UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI SCIENZA AZIENDALI - MANAGEMENT & INNOVATION SYSTEMS

Dottorato di ricerca in Big Data Management

XXXV Ciclo

Tesi di Dottorato

Towards Responsible AI Systems: A Methodological Approach for Integrating Explainable AI in Decision Support Systems

Tutor:

Ch.mo Prof.

Carlo Blundo

Candidato:

Roberto Abbruzzese

Mat. 8801500029

Coordinatore:

Ch.mo Prof.

Valerio Antonelli

ANNO ACCADEMICO 2021/2022

EXECUTIVE SUMMARY

Decision-making processes are critical to the success of any organisation, as they directly affect the company's ability to achieve its goals. Such processes over the years have become extremely "Knowledge Intensive", i.e., they have become collections of activities that require a significant amount of specialised knowledge or expertise. Moreover, they rely heavily on experience, intuition, and expert judgement, built on large volumes of information that, once handled, processed and analysed, turn into valuable knowledge, useful, precisely, to promote effective and efficient decisions for the company. The implementation of support technologies to optimise these types of processes is certainly represented by Decision Support Systems (DSS). Decision support systems are information systems that help users make decisions based on analyses of varying complexity performed on company and non-company data using various technologies. Over the years, DSS have undergone important evolutions, marked in particular by technological advances in the Information Technology domain, with particular emphasis in recent years on the progress made in the Artificial Intelligence (AI) domain. The integration of AI in the DSS stems, in fact, from the need to manage the growing complexity and volume of available data, linked, above all, to unstructured information. AI can rapidly process large amounts of data, identify patterns and trends, and provide valuable insights that can improve the efficiency and effectiveness of decisions.

The growing technological advancement of Natural Language Processing has also enabled the construction of AI-based services that are increasingly advanced and increasingly complex to manage and integrate into processes. One of the goals AI has demonstrated in recent years is certainly related to reducing the cognitive load on human decision-makers, allowing them to focus on more strategic and creative aspects of decision-making. However, the integration of AI into DSS presents various challenges. Problems related to prediction errors, discrimination, privacy, security, transparency and over-dependence on technology have been addressed in detail and summarised in this thesis work in order to be able to build a clear and precise overall view of the phenomenon and to enable the construction of tools capable of assessing and improving the social and ethical impact of the inappropriate use of AI. This in-depth literature study also included an analysis of the taxonomies, principles, and concepts of Explainable AI, which is also useful for constructing a common language. In order to alleviate and face the challenges promoted by the introduction of AI in decision-making processes, the main objective of this work is represented by the construction of a methodological framework, XAI4DSS, capable of supporting all the stakeholders of the process to promote the integration of AI technologies and systems in the processes, in compliance with the principles, regulations, while preserving an extreme added value in terms of performance. The methodological framework is described, elucidating all its main components, represented by phases accompanied by tasks and possible deliverables produced.

Specifically, the framework is divided into 6 stages, respectively:

1. **Context Understanding**: analysing the context and identifying priorities among different possible explanations considering different perspectives and factors.

2. **Requirement for data collection**: Define the useful requirements of the data collection process to improve the quality of the explanation provided to the various stakeholders.

3. **XAI Implementation**: Based on the requirements defined in the previous steps, the appropriate XAI techniques are developed for the appropriately selected AI models.

4. **XAI Identification**: Understanding the statistical rationale behind the models used.

5. **Design and Development of User Explanation**: Constructing and presenting information in a clear and accessible manner considering the requirements expressed by the previous phases.

6. **User Learning**: When human decision-makers are significantly involved in an AI-assisted outcome, they must be adequately trained and prepared to use the results of the model responsibly and fairly.

The research methodology for developing this thesis work contemplated a hybrid and integrated top-down and bottom-up approach. If, on the one hand, the study of Explainable AI started with a top-down phase, defining theories and general principles of XAI based on existing literature and consolidated knowledge, subsequently and simultaneously, a structured path of experimentation and extension of the XAI tools currently present in the literature was undertaken. These experiments, appropriately analysed, contributed to identifying specific models, techniques, and insights that extended and complemented the initial theories, creating an iterative cycle of learning and refining the understanding of XAI. This work reports three use cases that contributed to the construction of the methodological framework. The first scenario concerns using XAI to identify and explain pump-and-dump events in the cryptocurrency sector, using a tabular data analysis. The study demonstrated the feasibility of identifying Pump and Dump early, highlighting the superiority of financial exchanges in identification due to richer and semantically more meaningful data. Regarding the application of XAI techniques, both Feature Contribution Analysis and Single Instance Explanation demonstrated the importance of the new social feature introduced in the new approach. Furthermore, Single Instance Explanation aims to promote a concrete

explanation for each single prediction, providing an important and necessary tool for adopting the new model within a decision-making process. In the second scenario, transformer-type neural networks were used to address the issue of text segmentation in legal documents. The proposed approach proved effective, demonstrating excellent performance due to three fundamental components: a transformer-based architecture, a complex and important text pre-processing phase, and the introduction of a "spatial" feature on the alignment of rows. The study also highlighted the importance of generalisation in processing real, often perturbed or noisy data by analysing the robustness of the model. The results achieved are of particular relevance to integrating XAI models and techniques in the different development phases of the proposed textual segmentation model. The validation of the labelling process, Bias Labelling, which ensured the accuracy and reliability of the labels, was decisive in the performance evaluations since the errors committed during the labelling phase were detected in the early stages of development, leading to sub-optimal results. Furthermore, the demonstration of model robustness ensured that the model was resistant to various conditions and unexpected inputs. Last but not least, the XAI for Model Explainability application for the end user was created to ensure the comprehensibility and transparency of the results by clearly illustrating the model's decision-making process. In conclusion, the third proposed scenario stems from an idea developed to address the problems and performance observed in the second use case using the XAI. Situated in the context of the automatic classification of legal documents and the effect of anomalies caused by optical character recognition (OCR) systems, the proposed intuition aims to use XAI to refine the performance of a legal text classifier. This method uses a retroactive interpretation of the results of an XAI module to correct anomalies in OCR-scanned texts that adversely affect the performance of text classifiers. In particular, the approach provided by the framework, called REMOAC, involves the use of XAI approaches and techniques not only to make the interpretation and explanation of the system accessible but also to "retroactively" employ the results of XAI approaches (such as SHAP) in

order to improve the performance of a text classifier. This allows for accurate results even in real-world contexts, promoting targeted and specific correction of tokens within a sentence.

The present work, therefore, has addressed the ability of XAI techniques to meet various requirements for reliable AI in the context of DSS in the Enterprise domain, rationalising it all and synthesising what has been analysed in a methodological framework, capable of scanning the phases and tasks of an XAI integration process, not only to address the choice of the optimal tool based on performance and technological advances but to manage complexity and promote the importance of the principles of responsible AI, transferring the need for a deep analysis of the context and its processes. The main approach adopted in defining the framework was to promote the construction and integration of the XAI through human-centred co-design principles. Co-design is also seen as a participatory process that involves end users by making them active participants in the design process and facilitating effective and useful human-machine interactions. The experimentation and application of the tools in the various use cases identified also highlighted the operational and tangible importance of the XAI, particularly during the construction and development phases of machine learning systems. The retroactive use of XAI results has outlined a promising new approach to integrating XAI into AI systems through a proactive, real-time scenario, whereby the results of particular XAI techniques contribute directly to improving the performance of the analysed AI systems. The research directions outlined in this work imply the possibility of pursuing research studies in two directions. With regard to the methodological framework, future studies could focus on the construction of ad hoc tools to be correlated with the methodology to increase the framework's operability. For example, we could design and build tools such as decision matrices, evaluation checklists, compliance tools, and automatic generation modules to produce supporting documentation, all designed and developed with the aid of AI. On the other hand, as far as REMOAC is concerned, we could investigate new applications of the method and, more generally, explore the proactive use of

XAI techniques to improve the performance of AI models with an automated approach.