POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Large Language Models in Data Preparation: Opportunities and Challenges

Tesi di Laurea Magistrale in
Computer Science Engineering - Ingegneria Informatica

Author: **ANNA BARBERIO**

Student ID: 992665
Advisor: Prof. Cinzia Cappiello
Co-advisors: Dott. Ing Camilla Sancricca
Academic Year: 2022-23

Abstract

In the realm of data preparation and analysis pipelines, the crucial role of providing explanations emerges as the keystone for data analysis. As data-driven decision-making continues to gain significance, instilling confidence in tools that recommend optimal steps and techniques for enhancing datasets becomes dominant. While these tools assist users in improving datasets through automated decision-making processes, they often present enhanced datasets without adequate explanations. This lack of transparency leaves users unaware and uninformed about the changes made, potentially hindering trust in the suggested output.

To address this gap, various studies, highlight how incorporating explanations can offer users guidelines, reducing the opacity of automated machine learning processes. Consequently,our thesis work positions itself as the initial endeavor to provide explanations in data preparation, aiming to furnish users with valid explanations to enhance their understanding of the presented information.

The second crucial aspect of our work, connected to the initial point, revolves around the format of explanations. According to additional studies comparing diverse forms of explanation, natural language explanations have proven to be the most effective. For this reason, the thesis work involves delivering textual explanations using the large language model. Thus, in the proposed tool, the structure of explanations will be delegated to a Natural Language Processing (NLP) tool like ChatGPT, which is proficient in presenting explanations in a user-friendly format.

In addition to researching and assessing a methodology to integrate explanations into a standard data preparation tool, our contribution also encompasses an investigation into ChatGPT and its potential to offer explanations for a data preparation pipeline.

Keywords: Data Preparation,Description and Explanation ,Large language Model

Abstract in lingua italiana

Nel contesto delle pipeline di preparazione e nel campo della preparazione e analisi dei dati, emerge il ruolo cruciale di fornire spiegazioni come la chiave di volta per l' analisi dei dati.

Con l'aumentare dell'importanza delle decisioni basate sui dati, infondere fiducia negli strumenti che raccomandano passaggi ottimali e tecniche per migliorare i set di dati diventa predominante. Sebbene questi strumenti assistano gli utenti nel migliorare i set di dati attraverso processi decisionali automatizzati, spesso presentano set di dati migliorati senza spiegazioni adeguate.

Questa mancanza di trasparenza lascia gli utenti all'oscuro e non informati sulle modifiche apportate, potenzialmente ostacolando la fiducia nell'output suggerito. Per colmare questa lacuna, diversi studi evidenziano come l'inclusione di spiegazioni possa fornire agli utenti linee guida, riducendo l'opacità dei processi automatizzati di apprendimento automatico. Di conseguenza, il nostro lavoro di tesi si configura come il primo tentativo di fornire spiegazioni nella preparazione dei dati, con l'obiettivo di fornire agli utenti spiegazioni valide per migliorare la loro comprensione delle informazioni presentate.

Il secondo aspetto cruciale del nostro lavoro, collegato al punto iniziale, ruota attorno al formato delle spiegazioni. Secondo ulteriori studi che confrontano diverse forme di spiegazione, le spiegazioni in linguaggio naturale si sono dimostrate le più efficaci. Per questo motivo, il lavoro di tesi prevede la fornitura di spiegazioni testuali utilizzando il modello linguistico di grandi dimensioni. Così, nel tool proposto, la struttura delle spiegazioni sarà affidata a uno strumento di Elaborazione del Linguaggio Naturale (NLP) come ChatGPT, capace di presentare spiegazioni in un formato comprensibile agli utenti.

Oltre a studiare e valutare una metodologia per integrare spiegazioni in uno strumento standard di preparazione dati, il nostro contributo include anche un'indagine su ChatGPT e la sua potenziale utilità nell'offrire spiegazioni per una pipeline di preparazione dati.
Parole chiave: Preparazione dei dati,Descrizioni e Spiegazioni,modello linguistico avanzato

Contents

1 | Introduction

In today's data-driven world, the significance of high-quality data for well-informed decision-making cannot be stressed enough. It is imperative that data enhancement tools not only provide recommendations but also offer user-friendly guidance, allowing even those who lack the expertise to comprehend and act on these suggestions. The sophistication and complexity of data enhancement techniques have grown exponentially, making it challenging for users, particularly those without a solid background in data science, to grasp the nuances of the suggested improvements. While data enhancement tools can offer valuable insights and recommendations, these insights are often underutilized due to a lack of user comprehension.

Our research endeavors to fill this gap by constructing a framework that integrates descriptions and explanations into data enhancement tools. This integration serves as an educational resource, furnishing users with in-depth information about the recommended techniques and steps for enhancing datasets. Through descriptions, users gain a deeper understanding of the strengths and weaknesses of the datasets they input. Descriptions allow them to better interpret graphs and tables that highlight dataset characteristics. On the other hand, explanations guide users in the later stages of the dataset enhancement pipeline, elucidating the underlying reasons behind machine learning techniques that led to the recommendation of specific preparation methods.

Both descriptions and explanations are extracted with the assistance of an external natural language processing tool, which strives to present the explanations in the most descriptive and comprehensible manner possible.

The advantages of this approach are manifold. Empowering users with detailed descriptions and explanations makes data enhancement tools more accessible and less intimidating, even for those with limited data science skills. Users can follow the logic behind each recommendation, comprehend the objectives of each technique, and ultimately make informed decisions regarding dataset improvement.

Furthermore, user comprehension enhances trust. When users can see the solid logic and clear objectives behind the recommendations, they are more likely to rely on them

and take action accordingly. This not only bolsters data quality but also improves the effectiveness of the tool itself.

In an era where data-driven decision-making is omnipresent, this work has the potential to make every user, regardless of their level of expertise, capable of fully understanding how to enhance and work with the data at their disposal.

Thesis Structure

Chapter 2 Provides an overview of the essential data quality concepts and outlines the necessary steps for achieving higher data quality, introducing the innovative element of explanations in a way that aids users in better comprehension.

Chapter 3 Explains the approach taken in this study, emphasizing its objectives and the multiple stages that steer the user throughout the complete data Analysis process.

Chapter 4 The enumeration of technologies and the exploration of the setup choices used during the implementation. In this section, the preliminary research conducted on ChatGPT is presented.

Chapter 5 Shows the workflow of the web application, providing visual representations of each section's appearance.

Chapter 6 Summarizes the work accomplished and outlines potential future developments.

2 State of the art

2.1. Data Quality

In today's data-driven world, data analysis stands as a foundational skill that guides business strategies, fuels scientific research, influences policy decisions, and much more. It grants both organizations and individuals the ability to make informed choices supported by data, streamline operations, and foster innovation.[30] However, to ensure that the analysis is based on reliable and accurate data, one must address a crucial aspect, which is data quality. As data sources continue to expand in complexity and volume, data quality remains a paramount consideration at the forefront of extracting knowledge and value from data.

The term "data quality" can indeed be succinctly defined as "fitness for use."[28] This concise definition emphasizes the core principle that data quality is determined by how well data serves its intended purpose or meets the specific needs of its users. Data is considered of high quality when it aligns with the requirements and expectations of the tasks or analyses it is meant to support. This definition underscores the idea that data quality is not an absolute measure but a relative one, as data quality requirements may vary depending on the context and use case.

Now, a crucial aspect to bear in mind in achieving data quality is to perceive it as an ongoing improvement process. This journey necessitates a clearly defined pipeline consisting of multiple stages to maintain data's accuracy, reliability, and suitability for its intended use.

Figure 2.1: Data quality management process

First, data collection is the initial step. It is essential to comprehend the context and the objectives of data collection to lay a robust groundwork for data quality. Following that, data profiling becomes crucial as it aids in early issue identification, paving the way for more effective data quality enhancements. Subsequently, data cleaning and preprocessing are necessary prerequisites for the subsequent validation step. Lastly, measuring the actual quality and tracking the progress of this refined dataset are essential. This is achieved by applying data quality metrics.

2.2. Data Preparation

Before analyzing it, data needs to be organized into a suitable structure. Data preparation encompasses the procedures for manipulating and structuring data in preparation for analysis.[1]
Data preparation is generally an iterative procedure involving the refinement of raw data, often characterized by its unstructured and untidy nature, a more organized and usable format, primed for subsequent analysis. This comprehensive preparation process comprises several key activities, such as data profiling, cleaning, integration, and transformation.
It is widely acknowledged that data preparation is often the most arduous and time-consuming component of data analysis. A recent article featured in the New York Times [16] highlighted that data wrangling alone could consume as much as 80% of the entire analysis cycle's time. In essence, this leaves only a small fraction of time for data analysts and scientists to engage in actual analysis work. A 2015 data science report , published by Crown, reaffirms the challenges, noting that messy and disorganized data constitute the foremost obstacle hindering data scientists.[9] The same report indicates that a substantial 70% of a data scientist's time is dedicated to data cleaning and preparation.
The approach to data preparation varies significantly based on the intended analytical goals and the particular learning techniques and software that will be used for analysis.

Before discussing data profiling, it is essential to introduce a study presented at the SDMS conference in 2021[2] that focused on data readiness. The data readiness report serves as accompanying documentation for a dataset, enabling data consumers to gain detailed insights into data quality, as well as serving as a record of all data assessment operations. The term "data readiness" can also be defined as an umbrella term related to data profiling and cleaning. Existing tools often lack the capability to provide data quality parameters or their explanations, relying on users to guide their own machine learning processes to generate and interpret views.[2]

2.2.1. Data Profiling

Reviewing existing data, evaluating their appropriateness for the intended purpose, and exploring the feasibility of acquiring new data tailored to the task are essential steps. Data preprocessing primarily serves the purpose of ensuring and enhancing data quality. Evaluation of data quality revolves around common criteria, primarily accuracy and uniqueness.
Accuracy in data quality is an aggregated metric derived from the assessment of integrity, coherence, and density. It reflects how well data represent the real-world aspects they describe.

- Integrity consists of two key aspects: completeness and validity. Completeness ensures data are comprehensive and free of missing values, while validity enforces adherence to constraints.
- Consistency pertains to syntactic anomalies and contradictions, with a challenge in choosing reliable data sources when discrepancies arise.
- Density assesses the quotient of missing values in data and represents the presence of non-existent values or properties represented as null values.

Uniqueness is another vital quality measure, ensuring data contain no duplicates. Timeliness is also considered, reflecting data currency and its relevance in keeping information up to date.

2.2.2. Data Cleaning

In cases where data contains noise or anomalies, it becomes necessary to detect and address outliers and other questionable data points, often through remedial actions.[15] The data cleansing process can be outlined in four stages:

- Define and Identify Errors: Recognize and characterize errors in the data, including issues related to incompleteness, incorrectness, inaccuracy, or irrelevancy.
- Clean and Rectify Errors: Take corrective measures to address identified errors, which may involve replacing, modifying, or deleting erroneous data entries.
- Document Error Instances: Keep a record of instances and types of errors encountered during the cleansing process for reference and analysis.
- Measure and Verify: Assess the effectiveness of data cleansing by measuring and verifying whether the cleaned data align with user-specified tolerance limits regarding

data quality and cleanliness.

Addressing missing values is a significant aspect of data cleaning. It is crucial to identify the presence of missing values in the data and take the necessary steps to enable the learning system to manage this scenario effectively. This challenge is prevalent in real-world datasets, where missing values often occur due to various reasons, such as data not being recorded or data loss during the recording process.[26]
Outliers represent another form of data anomaly that requires attention during the cleaning process. These are data points that deviate from the overall data distribution. Outliers can be viewed from two distinct perspectives: they may be considered as data anomalies or, alternatively, as intriguing elements that could potentially signify significant phenomena within the data. For instance, outliers in a store's sales records might indicate a marketing campaign's success.

To classify data as outliers, it is imperative to establish what constitutes the normal behavior of the data and, consequently, how different or significant the outlier is in relation to this normal behavior. Different categories of outliers may emerge based on variations in what is considered normal for the data. Therefore, formalizing both the concept of normality in the data and the inconsistency exhibited by outliers is necessary to detect various classes of outliers.

The main aspect that interests our research the most is always related to the connection between data cleaning and explanations. A study conducted by Laure Berti-Equille and Ugo Comignani in Marseille, France, has enabled the definition of a process to study the explainability of automated data cleaning pipelines and propose CLeanEX [4], a solution capable of automatically generating explanations for the pipelines selected by an automated cleaning system, provided it can access its corresponding cleaning pipeline search space. The goal of this study is actually very similar to our research objective, which is to automate the presence of contextual explanations, but focusing on a different step of the pipeline, namely data preparation, which involved the step of data cleaning process studied in their research.

They proposed CLeanEX, a framework for generating explanations for automated data cleaning. The key advantages of their approach are presented in the following Table 2.1:

Characteristic	Description
Model-agnostic	It can be applied to any machine learning (ML) model, any set of data cleaning tasks, and any automated cleaning agent that can provide information about its cleaning pipeline search space.
Logic-based	Explanations are designed to be understandable to individuals with varying levels of expertise and can be extended to handle causal reasoning.
Both local and global	It can provide explanations for either a portion or the entirety of the cleaning pipeline.
Model quality-independent	It offers a reliable set of explanations regardless of the ML model's quality performance metric used by the automated cleaning agent to select the optimal cleaning pipeline.
User-defined	Optimization is based on the objectives defined by the user.

Table 2.1: Characteristics of the Cleaning Pipeline

In summary, their approach, CleanEX, centers around providing versatile and understandable explanations for automated data cleaning, regardless of the ML model used and tailored to user-defined objectives. In our work, we have used the explanations provided by CleanEX as guidelines, linking the presentation of explanations to the format generated by a large language model. This has allowed us to establish comprehensive guidelines for the entire data preparation tool, covering all phases of data preparation, not only the cleaning phase.

2.2.3. Data Integration and Trasformation

Data integration is a pivotal undertaking within data preparation. The task involves merging data from diverse sources, which becomes particularly challenging when dealing with vast datasets and heterogeneous sources. Data often exist in various formats, such as structured, semi-structured, or unstructured, and originate from different sources, which could be local or distributed. Furthermore, even structured data from a single source may exhibit distinct schemas. Therefore, the need frequently arises to transform data from one representation to another. Several motivations drive these transformations:

- To create symmetric distributions instead of the original skewed distributions.
- Transformation enhances data visualization, especially when data is densely clustered relative to a few outliers.
- Data transformations aim to improve interpretability.
- Transformations are often employed to align data with the assumptions underlying a modeling process, enhancing data compatibility.

Data integration and the associated transformations are essential steps in harmonizing and preparing data for analysis, especially when dealing with diverse and complex data sources.

2.3. Explainability

In recent years, Artificial Intelligence (AI) and Machine Learning have taken significant leaps forward in automating complex processes and making decisions across a wide array of applications, ranging from medical diagnoses to financial analysis. Nonetheless, this growing embrace of AI has brought forth a pivotal question: how can we place our trust in decisions made by algorithms that seem to function in an opaque and impenetrable manner? This is where the concept of "explainability" enters the scene.[11]

Explainability refers to the ability to provide a clear and understandable explanation of how a decision or outcome was reached by a complex system or algorithm. [23] It involves making transparent the reasoning and factors that contributed to a particular result in a way that can be easily understood by humans. An authoritative definition of explainability was proposed by the European Commission's High-Level Expert Group on AI (AI HLEG) in their Ethics Guidelines for Trustworthy AI: "Explainability refers to the ability of an AI system to provide an explanation for its decision-making, to understand why

a specific outcome was produced, to enable users to understand and verify how the AI system works, and to enable recourse in case of erroneous, biased, or unfair outcomes."[19]

Category	Techniques
Model-Agnostic Explainability	• Local Interpretable Model-Agnostic Explanations (LIME) • Counterfactual Explanations
Visual Explanations	• Visualizations • Heatmaps • Activation Maps
Textual Explanations	• Text Generation • Rule-Based Explanations
Human-in-the-Loop Explainability	• Expert Feedback • Interactive Interfaces

Table 2.2: Explainability Categories and Techniques[22]

In the vast field of AI and ML, we have multiple approaches for offering explanations. These diverse modalities, as presented in Table 2.2, are designed to effectively cater to user requirements.

- Model-Agnostic Explainability:
    1. Local Interpretable Model-Agnostic Explanations (LIME)[31]: LIME creates locally interpretable models around specific predictions, making it model-agnostic and suitable for explaining complex AI models.
    2. Counterfactual Explanations[12]: These generate alternative input instances that would lead to different model predictions, allowing users to understand what changes could affect outcomes.
- Visual Explanations[6] :

1. Visualizations: Graphs, charts, and plots visually illustrate the relationship between inputs and predictions, providing an intuitive understanding.
2. Heatmaps: Color-coded representations reveal feature importance or attention weights, particularly useful in image-based AI.
3. Activation Maps: Indicate which parts of an input (e.g., an image) the model focuses on during decision-making, aiding image classification tasks.

- Textual Explanations:
  1. Text Generation: Converts model predictions into natural language explanations, making them accessible to non-technical users.
  2. Rule-Based Explanations[17]: Offer decision rationale in human-readable rules or statements, enhancing transparency in decision-making.
- Human-in-the-Loop Explainability[21]:
  1. Expert Feedback: Involves human experts in the explanation process to validate or improve model interpretations, ensuring their accuracy and relevance.
  2. Interactive Interfaces: Allow users to interact with and manipulate explanations, facilitating deeper understanding and trust.

It is essential now to determine which among the aforementioned explanation modalities can be the most user-friendly and, above all, the most reliable for the user. According to research conducted by the Georgia Institute of Technology, a comparison was made among seven broad categories of explainable AI (xAI) methods. These categories encompassed case-based reasoning, decision trees, feature importance, probability scores, counterfactuals, natural-language explanations, and crowd-sourced explanations. The primary objective was to determine which of these forms is considered the most user-friendly, interpretable, and trustworthy.

In the final evaluation of the reaserach, ROAR (Relative Opinion Aggregates Rank) was employed to assess feature importance, and ERASER (Explainability Ranking for Algorithm Selection in Explanation Resources) was used to evaluate natural language explanations. Notably, research suggests users typically prefer simpler explanations, with natural language explanations provided by recommender systems being particularly well-received.

Key findings from the study include the correlation between participant trust and agent explainability. It might be expected that using a more user-intuitive natural language explanation method would be more reliable. Furthermore, the research reveals a correlation

between the social competence of the agent and its explainability. The results demonstrate that any agent perceived as more explainable is also perceived as more socially competent. Notably, counterfactual explanations and the simplest or clearest explanations tend to receive the highest scores in terms of social competence.

It was observed that both simple language-based explanations and case-based explanations were significantly perceived as more explainable. This observation can be attributed to the fact that explanations based on simple language and concrete examples offer a more accessible and understandable context for users. Simple language-based explanations are often clearer and more direct, making it easier for users to grasp the rationale behind a decision or model outcome. On the other hand, case-based explanations provide a more tangible connection between the model's decision-making process and real-world situations, allowing users to intuitively visualize the model's operation compared to explanations based on class-wise probability scores. The latter may be more abstract and challenging to interpret without a clearer context.[25]
According to Miller, various approaches have been explored to enhance trust and decision-making in artificial intelligence systems. Some of these approaches involve using transparent algorithms, offering visualizations or explanations of the decision-making process, and incorporating human feedback into the system's learning process.[20]

Research conducted by Miller at The University of Melbourne has revealed that people frequently disregard recommendations because of a lack of trust in them. Even more concerning is the tendency for individuals to blindly follow recommendations, even when they are incorrect. Explainable artificial intelligence plays a crucial role in addressing this issue by helping individuals comprehend how and why AI models generate specific recommendations.

In the context of Evaluative AI, there is a distinction between Contrastive explanation paradigms and Evaluative AI itself. Contrastive explanation paradigms aim to persuade users to accept a machine recommendation. They present evidence that supports the recommendation while refuting all other options. In contrast, Evaluative AI focuses on providing evidence for or against each option, not necessarily giving recommendations. It aims to explain why option A was chosen over option B, fostering a deeper understanding of the decision-making process.[27]

Regarding the idea of creating more mockups of screen designs for various types of decisions, can be highly beneficial. Mockups can serve as visual aids to clarify concepts and decision-making processes, making them more accessible to users. These mockups can be presented at different stages of the decision-making process, and users should have the

option to request explanations or view distributions from past cases. This approach, often referred to as stepwise or progressive decision-making, empowers users to backtrack and explore the components where judgments are being made, enhancing transparency and user control.[19]
An additional research study presented at the Conference on Human Factors in Computing Systems in April 2023 mapped existing findings by conducting a detailed scoping review of 48 empirical studies evaluating interactive explanations with human users. The aim was to create a classification and categorize these explanations based on their role in the cognitive process of explanation, distinguishing them as "selective," "mutable," or "dialogic."

In particular, interactive explanations, which are modeled after human dialogue, were found to be more effective for users. Although the term "interactive" carries various meanings in the XAI (Explainable Artificial Intelligence) community, it is primarily viewed as a form of communication between the user and the system. These interactions were further categorized into seven interaction types: selection, exploration, reconfiguration, encoding, abstraction/processing, filtering, and connection.

It is essential for explanations to be personalized and adapted to the context, audience, and purpose of the explanation. The taxonomy was organized into three distinct categories corresponding to the type of support they provide to the human cognitive explanation process: selective, mutable, and social:

- Selective: These allow users to access and select desired information through the explanation interface.
    1. Clarify: Enables users to access desired information on demand through hyperlinks, menus, or tooltips.
    2. Arrange: Allows users to organize the explanation space by hiding or reducing explanations and selecting the type of explanation to display.
    3. Filter/Focus: Permits users to filter and focus on specific AI model inputs or subsets of the dataset.
- Mutable: These empower users to "mutate" causes, meaning they can test hypotheses by simulating or comparing different scenarios.
    1. Reconfigure: Allows users to modify AI model parameters, such as dataset, model type, or model parameters, to observe changes in the explanation.
    2. Simulate: Enables users to test how changes in inputs affect local explanations

and model outputs.

3. Compare: Includes interaction techniques used to compare explanations for different inputs or input groups or for different predictions. Comparative explanations illustrate connections, similarities, and differences between selected inputs or outputs.

- Social (Dialogue With):

1. Dialogic: explanations can be progressively or iteratively provided to the user, who may also ask questions or provide feedback.

2. Progress: This interaction style delivers information in different stages, allowing users to navigate through the explanation using "next" and "previous" commands. However, it doesn't enable users to ask specific questions to the system.

3. Answer: It can be reversible, where users provide feedback or corrections to the system.

4. Ask: Represents the highest level of interaction, akin to a conversation where users can ask any question.

Furthermore, interactivity can introduce an additional layer of explanation, known as meta-explanation, which can become overwhelming in complexity. The question of the quantity of explanations is critical and necessitates further research to determine the appropriate level of explanation for each user. Understanding individual factors is important for managing cognitive load and increasing user trust in the model.

One reason for this is that people attribute human traits to XAI agents and thus expect them to follow social conventions. However, the presence of human traits in a conversation with an XAI agent may decrease user trust by giving them the impression of manipulation. Additionally, users may prefer robotic and "logical and clear" explanations.[5]
Textual explanations facilitate a human-machine dialogue structured similarly to typical human-to-human conversations. To achieve greater accuracy and credibility, the interaction could be helpful.

2.3.1. Adaptive Explenability

When discussing the need for more user-friendly explanations that can adapt to user needs, it is crucial to enhance the quality of interaction, and this is where adaptive explainability comes into play. The goal is to assess whether system explanations result in

differences in the user's subjective assessment of the system and whether they contribute to greater transparency and comprehensibility in AI systems. To successfully integrate explainability, a user-centered approach is essential to ensure that explanations align with the end user's requirements and achieve a sufficient level of causality.[10]
According to additional research, the Self-Adaptive System (SAS) method is deemed essential for furnishing the necessary capabilities to elucidate why a system demonstrates specific behavior. It is within this context that the second explanation of our thesis work unfolds, focusing on explaining why and which characteristics have had a more significant impact on the implementation of a specific imputation method that could enhance the dataset.[29]

2.4. Text-based AI assistant

Artificial intelligence (AI) is a rapidly advancing discipline within computer science, dedicated to the development of intelligent machines capable of mimicking human cognition and behavior.[24] Furthermore, AI can be seamlessly integrated with another cutting-edge technology, the Internet of Things (IoT), giving rise to a novel amalgamation known as AIoT (Artificial Intelligence of Things). Among the most promising AI innovations is ChatGPT, a natural language processing (NLP) system with the capacity to produce human-like conversational interactions.[8]

Before delving deeper into what ChatGPT entails, let is first clarify what NLP is. Neural networks represent a category of machine learning systems engineered to emulate the structure of the human brain. They consist of a sequence of interconnected units known as nodes, organized into layers. The initial layer receives incoming data, which is subsequently processed by intermediate layers before being emitted from the output layer. Each connection linking nodes is assigned a weight value, dictating the magnitude of the connection's influence. Inputs are multiplied by these weights, summed at each node, and then subjected to transformation via an activation function.

ChatGPT is an advanced Natural Language Processing (NLP) system crafted by OpenAI. Its primary purpose is to generate conversations that closely resemble human interactions. This is achieved through its ability to comprehend the contextual nuances of a conversation and subsequently generate suitable responses. ChatGPT is built upon the foundation of a deep learning model known as GPT-3, which has been meticulously trained on an extensive dataset comprising various conversational contexts.[13]

ChatGPT plays a pivotal role in streamlining operations by automating conversations, leading to significant time and resource savings as manual interactions become unneces-

sary. Moreover, ChatGPT boasts rapid response generation capabilities, facilitating swift and efficient conversations. The ChatGPT Improved Accuracy (CGA) model represents a potent Natural Language Processing (NLP) system that harnesses the prowess of a deep learning-based artificial intelligence (AI) architecture to yield precise and contextually meaningful dialogues. Leveraging a pre-trained model derived from OpenAI's GPT-3, CGA excels at crafting lifelike and engaging conversations based on input.[14]

However, it is important to acknowledge that ChatGPT has its limitations. A major constraint is its reliance on the input it receives; it lacks access to external information sources or internet browsing capabilities. Consequently, ChatGPT cannot furnish accurate or up-to-date information on a wide spectrum of topics, and it may struggle with generating responses to intricate or unconventional queries. Additionally, as ChatGPT is trained on an extensive dataset of human language, it may occasionally produce responses containing biased or offensive language.[3]

These limitations were also identified in our research. We uncovered both strengths and weaknesses of the tool, aiming to delineate the role that OpenAI's offerings could play in user explanations. Undoubtedly, the tool excels in the textual transformation context, making content more user-friendly.

3 | Methodology

Today, when any user encounters a tool that enables them to perform data preparation, they are presented with an output intended to meet their initial goal: obtaining a clean dataset. However, this user often lacks clear and specific guidance to achieve this objective. This lack of direction can lead to a lack of confidence in the user and an incomplete understanding of the necessary steps to achieve the desired output and the final result. In many cases, the choice of preprocessing techniques is left to the users, who make decisions based on requirements that may not be optimal for achieving the best possible result. This poses a challenge, especially for less experienced users, who may struggle due to their limited knowledge of the field and the available options.

This thesis work addresses precisely this scenario. The goal is to provide guidance through various forms of explanations that, step by step, assist in selecting the most suitable data preparation activities to obtain the most credible output possible. These explanations will be derived from a previously tested knowledge base, containing essential information for suggesting the optimal activities at various stages of the preparation process. Importantly, this approach always preserves users' autonomy, allowing them to make informed decisions at each step.

In Section 3.1, we will provide an overview of the methodological approach, outlining the primary objectives of the thesis project, which aim to address the gaps in the existing tools known up to this point. In the subsequent Section 3.2, we will initially present the overall architecture and then delve into each element of the presented pipeline in detail within the subsections.

3.1. Method purposes

As previously mentioned, the main objective of this thesis study is to provide explanations that help the user better understand decisions and make choices in order to achieve a more comprehensible and accurate output. Within the tool that will be presented to the user, even if they are not an expert in the field, they will be guided through each stage

by a description or explanation to assist them in making the best choices. Therefore, it is crucial to define the clear and significant difference between a description and an explanation.

A "description" is a detailed representation that provides clear and factual information about an object, process, or concept. It is an accurate illustration that seeks to present the facts objectively without necessarily explaining the reason or rationale behind what is being described. "Descriptions" are useful for providing an essential overview or context on a specific topic.

On the other hand, an "explanation" is a form of communication that aims to provide a deeper understanding or a clear rationale behind a particular event, process, or phenomenon. It also explains the why of a situation or a series of facts and seeks to clarify the mechanisms or causes behind what is observed. "Explanations" are usually more detailed and attempt to illuminate the connections between various parts of a concept or process.[18]

In summary, a "description" offers an objective overview, while an "explanation" seeks to bring out understanding and the rationale behind what is being described. Both play important roles in communication, depending on the type of information you want to convey.

Once the difference between "description" and "explanation" is understood, we can move on to the presentation of the two main objectives of the thesis work. The first objective is to guide the user in the initial phase of data profiling and data quality assessment to better understand the characteristics of the dataset loaded into the tool. To enhance comprehension during this phase, graphs and tables are supplemented with descriptions that are derived from responses generated by a large language model.

The second aim of this research is to support the user's decisions and interactions with the tool during the subsequent data preparation phase by presenting explanations. This ensures that irrespective of the user's familiarity with the context they are working in, they can comprehend and attain a satisfactory outcome. In this case, the text presented will also be generated from questions posed to a natural language model.

3.2. Architecture

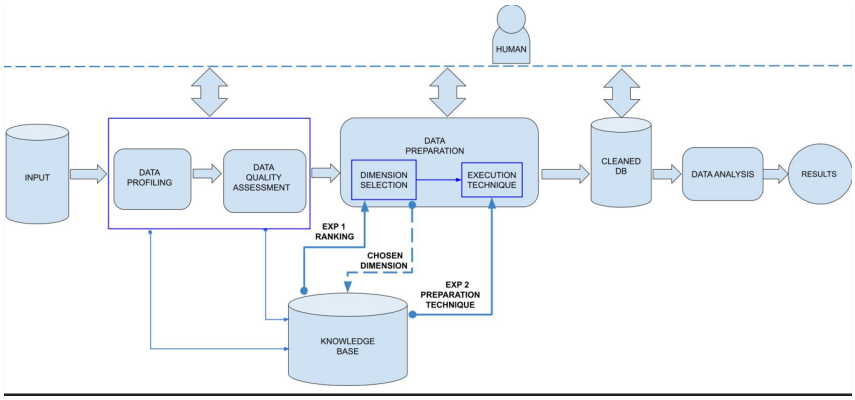

Figure 3.1: Data Analysis Pipeline

In this section, we will walk through the step-by-step process that guides the user toward creating a comprehensible output dataset. To begin, Table 3.1 provides a list of the notations used in the following text.

As depicted in Figure 3.1, the process commences with the selection of a dataset, $\boldsymbol{d}$, upon which necessary modifications will be applied. Once $\boldsymbol{d}$ is loaded, the data profiling and quality assessment phases follow. Within these phases, information entered during the dataset loading stage and information stored in the knowledge base will be used to present the user with data characteristics in both tabular and graphical formats. Here, the first objective of the thesis project will be fulfilled. Alongside these visualizations, text generated through a connection with an external natural language model will be provided. This generated text will offer textual descriptions to enhance the user's understanding of the graphs and tables.

The next phase is the central one in data preparation. In this phase, the interaction between the user and the tool is guided by two different types of explanations. Within this phase, the first step involves generating a ranking of quality dimensions ordered by which one has the highest urgency for improvement. In this context, the first explanation ***exp1*** will be displayed to the user to help them understand how the dimension ranking was generated.

After***exp1*** is generated, the user is presented with a series of solutions to improve the $\boldsymbol{d}$, categorized according to the type of dimension selected for improvement. In this context, the second explanation ***exp2*** comes into play. Through methods contained within the knowledge base, in combination with the characteristics of $\boldsymbol{d}$ in question, we can determine

what influenced and why certain solutions were suggested. Therefore, ***exp2*** will also be generated as text from an external model, providing a clear and explanatory description of the best choice to make regarding the preparation technique.

After applying the suggested techniques for all the dimensions of the ranking, we will have the opportunity to have a clean dataset.

Notation	Description
d	Dataset
A = {a1, a2, . . . , am}	Applications
exp1	Explanation of Dimension Ranking
exp2	Explanation of Preparation Techniques
NLP	Natural Language Processing

Table 3.1: Methodology Notation

3.2.1. Input

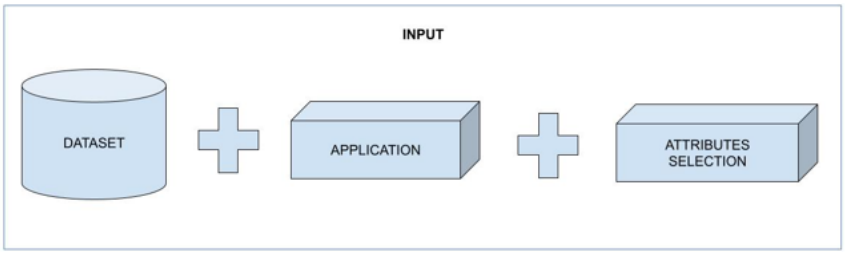

Figure 3.2: Input

The process initiates with the input phase, during which the user is required to provide three essential elements: the dataset ***d*** the chosen application ***am*** and, optionally, a selection of attributes to be included in the analysis. Figure 3.2 visually represents these input elements.

The ***d*** serves as the subject of analysis. The ***am*** represents the user's intended execution of a machine learning algorithm, with options including Decision Trees, K-Nearest Neighbours, and Naive Bayes.

In practice, not all attributes within a dataset are relevant for a given operation or machine learning algorithm. Hence, the user can select from ***d*** a list of attributes they consider most pertinent for the analysis.

These three elements will serve as fundamental data inputs for subsequent analyses.

3.2.2. Data Profiling and Data Quality Assessment

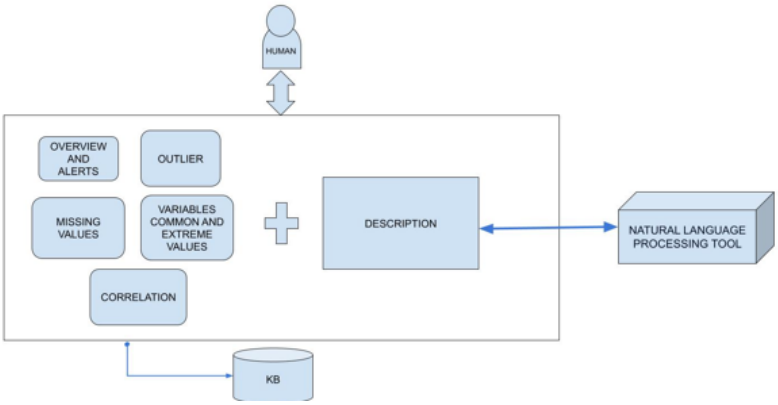

Figure 3.3: Data Profiling and Data Quality Assessment

During the Data Profiling and Data Quality Assessment phase, we focus on three key elements entered during the initial input stage: the ***d***, the***am***, and the choice of dataset attributes. Once these selections have been processed, a comprehensive presentation of the ***d*** is provided to the user.

The initial step involves conducting data profiling on the dataset ***d***, which generates a profiling report. A profiling report is a document that summarizes the insights gained from analyzing a dataset. It provides a detailed overview of the dataset's characteristics, structure, and patterns. This report typically includes information about data types, distributions, outliers, and relationships between variables. The goal is to offer a comprehensive understanding of the data, enabling data scientists, analysts, or decision-makers to make informed decisions and draw meaningful conclusions from the analyzed information. Following data profiling, we present various graphics and tables pertaining to the ***d*** to the user. Additionally, we generate a set of initial warnings and informative descriptions derived from the profiling report and the knowledge base.

To enhance users' understanding of the characteristics listed in the tables, we augment this information with inquiries to a ***NLP*** tool. The primary objective is to furnish users with a response that includes a textual description that contextualizes the dataset's features. Furthermore, the ***NLP*** response includes a brief summary to facilitate comprehension of the dataset's type and the data it contains.

In essence, during this stage of the pipeline, the elaborated descriptions, serving as responses to questions posed to the external tool, play a vital role in helping users gain a more thorough understanding of graphs, relationships between ***d*** features, and the overall picture and tables presented. Once this step is completed, users are made aware of

both the positive and negative characteristics of the ***d*** that require improvement in the subsequent phase.

Highlighting the novel aspect of this thesis project, the following list briefly outlines the descriptions that will be presented to the user, categorized according to subsequent sections of the tool.

- Overview and Alerts: The initial section, known as "Overview," involves scrutinizing several significant attributes of ***d***. The second section, labeled "Alerts," compiles a set of warnings intended to underscore potential data quality issues within ***d***. The descriptions provided to users aim to offer a response that incorporates a textual narrative to contextualize the dataset's characteristics. Furthermore, the ***NLP*** response will include a concise summary to enhance the understanding of the dataset's type and its contained data.
- Variables: In the "Variables" section, each variable is subject to a personalized analysis. This analysis is tailored to provide comprehensive insights into the content of each variable and comprises both statistical and metadata examinations along with distribution visualization. Additionally, a section dedicated to common or extreme values is included. In the case of categorical variables, the ***NLP*** tool's response will offer an explanatory description of the density curve graph. A textual description of the density curve's behavior concerning the loaded ***d*** will be displayed on the screen, closely related to the characteristics outlined in this section.
- Correlation: The Correlation section serves a distinct purpose by enabling users to examine the correlation matrix of their ***d***. In this context, the response generated by the ***NLP*** tool will provide a textual description of the correlation matrix, highlighting the relationships and the strength of each correlation among the attributes of ***d***.
- Missing Values: The Missing Values section furnishes users with information regarding NULL cells in the ***d***. In this scenario, an external ***NLP*** tool will be queried to provide textual descriptions of the graphs. This approach enhances the comprehension of the relationship between missing data within the***d*** and its visual representation.
- Outlier: The Outlier Inspection section provides a means to examine numerical variables using box plots, offering a straightforward way to detect outliers within the ***d***. Similar to the previous section, user interaction will result in textual descriptions of the box plots, articulating the number of outliers for each attribute.

3.2.3. Data Preparation

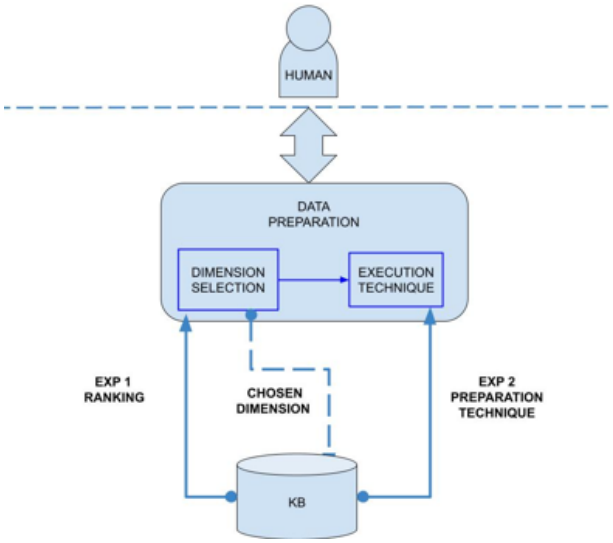

Figure 3.4: Data Preparation

Data preparation often considered the cornerstone of data analysis and machine learning, is a meticulous process encompassing raw data collection, cleaning, and transformation into a structured and usable format. This critical phase serves as the bridge between the characteristics of data collection and the challenges presented by the ***d*** itself, making it more comprehensible to users through detailed descriptions derived from external ***NLP*** tool responses.

In this phase of the pipeline, the focus shifts to understanding how to address and enhance the dataset's imperfections. Here, we delve into strategies to mitigate the dataset's shortcomings, thereby setting the stage for meaningful insights and accurate modeling. Effective data preparation is indispensable, as it ensures that the data attains a level of high quality, consistency, and relevance. Without it, the journey from raw data to actionable insights can be fraught with challenges and inaccuracies.

Within this section, we can divide our discussion into two distinct parts, each contributing to a comprehensive understanding of two pivotal aspects: ranking data quality dimensions and selecting the most suitable data preparation techniques.

Explanation 1

Firstly, we delve into the evaluation and ranking of data quality dimensions. This critical step involves a meticulous assessment of various facets of data quality, including completeness, accuracy, and uniqueness. To facilitate this process, we assign priorities and rankings to these dimensions. This ranking clearly explains which aspects demand the most attention and improvement during the data preparation process.

During the profiling phase, a Quality Assessment Score is generated, specifically pertaining to the quality of the ***d***. It present the quality dimensions in ascending order of score, with the lowest-ranking dimension appearing first. Concurrently, the Quality Assessment Profile is derived from our knowledge base and is closely tied to the chosen ***am*** during the input phase. It ranks the quality dimensions based on their influence on the performance of the selected ***am***, positioning the most impactful dimension at the top.

Moving forward, by incorporating the input from both the typology of orders, the quality assessment phase combines these two rankings while assigning them varying weights. The outcome is a unified ranking that orchestrates the dimensions in a manner that reflects both the dataset's intrinsic characteristics and the application's specific requirements. This final ranking serves as a cornerstone, guiding subsequent actions in the data profiling phase and ensuring data preparation efforts align with our overarching objectives.

At this juncture, the first explanation comes to the forefront. In ***exp1***, the initial form of explanation will embrace a distinctly static approach. Its purpose will be to elucidate which elements have wielded the most profound influence in shaping the ultimate output ranking.

To achieve this, a textual description will accompany the quality dimension ranking. This text will serve to instill confidence in the final ranking, encouraging the user to place trust in the prioritized order for enhancing these dimensions. Consequently, users will be inclined to follow the sequence of improvement for the dimensions as prescribed.

Explanation 2

Secondly, we embark on an exploration of the process of selecting the optimal data preparation technique. This endeavor involves making choices from a spectrum of techniques, including data cleaning, transformation, feature engineering, and normalization, with the aim of addressing the specific quality issues identified earlier. The selection of the right technique assumes pivotal significance as it plays a crucial role in molding the $\boldsymbol{d}$ into a format that aligns seamlessly with the objectives of analysis or machine learning.

The data preparation functionality is realized through a Human-in-the-Loop approach, wherein the user is actively engaged in constructing a data preparation pipeline by selecting from a curated set of suggested actions.The possible suggested actions are listed in table 3.2. This pipeline can either be guided by the system itself, offering a structured path through the preparation process, or crafted by the user's discretion.

Data Quality Dimension	Data Preparation Activities
Uniqueness	Remove duplicates
Completeness	Imputation (0/Missing) Imputation (mean/mode) Imputation (standard deviation/mode) Imputation (mode) Imputation (median/mode) Imputation using KNN Imputation using Mice Drop rows with missing values
Accuracy	Outlier correction Outlier correction with imputation (0/Missing) Outlier correction with imputation (mean/mode) Outlier correction with imputation (STD/mode) Outlier correction with imputation (mode) Outlier correction with imputation (KNN) Outlier correction with imputation (Mice) Outlier correction with drop rows

Table 3.2: Data Preparation Activities

Before exploring the generation of explanations for the recommended data preparation technique, it's crucial to understand how these recommendations are formulated. The refinement of the suggested best technique is closely tied to the use of a classifier. The specific code takes a trained classifier and a specific example from the dataset as input, creates perturbed versions of the example, assesses them using the classifier, and logs predictions. Lime then steps in to offer an explanation regarding which features had a more significant impact on the selection process. As we delve into the second phase of

data preparation, a pivotal development takes center stage - the emergence of ***exp2***. This explanation sets itself apart from its predecessors due to its close association with a computational marvel known as LIME.

LIME, an acronym for "Local Interpretable Model-Agnostic Explanations," [7] represents a powerful tool tailored to shed light on the intricate workings of complex machine learning models. In our current context, LIME assumes a fundamental role in unraveling the enigma of which factors have exerted the most profound, either positive or negative, influence on the optimal data preparation technique recommendation.

What sets this apart is how we translate LIME's revelations into a format that's lucid and accessible to our users. We transform LIME's output into a comprehensible textual representation to accomplish this. This representation finds its place within a query posed to an external ***NLP*** tool. The response elicited from this ***NLP*** tool is then thoughtfully presented on the screen for our users.

This process is meticulously designed to ensure that our users receive an informative and easily digestible explanation. It equips them with a comprehensive understanding of the factors that underpin the recommendation for the most suitable data preparation technique. This knowledge empowers users to make choices that harmonize seamlessly with their specific objectives and goals.

3.2.4. Data Analysis

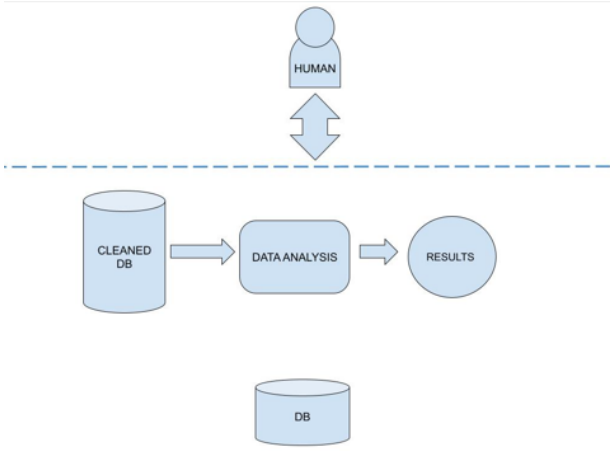

Figure 3.5: Data Analysis

Once the data preparation technique has been selected to make improvements that enhance the dataset's quality, a significant milestone is achieved. These choices have been suggested, contextualized, and explained to the user in the preceding phases, thanks to

the ***exp1*** and ***exp2***.

The enhanced CSV file can be downloaded.

3.2.5. Knowledge Base

In the context of data analysis and machine learning, a well-organized Knowledge Base plays a pivotal role in facilitating informed decision-making. This KB acts as a comprehensive repository, housing invaluable information necessary for two critical stages in the data analysis process: data profiling and data quality assessment. Moreover, it becomes increasingly vital as it aids in formulating a strategy for dimension ranking and the selection of the most appropriate data preparation techniques to enhance ***d*** quality.

The Knowledge Base remains an invaluable resource after a thorough data profiling and quality assessment. It serves as a guiding compass for the subsequent phase, which involves ranking the dimensions within the ***d***. Leveraging the insights gathered during data profiling, the KB helps identify the most essential dimensions based on their relevance and data quality.

Once the dimensions are ranked, the KB reenters the scene, offering recommendations for suitable data preparation techniques. These recommendations are customized to align with the specific requirements of the ***d*** and its intended analysis.

In summary, the Knowledge Base for the data processing pipeline stands as an indispensable asset in the realms of data analysis and machine learning. It provides the foundational knowledge necessary for informed decision-making, ensuring that data profiling, quality assessment, dimension ranking, and data preparation are executed with precision. By harnessing the wealth of information contained within the KB, analysts can make well-informed choices at every stage of the analysis process, ultimately leading to more accurate and valuable insights derived from the dataset.

The KB includes a comprehensive mapping of preparation activities to quality dimensions, data preparation clusters linked to quality dimensions, assessment metrics, dependencies between data preparation activities, pipelines for each quality dimension, data preparation activities impacting machine learning algorithms, data preparation methods, ranking of quality dimensions for machine learning algorithms, ***d*** suitability for machine learning algorithms, and much more.

4 | Experimental setup

In this chapter, we will enumerate the technologies and delve into the setup choices employed during the implementation of the thesis project tool. In Section 4.1, we will describe the technologies used, while in Section 4.2, we will provide detailed insights into the Python libraries utilized. Section 4.3 presents the research conducted on ***NLP*** tools, considered a reference for generating descriptions and explanations. This section comprehensively outlines all the research steps undertaken prior to the development of the tool. This extensive analysis aims to discern the strengths and weaknesses of ChatGPT, thereby gaining insights into its potential utility and how to leverage this tool effectively within the scope of the thesis work.

4.1. Technologies

Python The backend is primarily constructed using Python, a versatile programming language renowned for its extensive library support, specifically tailored for exploring and interpreting data within datasets. This choice provides the tool with robust capabilities for data manipulation and analysis.

On the other hand, the frontend, along with the communication layer facilitating seamless interaction between the frontend and backend components, is realized through a combination of HTML, CSS, and JavaScript. This front-end technology stack ensures an engaging and responsive user interface while ensuring efficient data exchange and real-time interactivity with the Python-based backend. ***Flask*** The web framework employed for this web application is Flask, which is written in Python. Flask applications follow a specific structural pattern, wherein HTML files are organized within a "templates" directory, while other assets, such as CSS and JavaScript files are kept in a "static" directory. Flask is known for its scalability, offering flexibility without imposing specific tool or library requirements. It allows developers to install libraries and Flask extensions on an as-needed basis, ensuring a lightweight yet adaptable environment for web development.

As required, data communication between the client and the server is facilitated through

the following methods: ***HTTP Methods*** The POST method is employed to transmit HTML form data from the client to the server, while the GET method is utilized to request data from the server. These HTTP methods are fundamental for exchanging information and performing various operations in the web application. ***Ajax POST requests*** In addition to traditional HTTP methods, Ajax POST requests are harnessed to enable asynchronous communication between the client and server. These requests, powered by JavaScript and the XMLHttpRequest object, allow for dynamic data retrieval and manipulation without the need for full page reloads, enhancing the user experience and responsiveness of the web application.

In the realm of cutting-edge technology and natural language processing, OpenAI and LIME have emerged as pivotal tools, revolutionizing the way we interact with and understand complex datasets. In the pursuit of enhancing user comprehension of dataset characteristics and decision-making, these two technologies played a central role in my thesis project.

OpenAI, a prominent player in the field of artificial intelligence, is synonymous with innovation and state-of-the-art NLP capabilities. This AI powerhouse has been instrumental in enabling machines to understand and generate human-like text, making it a powerful tool for various applications. OpenAI's GPT (Generative Pre-trained Transformer) models can generate human-like text, answer questions, and even hold conversations in a remarkably coherent and contextually relevant manner. In my thesis, I harnessed the capabilities of OpenAI to bridge the gap between complex datasets and end-users, simplifying information delivery and making it more accessible.

Local Interpretable Model-agnostic Explanations ***LIME*** is a critical component of my project that enhances transparency in machine learning models. LIME's primary function is to provide simple, interpretable explanations for machine learning predictions. It does this by approximating complex models with locally faithful interpretable models, which can be easily understood by non-technical users. This tool played a pivotal role in ensuring that the answers provided by the NLP tool in my project were simpler and transparent, allowing users to grasp the reasoning behind dataset characteristics and proposed choices.

The thesis project integrates OpenAI and LIME with an NLP tool, working in synergy to respond to user queries with simplified, descriptive answers. This integration not only improved the accessibility of information but also ensured that users could better understand the intricacies of the dataset and the options presented to them. The combination of these advanced technologies played a role in improving user experience, facilitating more informed and effective decision-making.

4.2. Libraries

Within the framework of this research thesis, ***Python libraries*** form the cornerstone of the toolkit employed for data analysis, visualization, and machine learning. These libraries offer a comprehensive suite of tools to effectively tackle the challenges the research problem presents. Let us explore each of them in detail, understanding their specific roles and contributions to the research endeavor.

- ***Pandas*** is an indispensable library for data manipulation and analysis. It furnishes data structures like DataFrames and Series, which facilitate efficient data handling. Pandas streamlines data loading, cleaning, transformation, and aggregation, thereby playing a pivotal role in managing and extracting insights from the research dataset.
- ***YData Profiling*** is a specialized library or tool tailored to the research project's requirements. It offers data profiling functionalities, including summarizing and analyzing the dataset, identifying patterns, and detecting anomalies. This tool contributes to the initial exploratory phase of the research.
- ***Plotly*** is a versatile library for creating interactive and visually appealing data visualizations. It offers a wide range of chart types and customization options, enhancing the ability to communicate research findings effectively. The library's interactivity enables dynamic exploration of data within the research context.
- ***Missingno*** is a library that focuses on visualizing missing data within the dataset. It provides visual summaries of missing values, helping to identify data gaps and assess their impact on the research. This is particularly important for ensuring data integrity and the reliability of analytical results.
- ***Scikit-Learn***stands out as a comprehensive machine learning library, replete with tools for data modeling. It covers a spectrum of machine learning tasks, including classification, regression, clustering, and dimensionality reduction. The library streamlines the creation and evaluation of machine learning models, which is essential for predictive analysis in the research.
- ***NumPy*** short for Numerical Python, is a foundational library for scientific computing. It furnishes support for arrays and matrices, making it a cornerstone for efficient numerical operations and mathematical computations. This library is fundamental for data manipulation, analysis, and various numerical tasks integral to the research.
- ***LimeTabularExplainer***is a specific component within the LIME (Local Inter-

pretable Model-agnostic Explanations) framework tailored for tabular data. LIME is a powerful tool used for interpreting the predictions made by machine learning models. The LimeTabularExplainer class specializes in providing locally faithful and interpretable explanations for predictions in scenarios where the data is organized in tabular formats, such as spreadsheets or structured databases.

This component is designed to tackle the challenge of explaining complex model predictions on tabular data, making it more transparent and understandable, even for non-technical users. The LimeTabularExplainer does this by approximating the behavior of a black-box machine-learning model within a local neighborhood of a data point. It builds simpler, more interpretable models that capture the model's decision-making process around the specific instance of interest.

In essence, LimeTabularExplainer creates a bridge between the intricate nature of machine learning models and the need for human-friendly explanations, making it a valuable tool for ensuring model transparency and aiding in model debugging and user trust.

Each of these Python libraries is instrumental in addressing distinct aspects of the research problem, from data preparation and exploration to modeling and visualization. Together, they provide a robust toolkit for the successful execution of the research thesis.

4.3. ChatGPT

ChatGPT is a type of AI model that belongs to the broader category of generative language models. It has been trained on a vast dataset containing text from the internet, books, articles, and more. This extensive training enables ChatGPT to comprehend and generate human-like text in a coherent and contextually relevant manner. It can understand and respond to text inputs in natural language, making it highly valuable for a wide range of applications.
In this section, we present the additional contribution of our thesis work. The intention is to emphasize the understanding of ChatGPT's potential in providing explanations. To effectively integrate this large language model into the thesis project, it is essential to grasp the ways in which it can be helpful, whether simply in terms of processing the textual aspect of explanations or in a different capacity. Our initial focus was on exploring the capabilities of this large language model to evaluate its suitability as a reference in analyzing a specific dataset during the preparatory phase.Then, we aimed to determine whether it could provide explanations for the rationale behind specific actions taken to enhance the dataset, particularly benefiting a less experienced user in the field.

We thus experimented with the model's capabilities by posing various questions to it using a specific sample dataset.The dataset selected for reference is a file named "beers.csv," encompassing over 1300 rows and providing information on IBU (International Bitterness Units), name, brewery ID, and ABV (Alcohol By Volume). It includes a diverse range of characteristics and detailed information related to beers. Our research unfolded across multiple phases, assessing the advantages and disadvantages of utilizing this innovative large language model.
All the phases were characterized by the ten questions posed to the model regarding the different stages of the data quality pipeline and how the model could address and, most importantly, explain the rationale behind specific choices made to improve the dataset.

During the initial phase, questions were directed to the online tool associated with ChatGPT-3. One of the first critical aspect encountered was the inability to directly input the dataset in CSV format; it had to be converted into a JSON file. With the assistance of Python code, the dataset was successfully transformed into JSON format. Initially, all 2431 rows of the entire dataset were loaded into the chat model's input box, but the submitted message exceeded the length limit. So, the number of rows was reduced to 551. It is important to note that during this phase, the selected rows are not representative of the entire dataset.So, in this initial phase, the first critical issues that arise are the limited length of the question and the format of the dataset.
When requesting the model to perform data profiling on those mentioned rows, the result consisted of a detailed list of dataset characteristics that could be further explored with additional questions. This level of accuracy and precision in responses was not observed when the model was tasked with generating responses for data cleaning or outlier detection requests. An additional critical aspect that emerged is that the online tool, seemingly hesitant to recommend the most suitable method or technique for a specific dataset, offered ambiguous responses and explanations. It presented generalizations and described various options without clearly indicating a preferred one.
In the second phase, we compared the responses obtained in the first phase with those obtained by integrating the OpenAI APIs into Python, effectively creating a local tool. After creating a Python script and associating it with a ChatGPT account key, we were able to pose the same 10 standard questions to this local tool. Again, it was necessary to convert the format into JSON. An additional limitation of this model is its capacity to process rows, limited to 4097 tokens, which corresponds to approximately 400 rows—a lower limit compared to the online tool. For this local tool, the same critical issues as those encountered in the initial phase where the online tool was tested are evident. Despite

there are positive features, such as providing recommendations for effective data cleaning techniques based on the dataset's characteristics and displaying the cleaned dataset afterward. One notable downside emerged is that this tool requires more information compared to the online tool to generate a response as comprehensive as the one obtained in Phase 1.
To further enhance the reliability and improve the research conducted in the previous phases, we found it necessary, given the space constraints imposed by CHATGPT, to no longer select the first 400 rows from the JSON file. Instead, we used a Python code to perform a dataset sampling, selecting the 400 most significant rows that were more representative. We considered the JSON file generated as the reference dataset for posing the usual 10 questions, this time using the online tool. It is worth noting that this online tool was linked to ChatGPT 3.5, which is different from the ChatGPT 3 version used in the initial phase. As a result of this change, the responses obtained for all inquiries were more detailed, but the strengths and weaknesses mentioned in the previous phases remained consistent.
We were able to observe something more interesting in the fourth and final phase of our research when questions were posed to the local tool utilizing ChatGPT 3.5. At this stage of the research, we were able to address some of the criticalities encountered in the earlier phases. An example of this improvement was the dataset sampling, which was representative of the proposed data collection. One noteworthy positive characteristic that stood out to us, and proved helpful for our objective, was a higher level of accuracy and contextualization in responses. In most requests, though not all, the tool would provide examples that helped users better understand the dataset's characteristics. In some cases, it even managed to suggest more appropriate techniques for dealing with missing values and data cleaning, contextualizing them for the research process.

However, there were some negative aspects to consider. For instance, when the request was about visualizing data, in the initial response, it would provide a brief description of why an image was necessary, along with a link to a non-existent page. Another unfavorable aspect is that it couldn't provide explanations in certain cases, as it claimed it couldn't identify the presence of missing values, given that the proposed data collection was just a sample.
In conclusion, it can be asserted that the use of ChatGPT for our objectives can be partially satisfying. Throughout the various phases of our research experimentation, both positive and negative aspects of this tool have emerged.

Certainly, as a supportive tool for generating textual explanations that provide a better and more accessible understanding of dataset characteristics and the reasons behind spe-

cific improvement choices, ChatGPT proves valuable. On the flip side, our research has also revealed that often, these suggestions and explanations were not directly presented to the user making the request. Instead, they remained somewhat generic, lacking thorough justification and contextualization. Additionally, in most cases, the visualization was not immediate, and the results were based on formulas, functions, and libraries that we were already familiar with. One of the major limitations encountered is the inability to input the entire dataset for a more comprehensive analysis.

Therefore, for our specific objectives, ChatGPT can indeed be considered useful in enhancing the textual aspects of results obtained through formulas and functions that we ourselves are proficient in.

5 | Implementation

After enumerating the essential technologies and components needed for the development of the thesis project tool, the subsequent section will showcase the tool and elucidate the innovations implemented to augment user engagement and enhance the user's ability to understand the choices they make. All the images presented in this chapter are aimed at providing a better understanding of what users will encounter in the tool refer to the loading of the sample dataset 'beers.csv'.

5.1. Dataset Input

The workflow begins with the user uploading the data source, which can be selected from the computer's folders or dragged and dropped into the designated area. The data source is only accepted if it is a .csv file. Once the dataset is uploaded, the user is directed to a new page. In this new view, a tabular representation of the dataset is displayed alongside sections related to the user's choices.

The user is prompted to select the features they are most interested in, choose the machine learning algorithm they intend to use in the analysis phase, and indicate whether they would like assistance in designing the data preparation pipeline. The novelty of this page is linked to the initial description of the tool. Just as with all the descriptions and explanations in this chapter, there is a 'Summary Dataset Description' button. When the user clicks on this button, they can directly view a response generated by ChatGPT, our external NLP tool. This response enables the user to better understand each attribute of the uploaded dataset and contextualize the data within it.

Additionally, there is another button below this one that provides a detailed description of the initial response, making the most relevant characteristics more clear. For a clearer visual representation of this view, we can refer to Figure 5.1 to see what the user experiences on this description page.

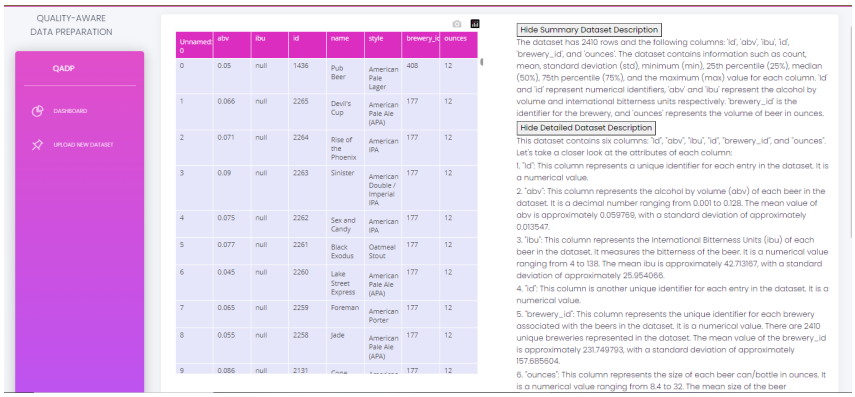

Figure 5.1: More Detailed about the dataset

5.2. Dataset Exploration

When the user has a clear understanding of the data they will be working with after reading the initial description, they will then be able to comprehend what will be presented on the subsequent page. The elements described in this section will enable the user to explore their dataset through statistical analysis, metadata examination, and graphical representations, all conveniently accessible on a single web page.

These pieces of information all pertain to the data profiling and data quality assessment phase of the previously described data analysis pipeline. Data profiling serves the purpose of assisting the user in comprehending the content and potential issues within the dataset they've uploaded. Data exploration also encompasses the implementation of data quality assessment, where the system evaluates the various dimensions of data quality.

Overview and Alerts

The initial section that users will encounter is the 'Overview and Alerts' segment. Within this informative portion, users will find a comprehensive description of both the positive and negative aspects of the uploaded dataset. As an integral part of the advanced features incorporated into the tool, the 'Alerts' subsection offers users the ability to access a 'Detailed Alerts Description' button. When this button is clicked, it initiates the presentation of a textual explanation, shedding light on the criteria used to flag specific data points within the dataset as noteworthy.

The generation of this textual explanation relies on interactions with ChatGPT, thereby ensuring that the presentation of alert information is comprehensible and designed to facilitate a deeper understanding of the underlying reasons behind these alerts. In essence, it serves a dual purpose: to present the information within the flagged data points and

empower users with the knowledge required to comprehend the basis for the alerts.

For a visual representation of what users will encounter within this section, Figure 5.2 provides an illustrative depiction of the user interface

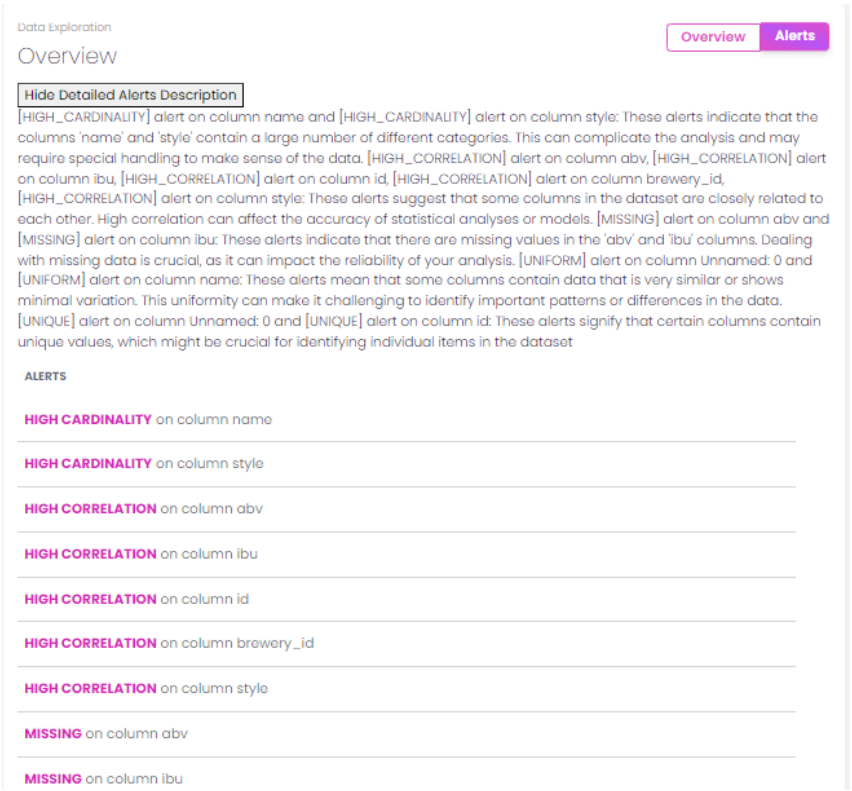

Figure 5.2: Detailed Alert Description

Variables

In the subsequent 'Data Exploration' section, a detailed examination of each column selected by the user on the previous page is provided within the 'Variables' segment. Statistical data is made available for each variable, individually. Additionally, the columns are exemplified with samples that can be visualized through an interactive area. This interactive component displays the most and least common values, characters, and words, as well as the highest and lowest values within the dataset.

The objective of the various graphical representations is to offer users a clear and accessible means of gaining insights into the dataset from multiple perspectives. The plots featured in the 'Variables' section primarily represent the distribution of the data. This distribution is visualized for numerical variables through a hist plot and a rug plot, while categorical variables are represented using a bar plot. All these plots can be interacted with in

numerous ways, allowing users to zoom in and out, move around, and access information about the values represented by simply hovering over the plot.

To further assist users in navigating these charts, there is an option within this section to click on the 'More Information about Zoomed Graphs' button. This action directs the user to a new page where they will find a distribution graph for each attribute of the dataset, enhancing their ability to delve deeper into the dataset's characteristics. Through user interaction with these graphs, a textual description is generated and presented below each graph, helping users gain a more comprehensive understanding of the specific area they have chosen to zoom in on.

For a more detailed understanding of what users will encounter, Figures 5.3 and 5.4 represent what users can see on the page both before and after zooming in.

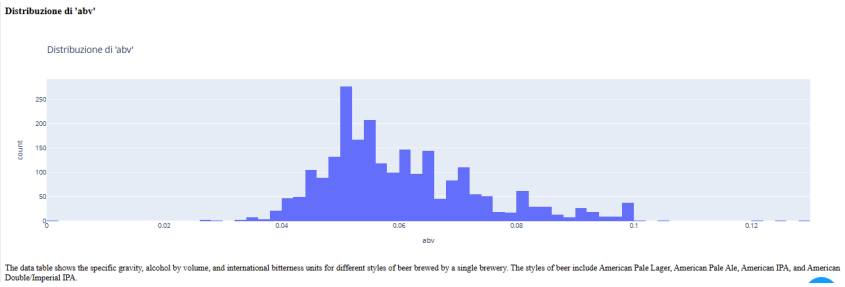

Figure 5.3: More Information about zoomed graph before zoom

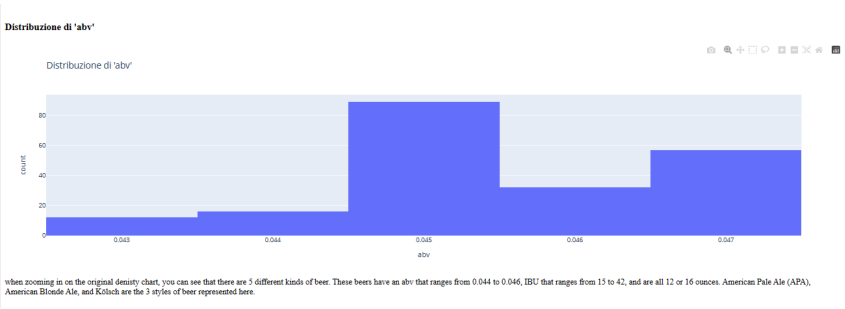

Figure 5.4: More Information about zoomed graph after zoom

Correlation, Missing Values and Outliers

In this final section of the exploratory part of the tool, users will be able to visualize the correlation, missing values, and outliers within the dataset using three different types of graphs.

A heatmap illustrating the correlations between attributes is provided for correlation analysis, offering insight into the relationships between all pairs of numeric variables.

To address missing values, users can access a bar graph that displays the total number of non-null cells within each variable. Additionally, a heatmap is available to identify correlations and common patterns between attributes with missing values. A visual representation of the overall distribution of missing cells in the table can be found in the 'Matrix' section.

For outlier inspection, interactive box plots are provided for each numerical feature to help users visualize and identify out-of-range values.

In each of these three sections, users will be guided to better understand these graphs. Each section features a title showcasing the's response generated by ChatGPT, which serves as a textual description of the information in the graphs. Users will find the most significant characteristics that need to be noted in relation to what is displayed in the graphs. Figure 5.5 provides an illustrative representation of the response for the missing values graph.

Figure 5.5: More Details

5.3. Dataset Improvement

After the user, with the assistance of the descriptions, has gained a comprehensive understanding of the dataset's strengths and weaknesses, they can identify the aspects that require improvement to achieve a better output than the initial dataset.

The workflow then proceeds to the 'Data Preparation' phase, where the final ranking of data quality dimensions is computed. Additionally, the user is provided with guidance in designing a data preparation pipeline

Exp1:Dimension Ranking

In this new phase, the user is guided, as before, on how to enhance three quality dimensions: Accuracy, Completeness, and Uniqueness. The tool also provides a ranking closely tied to the priority of improving a particular quality dimension. The user is presented with a ranking ordered according to a specific rationale. Once again, the reasoning behind this seemingly mysterious classification is unveiled by clicking a button that reveals a textual explanation. Unlike the other textual explanations, this one is generated statically since the ranking is calculated in the same manner for all datasets, regardless of the dataset being uploaded. The ranking of quality dimensions involves two initial rankings: RAS (Rnking from Assessment), reflecting user evaluation of importance, and RAP(Ranking from Application), representing application significance. These rankings determine the order of quality dimensions, with RAS prioritizing user preferences and RAP focusing on application requirements. The combination of these weighted rankings produces a final comprehensive ranking that considers both user and application needs, determining the order of importance for quality dimensions. Figure 5.6 provides an illustration of what will be explained to the user.

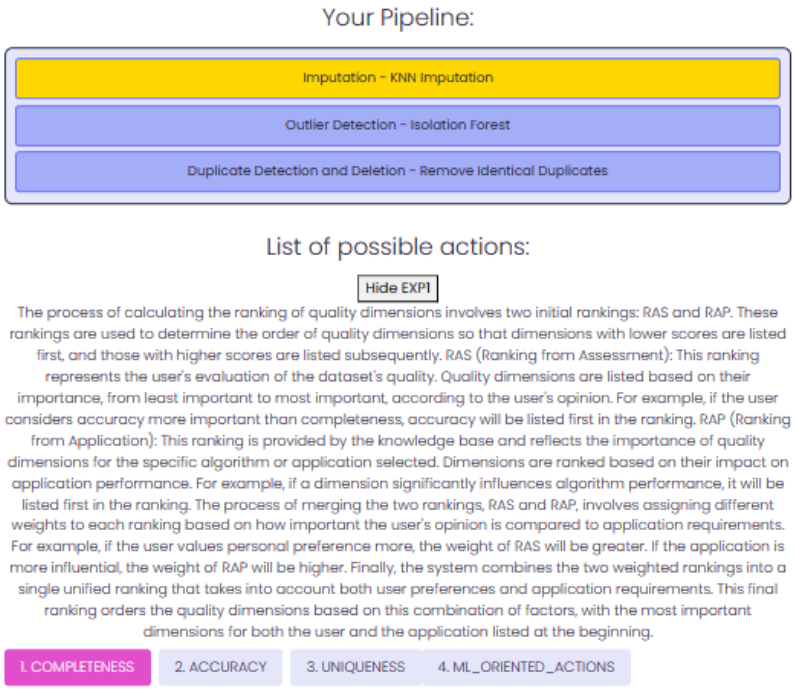

Figure 5.6: EXP1

Exp2:Preparation Techniques

In this final section, after the user's previous direction towards prioritizing one of the three dimensions, the focus shifts to how to improve them. Here, with the assistance of a previously trained classifier, the user is presented with the best technique advice for enhancing that specific quality dimension. The user can choose to accept or decline this advice. To further support the recommendation, a second explanation is provided, appearing upon clicking the 'exp2' button and generated by ChatGPT.

To provide the most suitable response, we have utilized LIME, which stands for Local Interpretable Model-Agnostic Explanations. LIME leverages a model and a sample on which the machine learning algorithm has been trained to determine which features were most relevant in selecting the best improvement solution offered to the user. ChatGPT, as always, plays a role in translating the explanations derived from LIME into user-friendly text, even for non-expert users. An example of summarized 'exp2' is depicted in Figure 5.7.

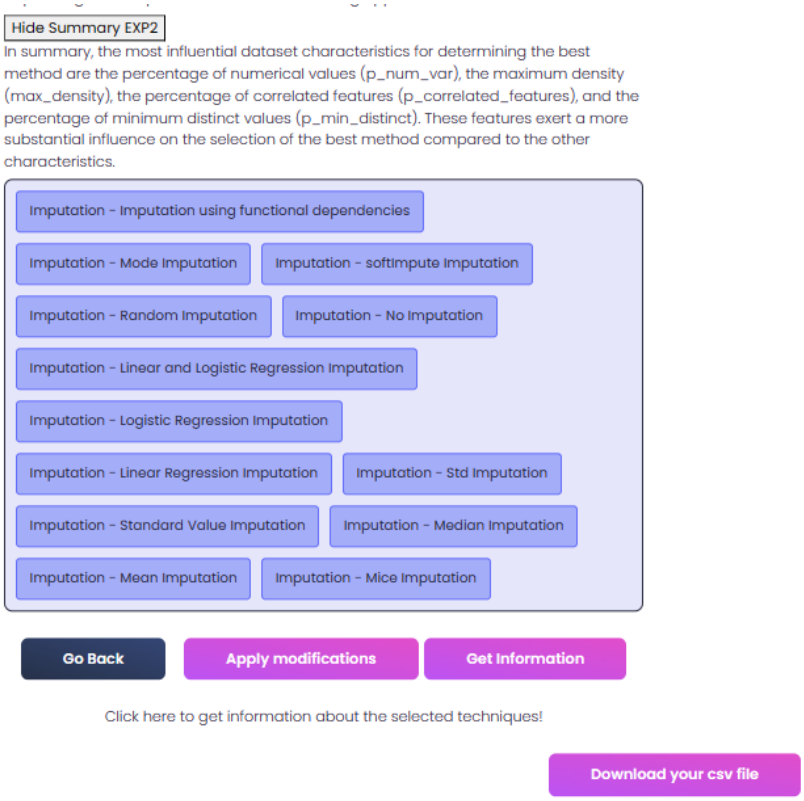

Figure 5.7: EXP2

After reading and being guided by the explanations listed above, the user will be free to construct the data preparation pipeline, following or deviating from the suggestions.

6 | Conclusions and future developments

In conclusion, let's underscore the innovative aspects of the tool just presented, which are set to bring new perspectives to the landscape of existing tools in data preparation context.

In a world where data-driven decision-making is prevalent, numerous tools have emerged to assist users in analyzing data, with the goal of achieving optimal results for decision-making. While these tools enhance datasets, they often lack explanations, leaving users uninformed about changes made and potentially eroding trust in the system.

The tool proposed introduces several innovations in this context. By providing detailed descriptions and explanations, it empowers users to fully understand the suggested data preparation tasks. The complexity of designing data preparation processes arises from managing diverse data sources, formats, and structures, requiring expertise to navigate and transform data for reliable, high-quality datasets.

This proposed solution stands out as an interactive, user-friendly platform designed to streamline the data analysis process for individuals with varying levels of expertise. Tailored for users optimizing data sources for machine learning analyses, its primary goal is to provide a user-friendly platform that empowers even those without expertise to effortlessly enhance dataset quality.

A significant innovation is the integration of pipeline knowledge into data preparation, accompanied by explanatory elements addressing a previously overlooked aspect. Users, who were previously uncertain about the tool's outcomes, are now guided in understanding the formerly opaque steps behind these enhancements. Textual explanations enable users to make informed decisions based on the tool's improvements.

During the data exploration phase, users receive detailed descriptions highlighting key dataset features, emphasizing strengths and weaknesses. These descriptions are complemented by explanatory graphics. In the data improvement phase, essential explanations

clarify the reasoning behind proposed changes. Importantly, the tool is complemented by generative AI, such as ChatGPT, simplifying complex results into easily understandable text.

6.1. Future Work

Looking ahead to the future development of this project, there are promising opportunities for growth that aim to enhance the user experience and understanding in the realm of data preparation.

A preliminary idea could involve expanding the functionality of the tool to support more advanced and conversational interactions. This could entail integrating natural language conversational agents that guide users through explanations and respond to questions in a more interactive manner. Additionally, collaborating with advanced data visualization tools to enhance the graphical representation of explanations, facilitating a more immediate visual understanding of data preparation processes. Drawing inspiration from other large language models that may offer support not only in textual aspects but also in more specific areas to aid in more comprehensible explanations.

Bibliography

[1] Z. S. Abdallah, L. Du, and G. I. Webb. Data preparation. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 318–327. Springer, 2017. doi: 10.1007/978-1-4899-7687-1_62. URL https://doi.org/10.1007/978-1-4899-7687-1_62.

[2] S. Afzal, C. Rajmohan, M. Kesarwani, S. Mehta, and H. Patel. Data readiness report. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, pages 42–51. IEEE, 2021.

[3] A. Azaria. Chatgpt usage and limitations. 2022.

[4] L. Berti-Équille and U. Comignani. Explaining automated data cleaning with cleanex. In *IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI)*, 2021.

[5] A. Bertrand, T. Viard, R. Belloum, J. R. Eagan, and W. Maxwell. On selective, mutable and dialogic XAI: a review of what users say about different types of interactive explanations. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, and M. L. Wilson, editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 411:1–411:21. ACM, 2023. doi: 10.1145/3544548.3581314. URL https://doi.org/10.1145/3544548.3581314.

[6] E. Bobek and B. Tversky. Creating visual explanations improves learning. In P. Bello, M. Guarini, M. McShane, and B. Scassellati, editors, *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014, Quebec City, Canada, July 23-26, 2014*. cognitivesciencesociety.org, 2014. URL https://mindmodeling.org/cogsci2014/papers/046/.

[7] S. Bramhall, H. Horn, M. Tieu, and N. Lohia. Qlime-a quadratic local interpretable model-agnostic explanation approach. *SMU Data Science Review*, 3(1):4, 2020.

[8] J. Deng and Y. Lin. The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83, 2022.

[9] S. García, J. Luengo, and F. Herrera. *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*. Springer, 2015. ISBN 978-3-319-10246-7. doi: 10.1007/978-3-319-10247-4. URL https://doi.org/10.1007/978-3-319-10247-4.

[10] J. Graefe, S. Paden, D. Engelhardt, and K. Bengler. Human centered explainability for intelligent vehicles - A user study. In *AutomotiveUI '22: 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Seoul, Republic of Korea, September 17 - 20, 2022*, pages 297–306. ACM, 2022. doi: 10.1145/3543174.3546846. URL https://doi.org/10.1145/3543174.3546846.

[11] T. D. Grant and D. J. Wischik. *On the path to AI: Law's prophecies and the conceptual foundations of the machine learning age*. Springer Nature, 2020.

[12] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[13] H. N. Hai. Chatgpt: The evolution of natural language processing. *Authorea Preprints*, 2023.

[14] A. Haleem, M. Javaid, and R. P. Singh. An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil transactions on benchmarks, standards and evaluations*, 2(4):100089, 2022.

[15] I. F. Ilyas and X. Chu. *Data Cleaning*, volume 28 of *ACM Books*. ACM, 2019. ISBN 978-1-4503-7152-0. doi: 10.1145/3310205. URL https://doi.org/10.1145/3310205.

[16] S. Lohr. For big-data scientists,'janitor work'is key hurdle to insights. *New York Times*, 17:B4, 2014.

[17] D. Macha, M. Kozielski, L. Wróbel, and M. Sikora. Rulexai - A package for rule-based explanations of machine learning model. *SoftwareX*, 20:101209, 2022. doi: 10.1016/j.softx.2022.101209. URL https://doi.org/10.1016/j.softx.2022.101209.

[18] D. L. Miller. Explanation versus description. *The Philosophical Review*, 56(3):306–312, 1947.

[19] T. Miller. Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 333–342. ACM, 2023. doi: 10.1145/3593013.3594001. URL https://doi.org/10.1145/3593013.3594001.

[20] T. Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support. *arXiv preprint arXiv:2302.12389*, 2023.

[21] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, 56(4):3005–3054, 2023. doi: 10.1007/s10462-022-10246-w. URL `https://doi.org/10.1007/s10462-022-10246-w`.

[22] G. Ras, N. Xie, M. Van Gerven, and D. Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396, 2022.

[23] A. Rosenfeld and A. Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33:673–705, 2019.

[24] Z. Shi. *Advanced artificial intelligence*, volume 4. World Scientific, 2019.

[25] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, and M. C. Gombolay. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *Int. J. Hum. Comput. Interact.*, 39(7):1390–1404, 2023. doi: 10.1080/10447318.2022.2101698. URL `https://doi.org/10.1080/10447318.2022.2101698`.

[26] I. Stanimirova, M. Daszykowski, and B. Walczak. Dealing with missing values and outliers in principal component analysis. *Talanta*, 72(1):172–178, 2007.

[27] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[28] G. K. Tayi and D. P. Ballou. Examining data quality - introduction. *Commun. ACM*, 41(2):54–57, 1998. doi: 10.1145/269012.269021. URL `https://doi.org/10.1145/269012.269021`.

[29] J. M. P. Ullauri, A. García-Domínguez, and N. Bencomo. From a series of (un)fortunate events to global explainability of runtime model-based self-adaptive systems. In *ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion, MODELS 2021 Companion, Fukuoka, Japan, October 10-15, 2021*, pages 807–816. IEEE, 2021. doi: 10.1109/MODELS-C53483.2021.00127. URL `https://doi.org/10.1109/MODELS-C53483.2021.00127`.

[30] R. Y. Wang, V. C. Storey, and C. P. Firth. A framework for analysis of data quality

research. *IEEE Trans. Knowl. Data Eng.*, 7(4):623–640, 1995. doi: 10.1109/69.404034. URL https://doi.org/10.1109/69.404034.

[31] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell. " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*, 2019.

List of Figures

List of Tables