Sara Pido’ – XXXV Cycle

Advisor: Prof. Stefano Ceri

Thesis Title: Exploiting AI and NLP methods for Empowering Naïve Users in Solving Data Science Problems

Data Science (DS) and Machine Learning (ML) have become critical tools for making informed decisions, predicting outcomes, and automating processes. The rise of big data and the availability of powerful computers, coupled with the development of new and more sophisticated ML algorithms, has led to a huge growth of interest in ML over the past decade. Despite the significant advancements in ML methods, building and training ML models can still be complex and time-consuming, requiring expertise in computer science, mathematics, and statistics; taking advantage of ML can still be challenging, especially
for people without these skills. The significant gap in a deep understanding of machine learning principles among IT and business professionals has led to incidents related to bias, privacy, security, transparency, and ethical concerns. The democratization of data science and machine learning aims to change this situation, by making ML technologies and techniques more accessible to a wider range of people. This thesis discusses the difficulties non-experts face in using ML tools. It explores various approaches to democratize data science and machine learning, such as developing user-friendly ML tools and platforms and educational initiatives to teach the necessary skills to non-experts. It discusses the potential benefits of democratizing data science, such as driving innovation, providing new insights into complex problems, and creating a more inclusive and diverse data science community.
In particular, my research introduces a progression of methods and underlying tools that make use of conversational agents, natural language, and autoML, with the objective of democratizing data science and make it more accessible to a wider range of people.
The thesis begins by presenting GeCoAgent and DSBot, two multi-modal conversational agents designed to facilitate data science processes starting from natural language input.
GeCoAgent and DSBot are two distinct conversational agents that serve different purposes in the context of data science automation. GeCoAgent takes a proactive approach by driving the conversation with the user, asking detailed and specific questions to better understand the user’s needs and goals. On the other hand, DSBot is a user-driven conversational agent, where the user provides a research question in natural language and the bot extracts the necessary information and executes the relevant data science processes.
However, the automation of data science processes raises issues such as the difficulty in formulating the research question and the importance of incorporating domain expertise into the pipeline. To address these challenges, the thesis then presents two additional tools: MLFriend and Zephyr. MLFriend enables automatic generation of prediction tasks, while Zephyr streamlines the integration of domain expertise and automated data science tools. We conducted empirical evaluations and user studies to illustrate the effectiveness of these tools in making machine learning more accessible and user-friendly.

By providing these four solutions, embodied within GeCoAgent, DSBot, MLFriend, and Zephyr, we show a progressive development of ideas, methods and tools towards the goal of improving the accessibility and usability of data science tools for non-experts. Our research contributes to the field of democratization of DS by providing new strategies that can be used to reduce the gap between experts and non-experts in the field. We trust that our results will contribute to address the remaining challenges and opportunities and make machine learning more accessible and user-friendly for a wider range of people.