LaraN01
herzra

Data Science Project

Drug’s risk evaluator

Conceptual Design Report

15 June 2024

Abstract

This project aims to assess the relationship between inactive ingredients in oral drugs and adverse events, offering potential insights into pharmaceutical safety. With drugs defined broadly as substances for treating or preventing disease, understanding the impact of both active and inactive components is crucial for ensuring medication safety. While extensive studies typically focus on active ingredients, this project shifts attention to the often-overlooked inactive ingredients, recognizing their potential influence on adverse reactions.

The methodology involves gathering and evaluating data from various sources, including the FDA and DailyMed, using techniques such as web scraping and data manipulation. The dataset comprises information on adverse events, drug properties, and active as well as inactive ingredients, which are processed and merged for analysis. The data extraction is followed by evaluation, unsupervised and supervised learning.

Challenges in data quality and complexity are addressed through meticulous preprocessing and feature engineering. Risk assessment, based on severity and outcome of adverse events, serves as the foundation for model development. The project focuses on documentation and risk management to ensure robustness and reproducibility

In conclusion, while this project represents a comprehensive effort to explore the role of inactive ingredients in drug safety, initial findings indicate the complexity of the relationship. Further research and data refinement are necessary to deepen understanding and enhance predictive capabilities.

Table of Contents

1 Project Objectives

With the word "drug" it is defined any substance or combination of substances presented as having properties for treating or preventing disease in human beings; Any substance or combination of substances which may be used in, or administered to, human beings, either with a view to restoring, correcting or modifying physiological functions by exerting a pharmacological, immunological or metabolic action, or to making a medical diagnosis (1) (2). Starting from this definition, a drug is the final product composed of one or more active ingredients, responsible of the effect achieved upon administration, and of one or more inactive ingredients, or excipients, inert pharmaceutical substances that are used in product formulation to perform a variety of functional roles. The properties of the final dosage form are, for the most part, highly dependent on the excipients chosen, their concentrations and interaction with both the active compound and each other (3). While extensive studies are performed to determine the excipient compatibility with the active ingredient/s and therefore the product safety, efficacy and quality, the occurrence of unexpected adverse events after product commercialization is not uncommon and a specific pharmaceutical field, pharmacovigilance was created with the target to track these unexpected events and react, through preventive or corrective actions, to ensure that any approved medication can be safely administered to patients in needs. Despite its existence since approximately 170 years, the bigger update occurred after 2012 when the Directive 2010/84/EU introduced major changes including, among others, the possibility for everyone to report an adverse event and the extension of the adverse event definition to drug misuse or administration errors.

While often the interest is to assess whether a particular drug, or drug class is more prone to exacerbate an adverse event occurrence, less attention is referred to the inactive ingredients or excipient combination, equally present in the mixture responsible for the side effect.

Starting from this knowledge, target of this project will be to

- Assess if exists any relationship between the list of inactive ingredient in a drug and the propensity to exacerbate an unexpected adverse event;
- Where a relationship exists, to create a model able to predict the risk category of newly developed drugs from the list of inactive ingredients and some selected properties of the active ingredient readily available.

The drugs will be classified in a risk category from 0 to 15, where 0 is no risk and 15 is death.

The project will be split in three main phases:

- Phase I: Data gathering and evaluation, explorative data analysis of the data acquired.
- Phase II: unsupervised learning and dimensionality reductions techniques.
- Phase III: classification by supervised learning.

Despite the pharmacovigilance data being public, the required information to evaluate a relationship between the label information (components) and the adverse event occurred are scattered and not readily available from the public. For this reason, the data were sourced from two different official sources:

- U.S. Food and Drug Administration
- Dailymed from U.S. National Institute of Health

Due to the various sourcing of data, common fields in different datasets will be used to obtain a final dataset suitable for unsupervised and supervised learning tasks. The results from this first preparation phase will be used in unsupervised learning techniques (i.e. clustering) and dimensionality reduction methods (i.e. PCA, t-SNE) to increase the understanding on the dataset and the relevant features, as well as to identify patterns for relationship between the data

Due to the complexity of the data sourcing and the computation intensity required batch learning algorithms will be preferred. By an evaluation of the available data, as well as field knowledge and frequency of update of the data sources, the frequency of the regular re-training, in order to prevent data drifting of a potential final model could be estimated as bi-yearly.

2 Methods

The computational part of the project for Phase I to Phase III will be performed using Python as programming language in Jupyter Notebook environment, through source code-editors (Visual Studio Code, Microsoft Corporation), python distributions (Anaconda, Anaconda Inc.) or hosted services (Google Colab, Google LLC). For joint coding and sharing GitHub VCS will be used. Large datasets, and any other material unable to be handled by GitHub, will be managed through Google Drive (Google LLC). Within the iPython environment at least the following libraries will be needed:

- Pandas: for data analysis and management.
- Numpy: for data analysis, data cleaning and preparation.
- Matplotlib for data plotting and exploration.
- Scikit-learn: for machine learning tasks.

- Tensorflow: for deep learning tasks.

3 Data

Information on adverse events were web scraped from the FDA API (4), as well as some drug active ingredient properties (moieties). Both sourced from the same API referenced above, the independent JSON files were extracted from different API endpoints as explained in the API documentation. Once downloaded, the JSON files were unwrapped in dataframes more easily visible and comparable. Due to the shape of the frames, subsequent manipulation were performed on arrays.

Information on inactive ingredients of the different drugs (label information) were downloaded from DailyMed (5). Despite the same information could potentially be available also through FDA API Drug API Endpoint, product labelling (6) the difficulty to extract unique names for the different excipient, listed with their common chemical name, directed towards DailyMed, where all the ingredients were listed using the unique UNII codes (7). The common chemical name can significantly change depending on the nomenclature in use, expanding the problem to multiple independent names and not only to a string formatting issue, an example of such variability can be observed in Table 3-1:

Table 3-1: Chemical name and UNII code

Chemical name	UNII
Benzoic acid Benzenecarboxylic acid Carboxybenzene Dracylic acid Phenylformic acid Benzeneformic acid Benzenemethanoic acid Phenylcarboxylic acid	8SKN0B0MIM

The data were publicly available and free of use. All the rights on the data in use belong to the U.S. Food and Drug Administration.

4 Metadata

In order to successfully run the notebooks a python environment with all required libraries should be used, this is available as a .venv in the same repository as the notebooks. Additionally, in the utility folder an additional python file containing all the custom functions called in the notebook is available.

Due to the dimension of the final dataset this cannot be saved in the same GitHub repository but can be shared by the authors upon request. Datasets are stored in a shared Drive folder for contemporaneous usage.

5 Data Quality

As also disclaimed in the main data source, U.S Food and Drug Administration, the data in use by the current project should be used as indication only but cannot be used as decision making tool:

"FAERS data does have limitations. There is no certainty that the reported event (adverse event or medication error) was actually due to the product. FDA does not require that a causal relationship between a product and event be proven, and reports do not always contain enough detail to properly evaluate an event. Further, FDA does not receive reports for every adverse event or medication error that occurs with a product. Many factors can influence whether or not an event will be reported, such as the time a product has been marketed and publicity about an event. Submission of a safety report does not constitute an admission that medical personnel, user facility, importer, distributor, manufacturer or product caused or contributed to the event. The information in these reports has not been scientifically or otherwise verified as to a cause and effect relationship and cannot be used to estimate the incidence of these events."

The target of the project is dual, first to evaluate if a relationship exists at all and just secondarily, upon existence of such relationship to built a model able to suggest if the probability of adverse event is increased by the specific choice of the inactive ingredients or whether it can be reduced by the choice of similar, but equally performant, inactive ingredients.

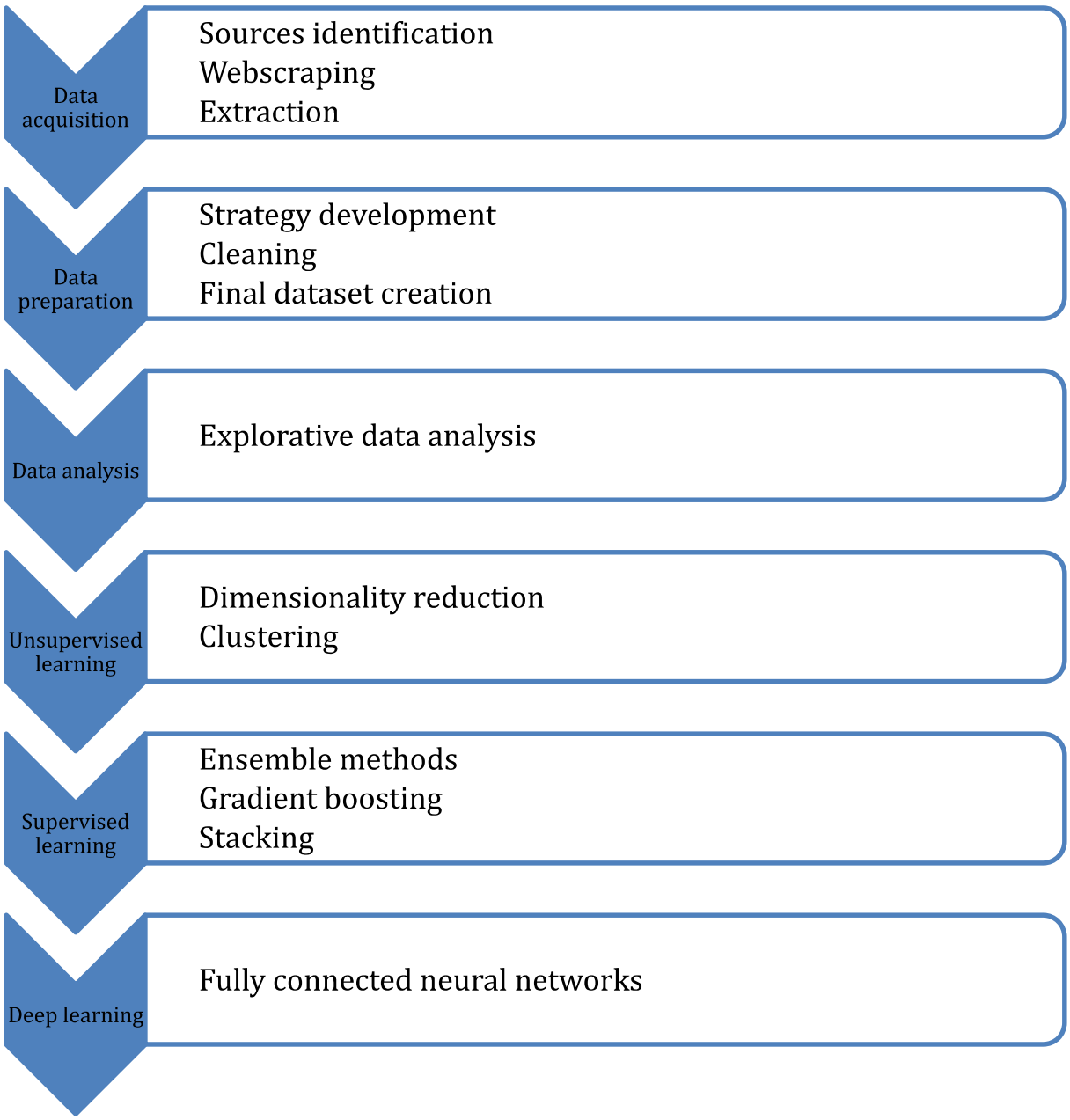
Starting from this target it is crucial to have a high number of different pharmaceuticals within the dataset and ideally the same active with different combination of inactive ingredients (i.e. generics). As the data available from FDA are not verified the target of this project is mainly to built a good strategy that can potentially be used as higher data quality are available, i.e. from EudraVigilance, the European Database, verified by experts or from Drugbank, but both available upon request and for dedicated and approved research only, therefore not accessible by these authors.

The current phase of the project embraces only few years of adverse events and fairly recent as better recorded than old events, the choice was due to the dimension of the Dailymed dataset, encompassing ca. 2000 drugs only, and therefore not able to cover many of the entries from the adverse events dataset. The use of more adverse events data would have resulted in duplicate entries once cleaned, without added value on the dataset.

The authors expect these drawbacks to have strong impact on the result of the project and will use the data to gain more understanding on possible relationship, on the valuable and cleaned data, to be used for a more rich and diversified dataset that will ideally include not only the list of the inactive ingredients but the relative amounts as well.

The latter, the availability of a better dataset with the amounts, as well as more drugs and generics, will be the measure to improve the data quality and the overall model performance.

6 Data Flow



7 Data Model

7.1 Conceptual:

7.1.1 Cleaning and Merging

After the unwrapping of the different JSON files, the following dataframes were obtained:

- Dailymed: of shape (19968, 8), containing information on the drug labelling;
- FDA moieties: of shape (134019, 16) containing the active ingredients properties;
- Adverse events: several dataframes divided per years or year quarters, with shape varying from several thousands to few million rows and 17 columns, containing information about the adverse event occurred, its severity and the event outcome.

Due to the adverse events dataset shape, and the computational resources available, it was not possible to handle the entire dataframe in a single instance. After cleaning and removal of the unused features, the Dailymed frame was merged with FDA moieties, using the UNII code of the active ingredient, in an intermediate dataframe named Dailymed_moieties of shape (17877, 17). This intermediate frame was then merged subsequentially with the several cleaned adverse event dataframes, using Set ID , a Globally unique identifier (8).

Each output dataframe was then cleaned from NaN and duplicates and its feature engineered.

7.1.2 Feature Engineering

The seriousness of the event, described within the adverse event dataset as several different seriousness type encoded with 1 (Yes) or 0 (No), was combined in a seriousness score using the following formula:

$$seriousness_{score} = Death + LT + H + Dis + CA + Other$$

Where:

Death=death contribution, 5 if Yes, 0 if No;

LT= life threatening contribution, 1 if Yes, 0 if No;

H= hospitalization contribution, 1 if Yes, 0 if No;

Dis= disability contribution, 1 if Yes, 0 if No;

CA= congenital anomalies contribution, 1 if Yes, 0 if No;

Other= other seriousness contribution, 1 if Yes, 0 if No;

The seriousness score was then combined with the overall classification of the event (serious / non-serious) and with the event outcome to obtain a risk score, the target feature that will be used for unsupervised and supervised learning. This score accounts for the kind of event

occurred, for its classification and for the outcome of the event on the patient and can be considered a suitable parameter to classify the drugs with respect to its propensity to cause severe events. The risk score was calculated with the following formula:

$$risk\ score = Se + seriousness\ score + Outcome$$

Where:
Se= serious classification, 5 if serious, 2 if non-serious;
Outcome = contribution of the reaction outcome, 5 if patient died, 4 if life threatened but not died, 3 if patient hospitalized or other severe outcome, 2 if the patient is in the process of recovering or resolving from the adverse event; 1 if the patient recovered from the adverse event or the adverse event resolved without any lasting effects

The risk score so calculated can assume values from 3 (non-serious event with other minor outcome), to 15 (serious event, deadly and causing patient death). The intermediate range (7-10) representing either non-serious event resulting nevertheless in the patient death, or severe event that resulted in outcomes other than patient death.

7.1.3 Concatenation and pre-processing

Final dataset was created from concatenation of the individual merged and feature engineered frames, the non-used columns were removed, route of administrations were combined on macro classes and correct data-types assigned to the remaining features. The data so obtained was saved as csv file for the different model training and testing.

Upon import , the final dataset was pre-processed prior to use in unsupervised and supervised learning tasks through specific pipelines to enhance the robustness, efficiency and manageability of the project, then divided in training and test set for model optimization and testing.

7.1.4 Unsupervised learning

Clustering techniques were used for a first evaluation of the data distribution in the multidimensional space. As the excipient encoding resulted in a highly dimensional object, the clustering techniques were tested on the pre-processed dataset itself and in a PCA dimensionality reduced dataframe and the results compared. Other dimensionality reduction techniques were tested, such as t-SNE alone or in sequence after PCA analysis. Clustering as described was performed on the complete dataset, on the dataset after grouping on the active component and average of the score, on subsets of the main dataframe per administration route. A basic visualization tool using dash was implemented for k-means and HDBSCAN.

7.1.5 Supervised learning

The task to be handled by the current project was a classification task, therefore classification algorithms were evaluated. Due to the dimension and the complexity of the task (no clear clusters or relationship observed through unsupervised learning) the simple model were not evaluated in favor of ensemble methods, more performant on complex relationships. Models were tested alone or in combination through stacking.

7.1.6 Deep learning

In a complexity increasing strategy, upon failure of machine learning model fully connected neural networks were evaluated upon an additional encoding of the target feature to fit the networks requirements. The different features were evaluated alone and as combination of features (more complex interactions)

7.2 Logical

Output column will be the risk score, as described in the previous paragraph, based on the seriousness of the event as well as it's outcome and effect. Composed of discrete values only will be treated as classes for a Classification task.

The input columns will be all those of the cleaned dataset with exception of "Route of administration" and "Product types" that are used to slice portions of the final dataset prior to preprocessing. All the dropped columns in the dataset preparation were not semantically relevant for the analysis (i.e. patient gender), with many missing values, or to be used only in the final dataset preparation (i.e. spl_set_id). The inactive ingredient column, as well as the other columns with discrete string values were encoded prior to be used in the models evaluation (both for supervised or unsupervised learning).

For phase III of the project only the relevant features resulted from Phase II will be used, in absence of clear prevalence of any feature all the encoded features will be used.

7.2 Physical

Minimum system requirements are:

- 110 GB RAM
- 50 GB Disk space

For time optimization, due to the computational intensity of the models the use of TPU or any other multiprocessing strategy is recommended.

8 Documentation

The project will be documented by GitHub version control system, release READMe document file as well as with code documentation created directly in the Notebooks that will decsribe the steps and results obtained.

Additionally, the project will be accompanied by the present report, describing background, choices and the strategy of the authors in the execution of the different steps.

9 Risks

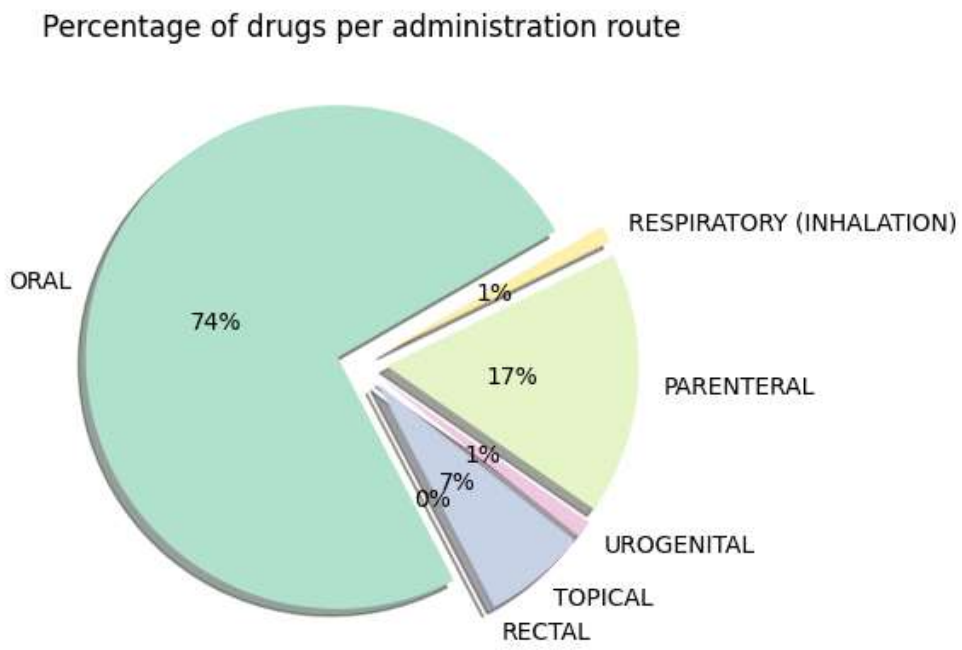
The greatest risk, addressed also in chapter 5 is the quality of the data not being high enough for a relevant model. While the quality, both in terms of dimension and diversity, is very high for some of the dataset to be merged (Adverse events) not the same could be stated for other dataset, like Dailymed, of too small dimension to address the complex task foreseen.

When this occurrence should happen more data will be extracted out of the Dailymed database or the other selected data sources will be used. A non-extensive list of other possible data sources identified is PubChem, Kegg or DrugBank. The preparation and pre-processing steps could be re-used, being built on general function on numpy arrays to allow the scaling on slightly different datasets or on other data-sources than the one used for the current version of the project.

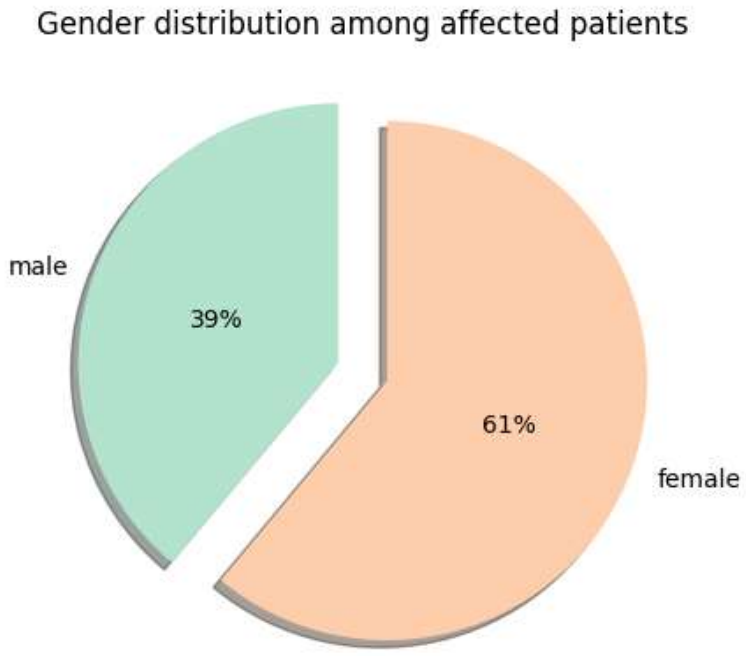
The overall impact will depend on the time necessary to acquire the data through the additional sources, all provided of API, while the preparation phase, as well as the pre-processing, being automized and standardized through pipelines, will not be relevantly impacted. All the data are free of use, with no impact on the project costs.

10 Preliminary Studies

The majority of the dataset (74%) comprises medicines for oral administration. Pharmaceutical for oral administration, including, among others, pills, capsules and granulates, are the most common type of pharmaceutical and often represent over-the-counter drugs (OTC). OTC do not require a medical prescription and can be sold by pharmacies over pharmacist advice or by patient request representing a category of pharmaceutical largely used. FDA Adverse event dataset is a collection of averse events reported from clinics, pharmacists but also from patients. The greatest prevalence of oral drugs can be therefore explained by their larger use compared with other type of administrations.



The majority of the patients who experienced adverse event in the datasets were female.



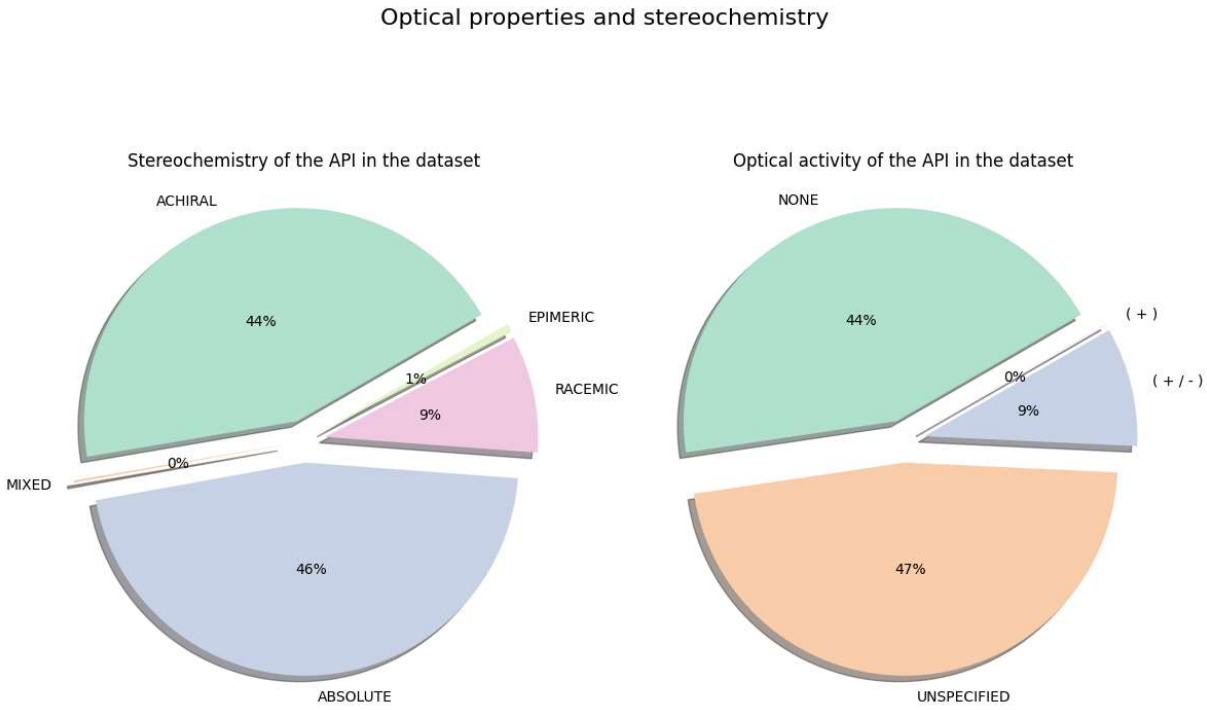
Stereochemistry and optical activity shows a similar distribution. Achiral molecules have no optical activity (green slice) as without any stereocenter.

Racemic mixtures are mixtures of equal quantities of two enantiomers. Each enantiomer rotates the plane of polarization of plane-polarized light through a characteristic angle, but, because the rotatory effect of each component exactly cancels that of the other, the racemic mixture is optically inactive (9%).

An epimeric drug represent instead the presence of only one of the two enantiomers, (+) enantiomer in the dataset.

Absolute indicate molecules with stereocenters, without specifying though, if only one of the enantiomers is present or a mixture of the two (racemate) and the optical activities of these molecules in unspecified.

Due to the big lack of data (the unspecified portion is approximately 50% of the data), despite the enantiomeric properties plays an important role in the risk of a molecule, famous is the case of Talidomide in the risk of one of the enantiomeric configutaions, it will be removed from the final dataset.

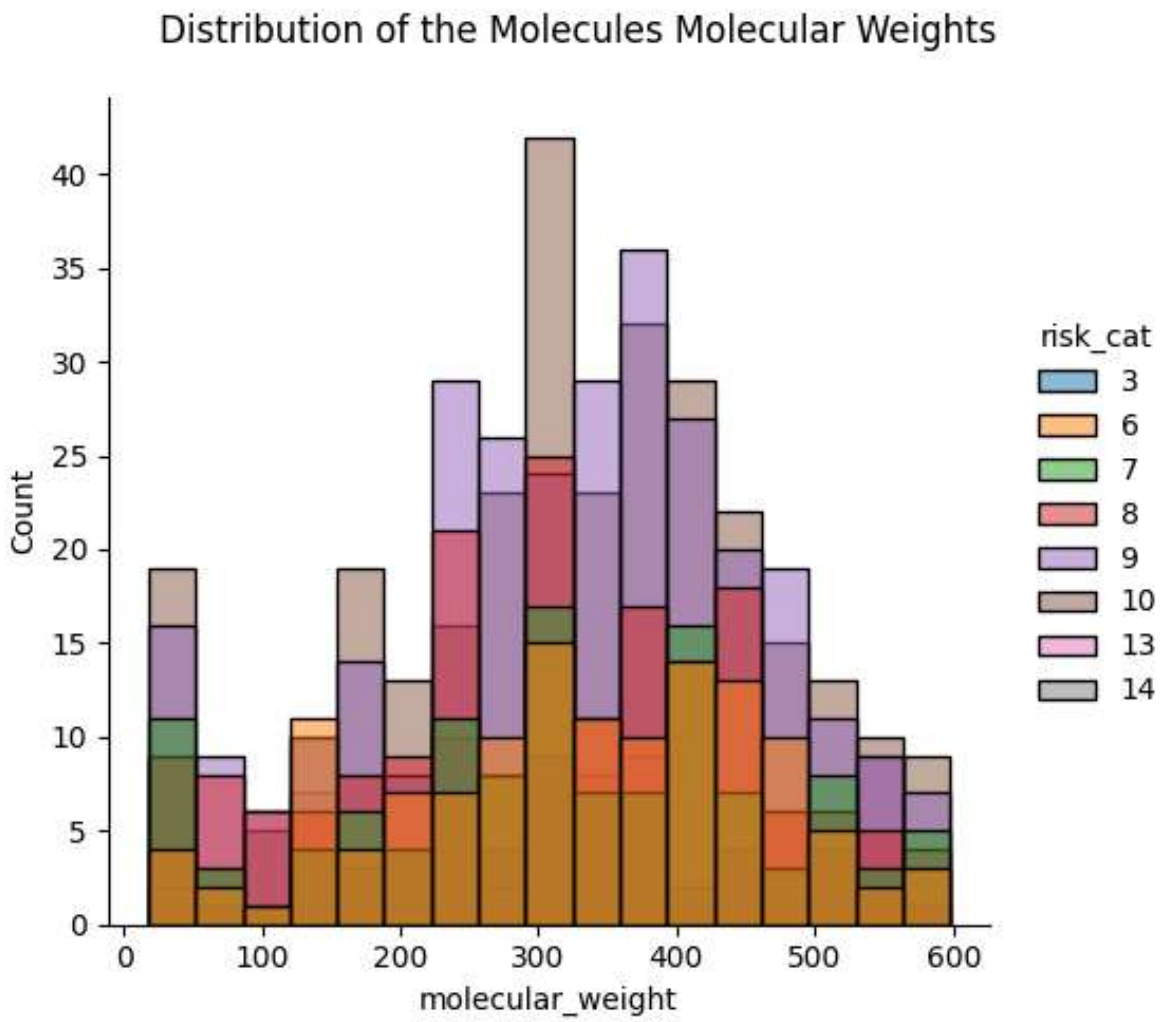


The molecular weights of the molecules range from 0 to about 600. The distribution is bimodal, with notable peaks around 100 and 300-400.

The highest concentration of molecules is around the molecular weight of 300, where the total count exceeds 40.Another significant peak is around 100 molecular weight, with counts around 20.

Risk category 9 (purple) and Risk category 10 (brown) are the most prevalent across most molecular weight ranges. Risk category 8 (red) and Risk category 14 (gray) are also prominent, particularly in the 200-500 molecular weight range. Risk category 3 (blue) and Risk category 6 (orange) appear less frequently compared to others. Risk category 7 (green) and Risk category 13 (pink) have the lowest counts overall but are present in various molecular weight ranges.

Molecules with lower molecular weights (under 100) tend to have a higher proportion of risk categories 9 and 10. Molecules in the mid-range (200-400) show a mix of risk categories, with categories 8, 9, 10, and 14 being significant.Higher molecular weights (above 400) still have contributions from multiple risk categories but with a noticeable presence of category 10.

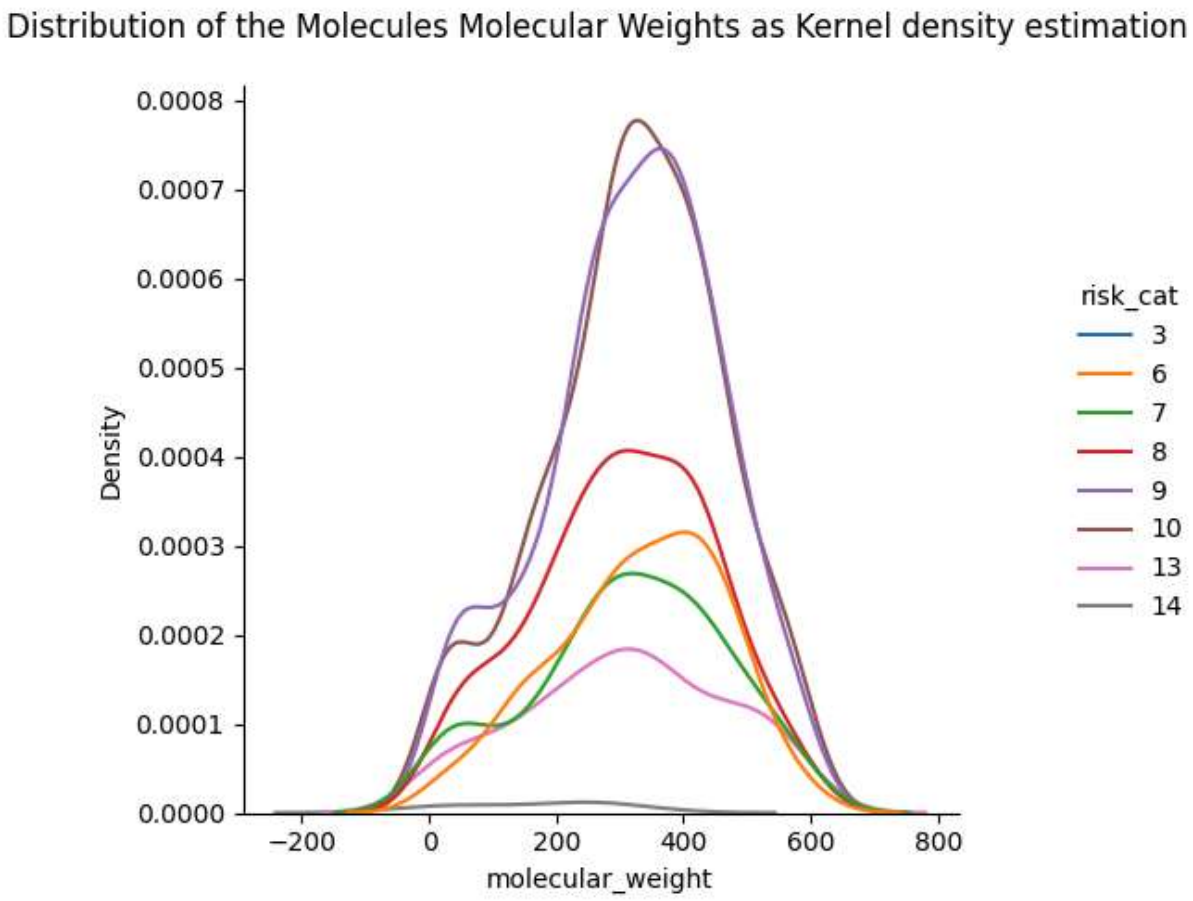


KDE plots provide a smoothed estimate of the probability density function of a random variable, offering a continuous view of the distribution.

Risk category 9 (purple) and Risk category 10 (brown) have the highest peak densities, indicating a higher concentration of molecules around 300-400 molecular weight. Risk category 8 (red) also shows a prominent peak but slightly lower compared to categories 9 and 10.Risk categories 6 (orange) and 13 (pink) have moderate densities with peaks around 200-300 molecular weight.Risk category 7 (green) has a relatively lower peak compared to the other categories but shows a broader distribution.Risk categories 3 (blue) and 14 (gray) have the lowest densities, indicating fewer molecules in these categories across the molecular weight range.

For most risk categories, the density increases rapidly up to around 300 molecular weight and then decreases more gradually. Categories 9 and 10 dominate the density in the range of 200-400, suggesting a high number of molecules in this range belong to these risk categories. Lower molecular weights (<200) and higher molecular weights (>400) have lower densities across all risk categories.

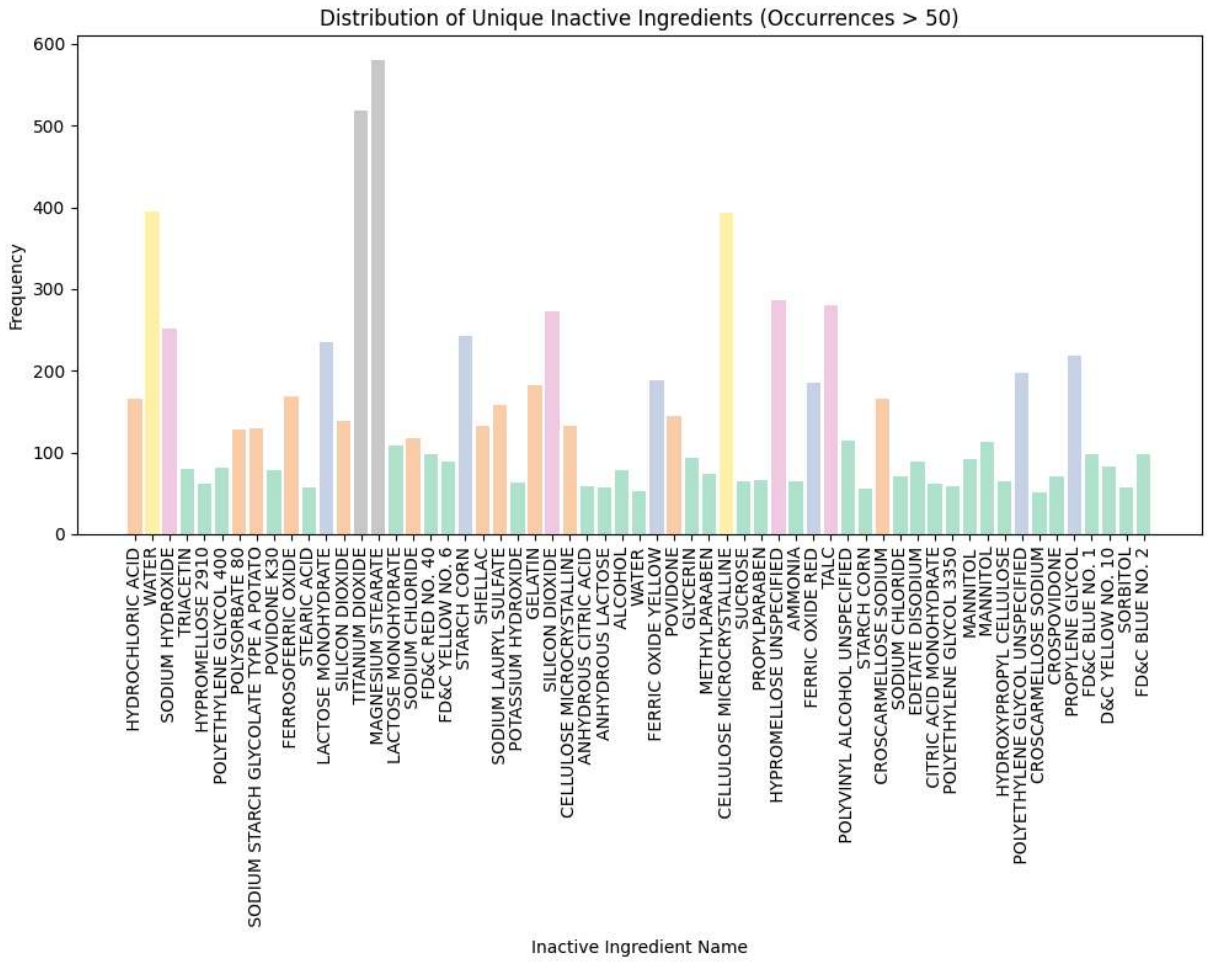
The KDE plot confirms the bimodal distribution observed in the histogram, with significant peaks around 100 and 300-400 molecular weight. The KDE plot provides a clearer, smoother view of the distribution, highlighting the dominant presence of certain risk categories in specific molecular weight ranges.



The most represented inactive ingredients are titanium dioxide and magnesium stearate. Magnesium stearate is a lubricant, widely used to prevent sticking of the capsules to each other and to the machine that creates them. Titanium dioxide is a coating agent often used in capsules and tablets to make opaque films, it produce pastel shades as a result of its whiteness, and maintain excellent light/heat stability.

Widely represented in the dataset are also Water, important component of several dosage forms, and cellulose microcristalline, the most widely used excipient for direct compression serving as a strong dry binder, tablet disintegrant, an absorbent, filler or diluent, a lubricant, and anti-adherent.

In accordance with the distribution of the route of administration of the drugs in the dataset, all the excipient mostly represented in the dataset are from oral dosage forms.



11 Conclusions

Part I of the project has shown no apparent relationship between the inactive ingredient of the drugs and the possibility to exacerbate a severe adverse event. A thorough interpretation of the unsupervised learning results (t-SNE) shows just minor clusters with different classes still mixed among them. This could be an indication of a complex relationship or of weak relationship only, therefore not of sufficient import to allow a proper classification. This conclusion is confirmed by the confusion matrix obtained as a result from the ensemble methods, where the assignment of the classes is random and biased by the most represented class.

References

1. *On Drugs*. **Baron, , Linton, and O'Malley, Maureen A.** 2023 Dec;, J Med Philos., pp. 48(6): 551–564.

2. **(MRHA), Medicines and Healthcare Products Regulatory Agency.** A Guide to What is a Medicinal Product: MHRA Guidance Note 8. London, United Kingdom : MRHA, 2020.

3. *12 - Stability Studies*. **Jessica Cha, Timothy Gilmor, Philip Lane, Joseph S. Ranweiler.** 2011, Separation Science and Technology, Vol. Volume 10, pp. Pages 459-505.

4. **U.S. Food and Drug Administration.** openFDA. *openFDA API.* [Online] https://open.fda.gov/apis/.

5. **U.S. Food and Drug Administration, .** Dailymed. *Dailymed Application Development Support: Web Services*. [Online] https://dailymed.nlm.nih.gov/dailymed/app-support-web-services.cfm.

6. **U.S. Food and Drug Administration.** Drug Labelling overview. *openFDA*. [Online] https://open.fda.gov/apis/drug/label/.

7. —. FDA's Global Substance Registration System. *G-SRS*. [Online] https://www.fda.giv/industry/fda-data-standards-advisory-board/fdas-global-substance-registration-system.

8. **U.S. Food and Drug Administration.** FDA Direct, CDER Direct and Cosmetic Direct. *FDA Direct*. [Online] https://direct.fda.gov/apex/f?p=100:80:0::NO::P80_QID:34.