NBNKLM

Fontys
University of Applied Sciences

Informa Pharmacy Data

Project Proposal

by

Class AI43 Group B

Date	Monday, March 15, 2021
Authors	Nour Majdalawi, Branimir Sandalski, Konstantin Mihaylov, Leon Matijevic, Noah Toulouse, Petar Mihaylov

Preface

This is a proposal document for the group project of Semester 4 ICT and Artificial Intelligence at Fontys University of Applied Sciences. Throughout this semester our team will be working on an exploratory project in cooperation with one of Fontys's Partners in Education companies - Informa. This will be a group effort in which we will experience all the different phases of the AI Project Methodology resulting in a final delivery for Informa. Using the provided data, our team should develop a business intelligence solution that could help Informa's clients to increase their business efficiency and customer satisfaction.

The document starts with a brief description of the domain of our project in the face of the pharmaceutical realm. Afterwards, we specify our project goal as well as the societal impact, which it would encompass. We specify our stakeholders and the specifics of the received data, we describe our working process and methodology, including our planning for each phase of the project, and we evaluate the potential risks involved in the process. Finally, we finish the document with a summarizing conclusion.

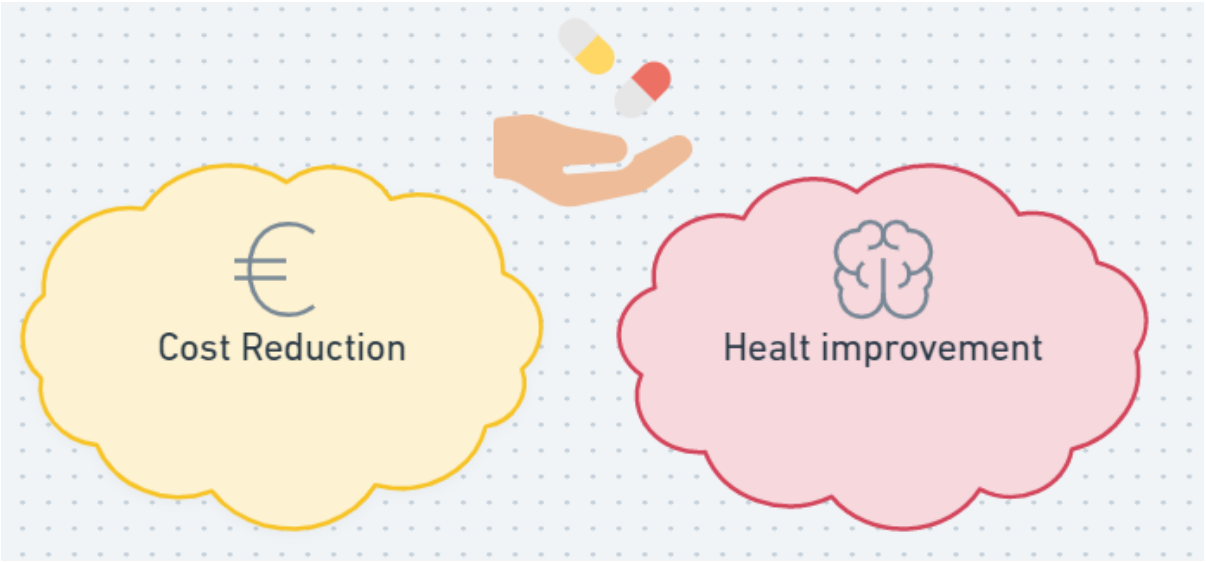
Fontys
University of Applied Sciences

Table of Contents

1. Project Domain

The pharmaceutical industry is part of the healthcare sector that deals mainly with medicines. While the industry consists of different subfields concerning the development, production, and medications marketing, these interdependent subfields consist of drug manufacturers, drug marketers, and biotechnology companies. The main goal of the pharmaceutical industry is to provide drugs that prevent infections, maintain health, and cure diseases. This industry directly affects the global population and therefore several international regulatory bodies monitor things like drug safety, patents, quality, and pricing.
It is important to mention that the pharmaceutical industry is one of the biggest and most important industries in the world and its companies place a particular emphasis on supply chain management. Any variation in the supply chain could lead to multiple disturbances in the system.

Highlighting this field is crucial as we are currently dealing with Informa, a company that provides services for retailers. In the scope of this project, we will be dealing with data relating to a chain of pharmacies situated in Belgium. They would like to implement some form of artificial intelligence as part of their business model but are still uncertain on how and where it could be implemented. Their main interests focus on two main areas:



2. Project Goal

Cost Reduction

The goal of our project is to help Informa with optimizing the costs reduction aspect for the chain of Belgian pharmacies they are affiliated with.

We can approach this by helping pharmacies make better educated purchasing decisions by predicting what types of medicine will be most demanded by the patients during different time periods. In addition to that, usually, for most health issues there are several different medications with similar effects and health improvement outcomes. Therefore, we may be able to analyse which medications, within their category, tend to be better preferred by the customers and thus allow the company to alter their purchases in favour of the medications with higher demand.

This could be highly effective for minimizing the costs of the pharmacies of Informa, since it would allow them to adjust their restocking based on customer demand and therefore minimize potential waste as well. In addition to that we can take into consideration which medications are in high demand during a specific time frame or circumstances such as a pandemic, as well as economic or weather changes during the year. There is valuable information that could positively influence the decisions taken by Informa.

3. Impact on society

Nowadays, when it comes to the effectiveness of machine learning, more data always leads to better results, and the healthcare sector is sitting on a data goldmine. The idea behind machine learning in pharma is not to replace the doctor but to enhance his medical expertise. With an increase in disease infections around the world, especially at present times, people's demand for medicine has increased drastically and currently has the potential to be one of the few sectors seeing tremendous growth. Hence, we must take a deeper look at this sector and think about whether we, as data analysts, can increase the pharmacy's profit by reducing their restocking costs. With the rise in drug prices, patients are searching for cheaper alternatives that their relatives or friends might have found effective. Our role comes here to minimize the costs spent to purchase medicines from manufacturers.

As the number of competitors increases, self-improvement is essential. Competition drives innovation and therefore, our goal is to launch our product to fulfill the purpose of cost reduction. Of course, it will do many more to improve the sector of pharmacy.

4. Technology reliability

This section handles the use of the technology and the issues related, such as privacy, transparency, stakeholders, target group and the technology sustainability.

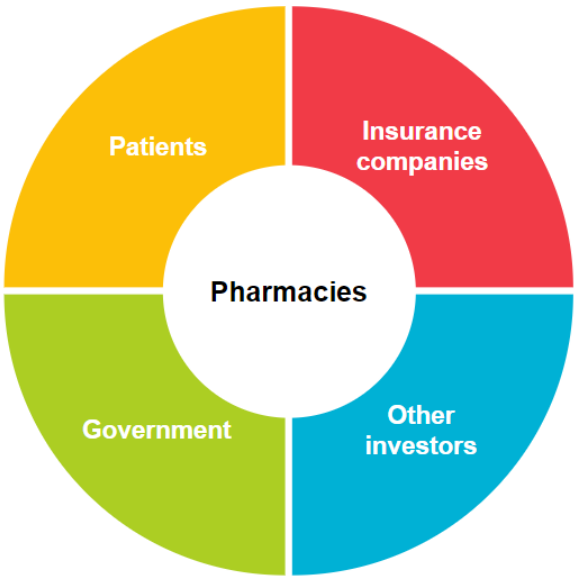
4.1 Privacy

We are not processing Informa's data in a way that could infringe privacy laws within the EU as well as following the guidelines signed through the non disclosure agreement.

4.2 Transparency

to be discussed ….

4.3 Stakeholders



With the rise of technologies, stakeholders are considering in what ways they could benefit from using them. By focusing on the pharmaceutical sector, a lot of stakeholders are playing an essential role. It is important to note that different stakeholders have different interests and might benefit from this technology in different ways:

Pharmacies - pharmacies will achieve great benefits in terms of financial profit if customers frequently visit and make a purchase in their pharmacy. Pharmacies always aim to reduce their restocking costs they spend to purchase sets of medicines from manufacturers. Many

pharmaceutical companies will be in control with this technology as they will be able to know in advance the trends of the most requested products and increase their stocks accordingly.

Patients - first and foremost, customers are primarily interested in their wellbeing; customers are pharmacies' patients and they expect the pharmacies to provide them with the medication they need at all times. Another important aspect that customers expect from pharmacies is transparency; patients do not want pharmacies to disclose any sensitive medical information in any shape or form.

Government - the Belgium government generally has a wide range of interests in the pharmaceutical industry. Apart from wanting their citizens to be healthy, they also want manufacturing companies and pharmacies to be compliant to not only their health & safety rules but also those set by the European Union. The government also cares about sustainability on different levels. Not just from an environmental standpoint but also in terms of the market; prices must be adjusted in a way that is affordable for the typical customer, ideally regardless of product availability.

Insurances - insurance companies are for-profit companies. They expect high RIOs and this is achieved in different ways in the pharmaceutical industry. The main way to reduce costs is to essentially keep their customers healthy. This can be done by mitigating risks through close monitoring. They want to be able to tell reliably what patients need in total transparency.

Investors - much like insurance companies, investors are interested in profits by reducing costs. By reducing pharmacies' restocking costs, money could be allocated elsewhere for other reasons. They also expect total transparency from the pharmacies as well as making sure that the restocking products are handled within the guidelines of the government.

4.4 Target Group

The core of the project is to create a reproducible research and development analysis for Informa that pours its intention to find any available solutions for the six Belgian pharmacies they work for. Additionally, those pharmacies are not located in a specific region of Belgium. Hence, they are spread all over the country and rank within the category of the 10% of the biggest pharmacies in Belgium. Each pharmacy has between 5 and 15 employees. However,

our client cannot provide us with sensitive information about the chain of pharmacies as it is not his responsibility.

4.5 Sustainability

The technology we have is fully sustainable since it is a prediction model which does not really use any materials or resources that could be a potential threat to the environment. So, it does not harm the earth by any chance.

5. Data Definition

Our team received pharmacy data from Informa, containing approximately 1.2 million rows. Here is a breakdown of the structure of that data by columns:

Feature name	Explanation
ID	Unique ID
Patient_token	A randomized token that can be used to identify specific patients in the data set. This token cannot be decrypted, so referencing back to the operating system of our relationship is not possible.
Year_of_birth	Year of birth of the patient.
Gender	Gender of the patient
CNK	A formal unique and nationwide specified identification of the product; National Code Numbers (CNK) system enables every package dispensed in pharmacies to be unambiguously identified. It allows digital data sharing between the stakeholders of the pharmaceutical sector.
Supply_hour	The time the product is supplied to the patient
Supply_date	The date the product was supplied to the patient (ex. hours)
Pharmacy	Pharmacy where the product was purchased
Name	The common name of the product
Price	The price of the product (This is the official price, not what the patient has to pay)
Quantity	The number of product units inside a package
ATC	Anatomical Therapeutic Chemical (ATC) Classification System is a drug classification system that classifies the active ingredients of drugs proved by the WHO and specifies on which organ or system the product works on, and how it works.

6. Objective

Having a proper forecast of the sales will be a huge reference for the business of Informa's clients, because it is going to allow pharmaceuticals to recognise patterns over the time in a way that will describe possible cost reductions compared to the total expected usage.

Hypothesis

As a starting point we are going to achieve it simply by using the supplied data provided by Informa. The forecast analysis will provide an accurate correlation with the unstable trends and client's needs in such a way that it will minimize the amount of the purchased units by pharmaceutical customers for a specific term. Along with that we tend to go further to the patients and make personalized recommendations based on their history and the calculated value of a medicine per specific category. It can help the pharmacies to develop strong connections with their clients when they have enough clarity for their future medical costs per patient.

Additionally, we can deliver a form of a Proof of Concept as a final solution to support the accuracy of the working model and represent the idea to the Informa's clients.

Analysis data:

The related data that we thought will help with promoting this idea and increase its efficiency with building the model from the dataset has:

Column	Usage
SUPPLY_DATE	We need the date supply in order to get accurate understanding about the time.
PHARMACY	We need the pharmacy name in order to distinguish them.
PRODUCT_NAME	We need the product name in order to classify them.
CNK	We need the National Code Numbers to identify the product
ATC	We can search through all of the ATC levels.
UNITS	We need to calculate the price of a single unit.
PRODUCT_PRICE	We need the product price in order to calculate the costs.

Ethical Considerations:

In the way as we focus on the supply chain of the pharmaceutical companies there is low-risk on data processing and privacy for all the patients of Informa's clients related to the General Data Protection Regulations from the EU. GDPR affects data science practices based on three main ways - processing and consumer profiling, automated decision-making and accountability for bias and discrimination for decision making; these rules supplement the requirements for data collection and management which in our case we will slightly relate to our data set as we are going to eliminate certain personal aspects, analyse or predict patient's performance at work, economic situation, health, persal preferences, interests, reliability, behaviour, location or movement.

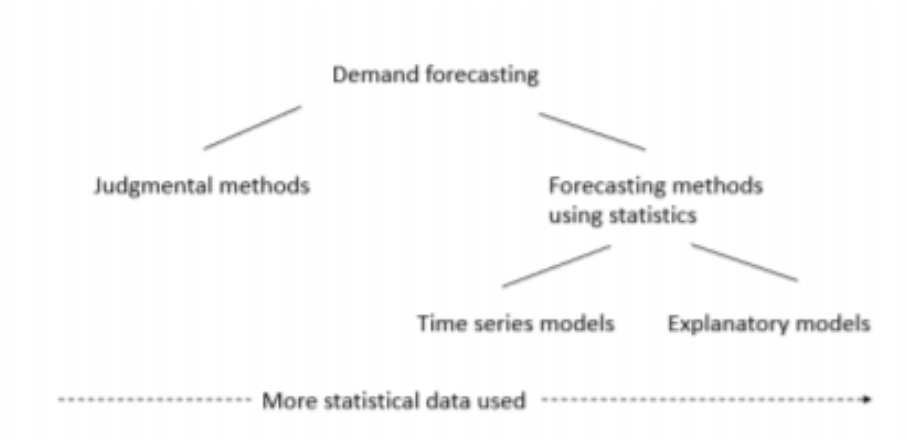
Value Added:

The purpose of this data exploratory analysis is to increase the company's understanding of its medications and reduce the cost per unit. This could be useful to the pharmacies because it can allow them to alter their order quantities for different medications based on live, current, and constantly updated data on the sales of these drugs in order to optimize profit and minimize waste.

This project, like any machine learning solution, fundamentally revolves around the idea that we can use past data to make predictions about the future. Informa, being involved in storing pharmaceutical information from a few different pharmacies, has data regarding medication categories and prices. With that knowledge it may be possible to establish specific analysis and therefore optimize the company's profit by minimizing their costs.

7. Forecasting methods

This section describes in general forecasting methods that might be applicable to the data structure. A common distinction for forecasting models is subjective or objective. At which a subjective forecasting method is made qualitatively based on human judgement, objective forecasting methods are made quantitatively based on data analysis.



7.1. Subjective forecasting methods

Judgmental forecasts are made by individuals or a group, based on knowledge and experience of the situation to forecast. Commonly used at the company, the “last like” rule is an example of judgmental forecasting. This means that they ordered the same quantity of products that was sold during a similar time in the past.

7.2. Objective forecasting methods

Objective forecasting methods are statistical methods that are used to deliver a quantitative forecast about the future which uses numerical measures and prior data in order to forecast future promotions. These techniques are based on mathematical models and are predominately objective. Three types of objective forecasting methods are described below:

Time series model

Time-series models examine historical data patterns and forecast the future based on these underlying data patterns.

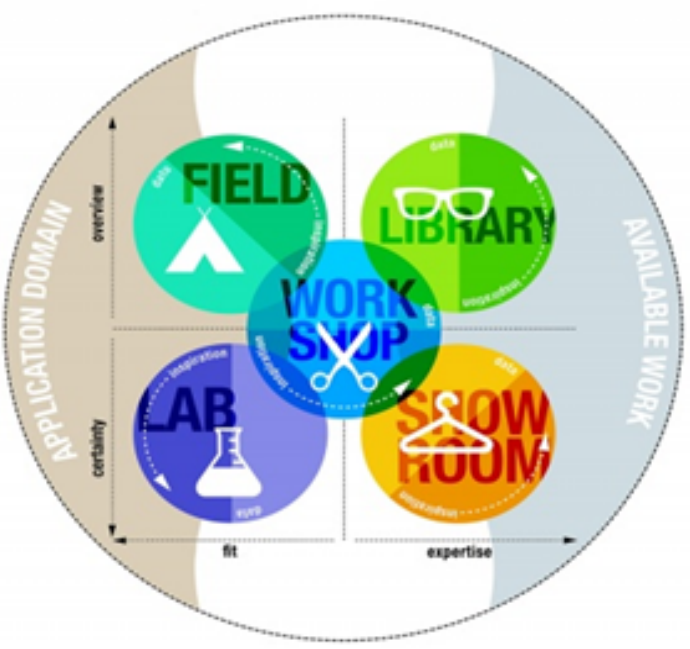
Explanatory model

Explanatory models are causal forecasting models that try to identify relations that were relevant in the past and then apply them in the future. Causal models assume that the variable that is being forecasted, the dependent variable, has a causal relation with the other independent variables. The forecasts of these models are based on this causalization. Linear regression is one of the simplest forms of a causal model. A regression line forecasts the dependent variables based on the selected number of independent variables.

Complex model (resource consuming)

Besides time series models and explanatory models there are more complex forecasting models, like support vector machines, neural networks for producing business forecasts. Neural networks have the advantage that they can approximate nonlinear functions. This however is both very time-consuming and relatively complicated.

8. DOT Framework



To understand the project by hand, we try to learn as much as we can about the context of the available work, the technology context and the innovation space. Following are the strategies we have followed during this research.

- **Library**

How can artificial intelligence help the pharmaceutical sector flourish*?*
○ Supply Chain: Identifying the most efficient supply system will become even more important as drugs are increasingly customized to small populations of patients with certain genetic profiles[3]. ○ Manufacturing: Drug companies are increasingly able to ensure consistent quality standards with the use of advanced analytics[3].

Where to find the weather data we need to give Informa's client the effective insight of what to buy and when?
○ The data has been collected through a free web API storing various weather data from stations all around the world[4].

- **Field**

  The end users of our technology are the company's employees who would take necessary steps corresponding to the restock requests.

- **Lab**

  In the prediction phase of the AI methodology, the final testing and conceptual of the product will be carried out

- **Showroom**

  Showroom research is done to test your ideas in relation to existing work. Showing your prototype to experts can be a form of showroom research or spelling out how your product is different from the competition. Also testing your product to general guidelines is a form of showroom research. The technology prototype will be tested once the proposal and provisioning phases of the AI methodology have been validated and accomplished.

9. Approach

9.1. Methodology

The methodology that is going to be used for this project is Agile. Daily standup meetings will be done, in order for the progress to be tracked. The team has decided to go with Scrum framework. Scrum increases the effectiveness of the team by including everyone and perceiving all members as equals. There is a frequent delivery of prediction data that helps to keep the client up to date with the progress and makes data analysis more flexible to change. The execution of the project will be in phases.

9.2. Phases

Part	Weeks	#Project Days	Phase(s)	Deliverable(s)
A	1- 12	12	1	Proposal, Notebook
B	13 - 14	10	2 and 3	Model with Descriptive Notebook
C	15 - 16	10	2, 3 and 4	Dashboard / API

For this project there will be 4 phases discovered. Phase 1 incorporates the proposal for the AI project. The main objective of phase 1 is to deliver a good document proposal in the end with a basic notebook that shows our data analysis. Then phase 2 and phase 3 are executed in repeating iterations until the project finalises, by which phase 4 will commence. In phase 2 and 3 we are going to dive deeper into the data analysis and in the end we have to come up with a reproducible, reliable model, which is described and jupyter notebook. The idea of making two iterations is that the second iteration most likely improves on the first one because the results of the first iteration give you insight in our project that makes it clear which data we need, what the quality is and how to improve it. In the final phase as a second

outcome, we have to deliver a better visualisation in the shape of an end point as an API or Dashboard, which doesn't need to be a full-stack application.

9.3. Deliverables

Feasibility	Deliverables
Must Have	• **Project Proposal** • **Findings Documentation** (aka resulting doc) • **Trained Model(s)** (Jupyter Notebook)
Should Have	• **Model Visualisation** (e.g. dashboard, Tableau, etc) • **EDA(s)** (data exploration)
Could Have	• **Working prototype** for data interaction and visualisation
Will Not Have	• **Online/Cloud Database** to store patient information for training new models (for security reasons)

Our team is working on developing an end product, which makes clear predictions that could certainly be of good use for the company to help reduce the restocking costs. The main deliverable of the project is a prediction model(s) using extended jupyter notebooks. The model(s) will be initially trained to be ready for prediction. The users of the application will add as much testing data as they want to run predictions on top. In order for users to have a better understanding of the process, a dashboard will be created for better visualization to give clear insight on how and why we made our decisions.

10. Communication

Most of the communication takes place in the form of online meetings via Microsoft Teams:

- Every Monday the Team gets a chance to meet the Semester Coach on Teams in order to demonstrate progress and receive feedback.
- The Team members meet twice a week (online or live) - every Monday and Tuesday. Starting each workday with a Daily Scrum meeting. In these meetings, the tasks and any difficulties are being discussed. Members also update each other on their progress, so everyone is in line with the current state of the project. Since the Scrum framework is being used to tackle the project in an agile manner, all the work is done in **10 iterations** in total, **6 iterations**, each lasting **2 weeks**, for the **first 12 weeks**, and **4 iterations**, each lasting **1 week**, for the last **4 project weeks**.
- Each sprint is completed with a Sprint Review (live or online) meeting where the Data Analysis Team demonstrates the work done and discusses what went well or could be improved. The outcome of this meeting is a revised Product Backlog.
- The Sprint is done after the Data Analysis Team has their Sprint Retrospective meeting (also live or online) where they share feedback reflecting on collaboration and overall process flow.

10.1 Team members

Our team consists of 6 people. Following are the names added with our study background, role in the group and our weekly availability.

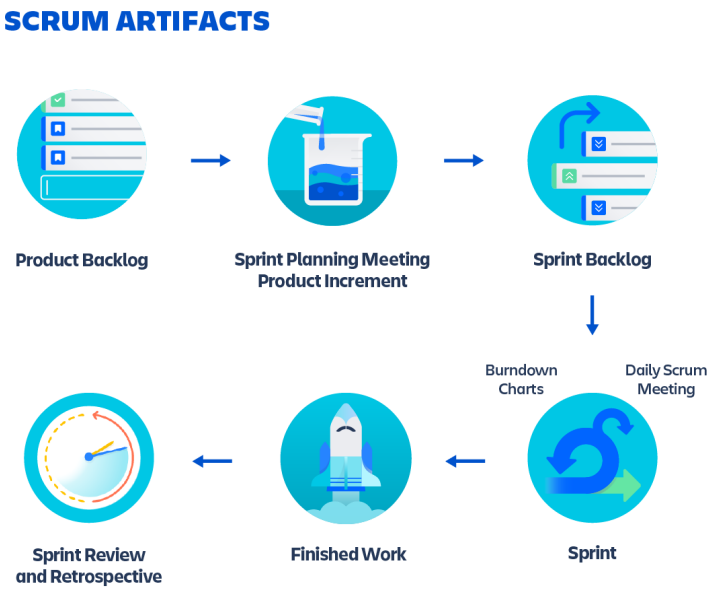
Name	Abbreviation	Study background	Role and functions	Availability
Nour Majdalawi n.majdalawi@student.fontys.nl	NM	Software	Communication person with the PO - Data Analyst	Monday - 9:00 - 16:00 Tuesday - 9:00 - 15:00 Wednesday - 09:00 - 16:00 Thursday - 09:00 - 16:00 Friday - 09:00 - 16:00
Kostantin Mihaylov k.mihaylov@student.fontys.nl	KM	Software	Data Analyst	Monday - 9:00 - 16:00 Thursday - 9:00 - 16:00 Wednesday - 09:00 - 16:00 Friday - 09:00 - 16:00
Leon Matijevic l.matijevic@student.fontys.nl	LM	Software	Data Analyst	Monday - 9:00 - 14:30 Tuesday - 9:00 - 14:30 Wednesday - 09:00 - 16:00 Thursday - 09:00 - 16:00 Friday - 09:00 - 16:00
Branimir Sandalski b.sandalski@student.fontys.nl	BS	Software	Data Analyst	Monday - 9:00-16:00 Tuesday - 9:00 - 16:00
Noah Toulouse n.toulouse@student.fontys.nl	NT	Software	Data Analyst	Monday - 9:00-17:00 Tuesday - 9:00 - 17:00 Thursday - 12:00 - 16:00 Friday - 12:00 - 16:00
Petar Mihaylov p.mihaylov@student.fontys.nl	PM	Software	Scrum Master Data Analyst	Monday - 9:00-16:00 Tuesday - 9:00 - 16:00 Wednesday - 10:00 - 18:00 Thursday - 10:00 - 18:00 Friday - 10:00 - 18:00

11. Artifacts Used

The beauty of a Scrum framework lies in its flexibility. Therefore, the framework has been customized to best meet the needs of the Data Analysis Team. The following artifacts have been chosen to be put into practice.

- **Product Backlog**, which is a list of all the requirements, models, and features for the whole duration of the project. Product Backlog is dynamic and will be changing with each iteration.
- **Sprint Backlog** - a forecast made by the Data Analysis Team about which Product Backlog items will be implemented in the next Increment. It is also dynamic, and the Data Analysis Team can make changes to it if they find any extra work deemed necessary to meet the Sprint Goal. It is a real-time representation of the work that is planned to be accomplished during the Sprint and is represented via a ***Scrum Kanban Board.*** The increment is the total of all Product Backlog items that have already been completed in the previous Sprints.
- **Sprint Goal** helps us to stay focused in each sprint by providing objectives to be met by implementing Product Backlog items. Sprint Goals are set by the Data Analysis Team to give a reason for each Increment.

These are the main artifacts used in this project to assist in getting the best result of data predictions.



12. Risk Analysis

Following is the risk analysis performed for being aware of the possible risks that may arise and how to avoid them appropriately.

RISK DESCRIPTION	RISK SEVERITY	RISK LIKELIHOOD	USER IMPACT	MITIGATIONS / SOLUTIONS
Poor Problem-Solution Alignment	UNDESIRABLE	IMPROBABLE	HIGH	Applying ML to the wrong problem. The purpose of integrating ML into a product or service is to add value.
Lack of variance in the model	UNDESIRABLE	POSSIBLE	HIGH	Reanalyze the ML model and exercise the different data types.
Low accuracy rate	UNDESIRABLE	PROBABLE	EXTREME	Selecting another ML algorithm that provides a better accuracy percentage.
Losing team member	TOLERANT	POSSIBLE	LOW	Redistribution of workload to cover

13. Conclusion

Nowadays, the pharmacy sector plays an essential role especially with the rise of pandemics. Our team, consisting of six data analysts, strives to find the most profitable and satisfying solution that will put a reasonable product to use by our target group. As cost reduction takes a valuable part in the field of business, we will make sure to deliver a meaningful solution in this regard to allow pharmacies to know where to put their money with highly productive models.

References

[1]. Fontys University of Applied Sciences, 24 March 2021,
https://fontys.edu/

[2]. Kerkdijk, Jelle. “Forecasting Promotional Demand Volume.” Utwente, 26 Sept. 2019,
https://essay.utwente.nl/79738/1/Kerkdijk_MA_BMS.pd

[3]. ML & AI in the pharmaceutical industry, 29 March. 2021,
https://blog.dataiku.com/ml-ai-pharma-biggest-gains

[4]. Visualcrossing. “Weather Forecast & Weather History Data”, 30 March. 2021,
https://www.visualcrossing.com/weather-data

Image Credits

1. © Pearson Education, Inc