UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Corso di Laurea Triennale in

Sicurezza dei Sistemi e delle Reti Informatiche

CORRELATION STUDY BETWEEN INTRINSIC CAPACITY AND PREDICTIVE ANALYTICS IN THE SMART BEAR PLATFORM

Supervisor:

Prof. Paolo CERAVOLO

Co-supervisor:

Dott.ssa Samira Maghool

Candidate:

Jacopo SCUDIERI

ID: 986645

Academic Year 2023/2024

To my softness, family and close friends,

for their endless encouragement and belief in me.

Table of Contents

1 Introduction

1.1 Overview of the Thesis

Individuals aged 65 and older represent a rapidly increasing demographic in Europe. Common health issues among the elderly include hearing loss, cardiovascular diseases, cognitive disorders, mental health challenges, and balance problems [1]. These conditions contribute to a decline in quality of life (characterized by inactivity, dependency, and loneliness) and pose significant health risks (such as physical injuries, disabilities, and hospitalization). Furthermore, managing these conditions imposes a heavy burden on healthcare systems, leading to high costs and deficiencies in quality, safety, and access.

Currently, with the advancement of digital technology, we are seeing a proliferation of smart healthcare devices. These devices utilize minimally invasive technology to continuously monitor patients and provide real-time health status updates. In this context, the European project SMART BEAR Horizon 2020 seeks to design and develop a platform that integrates data from smart monitoring devices with medical assessments, creating a unified intelligent medical support system. The objective is to deliver a secure and accessible service that respects the privacy of elderly individuals, promotes autonomy, and encourages a healthy lifestyle, thereby enhancing healthcare efficiency and reducing resource waste [2]. Additionally, the collected data not only offer a comprehensive view of the patient's current health status but can also be used for preventive medicine.

Unlike traditional medicine, preventive medicine focuses on early intervention to prevent or mitigate the onset of diseases by identifying risk factors when recovery chances are highest. Clinical data can be analyzed using advanced machine learning and artificial intelligence algorithms to assess the patient's history and current condition and predict future health outcomes, providing real-time feedback and facilitating personalized interventions.

Developing a predictive platform involves several challenges, such as the heterogeneity of collected data, which cannot be directly utilized. Despite the availability of various learning

tools today, it is crucial to identify the appropriate predictive strategy for each specific case study, particularly in the medical field, where incorrect predictions can have serious consequences on patient health. The goal is to establish an automated workflow within the platform that employs various evaluation and comparison tools to create the most effective predictive model for the specific issue at hand.

Following an initial study phase, patients' clinical data are extracted and subjected to pre-processing, where they are prepared for accurate interpretation by predictive analytics. This includes a cleaning phase to eliminate or correct erroneous data, followed by normalization and imputation phases to address missing values, a common issue in medical data. The processed data are then used to train different machine learning algorithms, with a subsequent model validation phase using k-fold cross validation, a strategy more suitable than traditional external validation for datasets with limited sizes and imbalanced classes.

1.2 Objectives and Scope

As previously said, this thesis is based on the European project Smart Bear, which leverages smart devices to monitor the clinical data of elderly patients in real time. The collected data is transmitted daily to a central cloud platform, where it is used for analysis and prediction of health risks. The focus of this thesis is double: first, it explores predictive analytics to assess cardiovascular risk based on patient data such as age, gender, blood pressure, BMI, smoking status, diabetes, cholesterol levels and more. Second, it investigates the concept of Intrinsic Capacity, a key measure of overall functional health in older adults, and its potential correlation with cardiovascular risk. By analyzing both predictive models and health capacity indicators, this thesis aims to contribute to a deeper understanding of how technology and data analytics can improve elderly care.

Hence, the primary objective of this thesis is to develop and evaluate a predictive model for cardiovascular risk using logistic regression and z-fold cross validation, based on data collected from the Smart Bear platform. The secondary goal is to study the correlation

between Intrinsic Capacity, previously established through a separate analysis, and the cardiovascular risk predictions.

1.3 Structure of the Thesis

To best illustrate the work carried out, this thesis is structured as follows:

- **Chapter 1**: Provides an Overview of the thesis.
- **Chapter 2**: Gives a Background Overview.
- **Chapter 3**: Discusses the Importance of Data in Healthcare.
- **Chapter 4**: Introduces The SMART BEAR Project
- **Chapter 5**: Explains the work done on cardiovascular risk prediction.
- **Chapter 6**: Illustrates the correlation between IC and predictive analytics

2 Background Review

2.1 Electronic Health Records (EHR) Systems

Electronic Health Record (EHR) systems enable the systematic collection of patient data in digital format through electronic devices and information systems. The adoption of EHRs brings numerous organizational and clinical benefits. Clinical decision support systems, computerized order entry systems, and health information exchanges can operate more efficiently with EHRs. EHRs also contribute to societal benefits by reducing medical errors, enhancing research capabilities, and improving information accessibility for patients and clinical staff.

The rapid aging of the population presents a global challenge for healthcare systems, prompting lawmakers in the EU, US, and other regions to establish guidelines and standards for EHR implementation to boost efficiency. This trend coincides with a rising interest in mobile health (mHealth) monitoring systems. Advances in hospital infrastructure and mHealth are paving the way for smart health ecosystems worldwide, leveraging data from mobile, wearable, and IoT devices both within and outside hospitals. These smart health ecosystems facilitate continuous data collection from daily life, which is analyzed to provide the evidence necessary for personalized interventions.

2.2 The Role of IoT in Smart Health

IoT devices play a crucial role in the Smart Health and EHR systems environment by enabling continuous and real-time monitoring of patients' health. These devices, which include wearables, sensors, and connected medical equipment, collect a vast array of health data such as vital signs, physical activity, and environmental factors. This data is seamlessly integrated into EHR systems, providing healthcare professionals with comprehensive and up-to-date information about patients' health status. The real-time data collection facilitated by IoT devices not only enhances the accuracy of health records but also allows for prompt

detection and response to potential health issues, thereby improving patient outcomes and reducing the risk of complications.

In the context of the SMART BEAR project, IoT devices are integral to creating a smart health ecosystem that supports personalized healthcare. By continuously gathering data from patients' everyday activities, these devices help in generating valuable insights that inform clinical decisions and personalized interventions. The integration of IoT data with EHR systems enables a holistic view of a patient's health, aiding in preventive care and chronic disease management.

2.3 Continuous Learning in Health Data Management

Continuous learning in health data management is essential for improving the accuracy and efficiency of healthcare systems. In the SMART BEAR project, continuous learning involves the iterative process of refining predictive models and algorithms using the latest data collected from patients. This approach allows the system to adapt to new information and evolving patterns in patient health, ensuring that the insights generated are current and relevant. By constantly updating the analytical models, healthcare providers can enhance their understanding of complex health conditions, leading to more precise and timely interventions.

The integration of continuous learning in health data management facilitates the development of more robust and personalized healthcare solutions. As new data is continuously fed into the system, machine learning algorithms can detect subtle changes and trends that might not be apparent through traditional analysis. Additionally, continuous learning helps in addressing data quality issues, such as missing or inconsistent data, by improving the system's ability to handle and learn from diverse datasets. This ongoing improvement cycle ultimately contributes to better patient outcomes and more efficient resource utilization in healthcare environments.

3 The Importance of Data in Healthcare

Clinical data, which consists of information contained in patient medical records, is essential for analysis and decision support in the healthcare sector. This collection of information allows for a comprehensive view of patients' conditions, personalization of care, and improved communication with doctors, thereby increasing the effectiveness of interventions and reducing costs (Figure 1).

Collecting observations from a sufficiently large and diverse number of patients enables the creation of comparable groups, in which prognostic characteristics are similar. These comparable data can be used to develop predictive models based on machine learning algorithms.

However, various issues can arise during the collection of clinical data that may compromise the effectiveness of predictive models. Therefore, it is crucial that the data provided to the machine learning algorithm for building the model be of high quality. Noisy, incomplete, inaccurate, or unclean data, the presence of missing values, and biases due to an imbalanced dataset can undermine the accuracy of predictions.

Data is a crucial element, and a lack of quality or quantity can compromise the functionality of the entire learning process. Therefore, data must be accurate, up to date, error-free, non-contradictory, complete, and relevant to the problem at hand.

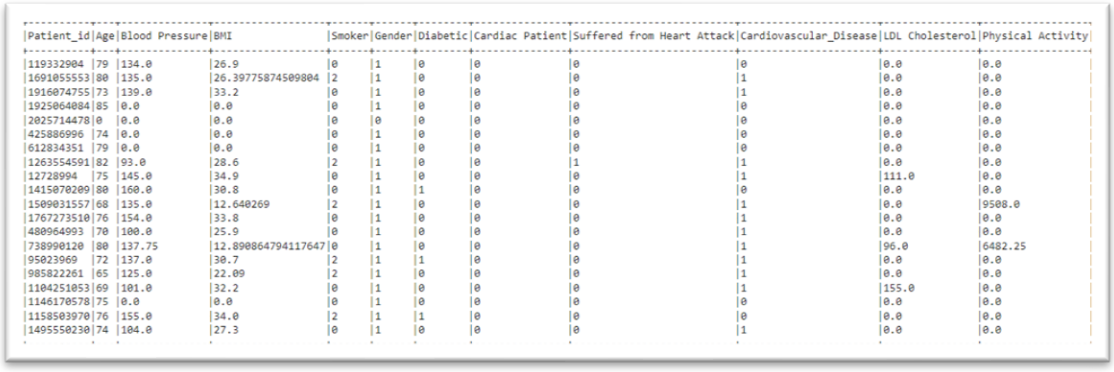
3.1 Quantitative and Categorical Data

Data can be classified based on the measurement method used, primarily divided into quantitative, derived from numerical evaluations, and categorical, identifiable through specific labels. Each type of data requires a specific approach for its management.

Quantitative data, such as height or weight, are expressed in numerical terms and can be either counted or measured. This type of data is essential for analysis and research, as it allows for calculations and statistical interpretations.

Categorical data, such as gender, marital status, and education level, emphasize specific qualities or attributes of an object or phenomenon. These data are descriptive in nature and are not suitable for direct mathematical calculations.

In the clinical context, we can further divide the data into time series data and static data. For example, values like blood pressure or heart rate are linked to specific moments in time, making them sequential and tied to the time of measurement. Conversely, data such as gender are typically unchanging over time. When working with time series, it is necessary to ensure that the data are relevant and up to date.



Patient_id	Age	Blood Pressure	BMI	Smoker	Gender	Diabetic	Cardiac Patient	Suffered from Heart Attack	Cardiovascular_Disease	LDL Cholesterol	Physical Activity
119332904	79	134.0	26.9	0	1	0	0	0	0	0.0	0.0
1691055553	80	135.0	26.39775874509804	2	1	0	0	0	1	0.0	0.0
1916074755	73	139.0	33.2	0	1	0	0	0	1	0.0	0.0
1925064084	85	0.0	0.0	0	1	0	0	0	0	0.0	0.0
2025714478	0	0.0	0.0	0	0	0	0	0	0	0.0	0.0
425886996	74	0.0	0.0	0	1	0	0	0	0	0.0	0.0
612834351	79	0.0	0.0	0	1	0	0	0	0	0.0	0.0
1263554591	82	93.0	28.6	2	1	0	0	1	1	0.0	0.0
12728994	75	145.0	34.9	0	1	0	0	0	1	111.0	0.0
1415070209	80	160.0	30.8	0	1	1	0	0	0	0.0	0.0
1509031557	68	135.0	12.640269	2	1	0	0	0	1	0.0	9508.0
1767273510	76	154.0	33.8	0	1	0	0	0	1	0.0	0.0
480964993	70	100.0	25.9	0	1	0	0	0	1	0.0	0.0
738990120	80	137.75	12.890864794117647	0	1	0	0	0	1	96.0	6482.25
95023969	72	137.0	30.7	2	1	1	0	0	1	0.0	0.0
985822261	65	125.0	22.09	2	1	0	0	0	1	0.0	0.0
1104251053	69	101.0	32.2	0	1	0	0	0	1	155.0	0.0
1146170578	75	0.0	0.0	0	1	0	0	0	0	0.0	0.0
1158503970	76	155.0	34.0	2	1	1	0	0	0	0.0	0.0
1495550230	74	104.0	27.3	0	1	0	0	0	1	0.0	0.0

Figure 1: Example of data in SMART BEAR

3.2 Missing Data

In clinical monitoring, the presence of missing data (Figure 1) [3] is a very common phenomenon. The causes can vary: from patients abandoning the course of treatment, to human errors during data collection, to hardware or software problems in data acquisition. In these circumstances, performing prediction activities becomes impossible because the learning model would not be able to determine the relationships between input and output to predict.

The imputation of missing data allows for the reconstruction of the entire behavior of the patient even in the presence of missing values, providing healthcare personnel with a complete view of the individual's state throughout their course of treatment. There are various

strategies to address the absence of information, each aiming to minimize the impact of missing data on analyses and predictions.

The absence of data can be divided into three main types [4]:

- MCAR (Missing Completely at Random)
    - This occurs when the absence of data is not related to either the present or missing data; the omission is entirely arbitrary and not connected to the information itself. This makes analysis less problematic, except for the pure loss of information.
    - Example: Blood pressure readings might be missing randomly due to user error or dead batteries in the device, without any relation to other variables.
- MAR (Missing at Random)
    - The absence of data can depend on the data that is already present. This category allows for the estimation of missing data based on existing information.
    - Example: In a survey, men might be more likely to report their weight compared to women, who might omit this information. This suggests a relationship between the propensity to respond and the gender of the subject, but not with the actual weight.
- MNAR (Missing Not at Random)
    - The omission of data is related to the missing value itself or other unrecorded information. This makes analysis more complex, as the absences can alter the data distribution.
    - Example: In an income study, people with extremely high or low incomes might avoid reporting their earnings for privacy or shame reasons, compared to those with average incomes.

The most immediate solution for dealing with missing data would be to delete them when they are incomplete. However, this is not the ideal strategy. In the medical field, the amount of missing data is often significant, and their deletion would significantly reduce the volume of information available for the learning model, compromising the accuracy of the training.

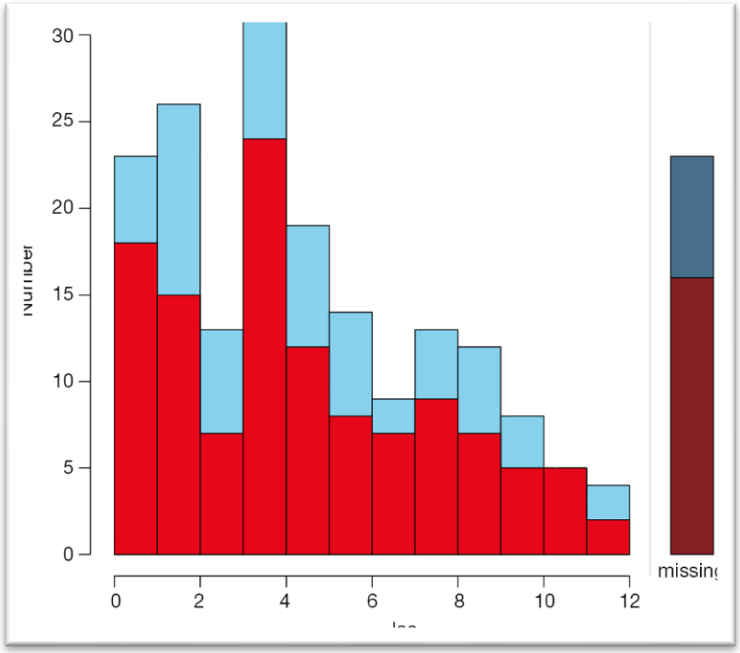


Figure 2: Missing Data Example

3.3 Data Normalization

Normalization is a technique often used during the data preparation phase for machine learning. In a dataset, it is common to have numerical data that vary significantly because they use different scales. This difference in scales prevents a direct comparison between variables and can create issues during model building. For instance, some algorithms might be influenced by data with larger values, thereby distorting the results.

The goal of normalization is to standardize the values of numerical columns, assigning all data a common scale without distorting the ranges or losing information. This process helps to avoid imbalances by generating new values that preserve the original proportions, ensuring that each variable retains the same importance within the model.

One of the most commonly used indices is the Z-score, which transforms the data to have a distribution with a mean of 0 and a standard deviation of 1:

$$Z = \frac{x - mean(X)}{stdev(X)}$$

For each distinct column in the dataset to be normalized, the mean and standard deviation are calculated. The values obtained are then used to apply the transformation to every single observation within the specific column.

3.4 Data Imputation

The better alternative consists of using data imputation techniques, which involve the process of replacing missing data with estimated values. Imputation allows for the reconstruction of the entire behavior of the patient even in the presence of missing data, providing healthcare personnel with a complete view of the individual's condition throughout their course of treatment.

There are various strategies to handle the absence of information. One of the most common imputation techniques is based on the use of numerical indices such as the mean or median. This method calculates the mean (or median, or other statistical indices) of the values present in a column and replaces the missing data with the obtained index. However, this technique does not consider the correlations between the data, making it less suitable for medical prediction. Another solution is the frequency method, which replaces the missing values with the most frequent ones within each column. This method is often used for categorical data, but it also does not take into account the correlations between columns and might introduce a certain degree of distortion into the data. These techniques fall into the category of "Single Imputation," where the missing data are replaced with a value determined by a specific rule.

There are more advanced methods, known as "Multiple Imputation," for handling missing values in a dataset. Unlike simpler techniques that replace each missing value with a single estimated value, multiple imputation (Figure 2) [5] fills in the missing data with several possible values based on the information already present in the dataset. This approach is not only more sophisticated than simply deleting missing elements, but it also provides estimates that are less susceptible to bias [6].

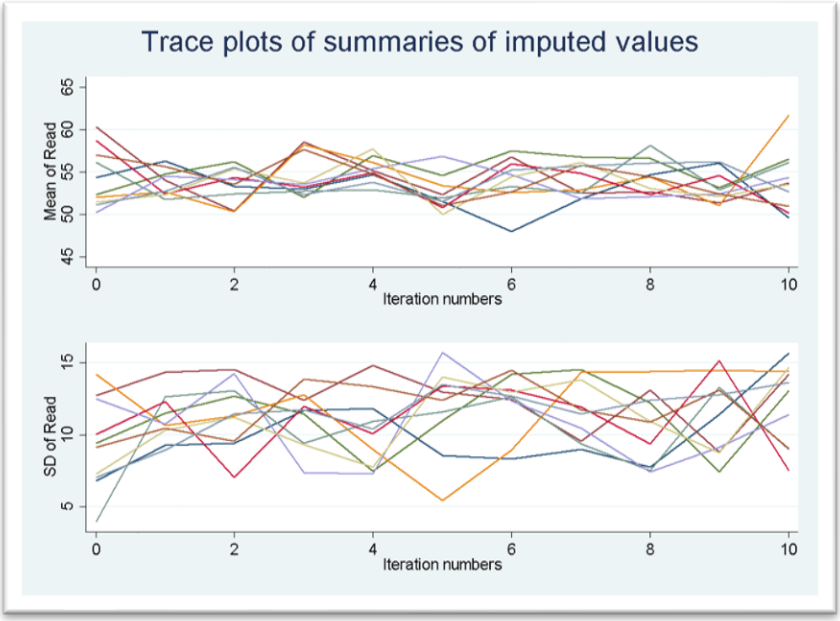


Figure 3: Multiple Imputation

To address missing data in analyses, it is essential to choose an imputation method that reflects the underlying mechanism of the missingness. For missing data that are completely random (MCAR), single imputation methods are often sufficient, as this type of missingness is not influenced by observable or unobservable variables. For missing data that are not random (MNAR), where the absence might be related to the values themselves, a single imputation method can still be used, but in cases of missing data that are random (MAR), multiple imputation methods are preferred. These methods allow for the use of relationships between observed variables to impute missing data more reliably [4].

In the imputation process, linear correlation plays a crucial role (Figure 3). This index, which ranges from -1 to 1, indicates the existence of a linear relationship between variables, either positive or negative. This approach is particularly useful for imputing missing values when there is known data available. For example, if the correlation between two variables is 1, it means that it is almost certain to derive the missing value from the other quantities. However, it is essential to note that this method is not effective for identifying nonlinear relationships

and cannot be applied to categorical data. Therefore, it is necessary to convert all categorical data into numerical format before proceeding with the imputation process.

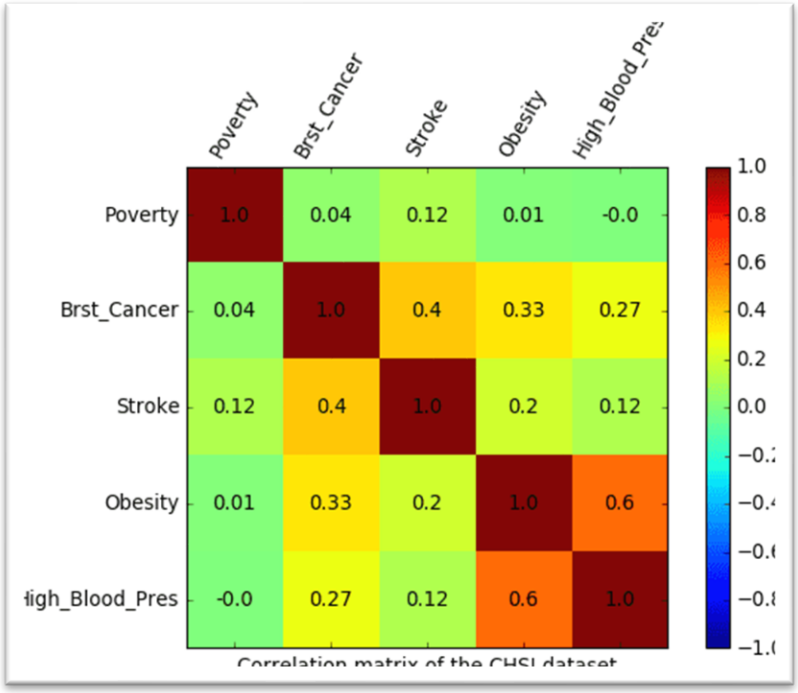


Figure 4: Correlation Matrix Example

4 The SMART BEAR Project

The SMART BEAR (SB) project, funded by the European Commission under the Horizon 2020 program, aims to design and develop an innovative platform that integrates cutting-edge devices, targeted at older adults, to provide medical support based on established studies, promoting healthy and independent living at home [7].

Figure 5: The SMART BEAR Project

The aim of the SMART-BEAR platform is to integrate heterogeneous sensors, assistive medical and mobile devices to enable the continuous data collection from the everyday life of the elderly, which will be analyzed to obtain the evidence needed in order to offer personalized interventions promoting their healthy and independent living. The platform will also be connected to hospital and other health care service systems to obtain data of the end users (e.g., medical history) that will need to be considered in making decisions for interventions.

SMART-BEAR will leverage big data analytics and learning capabilities, allowing for large scale analysis of the above mentioned collected data, to generate the evidence required for making decisions about personalized interventions. Privacy-preserving and secure by design data handling capabilities, covering data at rest, in processing, and in transit, will cover comprehensively all the components and connections utilized by the SMART-BEAR platform.

The SMART-BEAR platform will be tested and validated through six large scale pilots, spanning six different countries and 5.100 individuals: France, Greece, Italy, Spain, Romania and Portugal. The pilots will enable the evaluation of the platform in the context of healthcare

service delivery by private and public providers at regional, state and EU level, and demonstrate its efficacy, extensibility, sustainability, and cost effectiveness for the individual and the healthcare system.

The 27 partners of the consortium started working together on September 2019 to the achievement of this large-scale European project [8].

Figure 6: SMART BEAR Project Partners[8]

4.1 Project Overview and Objectives

The aging population represents a significant challenge in today's society. According to R. Suzman and J. Beard [9], 151 million people in Europe will be over 65 years old by 2060, with a particularly rapid increase in the number of people over 80. The physiological and

progressive decline due to aging, affecting physical abilities (osteoporosis, frailty) and cognitive abilities (memory problems and visuospatial disorders), leads to a reduction in independence and a necessary need for assistance. These conditions, in turn, cause a deterioration in mood and social participation of the individual. Hearing loss, cardiovascular diseases, cognitive disorders, mental health issues, and balance disorders are the most common health problems among the elderly population. Furthermore, managing these issues is costly for healthcare institutions, with high and rising expenses, as well as gaps in quality, safety, and access [2].

The World Health Organization (WHO) has recently made the promotion of "healthy aging" a priority [10], understood as the process of promoting and maintaining an individual's functional capacity. This includes managing one's basic needs, making decisions, staying active, building and maintaining a social life, and contributing to society [2].

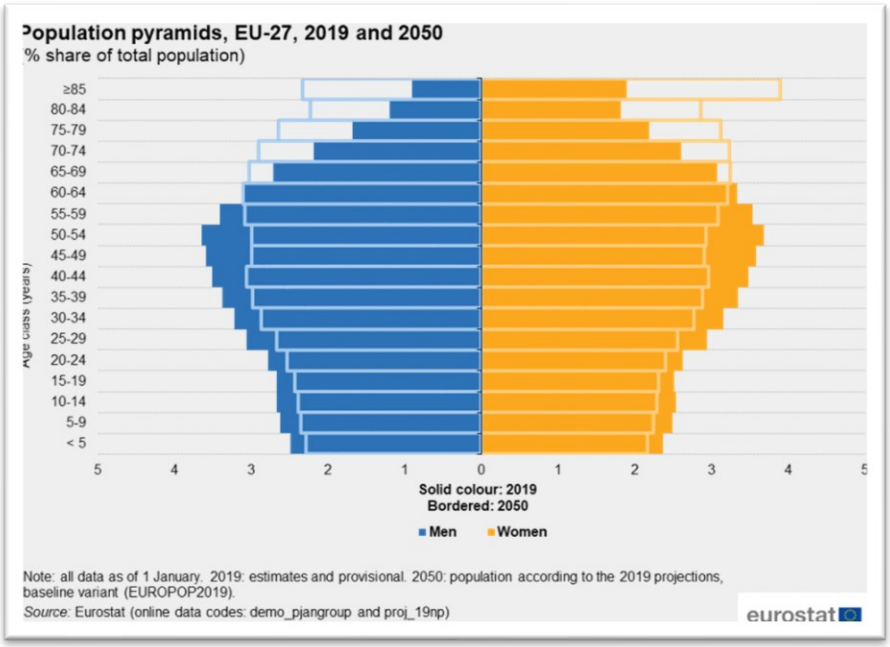


Figure 7: Forecast of European population distribution in 2050 [11]

4.2 Infrastructure and Components

The SMART BEAR platform includes a section dedicated to collecting data from monitoring devices used by patients, such as wearables or smart home technologies. The collected data is integrated with the clinical and demographic information of patients through interaction with the hospital system. This process defines the electronic health record (EHR), which is a systematic collection of individual or population medical information in digital format.

The data is then processed by a computing infrastructure, the core of the platform, which analyzes it and applies predictive metrics. Finally, the results are made available to medical staff through a monitoring platform, which alerts users to the presence of suspicious medical conditions, thus providing continuous feedback.

Figure 8: SMART BEAR Technologies

The selection of technologies for the computing engine of the platform was particularly important for executing data science and machine learning activities. The central component is Apache Spark, a unified analytics engine designed for large-scale data processing. Specifically, Pyspark was used, which is an interface of Apache Spark for Python, allowing applications to be written using Python APIs and providing a shell for interactive data analysis in a distributed environment. With Pyspark, the clinical data of patients is analyzed and manipulated, performing pre-processing tasks [12].

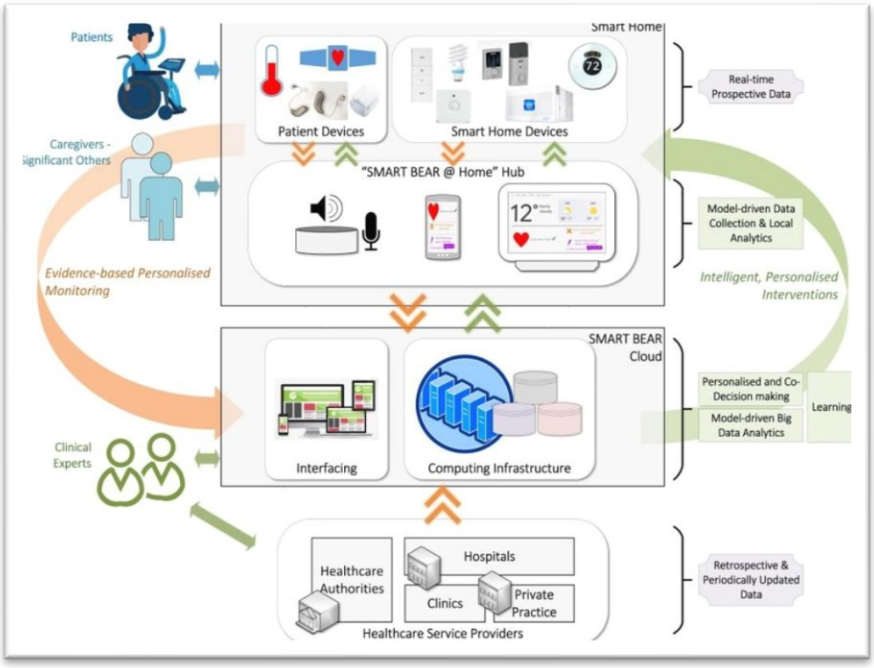


Figure 9: SMART BEAR Infrastructure

As discussed further along, the learning algorithms were manually written in Python using the ML libraries of Pyspark and scikit-learn. Finally, an external visualization library, Matplotlib, was used to display graphs, ROC curves, and Precision-Recall curves.

4.3 Data in SMART BEAR

The SMART BEAR project database is structured into 8 tables: Components, Observations, Encounters, Conditions, Patients, Medications, Research Subjects, Diagnostic Reports, and Questionnaire Responses.

The data used in this thesis is contained in 4 main tables:

- Patients: contains demographic information about the patients.
- Observations: includes medical observations of patients, such as test results and measurements taken by smart devices.

- Conditions: contains information regarding the medical status of patients, including the presence of diseases.
- Questionnaire Responses: includes the results of questionnaires administered to patients by medical staff.

The data representation follows the Fast Healthcare Interoperability Resources (FHIR) standard, a set of rules and specifications for the exchange of electronic health data. This is accompanied using a well-defined semantics captured using standards like the Logical Observation Identifiers Names and Codes (LOINC) and SNOMED CT, which are used to uniquely identify medical observations through universal numerical codes [13].

The handling of this information, as it pertains to sensitive personal data, must comply with all privacy requirements and obligations established by national legislation, as well as those imposed by data protection regulations (GDPR). Therefore, the information used has undergone an anonymization process to prevent the identification of the individuals to whom it belongs.

5 Predictive Analytics in SMART BEAR

This chapter focuses on the development of predictive analytics within the SMART BEAR project, aimed at assessing cardiovascular risk in elderly patients using data collected from smart devices. Leveraging features such as age, gender, blood pressure, BMI and more, a logistic regression model was designed to predict the probability of cardiovascular events. The process involved careful data preprocessing, including normalization and feature selection, followed by model training and evaluation.

By analyzing the relationships between various health factors and cardiovascular risk, this work tries to provide an interpretable, data-driven approach to risk prediction, offering valuable insights for preventive healthcare and real-time patient monitoring in the Smart Bear ecosystem.

5.1 Introduction to Cardiovascular Risk Prediction

Cardiovascular diseases (CVD) represent one of the leading causes of mortality worldwide, encompassing conditions such as heart attacks, strokes, and hypertension. As populations age, the burden of CVD continues to rise, particularly among the elderly, who are more vulnerable to heart-related complications. Early detection and prevention are vital for reducing both the incidence of CVD and the associated healthcare costs.

In this thesis, a predictive model was developed with cardiovascular disease as the target variable, a binary classification where 1 indicates the presence of CVD and 0 indicates its absence. The dataset used for this analysis comprises over 1,600 patients, providing a robust basis for training and evaluating the model. To test the prediction tool, the case study of cardiovascular disease was chosen due to its correlation with a variety of individual and environmental risk factors. Specifically, the following features were considered for building the learning model: age, gender, systolic blood pressure, BMI, smoking status, alcohol intake, physical activity, LDL cholesterol levels, diabetes condition, presence of coronary

arteriosclerosis, and presence of myocardial infarction due to atherothrombotic coronary artery disease.

Predicting cardiovascular risk based on these clinical and lifestyle factors enables the identification of individuals at higher risk, facilitating timely interventions that can significantly reduce the likelihood of life-threatening events and improve patient outcomes.

5.1.1 Relevance to Smart Bear Project

The Smart Bear project enhances the ability to predict cardiovascular risk by integrating smart devices that continuously monitor patients' health data in real time. This constant data flow allows for dynamic tracking and analysis of patients' health trends, offering an opportunity to identify risk factors for cardiovascular disease as they emerge. By using predictive models on this up-to-date data, healthcare providers can offer more timely and personalized interventions, reducing the chance of severe cardiovascular events. The seamless connection between patient data and predictive analytics makes the Smart Bear platform an innovative tool for improving preventive healthcare in elderly populations [8].

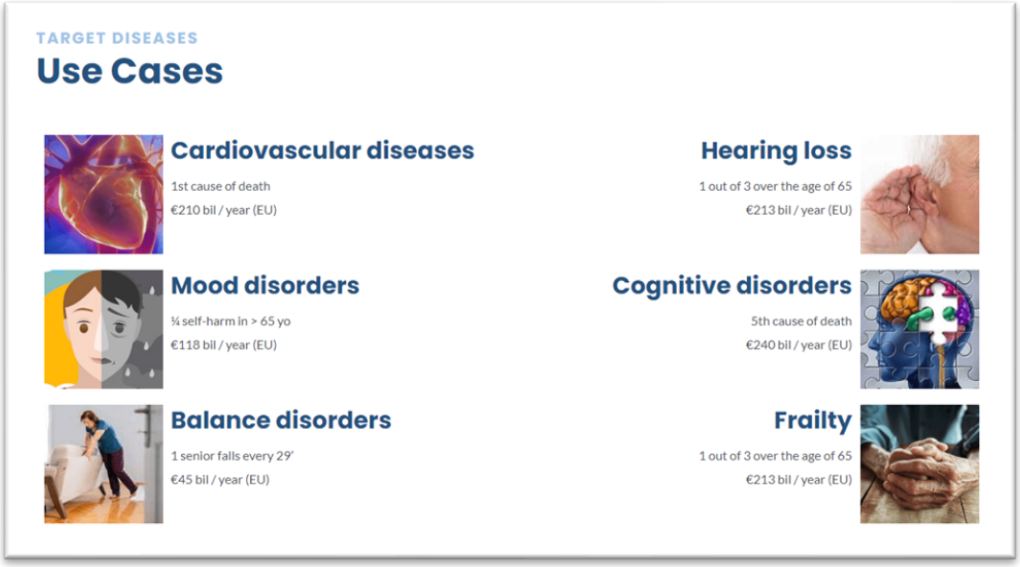


Figure 10: SMART BEAR Target Diseases

5.2 Dataset and Feature Selection

The data representation follows the Fast Healthcare Interoperability Resources (FHIR) standard, a set of rules and specifications for the exchange of electronic health data. This is accompanied using a well-defined semantics captured using standards like the Logical Observation Identifiers Names and Codes (LOINC) and SNOMED CT, which are used to uniquely identify medical observations through universal numerical codes [13].

The handling of this information, as it pertains to sensitive personal data, must comply with all privacy requirements and obligations established by national legislation, as well as those imposed by data protection regulations (GDPR). Therefore, the information used has undergone an anonymization process to prevent the identification of the individuals to whom it belongs.

5.2.1 Raw Data

The first phase of the process involves extracting all the necessary information for analysis and prediction activities. This data, referred to as "raw data," has not yet undergone any pre-processing.

Patient_id	Age	gender	Educational Status	Household Composition	Smoking Status	BMI	Alcohol Intake	effectiveDateTime	GDS Score
100	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
1001387968	80	Female	Graduated (completed) school	Lives with family	Never smoker	9.007305	NULL	2023-12-06 10:35:04	0
1001387968	80	Female	Graduated (completed) school	Lives with family	Never smoker	9.007305	NULL	2024-07-16 06:41:32	0
1001867813	66	Female	University	Lives alone	NULL	NULL	NULL	NULL	NULL
1003704144	73	Female	University	Lives with family	NULL	NULL	NULL	NULL	NULL
1008556145	75	Female	University	Lives with family	Former smoker	27.237846	NULL	2024-06-01 09:15:30	0
1010262872	67	Female	Elementary school	Lives with family	Never smoker	30.4	NULL	2024-06-18 07:21:19	0
1010332934	77	Male	Graduated (completed) school	Lives with family	Former smoker	30.585938	1	2024-02-08 10:32:20	2
1010863086	65	Female	Graduated (completed) school	Lives with family	Never smoker	28.3	NULL	2024-02-06 07:36:20	2
1011948368	79	Female	University	Lives alone	Never smoker	45.8	NULL	2024-01-30 12:02:02	2
1012320894	65	Male	Elementary school	Lives alone	Former smoker	24.72452	2	2023-11-14 09:39:19	11
1012320894	65	Male	Elementary school	Lives alone	Former smoker	24.72452	2	2024-07-18 10:00:31	1
1012587281	72	Female	Graduated (completed) school	Lives with family	Never smoker	NULL	NULL	2024-04-03 10:23:16	5
1013648307	78	Transgender female	University	Lives with family	NULL	NULL	NULL	NULL	NULL
1014898792	75	Female	Postgraduate	Lives alone	Former smoker	29.51856	NULL	2023-10-25 10:25:15	3
1015271899	76	Male	Elementary school	Lives with family	Former smoker	20.7	NULL	2024-01-26 15:31:51	1
1015271899	76	Male	Elementary school	Lives with family	Former smoker	20.7	NULL	2022-10-23 14:28:30	1
1015271899	76	Male	Elementary school	Lives with family	Former smoker	20.7	NULL	2023-10-26 14:55:07	4
1015623975	68	Female	Graduated (completed) school	Lives alone	Never smoker	18.716263	1	2024-07-03 08:10:50	1
1015741922	65	Male	Graduated (completed) school	Lives with family	NULL	25.428219	NULL	2024-07-16 14:07:13	0

Figure 11: Raw Dataframe

Each patient is uniquely identified by a numerical ID, which is used to link all their information across the various datasets. Each observation is associated with a set of details

following the FHIR schema, including medical identification codes according to LOINC and SNOMED standards. In the medical context, a distinction can be made between data extracted from time series and non-time series data. Time series data, such as blood pressure or heart rate, are always accompanied by the time at which the measurement was taken, as these are repeated over time. On the other hand, non-time series data, such as gender, tend to remain unchanged over time. When handling time series data, it is crucial to ensure that the extracted information is always up to date.

5.2.2 Selected Features

To test the prediction tool developed, the case study of cardiovascular disease was considered. This condition is linked to individual risk factors (genetic), such as gender, or environmental factors (social conditions, diet, smoking), which contribute to or determine the onset of the pathology. Specifically, the following features were considered for building the learning model: age, gender, systolic blood pressure, BMI, smoking status, alcohol intake, physical activity, LDL cholesterol levels, diabetes condition, presence of coronary arteriosclerosis, and presence of myocardial infarction due to atherothrombotic coronary artery disease.

5.2.3 Data Pre-Processing

After the extraction of raw data, the preparation phase begins, also known as pre-processing. During this stage, the raw data is cleaned and organized for the subsequent phases of processing and analysis. The process starts with a cleaning operation, aimed at identifying and correcting or removing any erroneous or inaccurate data. Following this, a reduction phase takes place, with the goal of eliminating non-significant redundancies in the collected values. Also, although some algorithms are capable of handling categorical values, most machine learning algorithms require datasets composed exclusively of numerical features to function correctly. Therefore, it is necessary to convert categorical features into numerical ones, also to apply the data imputation technique. The method used is label encoding, which assigns a corresponding increasing numerical value to each distinct categorical value of the

feature. The final step involves data imputation for replacing null values, as discussed in section 3.4.

Although there is the possibility of replacing missing data through data imputation techniques, if the amount of missing values is excessive compared to the overall size of the dataset, these are removed, as it would still be difficult to identify relationships between the data and perform proper replacement. In this work, features with more than 70% missing values compared to the total have been removed.

To build robust predictors from the raw clinical data, several transformations were applied. For instance, the gender feature was transformed into a binary variable, where male patients were assigned a value of 1, while female patients were assigned 1 only if they were older than 50 years, based on research indicating increased cardiovascular risk in post-menopausal women [18].

```
dfAgeGender = dfAge.join(dfGender, on="Patient_id")

# Creating column 'Gender_Risk' by following conditions
dfResult = dfAgeGender.withColumn(
    "Gender_Risk",
    when((col("Gender") == 'Male') |
         (col("Gender") == 'Transgender male') |
         ((col("Gender") == 'Female') & (col("Age") >= 50)) |
         ((col("Gender") == 'Transgender female') & (col("Age") >= 50)), 1)
    .otherwise(0)
)
```

Figure 12: Python Code – Converting to Numerical Feature

Given the availability of multiple daily data points for certain patients, averages were calculated for continuous features such as systolic blood pressure, physical activity, and LDL cholesterol. This aggregation ensured that each patient had a representative value for these features over time, minimizing variability due to daily fluctuations.

After preprocessing and merging all the patient data frames into a single, clean dataset, data imputation was employed to handle missing values.

```
# Merging all dataframes on 'Patient_id'
dfFinal = dfAge \
    .join(dfBP_Systolic_Risk, "Patient_id", "outer") \
    .join(dfBMI_Risk, "Patient_id", "outer") \
    .join(dfSmoker_Risk, "Patient_id", "outer") \
    .join(dfGender_Risk, "Patient_id", "outer") \
    .join(dfDiabetes_Risk, "Patient_id", "outer") \
    .join(dfCoronaryArteriosclerosis, "Patient_id", "outer") \
    .join(dfMyocardialConditions, "Patient_id", "outer") \
    .join(dfCardiovascularDisease, "Patient_id", "outer") \
    .join(dfCholesterol_Risk, "Patient_id", "outer") \
    .join(dfPhysicalActivity, "Patient_id", "outer")

# Drop Duplicates
dfFinal = dfFinal.dropDuplicates(['Patient_id'])
```

Figure 13: Python Code - Merging All Dataframes

For binary variables like gender, smoker status, diabetic status, history of heart attack and coronary arteriosclerosis, missing values were imputed with 0, representing the absence of these conditions, which aligns with the domain-specific logic of these features.

Initially, mean imputation was applied to continuous variables such as age, physical activity, BMI, blood pressure, and cholesterol, as the mean serves as a central tendency, helping to maintain the overall distribution of the dataset. However, due to the large proportion of missing data, this approach led to skewed results. Consequently, a decision was made to implement zero-imputation for these features as well, in order to minimize the impact of missing values on the model's performance.

```
def replace_null_with_zero(df):
    # Step 1: Identify numerical columns
    numeric_columns = [field.name for field in df.schema.fields if isinstance(field.dataType, (IntegerType, FloatType, DoubleType))]

    # Step 2: For each numerical column, replace Null with 0
    for col in numeric_columns:
        df = df.withColumn(col, F.when(F.col(col).isNull(), F.lit(0)).otherwise(F.col(col)))

    return df
```

Figure 14: Python Code - Data Imputation Function

The processed data, referred to as 'cleaned', are now ready to be used in building the learning model for the case study in question.

5.3 Model Selection: Logistic Regression

In predictive analytics, the selection of an appropriate model is crucial to accurately estimate outcomes and derive meaningful insights from clinical data. For the purpose of this study, logistic regression was chosen due to its effectiveness in binary classification tasks, such as predicting cardiovascular risk. Logistic regression models the probability of a binary outcome, in this case, whether a patient is at cardiovascular risk (1) or not (0), based on a set of input features. This approach offers a transparent framework, allowing for the interpretation of individual feature contributions through model coefficients. In subsequent sections, I will discuss the rationale behind this model choice, its design and architecture, as well as the training process. Moreover, I will explain the application of z-score normalization to prepare the input features, ensuring their comparability and improving the model's performance.

5.3.1 Model Rationale

Logistic regression was selected for this study due to several key advantages that align with the project's goals, particularly the need for model interpretability and transparency. One of the primary reasons for its selection is the straightforward interpretability of the model's coefficients. Each feature's contribution to the prediction can be directly examined, offering valuable insights into how variables like blood pressure, cholesterol levels, and BMI affect cardiovascular risk, while also showing how a good level of physical activity helps reduce it. This level of interpretability is essential when working with clinical data, where understanding the impact of each factor is critical for making informed medical decisions.

Furthermore, logistic regression provides a probabilistic framework, allowing the calculation of the probability of cardiovascular risk rather than a simple binary outcome. This offers a nuanced perspective, as clinicians can assess not only the risk classification (0 or 1) but also the likelihood associated with each prediction. The visual representation of logistic regression, such as in ROC curves or decision boundaries, is intuitive and easy to interpret, making it an ideal tool for communicating results to non-technical stakeholders. Additionally,

logistic regression's simplicity makes it computationally efficient, suitable for medium and large datasets like the one provided by the SMART BEAR project.

5.3.2 Model Design: Target and Standardization

Once the data is merged into a single dataset, containing all the relevant features for each patient. This allows normalization through Z-score, bringing numerical values to a common scale without losing information. The model was designed with cardiovascular disease as the target variable, a binary classification where 1 indicates its presence and 0 indicates its absence.

To ensure that all features contributed equally to the logistic regression model, standardization was applied using the StandardScaler function from the '*sklearn.preprocessing'* class [14]. StandardScaler normalizes each feature by removing the mean (centering) and scaling it to unit variance, which ensures that the dataset has a mean of 0 and a standard deviation of 1. This is essential for features with varying scales, such as age (measured in years) and physical activity (measured in thousands of steps). Standardizing the dataset allows the logistic regression model, which is sensitive to feature scales, to treat each feature comparably, thereby improving its performance and reducing bias.

```
class sklearn.preprocessing.StandardScaler(*, copy=True, with_mean=True,
with_std=True)
```

Figure 15: Scikit – StandardScaler Class

5.3.3 Model Training

To train the logistic regression model for binary classification, the LogisticRegression function from the '*pyspark.ml.classification*' class was employed [15]. This function calculates the probability that an observation belongs to a particular class and assigns it based on a threshold. Logistic regression uses the sigmoid function to convert a linear combination of the features into a probability. The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where z is a linear combination of the feature values and their respective coefficients. The sigmoid function outputs a probability value between 0 and 1. Logistic regression then classifies the observation based on this probability: if the probability exceeds a given threshold (typically 0.5), the observation is classified as belonging to the positive class (1). Otherwise, it is assigned to the negative class (0). The model's output includes both the predicted class (0 or 1) and the associated probability, offering a nuanced risk assessment. In the context of this study, the model predicts the cardiovascular risk for each patient, giving both a binary classification and a risk probability.

```
# Defining X for features and y for true label
X = dfFinal[['Physical Activity', 'Diabetic', 'Coronary Arteriosclerosis',
             'Suffered from Heart Attack', 'LDL Cholesterol', 'Blood Pressure',
             'BMI', 'Smoker', 'Gender', 'Age']]
y = dfFinal['Cardiovascular_Disease']

# Pipeline to combine Feature Standardization and Logistic Regression
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('model', LogisticRegression())
])
```

Figure 16: Python Code - Defining Features, Target and Pipeline

Given the moderate size of the dataset, consisting of approximately 1,600 observations, the k-fold cross-validation method from the 'sklearn.model_selection' class was used to robustly estimate the model's performance [17]. Instead of a single split into training and testing sets, cross-validation divides the data into k subsets (folds). The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, each time using a different fold as the test set, and the final performance metric is the average over all k iterations.

```
class sklearn.model_selection.KFold(n_splits=5, *, shuffle=False, random_state=None)
```

Figure 17: Scikit - KFold Class

In this study, the dataset was split into k=5 folds. For each iteration, the model was trained on 4 folds (approximately 1,200 observations) and tested on the remaining 1 fold (approximately 300 observations). This approach allows for a more reliable performance estimate, as each observation is used both for training and for testing. Performance metrics such as accuracy, precision, recall, and F1-score were averaged across the five iterations to provide a robust evaluation of the model's predictive power.

This process was implemented in Python using the scikit-learn library's KFold module [16], which takes as parameters: the number of splits into which the dataset will be divided; a Boolean value indicating whether the data should be shuffled before being split into folds; a random state that represent a seed for the random number generator used to shuffle the data, useful for ensuring the reproducibility of results. By using k-fold cross-validation, the model benefits from a balanced trade-off between training and testing data, ensuring a reliable assessment of its generalization capabilities.

Following the completion of these preliminary steps, the model was ready for training and subsequent predictions.

```python
# Setting k-fold cross-validation with k=5
k = 5
kfold = KFold(n_splits=k, shuffle=True, random_state=42)

# Model Training
model = pipeline.fit(X, y)

# Making predictions
predictions = pipeline.predict(X)
probabilities = pipeline.predict_proba(X)[:, 1]

# Final DataFrame with patient_id and results
results_df = pd.DataFrame({
    'Patient_id': patient_ids,
    'Predicted_Risk': predictions,
    'Probability_Percent': probabilities * 100,
})
```

Figure 18: Python Code - Model Training

5.4 Evaluation and Results

In this section, the performance of the logistic regression model developed to predict cardiovascular risk is presented and analyzed. A thorough evaluation is crucial to ensure the model's accuracy, reliability, and interpretability, which are essential for clinical applications. Several key metrics, such as coefficients of the features; distribution of predicted cardiovascular risk; k-fold cross-validation results; the Receiver Operating Characteristic (ROC) curve and the precision-recall curve are used to assess the model's predictive capabilities and its ability to provide meaningful insights.

	Patient_id	Predicted_Risk	Probability_Percent
0	119332904	1	68.204711
1	1691055553	1	69.354573
2	1916074755	1	72.678990
3	1925064084	0	8.920647
4	2025714478	0	7.241716
5	425886996	0	8.903382
6	612834351	0	8.911226
7	1263554591	1	94.296500
8	12728994	1	96.932036
9	1415070209	1	92.344375
10	1509031557	1	55.255350
11	1767273510	1	78.220588
12	480964993	1	52.229940

Figure 19: Prediction and Probabilities Dataframe

5.4.1 Coefficients of the Features

The first step in evaluating the logistic regression model involves analyzing the coefficients of the features to determine their relative impact on cardiovascular risk prediction. This analysis provides a clear interpretation of how each variable influences the model's

predictions, offering insights into the role of clinical factors in determining cardiovascular risk.

	Feature	Coefficient
0	Physical Activity	-0.163087
1	Diabetic	0.236180
2	Coronary Arteriosclerosis	0.867337
3	Suffered from Heart Attack	0.464029
4	LDL Cholesterol	0.935893
5	Blood Pressure	1.248094
6	BMI	0.269780
7	Smoker	0.019297
8	Gender	0.041637
9	Age	0.002529

Figure 20: Coefficients of the Features

In a logistic regression model, feature coefficients provide a quantitative measure of each variable's contribution to the predicted outcome. Positive coefficients indicate that as the value of a feature increases, so does the likelihood of the positive class—in this case, the presence of cardiovascular disease. For instance, if the coefficient for age is positive, it suggests that older individuals are more likely to be classified as at risk for cardiovascular disease.

Conversely, negative coefficients imply that an increase in the feature value reduces the likelihood of the positive class. For example, if physical activity has a negative coefficient, higher number of daily steps would be associated with a lower risk of cardiovascular disease.

The magnitude of the coefficient is also crucial: larger absolute values signify a greater impact on the model's predictions, while smaller coefficients suggest that the corresponding feature has a limited effect on the outcome. This allows for the identification of the most

influential predictors, such as blood pressure or LDL Cholesterol, and highlights which variables have a more modest effect.

This interpretability is a key strength of logistic regression, offering both predictive power and clear, actionable information.

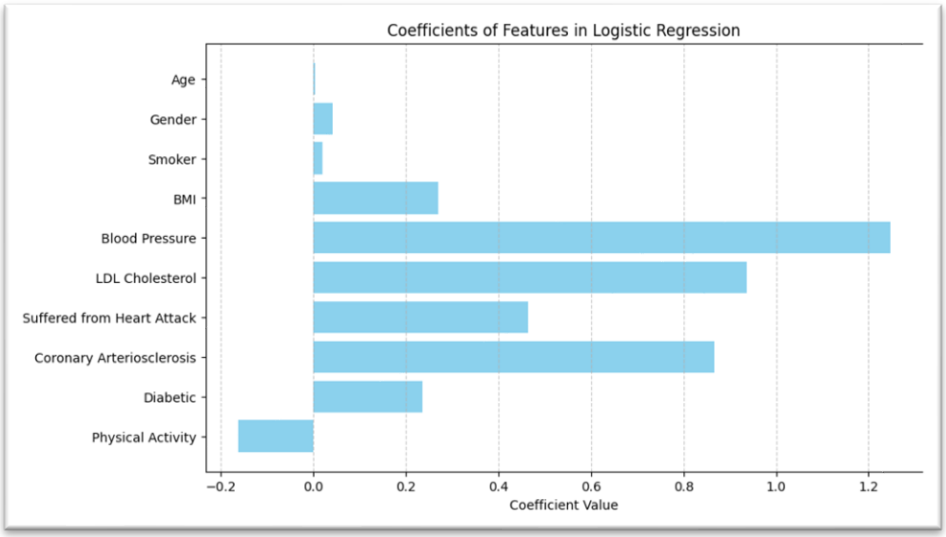


Figure 21: Coefficients Diagram

5.4.2 Distribution of Predicted Cardiovascular Risk Probabilities

Additionally, the distribution of predicted cardiovascular risk across the patient cohort is evaluated. This analysis is essential for understanding how effectively the model stratifies individuals into different risk categories, providing valuable insights into its ability to differentiate between low- and high-risk patients.

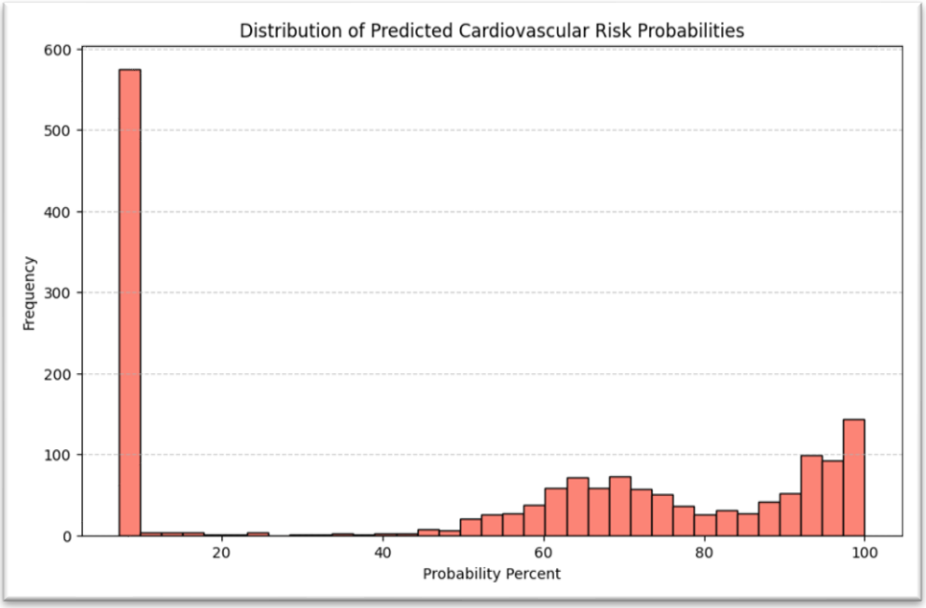


Figure 22: Distribution of Predicted Cardiovascular Risk Probabilities

The predicted probabilities represent the model's estimation of each patient's likelihood of developing cardiovascular disease, ranging from 0% (very low risk) to 100% (high risk). In this analysis, a histogram is used to visually represent the distribution of these probabilities across all patients. The x-axis of the histogram corresponds to the predicted probabilities, while the y-axis represents the frequency of patients falling into each probability range.

The shape of the distribution offers a clear indication of the model's performance. Ideally, a well-performing model will produce a bi-modal distribution, where one group of patients has low probabilities (indicating low cardiovascular risk) and another group has high probabilities (indicating high cardiovascular risk). This clear separation between risk categories reflects the model's confidence in its predictions and its ability to distinguish between healthy and at-risk individuals.

In contrast, a distribution where probabilities cluster around 50% or appear flat may suggest that the model struggles to confidently separate high-risk from low-risk patients. Such patterns could indicate that the model is either uncertain in its predictions or that the features used for prediction do not provide enough discriminatory power.

5.4.3 Cross Validation K-Fold Results

As previously described, to ensure the robustness of the model's performance evaluation, k-fold cross-validation was employed. This method mitigates the potential impact of random variations in data splitting by dividing the dataset into k equally sized folds. By doing so, it provides a reliable estimate of the model's overall performance and offers a more comprehensive assessment of its consistency across different subsets of the data.

The primary advantage of k-fold cross-validation lies in its ability to generate more robust performance metrics by training and testing the model on multiple data partitions. Each fold serves as a test set once, while the remaining k-1 folds are used for training, ensuring that all data points are used both for training and validation [17].

	Metric	Value
0	Mean Accuracy	0.831707
1	Accuracy Variance	0.008493

Figure 23: K-Fold Cross-Validation Results

5.4.3.1 Mean Accuracy (Average Performance Across Folds)

The mean accuracy represents the average accuracy score across all k folds, offering a general indication of the model's performance on different subsets of the data. For each fold, an accuracy score is computed based on the proportion of correct predictions. These accuracy scores are then averaged to obtain the mean accuracy, which reflects the model's overall ability to predict cardiovascular risk across the entire dataset.

This metric is crucial for evaluating the model's general performance. By averaging the accuracy over multiple folds, the model's performance is not biased by a particular test set or data partition, thus ensuring a more comprehensive evaluation. In the context of cardiovascular risk prediction, the mean accuracy provides a reliable indicator of how well

the model can classify patients into high- and low-risk categories, irrespective of the specific data used for training or testing.

5.4.3.2 Variance (Dispersion) of Accuracy

The variance or standard deviation of accuracy scores across the folds is another important measure. It quantifies the degree of fluctuation in performance between different subsets of the data. A lower variance indicates that the model's performance is stable across different folds, while a higher variance suggests that the model's predictions may be more sensitive to variations in the training or testing data.

A low variance in accuracy suggests that the model performs consistently across different patient groups, indicating greater confidence in its predictions when applied to new, unseen data. On the other hand, a high variance, even with a high mean accuracy, could signal that the model's performance is uneven and may not generalize well across different subsets of the population.

5.4.3.3 Interpreting the Results Together

A high mean accuracy combined with low variance indicates that the model performs both effectively and consistently across various data subsets. This is the ideal scenario, as it suggests the model can generalize well to new, unseen data. Conversely, if the model exhibits high mean accuracy but also high variance, it implies that, while the model can perform well, it may do so inconsistently, potentially overfitting to certain subsets of the data. In such cases, further investigation and refinement may be necessary.

A low mean accuracy with low variance signifies that the model consistently performs poorly, indicating the need for additional feature engineering, tuning, or possibly the selection of a different algorithm. The worst-case scenario is a low mean accuracy paired with high variance, which reflects a model that is both unreliable and performs inadequately. In this case, significant adjustments are required to improve performance.

In the cardiovascular risk prediction model developed for this thesis, the objective has been achieved, as both high mean accuracy, indicating effective risk stratification, and low variance, ensuring the model's consistency across diverse patient groups, have been attained.

5.4.4 ROC Curve

The Receiver Operating Characteristic (ROC) curve is also used to evaluate the model's overall discriminative ability. This curve illustrates the relationship between the true positive rate and the false positive rate, providing a comprehensive view of the model's performance across various classification thresholds. By analyzing the ROC curve, it is possible to understand how well the model differentiates between patients with and without cardiovascular risk.

The ROC curve is plotted with the false positive rate on the X-axis and the true positive rate on the Y-axis. The false positive rate represents the proportion of individuals without cardiovascular disease who are incorrectly classified as at risk, while the true positive rate reflects the proportion of actual cardiovascular patients correctly identified by the model. This comparison across different thresholds allows for an analysis of the trade-off between sensitivity and specificity.

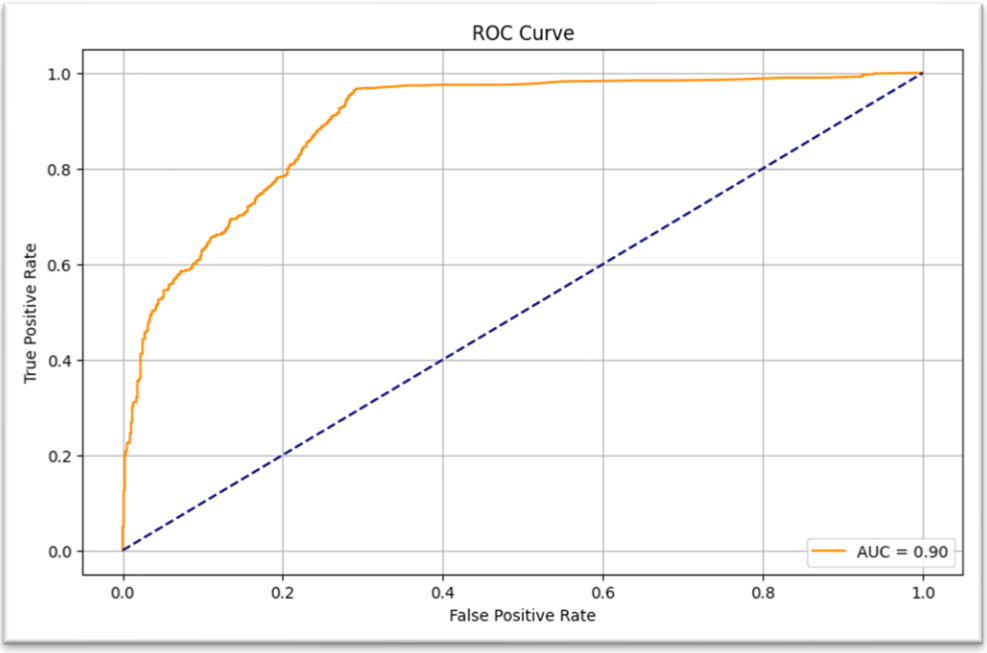


Figure 24: ROC Curve

A critical metric derived from the ROC curve is the Area Under the Curve (AUC), which serves as a summary statistic of the model's classification performance. The AUC provides a single value that quantifies the model's ability to correctly distinguish between positive and negative cases. AUC values range from 0 to 1, with higher values indicating better performance. In this case, a value of 0.9 suggests good/excellent discriminative ability.

The ROC curve and AUC score are particularly useful for evaluating the model's performance in scenarios where there is an imbalance between the positive and negative classes (e.g., fewer patients with cardiovascular disease).

5.4.5 Precision-Recall Curve

The precision-recall curve is employed to evaluate the model's performance, especially in the context of imbalanced datasets. This curve is particularly relevant when the positive class, such as patients at risk for cardiovascular disease, is less frequent in the dataset. By focusing on the relationship between precision and recall, the precision-recall curve provides valuable insights into the model's effectiveness in correctly identifying high-risk individuals.

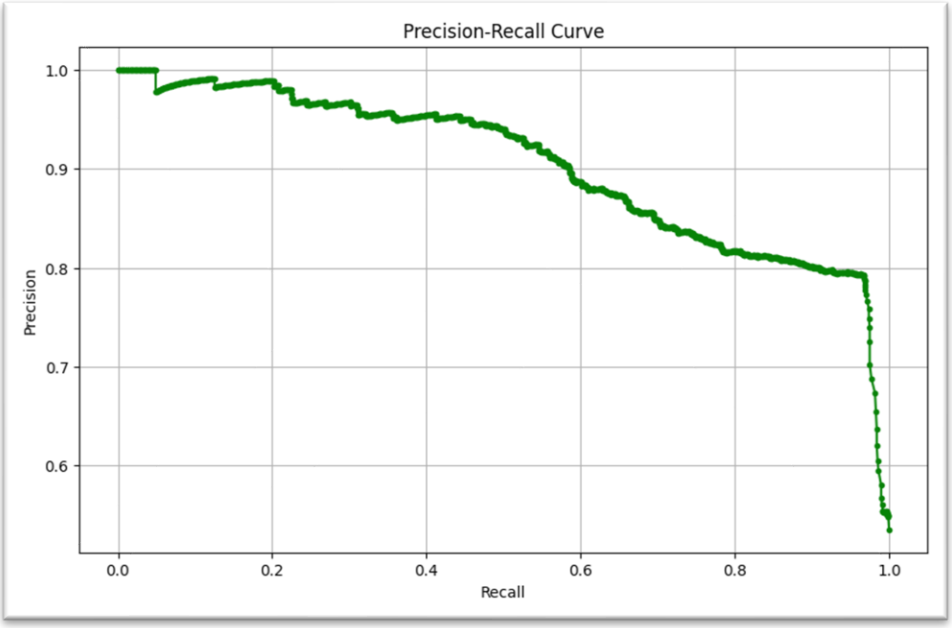


Figure 25: Precision-Recall Curve

The precision-recall curve plots recall on the X-axis and precision on the Y-axis:

- Recall (also known as sensitivity) represents the proportion of actual positive cases that the model correctly identifies. It reflects how well the model detects patients with cardiovascular disease.
- Precision measures the proportion of predicted positive cases that are actual true positives, indicating how reliable the model's positive predictions are.

The precision-recall curve is particularly useful when there is a class imbalance, as it emphasizes the model's performance on the positive class. In such cases, traditional metrics like accuracy might not provide an adequate picture of the model's performance, especially if the negative class (healthy patients) dominates the dataset as in this case. The curve helps highlight potential trade-offs between precision and recall, revealing situations where increasing recall might come at the expense of precision, or vice versa.

6 Intrinsic Capacity and Predictive Analytics

The objective of this section is to explore the association between Intrinsic Capacity (IC) and cardiovascular risk factors within the framework of the SMART BEAR project. While IC captures a person's physical and mental capacities, this study aims to determine whether there is a significant relationship between an individual's IC and their likelihood of developing cardiovascular diseases.

6.1 Introduction to Intrinsic Capacity

Intrinsic Capacity is a holistic concept in healthcare that refers to the composite of all the physical and mental capacities an individual can draw upon at any given time. It represents an individual's potential for healthy aging, encompassing a wide range of functional abilities such as cognitive, emotional, and physical capacities. The concept has been introduced by the World Health Organization as part of its framework for healthy aging, shifting the focus from disease-oriented care to maintaining and improving overall functional ability throughout the life course. IC is influenced by multiple factors, including genetics, lifestyle, and environmental conditions, and it plays a critical role in predicting health outcomes, particularly among the elderly. By assessing and optimizing IC, healthcare providers aim to prevent or delay the onset of age-related declines, enhancing quality of life and enabling more personalized interventions tailored to the individual's functional reserve. This approach underscores the importance of proactive, preventive strategies in modern healthcare, particularly as the global population continues to age.

6.1.1 Linking IC with Predictive Analytics

A correlation study is a type of research that explores the relationship between two or more variables to determine whether a systematic association exists between them. This approach does not imply causality but rather identifies if the variables tend to vary together in a consistent manner.

In this context of predictive analytics applied to cardiovascular risk and Intrinsic Capacity, the objective of the correlation study is to investigate how the overall IC index is associated with the predicted cardiovascular risk. Rather than focusing on individual features, the study aims to determine whether an individual's general IC level correlates with their likelihood of developing cardiovascular diseases, providing insight into how a patient's overall health status influences their risk profile.

6.2 The Relationship Between IC and Cardiovascular Risk

To thoroughly investigate the relationship between Intrinsic Capacity and cardiovascular risk, this study employs a combination of Pearson Correlation and Spearman Correlation. These methods are selected for their complementary strengths in analyzing both linear and non-linear relationships, providing a comprehensive understanding of how IC may influence cardiovascular risk.

6.2.1 Rationale for Correlation Study

The selection of Pearson Correlation and Spearman Correlation for this study is grounded in their complementary strengths in analyzing relationships between variables. Pearson Correlation is chosen for its ability to measure the strength and direction of linear relationships between continuous variables, making it suitable for initial exploration of the relationship between Intrinsic Capacity and cardiovascular risk. On the other hand, Spearman Correlation is included to account for potential non-linear monotonic relationships, providing a robust alternative that is less sensitive to outliers and does not assume normality. Together, these techniques offer a comprehensive approach to understanding the nature and strength of the relationship between the variables under study.

6.2.2 Methodology

The methodology for Pearson Correlation involves calculating the Pearson correlation coefficient ((r)), which quantifies the linear relationship between Intrinsic Capacity and

cardiovascular risk. This is achieved by standardizing the covariance of the variables by their standard deviations. Conversely, for Spearman Correlation, the ranks of the data points are used instead of their raw values, and the Spearman correlation coefficient ((\rho)) is computed to assess the monotonic relationship. This involves ranking the data, calculating the differences between the ranks, and then applying the Spearman formula.

In this study Pearson and Spearman Correlation have been achieved by using the 'pyspark.pandas.DataFrame.corr' method which accepts as parameters both 'spearman' and 'pearson'.

```
dfCorrelation = pd.merge(dfRisk_Final, dfIC_Final, on='Patient_id')

# Pearson Correlation
pearson_corr = dfCorrelation['Predicted_Risk'].corr(dfCorrelation['Intrinsic Capacity'], method='pearson')
print(f"Pearson correlation: {pearson_corr}")

# Spearman Correlation
spearman_corr = dfCorrelation['Predicted_Risk'].corr(dfCorrelation['Intrinsic Capacity'], method='spearman')
print(f"Spearman correlation: {spearman_corr}")
```

Figure 26: Python code - Pearson and Spearman Correlation

6.2.3 Challenges

Each of these techniques presents specific challenges. Pearson Correlation assumes a linear relationship and is highly sensitive to outliers, which can distort the correlation coefficient and lead to misleading conclusions if the data contains anomalies. Additionally, it requires the data to be normally distributed, which may not always be the case. Spearman Correlation, while more robust to outliers and non-normal distributions, only measures monotonic relationships and may lose detailed information about the actual values of the data points. This can result in a less precise understanding of the relationship's nature.

6.3 Results of Correlation Study

To better show the results of the Correlation Study, a table and a scatter plot with a regression line have been created.

	Correlation	Result
0	Pearson	-0.413919
1	Spearman	-0.370498

Figure 27: Correlations Results

6.3.1 Pearson Correlation

The Pearson correlation, which as illustrated before measures the linear relationship between continuous variables, returned a value of -0.414, indicating a moderate negative correlation. This result suggests that as Intrinsic Capacity increases, the predicted cardiovascular risk tends to decrease. The strength of this relationship is moderate, implying that while the two variables are inversely related, the connection is not particularly strong.

6.3.2 Spearman Correlation

Similarly, the Spearman correlation, which assesses monotonic relationships—not necessarily linear—yielded a value of -0.370. Like the Pearson result, this also points to a moderate negative correlation, confirming that individuals with higher Intrinsic Capacity tend to exhibit lower predicted cardiovascular risk. The slightly lower Spearman value suggests that while the relationship holds in terms of rank order, it may not be perfectly linear.

6.3.3 Graph Interpretation

The scatter plot created further illustrates these findings. Each point represents an individual patient, with Intrinsic Capacity plotted on the X-axis and Predicted Cardiovascular Risk on the Y-axis. The regression line on the plot, with its negative slope, reflects the inverse relationship between the two variables. The moderate spread of the points around the regression line corroborates the conclusion of a moderate correlation.

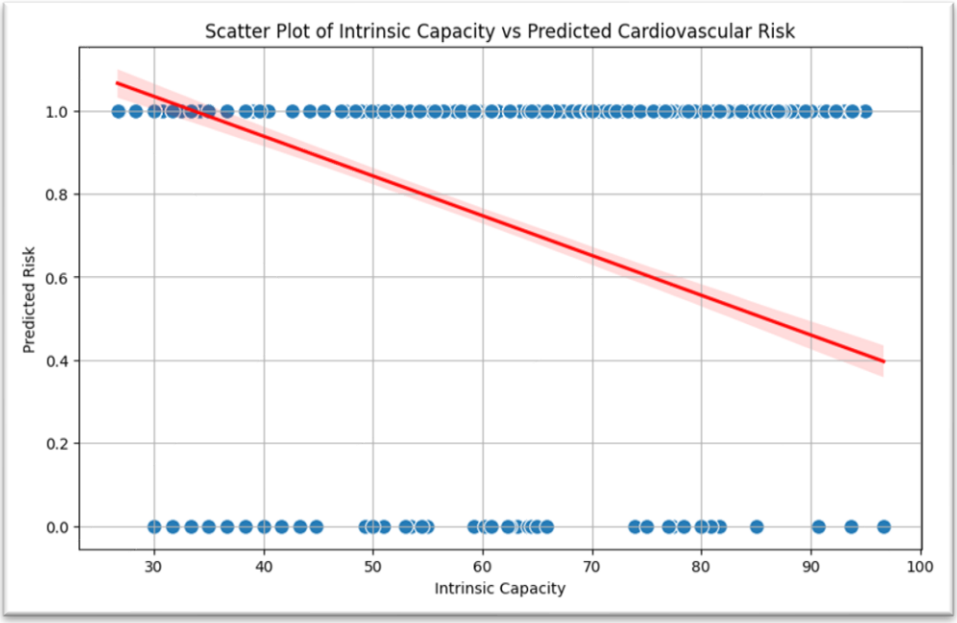


Figure 28: Scatter Plot with Regression Line

6.4 Discussion and Implications

6.4.1 Implications for Preventive Healthcare

The integration of an Intrinsic Capacity indicator, followed by predictive analytics on cardiovascular risk, and subsequently a correlation study between these two variables, holds significant implications for preventive healthcare. By establishing IC as a comprehensive measure of an individual's overall functional ability, healthcare providers can gain early insights into potential health declines before they manifest as clinical conditions. Predictive analytics on cardiovascular risk further enhances this approach by identifying individuals at heightened risk of cardiovascular events, enabling targeted interventions. The correlation study between IC and cardiovascular risk elucidates the relationship between general functional capacity and specific health risks, providing a nuanced understanding that can inform personalized preventive strategies. This multi-faceted approach allows for the early identification of at-risk individuals, the implementation of tailored preventive measures, and

the continuous monitoring of health outcomes, ultimately contributing to more effective and proactive healthcare management.

6.4.2 Future Integrations into the SB project

In future studies, Linear Regression could be applied to significantly enhance the field of preventive healthcare by providing a more detailed and quantifiable understanding of the relationship between Intrinsic Capacity and cardiovascular risk. By modeling the dependent variable (cardiovascular risk) as a function of IC and potentially other covariates, Linear Regression can help identify specific factors that contribute to increased risk, allowing for more precise risk stratification. This technique can also facilitate the development of predictive models that forecast individual health outcomes based on their IC scores, enabling healthcare providers to implement personalized preventive measures. Moreover, Linear Regression can help in evaluating the effectiveness of interventions by analyzing changes in IC and corresponding shifts in cardiovascular risk over time. By incorporating this approach, future research can offer deeper insights and more robust tools for early detection and prevention, ultimately improving patient outcomes and reducing the burden of cardiovascular diseases.

7 Bibliography

[1] Theo Vos, Christine Allen, Megha Arora, Ryan M Barber, Zulfiqar A Bhutta, Alexandria Brown, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Z Chen, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015.

[2] Alessia Cristiano, Sara De Silvestri, Stela Musteata, Alberto Sanna, Diana Trojaniello, Valerio Bellandi, Paolo Ceravolo, Matteo Cesari, et al. Iot platform for ageing society: The smart bear project. In eTELEMED 2021 The Thirteenth International Conference on eHealth, Telemedicine, and Social Medicine. IARIA, 2021.

[3] Zhang Z. Missing data exploration: highlighting graphical presentation of missing pattern. Ann Transl Med 2015;3(22):356. doi: 10.3978/j.issn.2305-5839.2015.12.28.

[4] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. When and how should multiple imputation be used for handling missing data in randomized clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology, 2017.

[5] Multiple Imputation in STATA. UCLA: Statistical Methods and Data Analytics.

[6] Lall Ranjit. How multiple imputation makes a difference. JSTOR, 2006.

[7] Valerio Bellandi, Ioannis Basdekis, Paolo Ceravolo, Matteo Cesari, Ernesto Damiani, Eleftheria Iliadou, Mircea Dan Marzan, and Samira Maghool. Engineering continuous monitoring of intrinsic capacity for elderly people. In 2021 IEEE International Conference on Digital Health (ICDH), pages 166–171. IEEE, 2021.

[8] Smart Bear. https://www.smart-bear.eu/project/

[9] R. Suzman and J. Beard. Global health and aging. NIH Publ., 2011.

[10] WHO. Active aging: A policy framework. 2002.

[11] Statistics on population development in Europe. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_statistics_on_population_developments. Accessed: 2024-01-10.

[12] Pyspark. https://spark.apache.org/docs/latest/api/python/#

[13] Vadim Peretokin, Ioannis Basdekis, Ioannis Kouris, Jonatan Maggesi, Mario Sicuranza, Qiqi Su, Alberto Acebes, Anca Bucur, Vinod Jaswanth Roy Mukkala, Konstantin Pozdniakov, et al. Overview of the smart-bear technical infrastructure. In Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health-ICT4AWE,, pages 117‑125. SciTePress, 2022.

[14] Scikit StandarScaler function: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[15] Scikit LogisticRegression function: https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html

[16] Scikit KFold function: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[17] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A.J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE, 14*.

[18] Dabla, Pradeep kumar, Vandana Dabla, Rajni Dawar and Sarika Arora. “APPROACH TO POSTMENOPAUSAL CARDIOVASCULAR RISK Review Article.” (2011).

Last but not least, I wanna thank me
I wanna thank me for believing in me
I wanna thank me for doing all this hard work
I wanna thank me for having no days off
I wanna thank me for, for never quitting

[Snoop Dog]