Business Intelligence Module
Case Study: Roche Challenge

GROUP 5
Kui Hong Lim (Charlene), Nasa Rizoanun, Marc Molina Van den Bosch,
Paul Tanner, Paul Stehberger, Revathi Ravada

FHNW, MSc. Business Information Systems, MSc. Medical Informatics
11th December 2022

Abstract

DigiMap is a repository of Roche digital solutions mapped by disease areas and customer journey which are at different initiative status. The first part of this project aims to unify different data sources and provide a clear and dynamic Tableau visualization of the whole repository so that one can easily filter and extract specific information. The second part consists of the description and implementation of a data mining pipeline which has the aim of identifying patterns and providing insights about whether a solution should be rolled out globally. This has been accomplished with text mining techniques that has enabled to process a semi-structured dataset in a structured format using WEKA toolbox.

1 Overview

1.1 Business Understanding

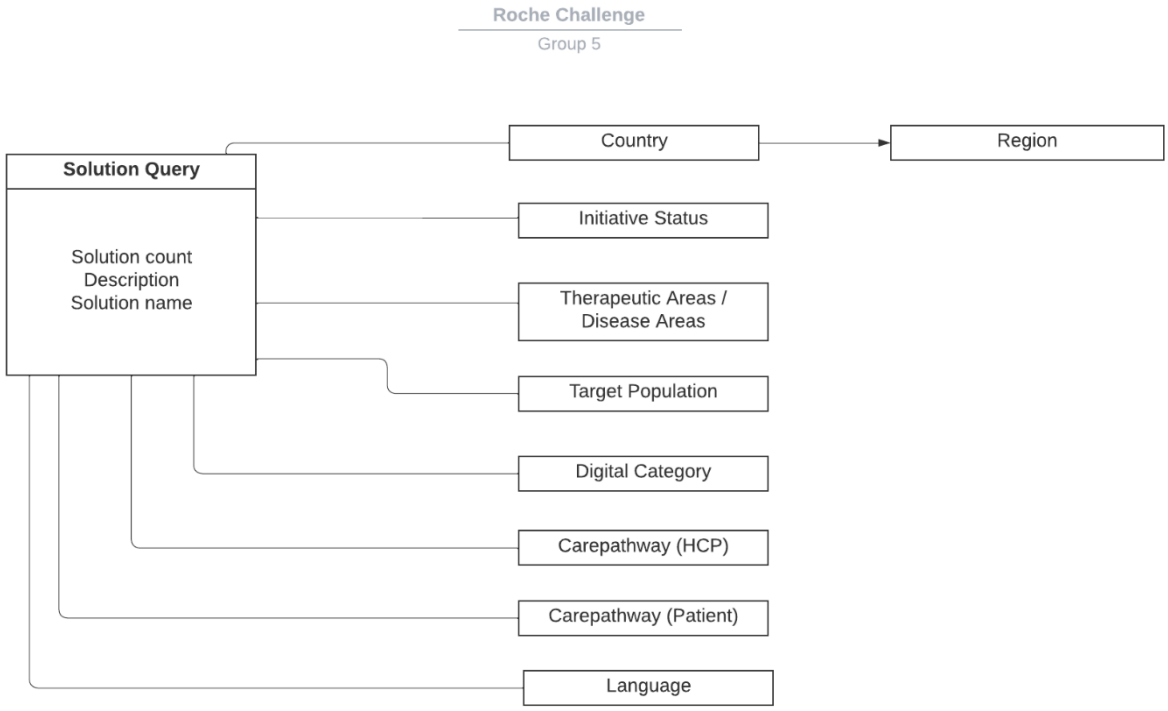
DigiMap is an internal Roche platform that enables its internal stakeholders and network affiliates across the globe to search for the right solutions at the right time with ease. The purpose of the unified dashboard is to connect its users and solutions available across six therapeutics and disease areas with visualization, filtering capabilities, and advanced analytics.

The expectation is that one can search for relevant information, and solution (particularly disease journeys) based on the different criterias and connect with the right contact while identifying what solutions are best suited for adoption by other countries on a unified dashboard. The following questions will be examined:

Target Population	Questions
PPOC / Medical Expert / Portfolio & Strategy manager	1. How many solutions are available in each therapeutic / disease?
	2. Which solutions (per disease area / per initiative status) are available / not available in which countries/regions?
	3. Which solutions are available in which languages?
	4. How many solutions (per disease area) are available for which stages in the patient care pathway?

Table 1: Analytical questions that will be examined for the target population

1.2 Multidimensional model



The multidimensional model of the DigiMap challenge consists of a model capable of searching and filtering a set of solutions from the whole repository. The output measures are the number of solutions meeting the defined criteria as well as the description and name of the solutions in a structured way.

2 Data Mining

2.1 Data Preprocessing

The initial dataset contains both structured and semi-structured data, which are mainly categorical data. One of the steps taken was to analyse what can be extracted from the data, including the text description. To identify the most frequent word on a data frame, some text cleaning was performed such as removing the whitespace, punctuation, numbers, etc). The words in English language that are used to make sentences flow but do not have much meaning on its own (i.e., the, and, but, etc) were removed during the text cleaning process. Once the text is "clean" (reformatted), "wordcloud()" was performed to identify the frequency of the words (mainly nouns and verbs) across the data frame and on the "*Description*" column (Figure 1 & 2).

Figure 1: Words appear more than 50 times on the data frame

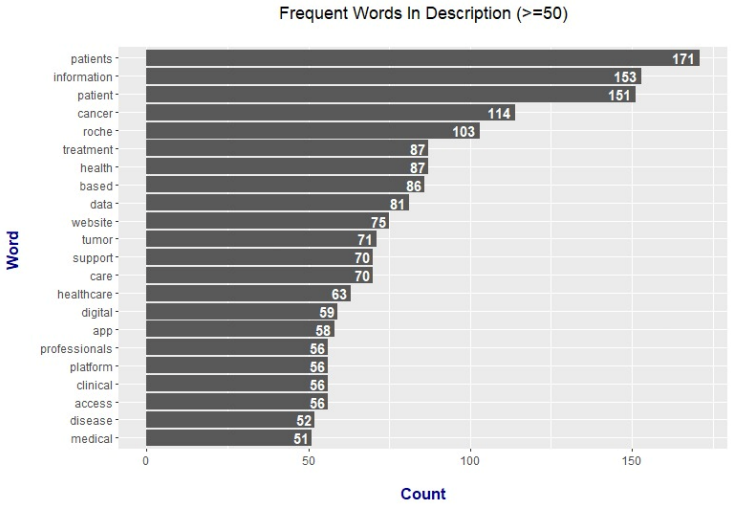


Figure 2: Words appear more than 50 times on the "Description" column

Based on the results, it infers that these words could provide a primitive overview of the overall dataset and might help paint a big a picture.

Data was cleaned using python before being visualised on Tableau. This included ensuring that there was uniformity in the data for variables such as country names and language. Capital letters were incorporated for the countries and German words contained in the "*Language*" columns were changed to their English equivalents. Some rows also contained cells with multiple data objects for a variable. For example, some cells contained multiple languages or country names. Tableau would therefore be unable to recognize that the solution could be counted in two categories and would instead categorise the solution as a unique case. There was therefore a need to modify the structure of the data to make it usable and readable by Tableau.

This meant that rows that contained cells with multiple data objects would need to be separated. First the python pandas library was utilised to split the data into different data frames based on specific variables which made it easier to handle. Separate "*Country*", "*Language*" and "*Patient care pathway*" data frames were created. Each data frame contained the solution ID as a key to connect the data frames to the main data base. In each data frame there was then a need to split the multiple data objects in the cells into new rows that contained the same solution ID. For example, if a solution had "USA, UK" in the country column then the row was copied and two separate rows were created one containing "US and the other "UK" but with the same index and solution ID. Before this was done the strings contained in the cells were first separated by a space and comma to distinguish them as separate objects. The pandas explode function was then used to transform the data into new rows. This enabled us to ensure that when filtering and counting solutions by a variable such as "Country" we would get an accurate count and not miss certain values. Of course, for an actual count of just solutions there was a need to contain the original number of solutions, and therefore the original cleaned data set was used for this purpose.

Once the data was split in this way the number of instances for certain variables increased. It also reduced the number of categories in the data itself and made it more succinct. The new files were then linked to the main data set in Tableau and gave an accurate count of variables such as "*Country*", "*Language*" and "*Patient care pathway*." Another hurdle that was encountered during this phase was quantity of missing values for certain key variables. In a count of solutions by disease area, there were 45 solutions that were not categorized. In the language category, there were 10 "nan" values that were returned when counted. This meant that numerous solutions could not be categorized or included in the analysis of languages. There were also problems with some categories of data that were provided. There was for example no explanation for the category "Ã" in the "*Patient care pathway*" category, which alone accounted for 23% of all instances. The decision was to relabel these "null" and "nan" value as "Not Categorized" or "Not Specified". It was important for us to include these new labels as it showed where there was a need for Roche's stakeholders to address a need for more categories. They are useful in identifying the gaps in the data that need to be addressed by Roche's stakeholders in the future. This will be further discussed in the Visualisation section of the report.

Attributes	Attribute Type	Description
Solution name	Nominal data	Solution name
Link	Nominal data	Web address reference to the solution
Description	Nominal data	Description for the solution
Therapeutic Area / Disease Area	Nominal data	Type of therapeutic area or disease area
Digital Categories	Nominal data	Type of solution
Point of contact	Nominal data	Contains name and/or email of the contact person
Initiatives	Ordinal data	Status of the solution
Country	Nominal data	Country involved
Other countries involved	Nominal data	Other countries involved
Language	Nominal data	Language available
Target Population	Nominal data	Target population for the solution
Care pathway (HCP)	Nominal data	Care pathway for healthcare professionals
Care pathway (Patient)	Nominal data	Care pathway for patient

Table 2: raw data set with all attributes

The tools used for data preprocessing are R, Python, and apps script google sheets.

3 Data Exploration and Visualisation

In order to help answer the analytical questions mentioned above, the preprocessed data was introduced to Tableau and visualizations were made using different worksheets. To create a unified dashboard, a list of all available solutions and the variable "*Global Rollout*" were added to the common representation of all visualizations. Empty fields in the data across most attributes which were shown by Tableau as “Null” or “nan”, were renamed to “Not categorized” or “Not Specified”, as it was not known if they represented missing data or a value which was intentionally left blank.

Analytical Question 1: How many solutions are available in each “Therapeutic / Disease Area”?

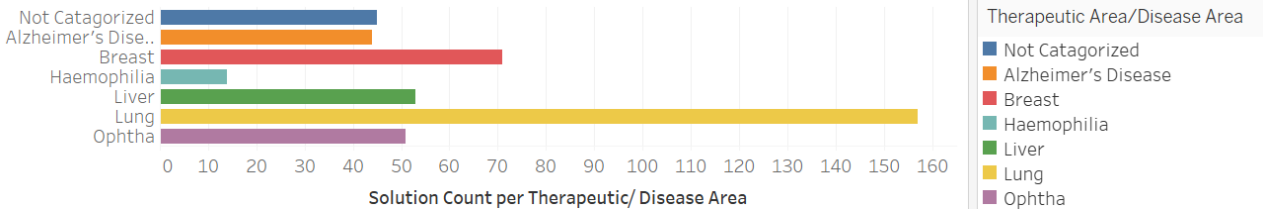


Figure 3: Absolute count of solutions by “Therapeutic / Disease Area”

Figure 3 shows the absolute count of solutions by “*Therapeutic / Disease Area*”. Most of the therapeutic and disease areas were covered by a similar number of available solutions between 44 and 71, with the

exceptions of the emphasis on solutions regarding pulmonary diseases and therapies with a count of 157, and the topic of Haemophilia for which only 14 solutions are available on Digimap.

Analytical Question 2: Which solutions (per disease area / per initiative status) are available / not available in which countries/regions?

In the original dataset, the provided geographical information was labelled as country and / or by region. Multiple countries per one solution were processed as described above to be shown on a map view in Tableau. Concerning the **Region** values, which were partially overlapping (i.e., "Europe" and "EU", "LATAM" and "CAC") and missing a unique definition (i.e., "APAC"), it was decided to not include them in the map view.

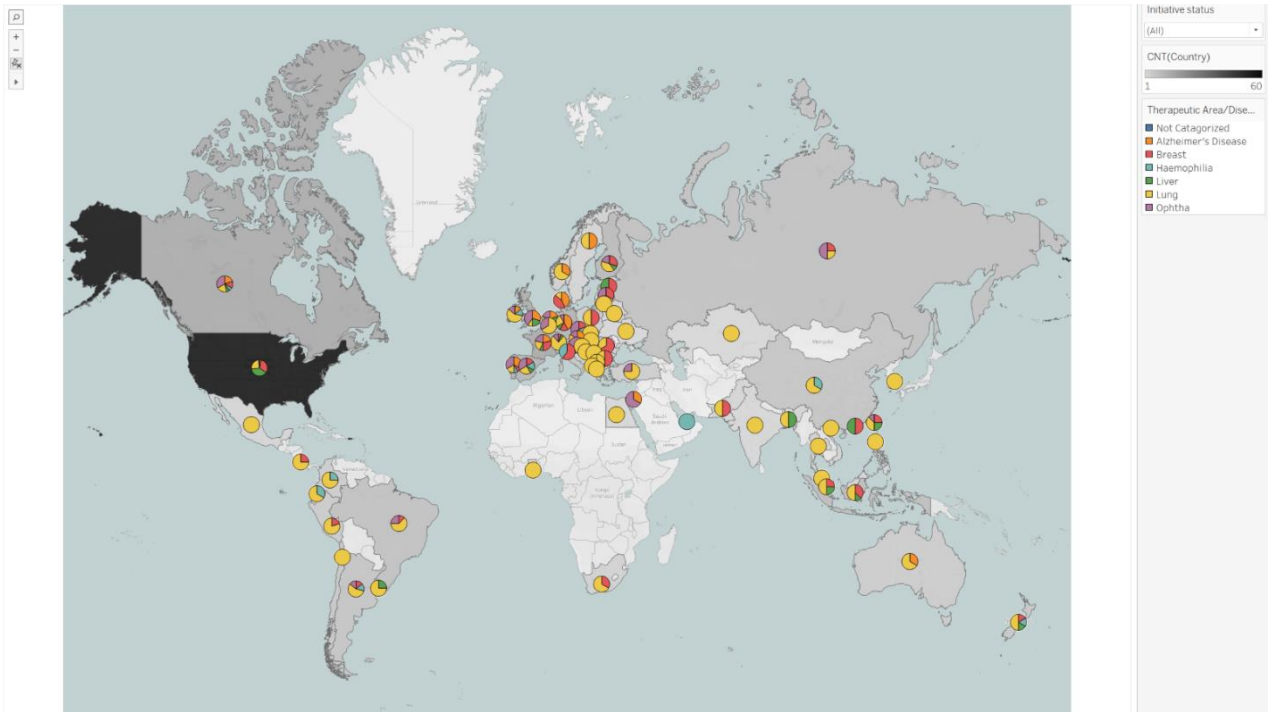


Figure 4: Map View based on Country Data

In order to answer the second analytical question, we included several different attributes in the visualization in (Figure 4). Only the solutions with one or more countries as geographical data are shown.

The grayscale of the country areas corresponds to the total number of available solutions in the respective country. In each country, a pie chart indicates the number of solutions per disease / therapeutic area. In addition, a filter by initiative status allows further insight.

The map shows a significant concentration of solutions provided by Roche Digimap in Europe and North America. The majority of solutions (197) are marked as ready to be used "Live" or "Launched". The

distribution of these solutions as well as the one of the solutions which are still in the different stages of development appear very similar to the distribution when considering the solutions in total.

When looking at the "*Therapeutic / disease areas*" column, it shows besides the concentration on Europe and North America, most are available in most continents. An exception is the field of "Alzheimer's disease", which is exclusively available in Europe, North America and Australia. This field is the only field without solutions in the "launched / live phase".

However, it must be pointed out that the data included in the map view is limited as described above.

Analytical Question 3: Which Solutions are available in which languages?

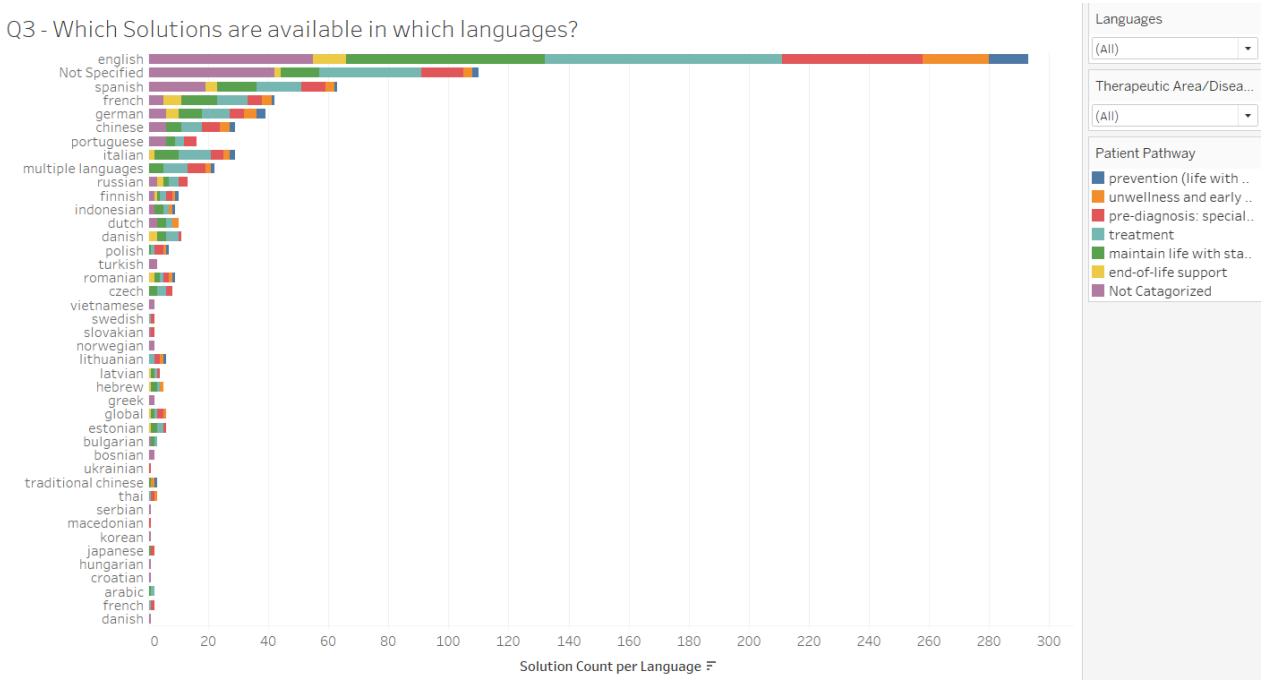


Figure 5: *Which solution is available in which languages*

In this visualisation (Figure 5), we wanted to show how which solutions were available for each language in the data set. We further decided to split each language into their specific "*Patient care pathways*". There are also filters for the "*Therapeutic/ Disease area"* that further enable one to disseminate more specific information. What stood out in this visualisation was the fact that "English" stands out as by far the most popular language for solutions in Roche with 293 instances. Important to note is that the second most common category of language was a "null" value. This was changed to "Not Specified" in order to demonstrate that there is missing data that needs to be completed. This meant that 110 solutions did not specify the language that they provided their solution in. It could be surmised that "English" is the DeFacto language that most solutions are created in. European languages in general are much more common place in solutions with the

closest non-European language being "Chinese" in 6th place with 29 solutions. What is interesting to note is the distribution of the "*Patient care pathways*" within the language count as seen on Figure 6.

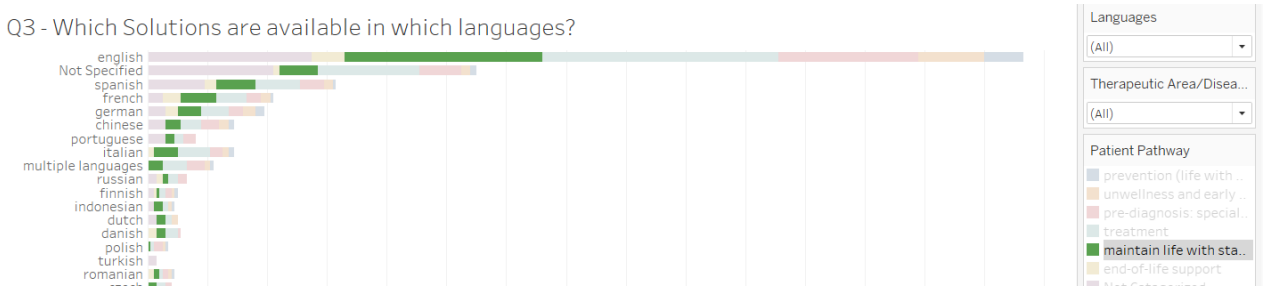


Figure 6: the distribution of "Patient care pathways" within the language count

When "maintain life with stabilised disease" under the "*Patient care pathway"* column is selected, it is apparent that the vast majority of solutions are available in "English". This is similar in all the "*Patient care pathway*". There were also six solutions classed as "global". This is unspecific but could be interpreted to mean that the solutions are available in all languages. When drilled down further it also appears that certain "*Patient care pathway"* for specific "*Therapeutic/Disease area*" are only available in small countries. On such case is shown below where the solution is only available in Lithuanian.

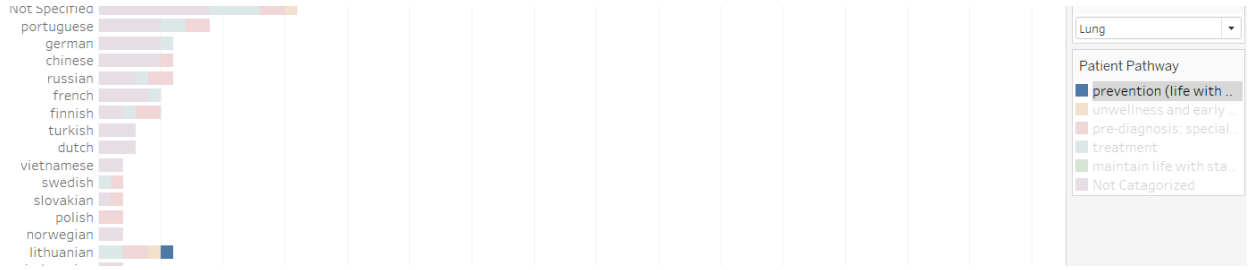


Figure 7: "Patient care pathway" for specific "Therapeutic/Disease area" count by "country"

Analytical Question 4: How many Solutions (per disease area) are available for which stages in the patient pathway?

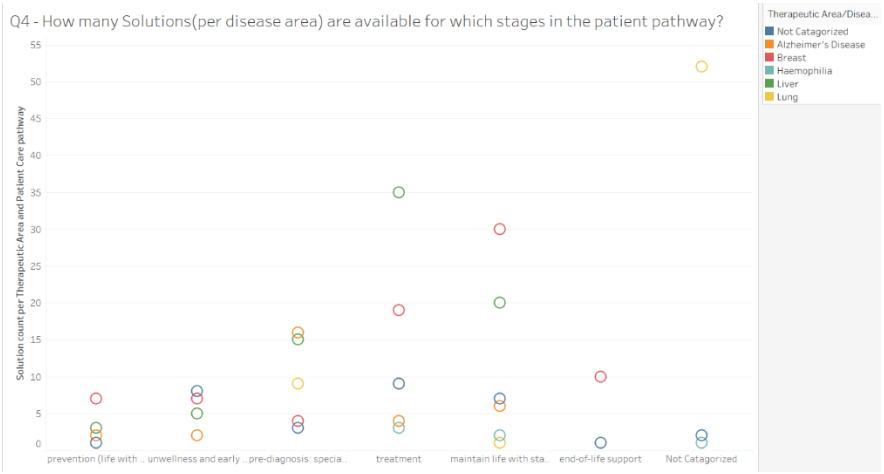


Figure 8: Solution (per disease area) available for which stages in the "Patient care pathway"

This question called for a visualization that included both the "*Patient care pathway*" and the "*Therapeutic/ Disease Area*" for each solution. There were two main options to display this data. One was to create simple matrix in order to count the instances where both categories were met and the other was to create a chart where the filter would enable the sure to make distinctions between different disease areas. In the chart there was the option to highlight the solution count for the different therapeutic areas based on the care pathway.

We can see from the graph that the most populated column is the solutions that are focused on general "treatment" with 174 instances with the subcategory "Lung" having the most instances. The second most instances were recorded in the "Not Categorized" category with 156 instances with 111 being in the sub category "Lung". From this visualisation we can see that there is a heavy emphasis on Liver, Lung, and Breast as therapeutic areas of interest within the company. Solutions in the category "Breast" were mostly categorized in the patient pathway "maintain life with stabilised disease" with 61 counts of 128 solutions. Solutions targeting liver were mainly focused in the "Treatment" category of the "*Patient care pathway*". When the "Not Categorised" variable is removed for "*Patient care pathway*" and "*Therapeutic Area/ Disease Area*", the solutions that fall into the Lung category drastically decrease in count as we will see in the second visualization. When looking further into the data we can see that the three top areas that have the most counts and therefore the most apparent importance are the following combinations; "Lung" and "Not Categorized", "Breast" and "maintain life with stabilized disease", and "Liver" and treatment accounting. These account for a total of 33% of the total solutions and indicate the that "Lung" and "Not Catagorized" account for 16% of all solutions. This is almost a 5th of all solutions in the database.

The second visualization (Figure 9) shows us the count represented as squares in a matrix. In this visualization the "Not Categorized" category has been dropped for both parameters. This has been done in order to provide us with a different more accurate view of the data we actually have. This representation makes it clear that the "*Patient care pathways*" and "Therapeautic areas" are targeted most by Roche are "treatment" with 35%, "maintain life with stabilized disease" with 26%, and "pre-diagnosis: specialists run tests and scans" with 21% of the "*Patient care pathway*" category. These three categories make up 82% of all the solutions.

The three areas with the most instances account for a total of 30% all the solutions within this subset of the data. These were "Breast" and "treatment" with 10 solutions, "Liver" and "treatment" with 11 solutions and "Breast" and "maintain life with stabilized disease" with 14 counts. This shows us that due to missing data the focus of the company seems to paint a different picture altogether. Instead of having "Lung" as an important area, "Breast" is now accounts for most of the solutions. This creates an issue when analysing the data as ideally we want the data to be categorised in order to have a better understanding of the results. It is

possible that a "null" value was imputed as there was no pre-defined category that suited the solution. In this case there would need to be a new category created.

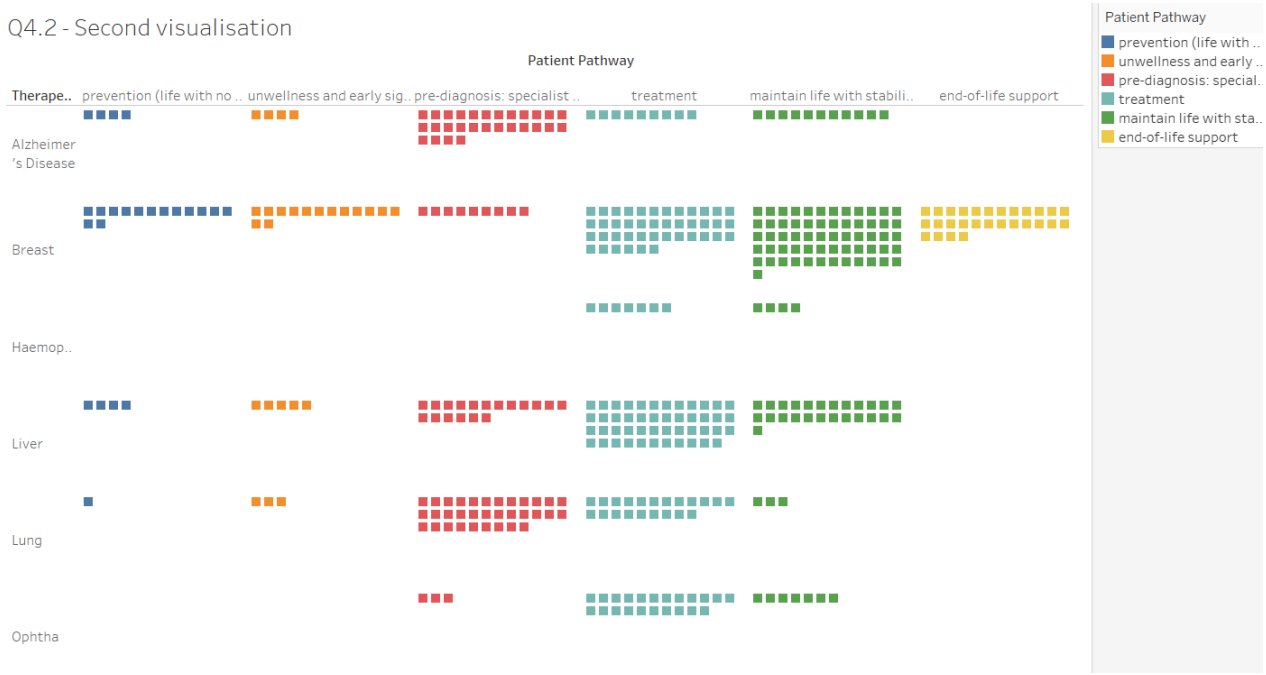


Figure 9: Count represented as squares in a matrix. There is a reduction in the number of "Lung" instances as predicted by dropping the "null" values

4 Data Modelling

4.1 Further Data Preprocessing

From the initial raw dataset which has been previously discussed in **2.1**, some data cleaning aspects are also applicable for the modelling section. Specifically, the "*Description*" of each solution has been cleaned so that non-informational characters (i.e., commas, full stops, "") have been removed.

Then, the non-important columns have been dropped because they would introduce more bias rather than provide information to the classification model. These columns are "*Solution name*", "*Link*" and "*Point of contact*".

After that, it has been assessed that the many values within a single cell (comma separated) prouved difficult to perform prediction on Weka, this applies for the columns "*Country*", "*Language*", "*Initiative status*", "*Patient care pathway*", and "*Healthcare provider care pathway*". The approach followed in this case has been to convert each of the columns into a one-hot encoded format where all the unique values of each column are now represented as a new column in a binary format (0 indicates non-existence, 1 indicates

existence). In the next table, a sample of the one-hot encoding process is presented for the "*Patient care pathway*" column.

Original CPP	Prevention	UnwellEarlySignal	Prediagnosis	Treatment	Maintain	EndOfLife
4	0	0	0	0	1	0
1,2,3,4	1	1	1	1	0	0
5,6	0	0	0	0	1	1
4,5,6	0	0	0	1	1	1
4,5,6	0	0	0	1	1	1

Table 3: One-hot encoding performed on the "Patient care pathway" column

After all these preprocessing steps, the target variable has been engineered, with the "*Country*" one-hot encoded column (one for each country). The heuristic approach is that a solution is considered global if it is present in three or more countries or if the country is a world region, being these EU (European Union), APAC (Asia-Pacific), LATAM (Latin America), and CAC (Central America).

The final target attribute results in a binary class, as described in the subsequent bar plot:

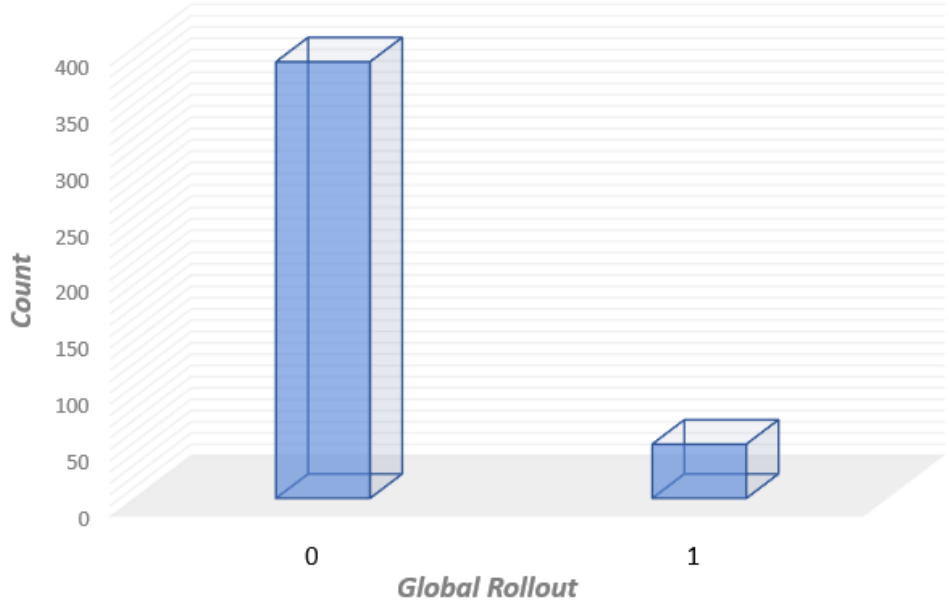


Figure 10: Global rollout (0 = Non-Global, 1 = Global)

Lastly, all features except the "*Description*" (which remains as string) has been kept as a "Nominal" data type and the ultimate dataset used in the consecutively presented classification benchmark is the following:

Features	Data Type	Description
Description	String data	Description for the solution
Therapeutic Area / Disease Area	Nominal data	Type of therapeutic area or disease area
Digital Categories	Nominal data	Type of solution
Initiatives	Nominal data	Status of the solution
Language	Nominal data	Language available in one-hot encoded format
Target Population	Nominal data	Target population for the solution
Care pathway (HCP)	Nominal data	Care pathway for healthcare professionals
Care pathway (Patient)	Nominal data	Care pathway for patient

Table 4: Selected features used for classification benchmark: In this table the term "in separated and one-hot encoded format" means that originally the column had multiple values in one cell and now it has been separated and one-hot encoded.

All in all, this sequence of steps enables to clean and structure the dataset. However, the "*Description*" remains as a *string* which would heavily shrink the number of models that can work with both categorical and text data. Therefore, after uploading the .csv to WEKA the "*Description*" has been processed using the unsupervised attribute filter names **String2WordVector**. This function converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings. Specifically, all data has been normalized, IDF (inverse document frequency used for weighting each item) and TF transforms (measure of how often an item appears) have been applied, with a Snowball stemmer (maps different forms of the same word to a common "stem"), Rainbow stop words (used to eliminate unimportant words, allowing applications to focus on the important words instead.) and N-gram tokenizer (breaks text down into words whenever it encounters one of a list of specified characters, then it emits N-grams of each word of the specified length. N-grams are like a sliding window that moves across the word).

After that the "*Description*" column is automatically dropped and the resulting columns of the **String2WordVector** are transformed from numerical to binary. The final dataset size that will be used for prediction has 435 rows and 1,423 columns.

4.2 Modelling

In this section, an assessment of the predictive element of the project is described. A thorough analysis of the Roche DigiMap solutions dataset shows no clear target columns to predict given all the other dataset attributes. Thus, a new column, the target variable, has been created to perform a hindsight analysis which can enable us to learn from previously deployed solutions (in whichever initiative stage). By these means,

we aim to develop a classification algorithm that can predict whether a solution has the main features to be globally rolled out.

To complete such task, there is the need to pre-process the raw dataset and create the target variable, perform an exhaustive benchmark of several classification algorithms and decide the final model through a critical evaluation procedure and finally, interpret the ultimately chosen model to extract business insights.

4.3 Classification benchmark

7 classifiers with selected accuracy metrics have been applied for predictive modelling:

1. **Naive Bayes Classifier:**
   Naive Bayes classifier is a statistical classifier. Values of attributes in the classes are assumed to be independent. This assumption is called class conditional independence between every pair of features given the value of the class variable. Naïve Bayes classifier is based on Bayes' statistical theorem, described as:

   $$P(C|X) = (P(X|C) * P(C))/P(X)$$

2. **RandomForest:**
   Random Forest is a meta estimator that fits a number of decision tree classifiers on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control overfitting.

3. **J48 tree:**
   J48 algorithm is one of the most widely used machine learning algorithms to examine data categorically and continuously. The C4.5 algorithm (J48) is mostly used among many fields for classifying data. The J48 implementation of the C4.5 algorithm has many additional features including accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In WEKA, J48 is an open-source Java implementation of the C4.5 algorithm allowing classification via decision trees or rules generated from them.

4. **JRIP rules:**
   JRip is a rule-based classifier that creates propositional rules used to classify instances, following Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is based on association rules with reduced error pruning (REP), where the training data is split into a growing set and a pruning set. First, an initial rule set is formulated over the growing set, using some heuristic method. This overlarge rule set is then repeatedly simplified by applying pruning operators. Then, iteratively, the

pruning operator chosen is the one that yields the greatest reduction of an error on the pruning set. The process ends when applying any pruning operator would increase the error on the pruning set.

5. **SMO:**

   Sequential Minimal Optimization (SMO) is one way to solve the SVM training problem that is more efficient than standard QP solvers. It has the lowest average error rate and is computationally faster by using heuristics to reformulate the training problem into smaller problems that can be solved analytically.

6. **IBk:**

   The IBk algorithm does not build a model, instead it generates a prediction for a test instance just-on time. The IBk algorithm, consisting of the k-nearest neighbor's classifier,uses a distance measure to locate k “close” instances in the training data for each test instance and uses those selected instances to make a prediction.

7. **PART (Rule-based-Classification algorithm)**

   The two dominant schemes for rule learning, C4.5 and RIPPER, operate in two stages First they induce an initial rule set and then they refine it using a rather complex optimization stage that discards (C4.5) or adjusts (RIPPER) individual rules to make them work better together. However, this algorithm consists of inferring rules by repeatedly generating partial decision trees thus combining the two major paradigms for rule generation creating rules from decision trees, and the separate-and-conquer rule learn Moreover, it operates efficiently and avoids post-processing, hence does not suffer from slow performance.

Parameters for measuring performance of classification techniques

Accuracy	The simplest intuitive performance metric is accuracy, which is the ratio of properly predicted observations to all observations. Accuracy = $\frac{TP+TN}{Total}$
Precision	The ratio of accurately predicted positive observations to total expected positive observations. Precision = $\frac{TP}{TP+FP}$
Recall	The ratio of accurately predicted positive observations to all observations in actual class Recall = $\frac{TP}{TP+FN}$
F1 Score	The weighted average of Precision and Recall. This score considers both false positives and false negatives. It is frequently more useful than accuracy, especially if the class distribution is unequal. F1 Score = 2*$\frac{Precision*Recall}{Precision+Recall}$
ROC	The ROC plots the true positive rate (TP) at y-axis of a classifier against its false positive rate (FP) at x-axis. TP = $\frac{TP}{TP+FN}$ and FP = $\frac{FP}{FP+TN}$

Table 5: parameters description

Several models are tested using 10-fold cross validation and the results are presented in the subsequent table:

Classifiers	**Correctly classified %**	**Precision Class 0**	**Precision Class 1**	**Recall Class 0**	**Recall Class 1**	**F-Score Class 0**	**F-Score Class 1**	**ROC**
naïve Bayes	83	0.92	0.29	0.88	0.42	0.9	0.36	0.74
Random Forest	90	0.90	**0.73**	**0.99**	0.17	**0.95**	0.27	**0.78**
J48 pruned	87	0.91	0.37	0.94	0.29	0.93	0.33	0.62
JRIP	**91**	**0.93**	0.64	0.96	**0.44**	**0.95**	**0.52**	0.70
SMO	87	0.91	0.39	0.95	0.25	0.93	0.30	0.60
IBk (k=1)	86	0.91	0.34	0.94	0.27	0.92	0.30	0.65
PART	88	0.92	0.43	0.95	0.31	0.93	0.36	0.59

Table 6: classifiers result based on 10-fold cross validation

To assess the difference between these models one must use the evaluation metrics provided above. By these means, the main scope of the predicting task is to identify which solutions can be rolled out globally. Therefore, the first metric to look at is the **Recall** *class 1* since it enables us to quantify how many True Positive predictions have been made out from all existing positive class instances. Secondly, **Recall** *class 1* must be maximized while preserving a proper precision of non-global solutions (all models have high **Precision** and **Recall** metrics for *class 0*). In conclusion, the trade-off is made by addressing the **F-score** for each class since we are dealing with a highly imbalanced dataset, ROC curve is not informative because it is biased towards the larger population when it comes to classification).

In these terms, out of all the models, **JRIP** has the **best F-score** for *class 1* and *class 0*. Moreover, **JRIP** is an acceptable final type of model since its output is a set of classification rules which, in a real-world scenario, are essential in order to draw interpretations and, in our case, explain why a solution needs to be rolled out globally.

The last step of the modelling part is to assess whether the model can be improved using a cost matrix. In the next table, a summary of all the evaluated models is presented:

JRIP Cost matrix Classifiers	**Correctly classified %**	**Precision Class 0**	**Precision Class 1**	**Recall Class 0**	**Recall Class 1**	**F-Score Class 0**	**F-Score Class 1**	**ROC**
[[0, 1],[1, 0]]	**91**	**0.93**	**0.64**	**0.96**	0.44	**0.95**	**0.52**	0.70
[[0, 1],[2, 0]]	88	0.92	0.46	0.93	**0.52**	0.93	0.49	**0.72**
[[0, 1],[1.5, 0]]	88	**0.93**	0.45	0.93	0.46	0.93	0.45	0.69
[[0,0.5][1.2, 0]]	89	**0.93**	0.51	0.95	0.44	0.94	0.47	0.67

Table 7: Cost Matrix benchmark and evaluation results

By changing the values of the cost matrix, we don't see any improvement in the **F-Score** in *class 1*, so then as a result, we might keep the first chosen model. Despite this fact, the model with a cost for the False Negatives of 2 shows the **best Recall** for *class 1* which could be insightful to analyse and interpret if we want to know which rules are the ones that enable to capture more global solutions.

5 Model interpretation

5.1.1 Best F-Score model

```
Classifier Model
JRIP rules:
===========

(trials_binarized = 1) => GlobalRollout_binarized=1 (18.0/8.0)
(Multiple Languages_binarized = 1) => GlobalRollout_binarized=1 (10.0/2.0)
(Initiative status = Research) and (Prediagnosis_binarized = 0) => GlobalRollout_binarized=1 (12.0/4.0)
 => GlobalRollout_binarized=0 (395.0/22.0)

Number of Rules : 4
```

Based on **4 JRIP Rules**, this model obtains the higher **F-Score** for *class 0* and *class 1*, which means that it has the best balance between **Precision** and **Recall** for both classes out of all the tested models. These 4 rules can be understood as:

- Solution description with word “trial” : out of 18 solutions that have the word “trial” in the description, 10 of them are global, which shows an association that could not be identified at first glance. Nevertheless, it is still difficult to precisely tell the word “trial” and its influence on *GlobalRollout*.
- Multiple languages association with *GlobalRollout*: out of 10 solutions 8 are global. Clearly a solution that contains multiple languages would ease the deployment in other countries.
- If the solution is at “*Research*” stage and its intended “*Patient care pathway*” at pre-diagnosis : out of 12 solutions, 8 of them are global. This leads us to think that in the last decade pre-diagnosis is being considered important for major stakeholders (i.e., PPOC, medical expert, etc) since it reduces patient morbidity and costs. It seems that this could be achieved by providing the relevant educational content (a.k.a information) and support to help the patient gain knowledge about his or her illness. As for the condition where “*initiative status*” is at “*Research*” stage, further analysis is required to be performed for further interpretation.

5.1.2 Best Recall model

```
Classifier Model
JRIP rules:
===========

(Initiative status = Research) and (Prediagnosis_binarized = 0) => GlobalRollout_binarized=1 (18.012422360248447/3.6024844720496896)
(TherapeuticArea = Liver) and (clinical_binarized = 1) => GlobalRollout_binarized=1 (17.111801242236027/2.7018633540372674)
(TherapeuticArea = Liver) and (digital_binarized = 1) => GlobalRollout_binarized=1 (5.4037267080745335/0.0)
(Multiple Languages_binarized = 1) => GlobalRollout_binarized=1 (12.608695652173912/1.8012422360248448)
(trials_binarized = 1) and (TherapeuticArea = Breast) => GlobalRollout_binarized=1 (8.1055900621118012/0.9006211180124224)
(and legal_binarized = 1) => GlobalRollout_binarized=1 (4.503105590062113/0.9006211180124224)
(target_binarized = 1) => GlobalRollout_binarized=1 (4.503105590062113/0.9006211180124224)
(monitoring_binarized = 1) and (triage_binarized = 1) => GlobalRollout_binarized=1 (3.6024844720496896/0.0)
 => GlobalRollout_binarized=0 (361.1490683229791/23.41614906832298)

Number of Rules : 9
```

The Best Recall model is based on 9 rules which lead to the correct classification of many global solutions. Although this is done by diminishing the overall model performance, it could be interesting to assess which rules can be used to identify positive class instances. The model applies to the following rules:

- If the solution is at “*Research*” stage and its intended “*Patient care pathway*” at pre-diagnosis: this rule was also identified in the **best F-score model**, which further corroborate its usefulness for the classification problem.
- Liver as “*Therapeutic/Disease Area*” and text description with “clinical”: fundamentally, this rule points out that if the solution is focused on liver diseases and is intended to apply it in the clinical environment the solution might be global.
- Liver as “*Therapeutic/Disease Area*” and text description with “digital”: similar to the previous rule, this one considers a solution to be global if it is applied in the liver disease can be applied with digital technologies.
- Multiple languages association with *GlobalRollout*: same rule identified in the **best F-score model**, it points out how important it is to have many languages made available for a solution to be considered global

- Text description with "trial" and Breast as "*Therapeutic/Disease Area*": similar to the first rule identified at the best F-score model where the use case of the solution is focused on clinical trials. Here, it applies to breast disease. This can be understood since breast cancer has a massive impact on our population and conducting clinical trials is of paramount requirement to design efficient and effective treatments.
- Text description with "and legal" : This could signal that global solutions tend to reflect in its description the regulatory approval needed to deploy a solution. Moreover, it could also reflect the importance that the stakeholders give to the legality of a solution.
- Text description with "target": At first glance, the fact of having the word "target" should not have a great impact on whether a solution is global or not since this word can be used to describe a variety of things. The next bullet points show an example where this rule applies and would fail to classify the instances.
    - ❖ Non-global: "Universal patient community app designed to improve digital care for people with breast cancer... patients and their care team, support and empower each other in a ***targeted*** …"
    - ❖ Global: Building a good database of ePermissions for affiliates to open up more opportunities to be in contact with your ***target*** audience via their preferred channel and time.
- Text description with "monitoring" and "triage": This rule can be understood as if the solution is focused on monitoring the patient in order to make accurate triages. However, the weight of this rule is not large which can be more a bias of the model due to our limited dataset.

There are a few features that have been interpreted at the **best F-score model** (refer **4.1.1**). It is noted that the words "monitoring", "triage", "clinical", "digital" "trial", "and legal", and "target" that appear on the description require thorough text mining for better interpretation and to improve the prediction analysis, i.e, the word "digital" could also be a classification under the "*Digital Category*" column, hence, it is hard to interpret and/or quantify the relation between these words and *GlobalRollout*.

To add further discovery to the association of both liver and breast "*Therapeutic/Disease Area*" have an association with *GlobalRollout*, this could imply that both liver and breast are perhaps one of the top selling diagnostics for Roche to have that GlobalRollout association. This is further supported by Roche first quarter revenues report that Roche's top selling oncology drug in 2021 was HER2-targeted breast cancer drug and one of the drugs further growths Tecentriq was said to come from the additional penetration of liver cancer as it launched around the world in adjuvant lung cancer.

6 Conclusion

According to the classifier results based on **10-fold cross validation** [Table 6], **JRIP** and **Random Forest** performed well, particularly **JRIP**, which demonstrated high **correctly classified instances** (91%), **high Precision Class 0** (0.93), **Recall Class 1** (0.44), **F-Score Class 0** (0.95), and **F-Score Class 1** (0.52). In addition, two models of **JRIP** have proven to perform the best: **Recall** (9 rules) and **F-Score model** (4 rules), with **Recall** showing instances with ratio of accurately predicted positive observations to all observations in actual class much higher than the misclassified instances. Nevertheless, it is difficult to interpret the words and its relation to *GlobalRollout*. One way to look at it, from the first findings with words appear more than 50 times on the data frame [figure 1], some words from the features (i.e., "*global*", "*liver*", "*breast*", "*digital*", "*English*", "*pre-diagnosis*") appear frequently together. Hence it is suggested that with thorough text mining, more data, and advanced analytics, it might provide more capacity to identify meaningful patterns and relationships and achieve better prediction. Finally, solution with multiple languages, solution at "*Research*" stage, and "*Patient care pathway*" at "*pre-diagnosis*" stage seem to have a strong association with *GlobalRollout*, same for liver and breast treatment.

These findings seem to correspond partly to what was displayed on tableau visualization where "Breast", and "treatment", "Liver" and "treatment", and "Breast" and "maintain life with stabilized disease" were the three areas with most instances. Same goes to "English", a language which is prevalent and required to meet the "*GlobalRollout*" criteria.

In a nutshell, data wrangling remains a huge challenge due to imperfect dataset; nevertheless, we've managed to connect the dots using the different tools and visualization to respond to the analytical questions and perform predictive analytics.

7 Acknowledgement

With the guidance, support, and coaching from Roche's representative Fabio Eglin and our professor Dr. Hans-Friedrich Witschel, we have managed to navigate through the challenges and complete the assignment.

8 References

1 Weka Classifiers functions: Classifiers rules: Class JRip[Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/JRip.html

2 Prof. Dr. Hans Friedrich Witschel. (October, 2022). A Short Primer of Predictive Analytics. FHNW. Lecture

3 Prof. Dr. Hans Friedrich Witschel. (September, 2022). A Short Primer of Multidimensional Modeling. FHNW. Lecture

4 Priya Pedankar. "Predictive Analytics vs Data Mining". EDUCBA. [online]. Available: https://www.educba.com/predictive-analytics-vs-data-mining/

5 Tan, Steinbach, Karpatne, Kumar. (September 8, 2021). Lecture Notes for Chapter 1, Introduction to Data Mining, 2nd Edition. [online] available: www.coursehero.com/file/112440997/Intro-Ch1-F21pdf/

6 Tan, Steinbach, Karpatne, Kumar. (September 9, 2021). Lecture Notes for Chapter 2, Introduction to Data Mining, 2nd Edition. [online] available: https://www.coursehero.com/file/48103559/chap2-datapdf/

7 C.Dèlènlè, M.U. Cuma. (2021) "Comparative Analysis with Rule Based Algorithms for Various Datasets", Artificial Intelligence Studies, vol.3, no.1, pp.01-09, 2021. Doi:10.30855/AIS.2021.02.01 [online] https://aistudies.org/index.php/ais/article/view/43/19

8 (April 25, 2022). "Roche Q1 Pharmaceutical Revenues Grow 5 Percent with Contributions from PrecisionCancer Drugs". Precision Oncology News. [online]. Available: https://www.precisiononcologynews.com/business-news/roche-q1-pharmaceutical-revenues-grow-5-percent-contributions-precision-cancer-drugs

9 William W. Cohen: (1995) "Fast Effective Rule Induction". In: Twelfth International Conference on Machine Learning, 115-123.

10 "Sklearn.ensemble.RandomForestClassifier scikit learn. [online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble

11 Jason Brownlee. (February 26, 2014). "How to Tune a Machine Learning Algorithm in Weka". Machine Learning Mastery. [online]. Available: https://machinelearningmastery.com/how-to-tune-a-machine-learning-algorithm-in-weka/

12 Weka Classifiers functions: Classifiers rules: Class lazy IBk[Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/lazy/IBk.html

13 Sunil Ray. (September 11, 2017). "6 Easy Steps to Learn Naïve Bayes Algorithm with codes in Python and R". Analytics Vidhya. [online]. Available: analyticsvidhya.com/blog/2017/09/naïve-bayes-explained/

14 "pandas.DataFrame.explode". pandas.[online]. Available : pandas.pydata.org/docs/reference/api/pandas.DataFrame.explode.html

15 "Finding the Most Frequent Words in Text with R". dkmathstats. [online]. Available: https://dk81.github.io/dkmathstats_site/rtext-freq-words.html

9 Annexes

1 Naïve Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         359               82.5287 %
Incorrectly Classified Instances        76               17.4713 %
Kappa statistic                          0.2475
Mean absolute error                      0.1748
Root mean squared error                  0.4072
Relative absolute error                 88.3377 %
Root relative squared error            129.9627 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area   PRC Area   Class
                 0,876     0,583     0,924       0,876    0,899       0,252    0,744      0,957      0
                 0,417     0,124     0,294       0,417    0,345       0,252    0,744      0,253      1
Weighted Avg.    0,825     0,533     0,854       0,825    0,838       0,252    0,744      0,880

=== Confusion Matrix ===

   a   b   <-- classified as
 339  48 |   a = 0
  28  20 |   b = 1
```

2 Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         392               90.1149 %
Incorrectly Classified Instances        43                9.8851 %
Kappa statistic                          0.2399
Mean absolute error                      0.166
Root mean squared error                  0.2846
Relative absolute error                 83.8851 %
Root relative squared error             90.8235 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area   PRC Area   Class
                 0,992     0,833     0,906       0,992    0,947       0,317    0,784      0,964      0
                 0,167     0,008     0,727       0,167    0,271       0,317    0,784      0,421      1
Weighted Avg.    0,901     0,742     0,886       0,901    0,872       0,317    0,784      0,904

=== Confusion Matrix ===

   a   b   <-- classified as
 384   3 |   a = 0
  40   8 |   b = 1
```

3 J48 Prunned

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         377               86.6667 %
Incorrectly Classified Instances        58               13.3333 %
Kappa statistic                          0.2527
Mean absolute error                      0.1599
Root mean squared error                  0.3358
Relative absolute error                 80.8238 %
Root relative squared error            107.1743 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,938    0,708    0,914      0,938    0,926      0,255    0,621     0,911     0
                 0,292    0,062    0,368      0,292    0,326      0,255    0,621     0,238     1
Weighted Avg.    0,867    0,637    0,854      0,867    0,860      0,255    0,621     0,837

=== Confusion Matrix ===

   a   b   <-- classified as
 363  24 |   a = 0
  34  14 |   b = 1
```

4 JRIP

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         396               91.0345 %
Incorrectly Classified Instances        39                8.9655 %
Kappa statistic                          0.471
Mean absolute error                      0.1437
Root mean squared error                  0.2747
Relative absolute error                 72.5924 %
Root relative squared error             87.6528 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,969    0,563    0,933      0,969    0,951      0,481    0,699     0,930     0
                 0,438    0,031    0,636      0,438    0,519      0,481    0,699     0,393     1
Weighted Avg.    0,910    0,504    0,900      0,910    0,903      0,481    0,699     0,871

=== Confusion Matrix ===

   a   b   <-- classified as
 375  12 |   a = 0
  27  21 |   b = 1
```

5 SMO

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         380               87.3563 %
Incorrectly Classified Instances        55               12.6437 %
Kappa statistic                          0.2378
Mean absolute error                      0.1264
Root mean squared error                  0.3556
Relative absolute error                 63.8912 %
Root relative squared error            113.4757 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,951    0,750    0,911      0,951    0,930      0,245    0,600     0,910     0
                 0,250    0,049    0,387      0,250    0,304      0,245    0,600     0,180     1
Weighted Avg.    0,874    0,673    0,853      0,874    0,861      0,245    0,600     0,829

=== Confusion Matrix ===

   a   b   <-- classified as
 368  19 |   a = 0
  36  12 |   b = 1
```

6 IBk

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         375               86.2069 %
Incorrectly Classified Instances        60               13.7931 %
Kappa statistic                          0.2269
Mean absolute error                      0.1496
Root mean squared error                  0.3697
Relative absolute error                 75.6056 %
Root relative squared error            117.977  %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,935    0,729    0,912      0,935    0,923      0,229    0,650     0,932     0
                 0,271    0,065    0,342      0,271    0,302      0,229    0,650     0,237     1
Weighted Avg.    0,862    0,656    0,849      0,862    0,855      0,229    0,650     0,855

=== Confusion Matrix ===

   a   b   <-- classified as
 362  25 |   a = 0
  35  13 |   b = 1
```

7 PART

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         382               87.8161 %
Incorrectly Classified Instances        53               12.1839 %
Kappa statistic                          0.2959
Mean absolute error                      0.1525
Root mean squared error                  0.3399
Relative absolute error                 77.0591 %
Root relative squared error            108.4727 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,948    0,688    0,918      0,948    0,933      0,300    0,587     0,901     0
                 0,313    0,052    0,429      0,313    0,361      0,300    0,587     0,224     1
Weighted Avg.    0,878    0,617    0,864      0,878    0,870      0,300    0,587     0,826

=== Confusion Matrix ===

   a   b   <-- classified as
 367  20 |   a = 0
  33  15 |   b = 1
```

8 JRIP [[1,0],[2,0]]

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         383               88.046  %
Incorrectly Classified Instances        52               11.954  %
Kappa statistic                          0.4228
Mean absolute error                      0.1711
Root mean squared error                  0.3075
Relative absolute error                 86.4658 %
Root relative squared error             98.1342 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,925    0,479    0,940      0,925    0,932      0,424    0,718     0,931     0
                 0,521    0,075    0,463      0,521    0,490      0,424    0,718     0,387     1
Weighted Avg.    0,880    0,435    0,887      0,880    0,884      0,424    0,718     0,871

=== Confusion Matrix ===

   a   b   <-- classified as
 358  29 |   a = 0
  23  25 |   b = 1
```

9 JRIP [[1,0],[1.5,0]]

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         382               87.8161 %
Incorrectly Classified Instances        53               12.1839 %
Kappa statistic                          0.3851
Mean absolute error                      0.1663
Root mean squared error                  0.3046
Relative absolute error                 84.0423 %
Root relative squared error             97.2128 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,930    0,542    0,933      0,930    0,931      0,385    0,693     0,928     0
                 0,458    0,070    0,449      0,458    0,454      0,385    0,693     0,383     1
Weighted Avg.    0,878    0,490    0,879      0,878    0,879      0,385    0,693     0,868

=== Confusion Matrix ===

   a   b   <-- classified as
 360  27 |   a = 0
  26  22 |   b = 1
```

10 JRIP [[1,0],[1.2,0]]

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         388               89.1954 %
Incorrectly Classified Instances        47               10.8046 %
Kappa statistic                          0.4121
Mean absolute error                      0.1574
Root mean squared error                  0.2936
Relative absolute error                 79.5374 %
Root relative squared error             93.701  %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,948    0,563    0,931      0,948    0,940      0,414    0,677     0,926     0
                 0,438    0,052    0,512      0,438    0,472      0,414    0,677     0,349     1
Weighted Avg.    0,892    0,506    0,885      0,892    0,888      0,414    0,677     0,863

=== Confusion Matrix ===

   a   b   <-- classified as
 367  20 |   a = 0
  27  21 |   b = 1
```