances and diseases. The most exposed occupational groups are particularly exposed to physical nuisances, such as postural and joint constraints, manual handling of heavy loads, repetitive movement, stranding and trampling, hand-arm vibration, as well as to certain chemical nuisances, such as asbestos and crystalline silica; nuisances typically expected in the construction sector. These groups are also heavily affected by musculoskeletal disorders like dorsopathies, nerve root and plexus disorders; diseases typically present in the construction sector. This work highlights the need to strengthen the preventive measures currently in place, for example by powered exoskeletons for reducing the risks related to the manual handling of heavy loads [18].

VI. CASE STUDY TWO: QUALITATIVE ANALYSIS

Our objective here is to measure the consistency of information contained in distinct databases in order to assess the complementarity of such sources. To do so, we created a consistency score for each nuisance and disease.

A. *Consistency score construction method*

This score is defined as the number of sources containing a value greater than 0 for a nuisance or a disease. The higher the number of sources, the greater the consistency between them.

B. *Example: focus on the occupational group "43_632_1"*

Following the above example on the occupational group "43_632_1", SUMER, C2P, MCP and RNV3P data sources confirm exposure to the nuisance "manual handling of heavy loads" with values greater than 0. The MatGene, COLCHIC/SCOLA and COLPHY data sources do not contain any information about this nuisance. They are therefore not taken into consideration when calculating the consistency score. The AT-MP, MCP and RNV3P data sources confirm the presence of disease "arthrosis" with also values greater than 0. The consistency scores for manual handling of heavy loads and for arthrosis will therefore be strong.

C. *Heatmap modelling of consistency score in the construction sector*

These scores are also represented in the form of a heatmap showing the groups for which the consistency of the information is strongest (Figure 6). The x-axis in this figure corresponds to that of the preceding heatmap. However, the y-axis differs slightly, as the values depicted here are derived from the consistency score rather than from the indicators. The upper portion of the y-axis indicates the nuisances or diseases for which the information is most coherent, while the lower portion reflects the nuisances or diseases associated with a smaller number of occupational groups demonstrating high consistency.

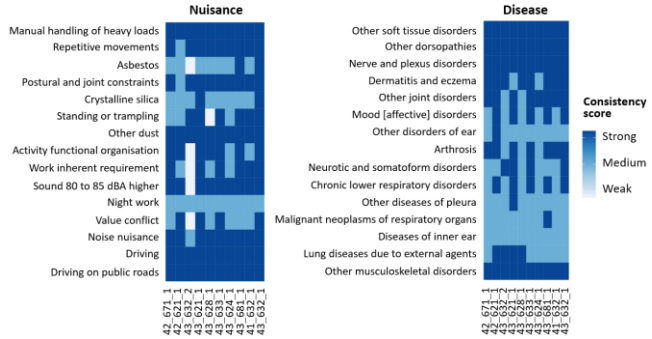

Figure 6. Heatmap of the consistency score by occupational group and by nuisance and disease: focus on 10 sickest occupational group and 15 nuisances and diseases.

For example, the "43_632_1" occupational group consistency scores for arthrosis and manual handling of heavy loads are high. We can also see that this group has a high representation of high scores for nuisances and a more moderate representation for diseases. The consistency scores are overall lower for diseases than for nuisances. This is due to the imbalance in the visibility of diseases. Some diseases that could be recognised as occupational are often not declared as such (hearing loss, for example). Information on these diseases are therefore not included in the main database on occupational diseases (AT-MP). However, these diseases may be attributed or reported in other databases. It highlights the complementarity of disease-related data sources and the need for integrating all data sources in order to get a broader picture.

VII. CONCLUSION AND FUTURE WORK

Structuring the data using the OBDI approach allowed us to implement two approaches to analyse the occupational health data. To the best of our knowledge, this study is the first attempt of integrating 10 heterogeneous data sources relying on four domain ontologies for the construction of indicators allows visualising information and highlighting occupational groups exposed to multiple nuisances and

victims of diseases. The consistency score is useful to check the validity of these indicators and the complementarity of the data.

The OExO ontology fits well with the concepts covered in the available data, although some modifications were applied to adapt it to our context. However, we could consider the reverse approach: adjust the data to the ontology. In this case, a data row in each source could be considered as a "nuisance event", such as defined in OExO. For this, we would need to explore further the ontology concepts related to nuisance events, in particular "assay" corresponding to all the features needed to assess the "nuisance event".

The two case studies presented here are examples of what can be done with data. We consider that our methodological proposal for data integration will enable us to integrate other data sources without any difficulties. Thus, a straightforward extension of the first use case would be to insert other data concerning the total number of workers per occupational group, in order to better assess the proportions of exposed and diseased workers.

The contextualisation of the data we propose opens new perspectives for their use and analyses for risk assessment purposes. For example, it would be possible to search for correlations between nuisance or co-nuisance indicators and disease indicators, in order to identify the main risks and subsequently create a tool for risk assessment [19]. Another possible perspective would be to use the indicators to establish worker nuisance profiles to centralise information on all possible nuisances or co-nuisances in a specific occupation [20]. These profiles could then be used to anticipate future occupational diseases and implement job-specific safety measures. Other work in the health sector is focused on the contextualisation of heterogeneous data to define a common semantic to facilitate knowledge sharing [21][22]. This generalisation would enable the development of responses adapted to current and future health concerns by means of tools and queries. It would also open the door to interdisciplinarity, and provide knowledge based on less theoretical situations.

We would like to generalise our methodology so that it can accommodate other data sources and subsequently facilitate data sharing. To achieve this, improvements are envisaged, notably in the mapping between data and ontologies, which remains complex, particularly regarding the occupations.

ACKNOWLEDGEMENT

The authors thank E. Algava and M. Duval (Dares), L. Meunier (CNAM), C. Pilorget, J. Chatelot and J. Homère (SPF), C. Nisse and A. Aachimi (Anses), L. Rollin and A. Leroyer (Evrest) for providing the data.

REFERENCES

[1] L. Rollin et al., "Complementarity of 4 data bases in occupational health", *Arch. Mal. Prof. Environ.*, vol. 82, no. 3, pp. 261-276, May 2021, doi: 10.1016/j.admp.2020.11.002.

[2] https://data.risquesautravail.be/fr, [last accessed Sept., 2024].

[3] R. Dandan, S. Despres, and J. Nobecourt, "OAFE: An Ontology for the Description of Elderly Activities", in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Nov. 2018, pp. 396-403. doi: 10.1109/SITIS.2018.00068.

[4] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati, "Ontology-Based Data Access and Integration", in *Encyclopedia of Database Systems*, 2017, pp. 1-7. doi: 10.1007/978-1-4899-7993-3_80667-1.

[5] D. A. Vallero, "A Draft Ontology for Occupational Exposure", Oct. 2016, doi: 10.13140/RG.2.2.16261.14564.

[6] C. J. Mattingly, T. E. McKone, M. A. Callahan, J. A. Blake, and E. A. C. Hubal, "Providing the Missing Link: the Exposure Science Ontology ExO", *Environ. Sci. Technol.*, vol. 46, no. 6, pp. 3046-3053, Mar. 2012, doi: 10.1021/es2033857.

[7] J. Bloch et al., "National Network for Monitoring Prevention of an Occupational Disease (RNV3P) - 2022 Annual Report", Oct. 2023.

[8] https://www.who.int, [last accessed Sept. 2024].

[9] J. A. Hirsch et al., "ICD-10: History and Context", *Am. J. Neuroradiol.*, vol. 37, no. 4, pp. 596-599, Apr. 2016, doi: 10.3174/ajnr.A4696.

[10] https://www.insee.fr, [last accessed Sept. 2024].

[11] https://www.nomenclature-pcs.fr, [last accessed Sept. 2024].

[12] https://entreprendre.service-public.fr, [last accessed Sept. 2024].

[13] G. Mater, C. Paris, and J. Lavoué, "Descriptive analysis and comparison of two French occupational exposure databases: COLCHIC and SCOLA", *Am. J. Ind. Med.*, vol. 59, no. 5, pp. 379-391, May 2016, doi: 10.1002/ajim.22569.

[14] https://www.inrs.fr

[15] J. Févotte et al., "Matgéné: A Program to Develop Job-Exposure Matrices in the General Population in France", *Ann. Occup. Hyg.*, vol. 55, no. 8, pp. 865-878, Sept. 2011, doi: 10.1093/annhyg/mer067.

[16] https://www.service-public.fr, [last accessed Sept. 2024].

[17] https://docs.posit.co/ide/user/, [last accessed Sept. 2024].

[18] Z. Zhenhua, A. Dutta, and F. Dai, "Exoskeletons for manual material handling – A review and implication for construction applications", *Autom. Constr.*, vol. 122, Feb. 2021, doi: 10.1016/j.autcon.2020.103493.

[19] A. J. Williams, J. C. Lambert, K. Thayer, and J.-L. C. M. Dorne, "Sourcing data on chemical properties and hazard data from the US-EPA CompTox Chemicals Dashboard: A practical guide for human risk assessment", *Environ. Int.*, vol. 154, pp. 106566, Sept. 2021, doi: 10.1016/j.envint.2021.106566.

[20] C. Fourneau et al., "The French 2016-2020 National Occupational Health Plan: a better understanding of multiple exposures", *Environ. Risques Santé*, vol. 20, no. 4, pp. 377-382, Aug. 2021, doi: 10.1684/ers.2021.1570.

[21] R. R. Boyles, A. E. Thessen, A. Waldrop, and M. A. Haendel, "Ontology-based data integration for advancing toxicological knowledge", *Curr. Opin. Toxicol.*, vol. 16, pp. 67-74, Aug. 2019, doi: 10.1016/j.cotox.2019.05.005.

[22] R. R. Rao, K. Makkithaya, and N. Gupta, "Ontology based semantic representation for Public Health data integration", in *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, p. 357-362, Nov. 2014, doi: 10.1109/IC3I.2014.7019

Evaluation of Property Filtering Algorithms Using Tags for a Property Rental Recommendation Application

Sean Lynch, Stephen Anokye, Oisín Fitzpatrick, Fenna Kadir, Charlotte Martin, Nishant Kapur, Brendan Tierney, Andrea Curley, Damian Gordon
School of Computer Science, Technological University Dublin, Ireland
e-mail: brendan.tierney@tudublin.ie, damian.x.gordon@tudublin.ie, andrea.f.curley@tudublin.ie

***Abstract*— Renting Made Easy is a project to create an easy-to-use, aesthetically pleasing rental listing application. The application provides a range of functionalities, including a comprehensive search experience, applying and managing rent applications, and property browsing. Initially based on the Zillow data set, the application has been expanded to include data sets with location-based services and crime-related data. A core feature is to provide a user-driven feature search based on property Tags. Property filtering algorithms were evaluated to determine which one provides suitable properties to the end-users. These algorithms included k-Nearest Neighbors (kNN) and Collaborative filtering. Qualitative research was performed to assess the usefulness and accuracy of the filtering algorithms.**

Keywords - Data Science; Recommendation System; Collaborative Filtering; User Evaluation.

I. INTRODUCTION

Renting Made Easy (RME) is a collaborative project involving students with a diverse range of skills including front-end development, back-end development, Application Development Interfaces (APIs) development, data science and machine learning. The resulting application reduces the anxiety of prospective tenants when searching for accommodation by streamlining the process. The system was developed with data sets including property listing information provided by Zillow [1], coordinate data from Google Maps, and various crime data from Open Baltimore [5]. With this data, a set of property services scores and crime safety scores were created [2]. Additionally, the application includes a recommendation system that leverages the collected data to provide property suggestions to users. This system employs content-based filtering to match properties with user preferences ensuring personalized and relevant recommendations.

The evaluation approaches included usability testing, accessibility testing, cognitive walkthrough, the think-aloud protocol, and expert feedback. The feedback indicated that the application delivered an intuitive and easy-to-use property rental website that displays standard and novel property details to users more clearly than other existing websites and provides users with property suggestions using a built-in recommendation system.

The recommendation system was built using a combination of a Property Scoring System, Property Tag selection and filtering algorithms. The filtering algorithms evaluated included kNN recommendation system and User Collaborative Filtering using Cosine Similarity. These different approaches were evaluated using subject matter experts. These were selected from various industry-related roles including property rental agents, estate agents, property owners and renters in the 20–35-year-old range.

II. RENTING MADE EASY

The Renting Made Easy project was designed to enhance the user experience by allowing tenants to filter properties based on lifestyle suitability in different Baltimore (USA) regions. Each property listing featured detailed scores that evaluated the availability of nearby services, and the safety levels based on local crime data. To enhance the personalized experience, user profiles included features such as saved searches, favorite properties, and a section for tracking property applications.

An integral part of RME was its recommendation system which utilized content-based filtering to suggest properties. The project aimed to surpass existing websites by providing clearer, more detailed property information and tailored property suggestions, thereby ensuring a more user-friendly and efficient property rental process.

The project team comprised two developers focused on front-end software development and User eXperience (UX), one developer dedicated to back-end software development, and three specialists responsible for building the recommendation engine and managing the data infrastructure. Figure 1 illustrates the core technical architecture of the project.

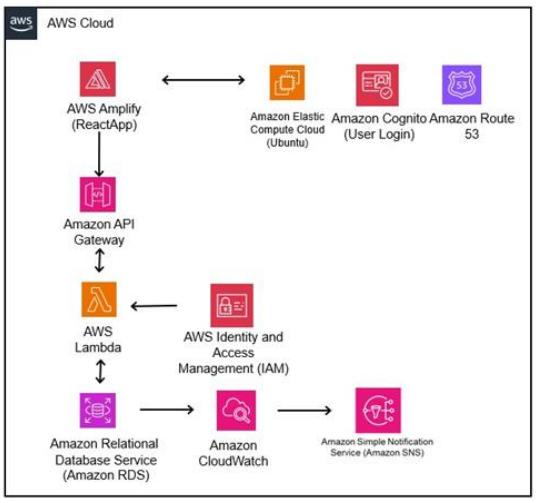

Figure 1. Basic System Architecture.

Architecture and deployment were cross-team responsibilities, as was project management. The technical architecture and technical flow (Figure 2) include: frontend, backend, data systems and data storage. Figure 2 illustrates the favourited properties and the recommendation engine process, including kNN and Cosine Similarity [3] [4].

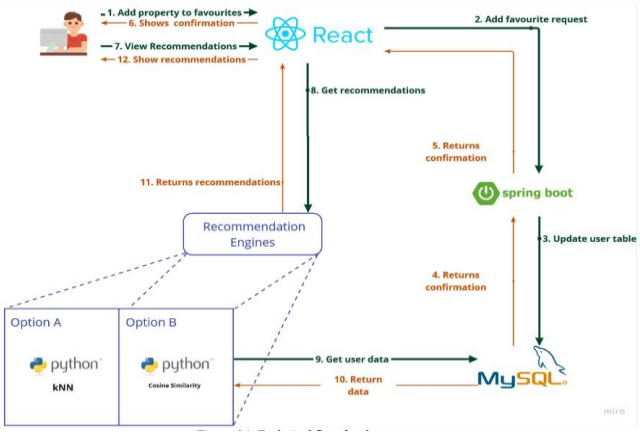

Figure 2. Technical Flow Incorporating Recommendation Engines.

In addition, data ingress from external APIs was incorporated. Due to legal limitations on storing data from Google's Map and Places APIs, a temporary caching strategy was required. This data is then combined with the Properties data set and used by the recommendation engines, scoring systems and tag categories. The entire application was deployed on Amazon Web Services (AWS) with AWS Lambda functions being used for the deployment of the recommendation engine.

III. RECOMMENDATION SYSTEMS

Property scores were generated for each property and displayed as values out of five to users. These scores provided a general overview of the area of each property. This first score displays the crime safety rating based on the neighbourhood of the property.

There was a data balancing issue when generating the safety score ratings. Some crime categories will naturally have higher counts than others due to their frequency which introduces a bias. This bias needs to be addressed to ensure that, when generating an overall crime score, these more common categories do not disproportionately influence the final result. Adjustments or normalization techniques should be applied to balance the impact of different crime categories, allowing for a more accurate and fair representation of crime levels. To solve this problem, a z-score standardisation was implemented. This generates a different score for each category in each neighbourhood. These scores were described in terms of their relationship to the mean, where their values are measured in terms of standard deviations from the mean. For consistency across the application the z-scores were mapped between one and five using sigmoid transformation. A sigmoid function was used over other mapping techniques following experimentation, including Min-max normalization, Winsorized min-max normalization, and Winsorized linear transformation. Using these mapping techniques resulted in unbalanced scores where the outputs were moving the values closer to either one or five. With a sigmoid mapping function, the values closer to the mean could be increased, preventing outliers from overpowering the results. Figure 3 displays the mapping function to map the z-scores. The sigmoid function was inverted, as the goal was to ensure that areas with a low crime rating have a higher safety rating.

The values closer to the mean were increased more than the values further away from the mean. This was because the extreme outliers with the higher summed z-scores were originally pushing these values closer to one after the mapping. This indicated that they were in safe areas. This was not the case. The values around the mean still had high levels of crime, but the distribution of the data was preventing this from being represented correctly. The sigmoid mapping function was shaped so that extreme outliers were increased marginally, and the values that lay before them were increased substantially.

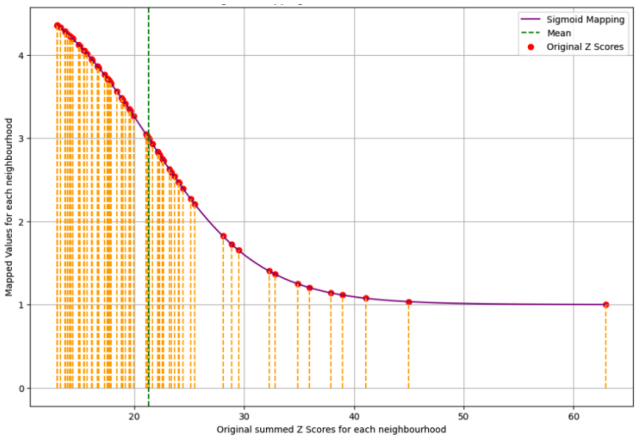

Figure 3. Sigmod Mapping of Z-Scores with Mean Life.

Finally, these z-scores were used to produce an overall crime safety score. The z-scores for each crime category were aggregated for each neighbourhood. Afterwards, the z-scores were standardized against each neighbourhood to generate the overall crime safety score. Using a sigmoid mapping function these z-scores were mapped between 1 and 5. The same approach was used to calculate the property nearby service score. These scores incorporated user weights into their calculation. Each weight represents a user's interest in each category.

A kNN model was incorporated within the system, including the steps for data pretreatment and feature selection. The kNN model obtained its data from a CSV file that included property listings and their corresponding attributes. The preprocessing steps included:

- **Dealing with Missing Values**: The absence of data can have a substantial effect on the accuracy of suggestions. Missing values in columns such as bathrooms, bathroomsFull, and bedrooms were imputed with zeros, based on the idea that the absence of a value may be adequately substituted by '0'.

- **Data Normalization**: This involved the use of MinMaxScaler. This ensured that these features had equal contributions in the distance calculations carried out by the kNN algorithm.
- **Categorical Encoding**: The OneHotEncoder was utilized to convert categorical variables into a format suitable for machine learning algorithms.

The feature selection process entailed selecting property qualities that have the greatest impact on a user's decision to rent. For example, considerations such as the number of bedrooms, cost, and accessible amenities were deemed crucial. The dataset has 72 features which were examined to determine their significance and influence on the recommendation results.

The kNN model was trained using the pre-processed and encoded dataset to select attributes closest in similarity based on their features. The training process entailed the following:

- Instantiating the kNN model using the NearestNeighbors class from the scikit-learn library.
- Applying the model to the dataset that has been both normalised and encoded.
- Evaluating the model using a certain attribute and obtaining the closest suggestions.

Once the model has been trained and validated, it was saved using *joblib* for persistence and reuse. The technical architecture incorporated the kNN recommendation system as a back-end service which is integrated into the wider RME platform. Upon completion of the analysis, the model generates property recommendations that are subsequently presented to the user.

The user collaborative filter system only incorporated click data. A click data point represents a user clicking on a property card. The model took a user as input and created a pattern for them. This pattern was represented as a vector with the number of properties in the database in it as its length and the number of clicks for each as the property's value. It took all of the other vectors for the other users in the database and compared the vectors to one another, looking for users with similar interactions. This was measured by finding the Cosine of the angles with values closest to zero between each vector. The formula used can be seen in equation (1) where A is the input vector, B is the comparison vector, n is the number of vectors to be compared, and i is the current index inside vectors A and B.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

The ten most similar users to the input user were retrieved and a new data set was made. This data set contained all the combined clicks for each property between all the recommended users. The properties were sorted in descending order by click count, and the first ten were displayed to the user. These were the ten most interacted properties between all recommended users. Once the Cosine similarity model was deployed to a lambda function, the data was stored temporality in the front-end. When the lambda function is called, the user can observe the recommendations presented to them. During the development of the Cosine similarity model other iterations of the model were created. These iterations were hybrid models that combined geocoordinate data as well as property data. The curse of dimensionality caused the model to struggle to find similarities upon initial testing. The Cosine of the angles between all users had a value close to one, this indicates that no similarity was found. For this reason, the model that incorporated click data was chosen to be deployed and evaluated during testing.

IV. EVALUATION

The recommendation system for both the kNN and User Collaborative Filtering models were evaluated using user feedback obtained from surveys. The recommendation systems gave recommendations based on content and user interactions. The recommendation engine's performance is based on user sentiment toward their recommendations. The surveys allowed for qualitative feedback that enabled the models to be adjusted based on user preferences. The evaluation of the recommendation system aims to determine:

- The suitability score for each feature in the recommended properties (based on user feedback).
- The overall suitability score of the recommended properties (based on user feedback).

Due to the interconnecting nature of the recommendation and tag systems, the evaluation of both was joined into a single questionnaire in conjunction with the scoring system. As such, the methodology and evaluation metrics are similar for the property tags and the recommendation system. As mentioned, property tags are designed to increase the usability of RME. They also confirm the recommended properties, for example, the user's favourite properties with 'secure' tags, and therefore secure properties should be recommended to the user. The two recommendation models were made available at different intervals with approximately half of the participants testing each model. Users did not know which recommendation system they were testing. The evaluation aimed to explore the tag contribution to the recommended properties, testing kNN and Cosine recommendation models.

Table 1 shows the results of Kendall's Tau which is used to measure the correlation coefficient between each of the suitability ratings of the kNN features and the overall suitability rating of the kNN model. The decision to measure the correlation between participant kNN feature suitability and participant kNN overall suitability scores was based on the actual values of the variables without assuming any specific underlying distribution. This approach focused on evaluating the similarity in the ordering of the data points.

Each feature defined a hypothesis as follows:

- Null hypothesis (H_0): Feature x does not impact the kNN overall suitability score.

- Alternative hypothesis (H_1): Feature x does impact the kNN overall suitability score.

A p-value of 0.05 was set for each test as this is a commonly used metric, where a p-value less than 0.05 is generally considered to be statistically significant and considered to be grounds for rejecting the null hypothesis.

TABLE 1. KENDALL'S TAU CORRELATION COEFFICIENT OF KNN FEATURES

Feature	P-value	Kendall's Tau	Correlation Strength	Failed to reject:
Available amenities	p = 0.191	r = 0.29	Weak correlation	H (0)
Nearby personal care	p = 0.069	r = 0.41	Moderate correlation	H (0)
Nearby banks	p = 0.049	r = 0.46	Moderation correlation	H (1)
Price	p = 0.007	r = 0.60	Strong correlation	H (1)
Nearby emergency services	p = 0.028	r = 0.50	Moderate correlation	H (1)
Nearby public transportation	p = 0.208	r = 0.29	Weak correlation	H (0)
Nearby leisure activities	p = 0.009	r = 0.59	Moderate correlation	H (1)
Nearby retail	p = 0.014	r = 0.56	Moderate correlation	H (1)
Nearby gyms	p = 0.060	r = 0.44	Moderate correlation	H (0)
Area safety	p = 0.210	r = 0.28	Weak correlation	H (0)
Number of bathrooms	p = 0.042	r = 0.46	Moderate correlation	H (1)
Number of bedrooms	p = 0.267	r = 0.25	Weak correlation	H (0)

Table 1 illustrates half of the input features were positively correlated with the overall recommended suitability score. Nearby banks, emergency services, leisure activities, retail, and bathroom count all appeared to be positively correlated with the overall suitability score, which was unexpected. Price appeared to have a strong correlation with the overall suitability score. The alternative hypothesis for each of these features failed to be rejected as they had p-values < 0.05. The relationship between the input features and the overall suitability score is not linear. The features with a positive correlation with the overall suitability score indicate there is a consistent but not constant relationship between the two variables.

When users were asked to rate their interest in these features, most results showed neutral and negative sentiment. Bathroom, retail, and price having a moderate–strong correlation was expected, as people expressed interest in these features when asked. It was expected that bedrooms, bathrooms, area safety, transportation, and available amenities would have a stronger positive correlation, with lower p-values due to the overall initial interest expressed by participants for these categories.

The only feature with a strong positive correlation with the overall suitability score was property price. The kNN model does not weight its parameters. This feature affects the result just as much as the other features inputted into the model do. It is possible for different property price ranges to be paired with similar values of the other popular categories that make up the majority input of the kNN's parameters. This could suggest the reason for the property price's strong correlation.

Before testing the user collaborative filtering model, it was presumed that user collaborative filtering would prosper when recommending properties to users with an interest in the property information. This information consists of price, property size, bathroom count, bedroom count, amenity, and tag information, all of which are displayed on the front of the property cards.

A click is intended to represent a genuine interest in a property. However, false signals of interest can occur if the information displayed on the property card highlights only certain 'key' points that attract the user's attention. For example, if a user is specifically searching for properties with a microwave, they might click on properties with an "amenities" tag, even if the property does not fully meet their other criteria. This can lead to misleading data about user preferences and interest levels.

Most users had a positive sentiment for the crime tags and nearby service scores. The crime tag is not based on weights. Changing the shape of the crime tag sigmoid function was the only way to modify the influence of the crime tag. The nearby service scores could be improved by changing the application of the weights with users specifying their preference using service weights.

V. CONCLUSION

The RME application was built to provide a better user experience for the end-user. A key component of this application was the recommendation system which included a combination of a Property Scoring System, Property Tag selection and filtering algorithms. The filtering algorithms evaluated included kNN recommendation system and User Collaborative Filtering using Cosine Similarity. These different approaches were evaluated using subject market experts. These were selected from various industry-related roles including property rental agents, estate agents, property owners and renters in the 20–35-year-old range. The outcomes from the initial testing demonstrated positive outcomes and feedback from the end-users with particular feedback pointing to the usefulness of the Tagging system and the inclusion of their preferences. As noted, the evaluation also revealed a relationship between the input features and the overall suitability score that was consistent but not constant. Future work will further develop the recommendation engines. This will involve expanding the dataset by incorporating data from different cities or geographic regions and increasing the number of subject matter experts to ensure a broader and more comprehensive analysis.

REFERENCES

[1] Zillow, Real Estate, Apartments, Mortgages & Home Values, www.zillow.com, date last accessed August 19, 2024.

[2] I. Borg, D. Hermann and W. Bilsky, "The perceived seriousness of crimes: inter-individual commonalities and differences." *Quality & Quantity*, vol. 57, pp. 765–784, 2023. https://doi.org/10.1007/s11135-022-01379-9

[3] A. Mucherino, P. J. Papajorgji and P. M. Pardalos, "*k*-Nearest Neighbor Classification." Data Mining in Agriculture. Springer Optimization and Its Applications, vol. 34, pp. 83-106, Springer, 2009. https://doi.org/10.1007/978-0-387-88615-2_4.

[4] B. Li, L. Han, "Distance Weighted Cosine Similarity Measure for Text Classification.", Intelligent Data Engineering and Automated Learning – IDEAL 2013, vol. 8206, pp. 611-618. https://doi.org/10.1007/978-3-642-41278-3_74

[5] Open Baltimore Datasets, Crime Data, https://data.baltimorecity.gov/search, date last accessed September 21, 2024.