BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

ClimaRisk

– MIRPR report –

Team members
Bernadett Hoszu, IE 933
Samuel Zirbo, IE 935

2023-2024

Contents

Chapter 1

Introduction

1.1 What? Why? How?

The problem we are addressing centers on forecasting the likelihood of individuals experiencing sudden and critical medical events such as heart failure, respiratory failure and cerebral infarction. This prediction is founded on two fundamental factors: the current environmental conditions and the individual's medical history, particularly their chronic diseases.

The significance of this issue lies not only in its potential to improve emergency medical response and healthcare outcomes, but also in the vital aspect of prevention. Predicting medical emergencies with precision and timeliness can result in faster interventions, with the potential to save lives and cut down on healthcare expenses.

Our basic approach involves the development of an AI model that leverages machine learning to estimate the probability of an individual encountering a sudden medical emergency based on their current health condition and current weather conditions. The model aims to provide an early warning system for both healthcare providers and individuals, enabling more proactive and targeted interventions.

Chapter 2

Scientific Problem

2.1 Problem definition

The problem revolves around predicting sudden and severe medical emergencies, such as heart failure, respiratory failure and cerebral infarction, based on current weather conditions and an individual's chronic diseases. Solving it demands the application of intelligent algorithms due to its intricate nature. The interplay between various factors like current weather conditions and individual medical history is complex and often nonlinear. Intelligent algorithms, particularly machine learning models, excel at processing vast datasets, identifying subtle patterns, and adapting to changing conditions. The complexity, need for precision, and various data characteristics underscore the essential role of intelligent algorithms in effectively addressing this problem.

From a scientific point of view, it is a multi-class classification problem with the added complexity of providing probabilities for each class, allowing for a more nuanced and probabilistic approach to predicting medical emergencies. For this classification, the inputs will be composed of prevailing weather condition: temperature, precipitation, atmospheric pressure, cloudiness, etc. and also the individual's details: age, gender, hypertension, heart and lung chronic diseases.

Its inherent complexity, the potential to save lives, the utilization of vast datasets, and the interdisciplinary fusion of healthcare, meteorology, and data science combine to make it an intriguing challenge.

Chapter 3

State of the art/Related work

1. *"HealthCare Problem: Prediction Stroke Patients" - Saumya Agarwal [Kaggle]*
   Using as features: gender, age, hypertension, heart disease, marriage status, work type, residence type, glucose level, bmi and smoking status, it was developed using Logistic Regression a stroke classifier.

2. *"EDA and Modelling - Heart Attack" - Dorian Voydie [Kaggle]*
   Using different ML approaches, such as Logistic Regression, Support Vector Machine, Gradient Boosting and Random Forest, a heart attack predisposition classifier was build, having around 90% accuracy. The features used for the classification are: age, sex, number of major vessels, chest pain, blood pressure, cholesterol, blood sugar as well as electrocardiographic results.

3. *"Machine Learning Analyzed Weather Conditions as an Effective Means in the Predicting of Acute Coronary Syndrome Prevalence" - Aleksandra Wlodarczyk, Patrycja Molek, Bogdan Bochenek, Agnieszka Wypych, Jadwiga Nessler, Jaroslaw Zalewski*
   Using data colected over a 10 year interval from around 100K patients from the Lesser Poland, a Random Forest Classifier was build to predict the number of Acute Coronary Syndrome (ACSs) based on the weather conditions.

The actual dataset used for this project: The dataset used for this project is an aggregate set of weather and medical conditions from the county Cluj and its neighbouring regions. The dataset contains entries for every day, starting from january 2007 to december 2016. Source of the weather data is Meteomanz.com and the medical data is collected from the County Emergency Hospital Cluj-Napoca.

The analysed features are as follows:

1. *Temperature ($^\circ C$)*
2. *Precipitation (mm)*
3. *Wind speed (km/h)*
4. *Cloudiness (on a scale from 0 to 8)*
5. *Hypertension*
6. *Atrial fibrilation*
7. *Chronic ischemic heart disease*
8. *Valvular insufficiencies*
9. *Chronic obstructive pulmonary disease*

The expected output refers to the probability of falling ill with one of 3 diseases common among dataset entries, that have been picked by relevance. These are the following:

1. *Heart failure*
2. *Respiratory failure*
3. *Cerebral infarction*

Chapter 4

Investigated Approach

4.1 Algorithm

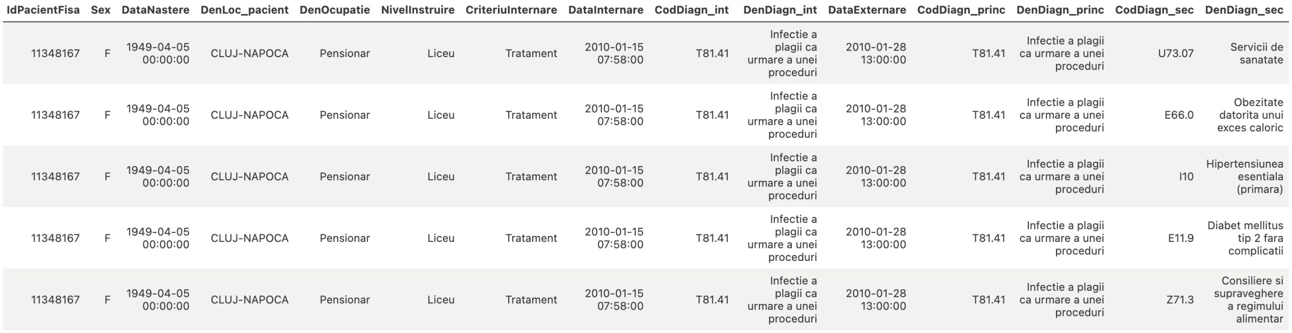
Gradient Boosting technique is an decision tree based ensemble method that leads the machine learning fields for regression and classification. In comparison to Random Forests which use Bagging algorithms, Boosting algorithms take a serial approach. Each model in the series trains upon its predecessor's mistakes, trying to correct them. The final model is a collection of weak learners trained on the residue of strong learners to form the final prediction. The intuition behind GB is that by subtracting the errors predicted by each tree from the actual prediction from the first learner, we can bring the value closer to the ground truth.

Specifically, XGBoost, which stands for Extreme Gradient Boosting follows the same algorithm as GB, but uses advanced regularization techniques to suppress weights, prevent overfitting, and enhance its performance in real-world scenarios. On top of this, the implementation allows the algorithm to cache data and utilize multiple CPU cores for speedy processing.

4.2 Data

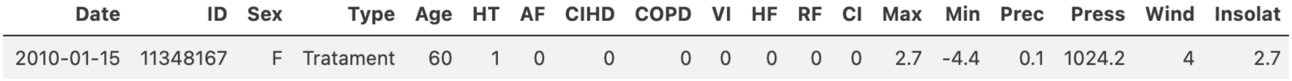
The Initial DataSet consists of 3: Patients, Hospital Admissions and Weather data.

4.2.1 Patients Data



IdPacientFisa	Sex	DataNastere	DenLoc_pacient	DenOcupatie	NivelInstruire	CriteriuInternare	DataInternare	CodDiagn_int	DenDiagn_int	DataExternare	CodDiagn_princ	DenDiagn_princ	CodDiagn_sec	DenDiagn_sec
11348167	F	1949-04-05 00:00:00	CLUJ-NAPOCA	Pensionar	Liceu	Tratament	2010-01-15 07:58:00	T81.41	Infectie a plagii ca urmare a unei proceduri	2010-01-28 13:00:00	T81.41	Infectie a plagii ca urmare a unei proceduri	U73.07	Servicii de sanatate
11348167	F	1949-04-05 00:00:00	CLUJ-NAPOCA	Pensionar	Liceu	Tratament	2010-01-15 07:58:00	T81.41	Infectie a plagii ca urmare a unei proceduri	2010-01-28 13:00:00	T81.41	Infectie a plagii ca urmare a unei proceduri	E66.0	Obezitate datorita unui exces caloric
11348167	F	1949-04-05 00:00:00	CLUJ-NAPOCA	Pensionar	Liceu	Tratament	2010-01-15 07:58:00	T81.41	Infectie a plagii ca urmare a unei proceduri	2010-01-28 13:00:00	T81.41	Infectie a plagii ca urmare a unei proceduri	I10	Hipertensiunea esentiala (primara)
11348167	F	1949-04-05 00:00:00	CLUJ-NAPOCA	Pensionar	Liceu	Tratament	2010-01-15 07:58:00	T81.41	Infectie a plagii ca urmare a unei proceduri	2010-01-28 13:00:00	T81.41	Infectie a plagii ca urmare a unei proceduri	E11.9	Diabet mellitus tip 2 fara complicatii
11348167	F	1949-04-05 00:00:00	CLUJ-NAPOCA	Pensionar	Liceu	Tratament	2010-01-15 07:58:00	T81.41	Infectie a plagii ca urmare a unei proceduri	2010-01-28 13:00:00	T81.41	Infectie a plagii ca urmare a unei proceduri	Z71.3	Consiliere si supraveghere a regimului alimentar

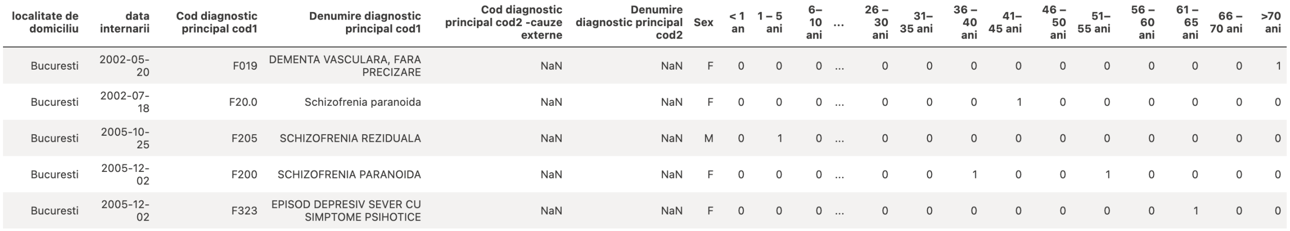
Figure 4.1: Patients Source



Date	ID	Sex	Type	Age	HT	AF	CIHD	COPD	VI	HF	RF	CI	Max	Min	Prec	Press	Wind	Insolat
2010-01-15	11348167	F	Tratament	60	1	0	0	0	0	0	0	0	2.7	-4.4	0.1	1024.2	4	2.7

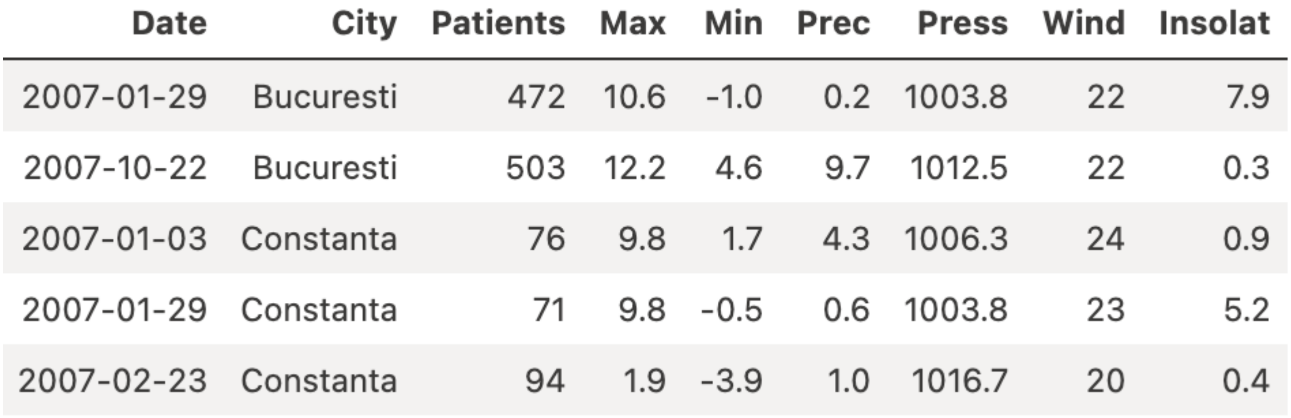
Figure 4.2: Patients Processed

4.2.2 Hospital Data



localitate de domiciliu	data internarii	Cod diagnostic principal cod1	Denumire diagnostic principal cod1	Cod diagnostic principal cod2 -cauze externe	Denumire diagnostic principal cod2	Sex	< 1 an	1 – 5 ani	6– 10 ani	...	26 – 30 ani	31– 35 ani	36 – 40 ani	41– 45 ani	46 – 50 ani	51– 55 ani	56 – 60 ani	61 – 65 ani	66 – 70 ani	>70 ani
Bucuresti	2002-05-20	F019	DEMENTA VASCULARA, FARA PRECIZARE	NaN	NaN	F	0	0	0	...	0	0	0	0	0	0	0	0	0	1
Bucuresti	2002-07-18	F20.0	Schizofrenia paranoida	NaN	NaN	F	0	0	0	...	0	0	0	1	0	0	0	0	0	0
Bucuresti	2005-10-25	F205	SCHIZOFRENIA REZIDUALA	NaN	NaN	M	0	1	0	...	0	0	0	0	0	0	0	0	0	0
Bucuresti	2005-12-02	F200	SCHIZOFRENIA PARANOIDA	NaN	NaN	F	0	0	0	...	0	0	1	0	0	1	0	0	0	0
Bucuresti	2005-12-02	F323	EPISOD DEPRESIV SEVER CU SIMPTOME PSIHOTICE	NaN	NaN	F	0	0	0	...	0	0	0	0	0	0	0	1	0	0

Figure 4.3: Hospital Source



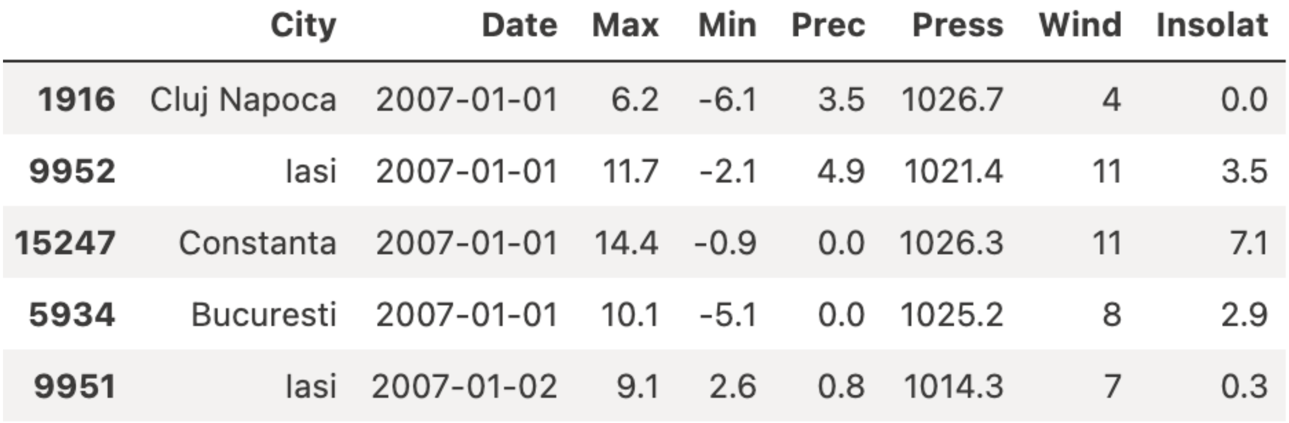
Date	City	Patients	Max	Min	Prec	Press	Wind	Insolat
2007-01-29	Bucuresti	472	10.6	-1.0	0.2	1003.8	22	7.9
2007-10-22	Bucuresti	503	12.2	4.6	9.7	1012.5	22	0.3
2007-01-03	Constanta	76	9.8	1.7	4.3	1006.3	24	0.9
2007-01-29	Constanta	71	9.8	-0.5	0.6	1003.8	23	5.2
2007-02-23	Constanta	94	1.9	-3.9	1.0	1016.7	20	0.4

Figure 4.4: Hospital Processed

4.2.3 Weather Data

City	Date	AveT	MaxT	MinT	Prec	Press	Wind dir	Wind sp	Cloud c.
Cluj Napoca	31/01/2013	3.80	7.80	-0.30	3.8	1012.6 Hpa	297(NW)	8	4/8
Cluj Napoca	30/01/2013	-0.70	1.10	-2.40	4.8	1017.1 Hpa	56(NE)	5	8/8
Cluj Napoca	29/01/2013	-1.80	1.00	-4.60	3.5	1021.2 Hpa	53(NE)	4	8/8
Cluj Napoca	28/01/2013	-3.10	-0.70	-5.50	0.6	1017.8 Hpa	47(NE)	4	7/8
Cluj Napoca	27/01/2013	-4.80	-1.40	-8.10	2.2	1017.4 Hpa	23(NE)	5	8/8
Cluj Napoca	26/01/2013	-4.20	-1.40	-7.00	0.0	1016.0 Hpa	62(NE)	5	6/8
Cluj Napoca	25/01/2013	-0.80	-0.10	-1.40	2.0	1008.7 Hpa	63(Âº)NE)	4	8/8
Cluj Napoca	24/01/2013	-0.10	1.80	-1.90	5.4	1009.4 Hpa	75(E)	5	8/8
Cluj Napoca	23/01/2013	0.80	2.20	-0.60	0.5	1008.5 Hpa	220(SW)	4	8/8
Cluj Napoca	22/01/2013	4.00	7.40	0.70	0.7	1000.9 Hpa	204(SW)	4	7/8

Table 4.1: Weather Source



	City	Date	Max	Min	Prec	Press	Wind	Insolat
1916	Cluj Napoca	2007-01-01	6.2	-6.1	3.5	1026.7	4	0.0
9952	Iasi	2007-01-01	11.7	-2.1	4.9	1021.4	11	3.5
15247	Constanta	2007-01-01	14.4	-0.9	0.0	1026.3	11	7.1
5934	Bucuresti	2007-01-01	10.1	-5.1	0.0	1025.2	8	2.9
9951	Iasi	2007-01-02	9.1	2.6	0.8	1014.3	7	0.3

Figure 4.5: Weather Processed

4.3 Metrics

4.3.1 Mobile Model

The mobile model is a multilabel classifier designed to predict the risk of individuals having life threatening conditions. To assess its performance, precision is chosen as the primary metric. Precision is particularly relevant in this context as it measures the accuracy of positive predictions, providing insights into the model's ability to correctly identify cases at risk.

4.3.2 Hospital Model

The hospital model aims to predict the number of hospital admissions, a regression task. For evaluating the performance of this model, Mean Absolute Error (MAE) is selected as the metric. MAE quantifies the average absolute difference between the predicted and actual number of admissions, offering a clear measure of the model's accuracy in predicting hospitalization numbers.

4.4 Model Selection and Hyperparameter Tuning

4.4.1 Mobile Model

To mitigate the class imbalance, a NearMiss undersampling technique was employed. This approach equalizes the representation of each target class by intelligently selecting instances from the majority class, ensuring a more balanced training dataset.

XGBoost was chosen as the primary modeling tool. Its ability to handle imbalanced datasets, along with its ensemble learning approach, contributed significantly to improving the model's performance. The combination of undersampling and XGBoost resulted in a more robust and accurate disease risk prediction model, capable of providing reliable assessments across most classes.

4.4.2 Hospital Model

In the pursuit of an optimal model for the mobile application's disease risk prediction, a thorough exploration was conducted with multiple algorithms, including MLPRegressor, Decision Tree, Random Forest, and XGBoost. Through empirical evaluation, XGBoost emerged as the most effective choice, exhibiting superior performance in terms of precision.

Hyperparameter tuning was performed to further enhance the XGBoost model's effectiveness. The fine-tuning process involved systematically adjusting key parameters to achieve the best balance between predictive accuracy and generalization. The final hyperparameter configuration for the XGBoost model in the mobile application's disease risk prediction is as follows:

4.4.3 Experiments

In the realm of hospital data preprocessing, a comprehensive exploration of multiple parameters has been conducted. Each stage in the data processing pipeline underwent rigorous testing through cross-validation, leading to the identification of optimal parameters. The process involved several key components:

DateTransformer

- **day**: Transforms the date into the day of the week or identifies whether it falls on the weekend.
- **month**: Introduces a new column representing the month.
- **holiday**: Indicates whether a given date falls within a holiday period.

CityTransformer

- **type of city encoder**: Utilizes either one-hot encoding or label encoding.
- **coordinates**: Incorporates geographical coordinates for each city.
- **population**: Appends population data for each city.

BasicExperiments

- **avg**: Replaces correlated min and max values with the average.
- **scaler**: Scales continuous columns for enhanced consistency.

Following an exhaustive cross-validation of all parameter combinations, the mean absolute error was aggregated, leading to the determination of optimal settings:

- **day**: dayofweek
- **month**: True
- **holiday**: True
- **type of city encoder**: ohe
- **coordinates**: True
- **population**: True
- **avg**: False
- **scaler**: False

Subsequently, a meticulous cross-validation for the XGBoost model involved experimentation with various parameters, such as learning rate, max depth, no estimators, subsample, min child weight, gamma, lambda, alpha. Following an exhaustive cross-validation of all parameter combinations, the mean absolute error was aggregated, leading to the determination of optimal settings: XGBoost Parameters

- **learning rate**: 0.2
- **max depth**: 5
- **n_estimators**: 50
- **subsample**: 1
- **min child weight**: 5
- **alpha**: 0.2
- **gamma**: 0
- **lambda**: 0

These parameter configurations were derived by aggregating mean absolute errors across all combinations, ensuring a robust and optimized preprocessing and modeling approach for the hospital data.

4.4.3.1 Data Transformations

MAE	encoding	coordinates	population	day	holiday	month	avg	scaler
16.36	ohe	False	True	dayofweek	True	True	False	False
16.39	le	True	False	dayofweek	True	True	False	False
16.39	le	False	True	dayofweek	True	True	False	False
16.40	le	True	False	dayofweek	True	True	True	False
16.40	le	False	False	dayofweek	True	True	True	False
16.41	le	False	True	dayofweek	True	True	True	False
16.43	le	True	True	dayofweek	True	True	False	False
16.44	ohe	False	True	dayofweek	True	True	True	False
16.44	le	False	False	dayofweek	True	True	False	False
16.45	le	True	True	dayofweek	True	True	True	False

Table 4.2: Top 10 Parameter Combination

Parameter	Value	MAE
avg	False	20.57
	True	20.63
coordinates	False	20.66
	True	20.53
day	dayofweek	17.60
	weekend	23.60
encoding	le	20.37
	ohe	20.83
holiday	False	20.99
	True	20.21
month	False	20.71
	True	20.49
population	False	20.80
	True	20.39
scaler	False	20.30
	True	20.90

Table 4.3: Overall Parameter Performance

4.4.3.2 Model Tuning

MAE	Learning Rate	Max Depth	No Estimators	Subsample	Alpha	Lambda	Gamma	Min Child Weight
15.49	0.20	5	50	1.00	0.00	0.20	0.10	5
15.49	0.20	5	50	1.00	0.00	0.20	0.00	5
15.49	0.20	5	50	1.00	0.00	0.20	0.20	5
15.49	0.20	5	50	1.00	0.20	0.20	0.00	3
15.49	0.20	5	50	1.00	0.20	0.20	0.20	3
15.49	0.20	5	50	1.00	0.20	0.20	0.10	3
15.50	0.20	5	50	1.00	0.00	0.20	0.00	3
15.50	0.20	5	50	1.00	0.10	0.20	0.10	5
15.50	0.20	5	50	1.00	0.10	0.20	0.20	5
15.50	0.20	5	50	1.00	0.00	0.20	0.20	1

Table 4.4: Top 10 Parameter Combination

Parameter	Value	MAE
Alpha	0.0	40.39
	0.1	40.38
	0.2	40.38
Gamma	0.0	40.38
	0.1	40.38
	0.2	40.38
Lambda	0.0	40.37
	0.1	40.38
	0.2	40.40
Learning Rate	0.01	52.56
	0.2	16.04
Max Depth	3.0	40.43
	5.0	40.31
	7.0	40.42
Min Child Weight	1.0	40.40
	3.0	40.39
	5.0	40.36
N Estimators	50.0	61.28
	100.0	39.47
	200.0	20.40
Subsample	0.8	40.41
	1.0	40.36

Table 4.5: Overall Parameter Performance

4.5 Results

Class	**Precision**	**Recall**	**F1-Score**
Heart Failure	0.75	0.73	0.74
Respiratory Failure	0.73	0.31	0.44
Cerebral Infarction	0.52	0.11	0.18
Weighted Avg	0.71	0.53	0.57

Table 4.6: Mobile Model Evaluation

Metric	**Value**
Training MAE	13.44
Test MAE	14.99

Table 4.7: Hospital Model Performance Metrics

Chapter 5

Application (Study case)

5.1 App's description and the main functionalities

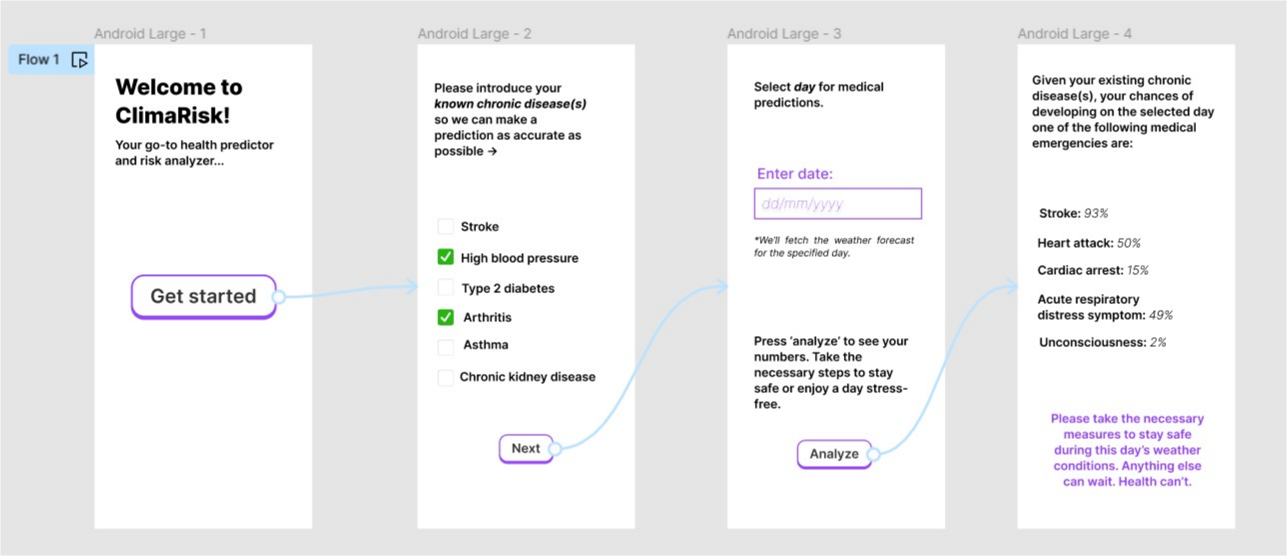
This application caters to two distinct user groups: medical professionals and the general public. The version designed for medical staff empowers users to forecast the influx of patients with various medical emergencies on a specified day. These predictions are intricately linked to the prevailing weather conditions. To initiate this prediction process, users select a day in the near future, for which existing weather forecasts are available, and click the 'estimate' button. The application seamlessly retrieves weather data and feeds it into the machine learning model, which then generates the requested predictions.

Conversely, the version developed for civilians, regardless of their professional background, predicts the likelihood of the user falling ill on a specific day. This prediction takes into account both the forecasted weather conditions for that day and the individual's info and chronic health conditions. The probability of developing severe symptoms related to a disease translates to an incoming patient count in the medical version of the app. To generate these predictions, users simply check their chronic diseases from a predefined list. The application automatically retrieves weather parameters in the background and combines them with the user's health profile. The integrated data is then passed to the model, which predicts the user's predisposition to certain symptoms on that day.

5.2 Implementation

In our implementation, the preferred programming language is Python, while the frameworks we are considering for our solution are Scikit-Learn, Tensorflow, PySpark. Adversely, for processing, analyzing and storing big data, we will be using different plaform from MSFabric such as Synapse

Data Engineering, Synapse Data Science and Power BI. To facilitate this development process, we leverage the benefits of containerization using Docker, enabling us to create consistent and reproducible environments for our machine learning projects. Additionally, we enhance our development workflow by integrating Visual Studio Code, Jupyter notebooks and Azure Data Studio.



Welcome to ClimaRisk!

Your go-to health predictor and risk analyzer...

Please introduce your ***known chronic disease(s)*** so we can make a prediction as accurate as possible →

□ Stroke
☑ High blood pressure
□ Type 2 diabetes
☑ Arthritis
□ Asthma
□ Chronic kidney disease

Select ***day*** for medical predictions.

Enter date:

dd/mm/yyyy

**We'll fetch the weather forecast for the specified day.*

Press 'analyze' to see your numbers. Take the necessary steps to stay safe or enjoy a day stress-free.

Given your existing chronic disease(s), your chances of developing on the selected day one of the following medical emergencies are:

Stroke: *93%*

Heart attack: *50%*

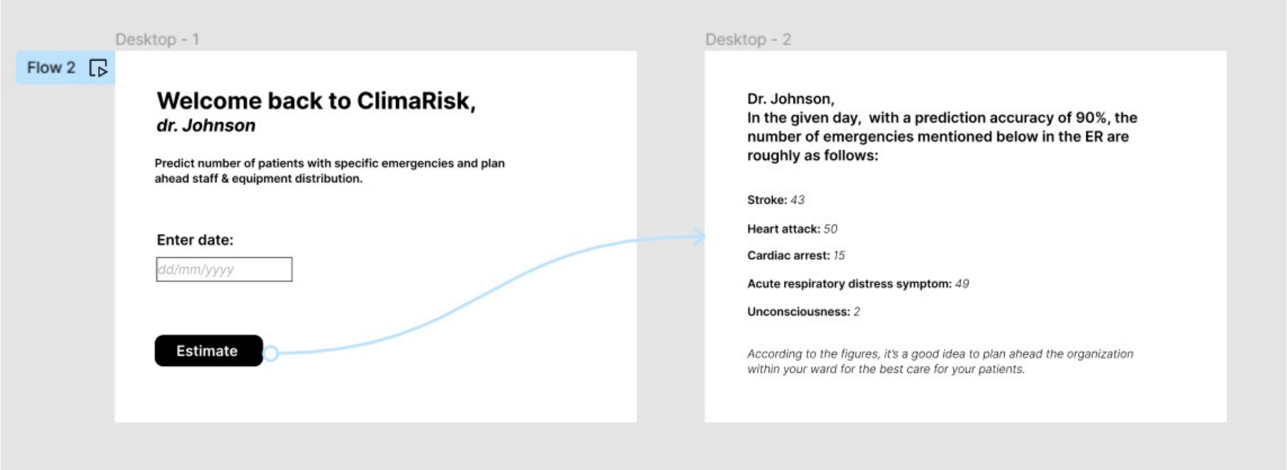
Cardiac arrest: *15%*

Acute respiratory distress symptom: *49%*

Unconsciousness: *2%*

Please take the necessary measures to stay safe during this day's weather conditions. Anything else can wait. Health can't.

Figure 5.1: Mobile Flow



Welcome back to ClimaRisk,
dr. Johnson

Predict number of patients with specific emergencies and plan ahead staff & equipment distribution.

Enter date:

dd/mm/yyyy

Dr. Johnson,
In the given day, with a prediction accuracy of 90%, the number of emergencies mentioned below in the ER are roughly as follows:

Stroke: *43*

Heart attack: *50*

Cardiac arrest: *15*

Acute respiratory distress symptom: *49*

Unconsciousness: *2*

According to the figures, it's a good idea to plan ahead the organization within your ward for the best care for your patients.

Figure 5.2: Desktop Flow