Spring 5-2024

Using Predictive Analytics to identify risk of Heart Disease based on lifestyle factors and health metrics.

Luiza Cavalcanti Albuquerque Brayner
CCT College Dublin

Edgard Pacheco
CCT

CCT College Dublin

Assessment Cover Page

Module Title:	Problem Solving for Industry
Assessment Title:	Capstone Pair Project
Lecturer Name:	Dr. Muhammad Iqbal
Student Full Name:	Luiza Cavalcanti Albuquerque Brayner & Edgard Pacheco
Student Number:	2020309 & 2020332
Assessment Due Date:	17th May 2024
Date of Submission:	17th May 2024

Dataset gathered from: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
Github Collaboration link: https://github.com/luizaalbuquerque/Strategic-Analysis.git

WORD COUNT: 5.008 words.

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

CCT College Dublin **1**
Introduction 3
What are Heart Diseases? 3
Project Plan 3
Overarching the area 4
Investigate data available 4
Understanding of data and concerns 5
Project Plan by the CRISP-DM framework 5
Resource and risk Analysis (Cost and Benefits) 11
Modelling performance evaluation - Diving deeper in the results 11
Diving deeper into prototype deployment phase 13
Roles and Responsibilities - division by team members 15
Conclusion 16
References 17
Appendix: 21
Group Reflection 21
Evidence of group work 22
Individual Contribution Report 22
Luiza 22

Introduction

Ref.[10]

In this part of the project, we will report an innovative application, for the healthcare sector usage, which basically is a health tracking and disease prevention application. The application will enable users to log their daily meals, exercise routines, and lifestyle habits, providing a comprehensive overview of the user's health status. By making use of Machine Learning and data analytics, our solution offers a personalised and automated insight and predictive analytics, which empowers users to proactively manage their well-being.

Through a detailed data analysis, users will gain valuable insights of potential diseases development and risk. This report will explore the development process, implementation of machine learning models, data visualisation techniques, and the transformative impact of our application over health management.

What are Heart Diseases?

Ref.[21]

Heart diseases encompass a range of conditions affecting the heart’s structure and function, including but not limited to: coronary artery disease, heart rhythm disorders, and congenital heart defects. Which are a leading cause of morbidity and mortality worldwide.

This project aims to address heart disease by developing a personalised health tracking and disease prediction application. By analysing lifestyle factors and health metrics, the application can identify individuals at risk of heart diseases early, which enables proactive interventions and improved management of cardiovascular health. (Mayo Clinic Staff)

Project Plan

Ref.[9]

1. **Week 1**:

The project required a few steps, which involved business understanding, where the team members were focused over gathering project requirements, defined objectives, and understanding stakeholder needs.

2. **Week 2**:

Then the team members moved forward to the data understanding phase, where the responsibilities included exploring available datasets, checking data quality, identifying potential sources of information, etc. This project phase took about 2 - 3 weeks of the project.

3. **Week 3**:

On the third step, it included the data preparation, which involved cleaning, transforming and integrating data to prepare it for analysis. After the data was ready to be used, we reached the modelling phase, where team members focused on developing models to predict health outcomes. This step took 1 - 2 weeks of implementation.

4. **Week 4**:

Over the evaluation phase, there was an assessment for model performance to ensure that the project was aligned with the goals. Last it was the deployment phase, which included deploying the solution and communicating the results in an effective way. This phase lasted 1 week, it was also here where we were able to perform the report, poster presentation and display solutions and findings.

Through the project, team members were able to collaborate to deliver a comprehensive presentation done through extensive research of the topic chosen, addressing design, content, layout, risk analysis, containing cost and benefits, inclusion of a section of judgement and conclusion on the application performance.

An section also for ethical and legal issues that it might face, and articulation of the full project, ensuring the delivery of such, inclusion of how it was developed, how is the project relevant for the sector in question (healthcare improvement/advancement), last how it could be improved and application of further development of the platform.

Overarching the area

Ref.[11]
In the healthcare industry, the technology proposed focuses on developing a prototype for a health tracking and disease risk prediction application. Making use of machine learning algorithms, we aimed to predict the likelihood of developing certain diagnosed diseases, in this case, we focused only on developing heart diseases detection based on lifestyle factors and health metrics, providing insights for preventative care.

Investigate data available

Ref.[12]
The dataset has various health-related attributes including age, gender, lifestyle factors like smoking habits, alcohol consumption, physical activity, and general health. Health metrics such as physical health, mental health, and the presence of diagnosed diseases like asthma and diabetic are also included. This dataset provides valuable insights into factors influencing health outcomes.

Understanding of data and concerns

Ref.[13, 14 & 15]

The dataset covers a diverse health aspect. However, it lacks information over family medical history and genetic predispositions, which could influence disease risk. The chosen dataset was gathered from online open source, see details in references, to make it more usable for the work in question, we had to perform an EDA (exploratory data analysis, and moving forward with data cleaning, visualisation and engineering).

Data consistency and accuracy was ensured, specially for self-reported variables like physical activity and general health. Missing values and outliers were also handled over the data gathering and cleaning process, which ensured a more robust analysis.

Data security and storage is also a concern in this project. Due to ethical, legal and privacy issues, as we will be gathering, storing and processing private and health data, we need to make the application compliant with all local regulations previous the deployment step, further research is attached over the references sections of this project.

Project Plan by the CRISP-DM framework

Ref.[1]

In this project phase, it was possible to dive deeper into the project planning, through the use of the CRISP-DM Framework (Cross-Industry Standard Process for Data Mining). (“Six Steps in CRISP DM - The Standard Data Mining Process”)

1. **Business Understanding**

- **Objectives**: To develop a prototype of a personalised health tracking and disease risk prediction application to assess the potential impact of the concept in society. The primary objective of this project was to identify and address the healthcare sector’s evolving needs for personalised health tracking and disease risk prediction.

  After some research to understand the challenges faced by healthcare professionals in disease prevention and management, we aimed to develop a solution that leverages machine learning and data analytics to provide actionable insights.

  By focusing on user-centric design principles, we aimed to create a user-friendly environment that promotes engagement and adherence to healthy lifestyle choices.

  Furthermore, we recognized the importance of aligning the project with broader healthcare industry trends towards preventative care and patient empowerment. Overall, our business understanding objectives aimed to develop an innovative application that addresses critical gaps in current healthcare practices and delivers tangible benefits to both users/patients and healthcare providers.

- **Stakeholders:** Healthcare sector, such as healthcare professionals, data scientists, application developers, end-users/patients.

- **Requirements:** Determine features, such as age, gender, smoking habit, physical activity, general health, asthma, and alcohol consumption for predicting the risk of diagnosed diseases (focusing in this case, heart diseases prediction).

- **User needs:** Conducted a user research to understand preferences for data input methods, visualisation of health insights, and recommendations done through that. (To be done on the user integration part)

2. Data Understanding

- **Exploration:** Identification and exploration of available datasets (as it is possible to see over the folder and several dataset files) relevant features for different health aspect, we decided to move forward detecting heart diseases, which was the most complete dataset (was needed to do an deep extensive research), such as lifestyle habits, and medical history and its relevance for heart disease development. Possibly merging two separate datasets for better usage, but keeping making most usage of available resources over the main dataset.

- **Quality check**: Evaluate the quality of the chosen dataset, and check for completeness (any missing values), accuracy, and data consistency. It was performed an Exploratory Data Analysis (EDA) to do the proper check. After that, also implemented feature engineering, see more details on the attached Jupyter notebook.

- **Data Visualization**: In this project, data visualisation played an important role, to extract complex health data. By the usage of charts, graphs, and dashboards, we illustrated patterns, trends, and correlations among various health factors/features. As an example, we made a visualisation over the distribution of risk factors, such as smoking, and physical activity, across different age groups and genders. Those visual representations enhance understanding and empower users to make informed decisions about their health, fostering a proactive approach to disease risk prevention and management.

- **Documentation:** the source of the data was inputted over the Jupyter file and include in such file any data preprocessing steps or transformations of the data applied. All code was correctly documented and explained, either through markdowns or through code commenting.

- **Relationship analysis:** the analysis of correlations between features and the target variable (heart disease) was done to identify key predictors of heart disease risk.

3. Data Preparation

- **Cleaning:** data cleaning was performed to handle missing values, which in the analysis, there were none, outliers, and duplicated values were also checked to ensure data quality. As previously mentioned, it was also performed an exploratory data analysis.

- **Transformation:** Scaling numerical features to a consistent range using methods such as Min-Max scaling and standardisation was performed.

- **Integration:** Checking and combining relevant datasets, to create a comprehensive dataset for heart disease risk prediction, to be analysed in further steps.

- **Splitting:** dividing the dataset into training, validation, and test sets to ensure that each set is representative of the overall data distribution.

4. Modelling

Ref.[2]

- **Technique Selection:** the selection of a suitable machine learning algorithm for binary classification tasks, such as evaluated (Logistic Regression, Decision Trees, and Random Forests). We finally chose the *Logistic Regression Model*, considering the task of predicting disease risk, the model's ability to correctly identify positive cases (heart disease) is crucial.

  In this context, the logistic regression model outperforms the other models in terms of precision, recall, and F1-score for the 'yes' class. It achieves a balanced performance in predicting both positive and negative cases, with a relatively high accuracy and ROC (Receiver Operating Characteristic Curve) AUC score (Area Under the Curve) which measures the model's ability to distinguish between positive and negative classes, with a relatively high accuracy and ROC AUC score, therefore, based on the findings, the logistic regression model appears to be the most suitable for the case studied.

  - **Logistic Regression**:

    Ref.[16] ("What is Logistic Regression?")

  Logistic regression is a statistical model used for binary classification tasks, where the result of the variable has two possible outcomes. In this specific case, we made use of such to check the feature/variable 'HeartDisease', where the dataset output 'yes' and 'no' answer, therefore, making possible the usage of binary classification.

  In the context of our heath tracking and disease risk prediction application, logistic regression is suitable because it allows us to predict the likelihood of being diagnosed with heart diseases based on lifestyle factors and health metrics, which aligns with our project objectives.

  Moreover, logistic regression is interpretable, which makes it easier to understand the relationships between predictor variables and likelihood of heart disease development.

  Therefore, providing actionable insights into the users health risks, which enables them to take proactive measures for disease prevention and early intervention. Additionally, this model is computationally efficient and scales well with large datasets, making it suitable for real-time prediction tasks in our application.

  - **Decision Tree**:

  Ref.[17]

  Decision trees were suitable for our health tracking and disease risk prediction application due to its recursively split the data into subsets based on the feature that best separates the data. Which created a tree-like structure decision, with possible consequences.

  Given our diverse dataset with both categorical and numerical features, decision trees could effectively capture complex relationships between lifestyle habits, health metrics, and disease risk. The understanding of decision trees also aligns with our goal of providing actionable insights to the uses, which allows them to understand the factors influencing their health outcomes.

  Additionally, to handle any missing values well, especially robust to outliers, that stage was addressed in our data pre-processing. Overall the decision trees offered a versatile approach to the application.

  - **Random Forest:**

    Ref.[18]

  Random forest is a learning method that operates by constructing decision trees during training and outputting the mode of the classes or mean prediction of the

individual trees, single result. In this project, this model was a suitable choice for predicting disease risk based on lifestyle factors and health metrics.

Random forest ability to handle both numerical and categorical features/data, handle missing values, and reduce overfitting, was the advantage. Moreover, Random forest's inherent feature importance evaluation can provide insights into the most influential factors, which contributes to disease risk, aiding in actionable insights for preventive care. Also taking into consideration the diversity of the dataset, makes random forest even more suitable for the task of predictive performance.

- **Model development:** the training of the predictive models using the prepared dataset, focused on predicting the likelihood of being diagnosed with heart diseases based on lifestyle factors and pre-processed health metrics. The modelling of logistic regression performance was considered satisfactory in terms of accuracy, achieving 91.61%. However, the model showed limitations in correctly predicting positive cases, as indicated by the low recall and F1-score for the 'yes' class, reflecting its poor ability to detect true positives.

- **Optimization of model:** Fine-tune model hyperparameters and evaluate different feature selection techniques to improve model performance, as well as, maybe dive deeper into another tested modelling might be an option.

- **Validation:** Assess model performance using evaluation metrics, such as accuracy, since data loss was already done over the previous steps, precision was another metric evaluated, as well as recall, and F1-score on the validation set. (F1-score: mean of precision and recall). The findings over the validation phase, found that the logistic regression model achieved an accuracy of 91.61%. However, it exhibited low recall and F1-score for the positive class, indicating its limitation in correctly identifying instances of the target disease (heart disease). This suggested the need for further evaluation and potentially the exploration of alternative models, as it was done in the above step.

- **Evaluation**

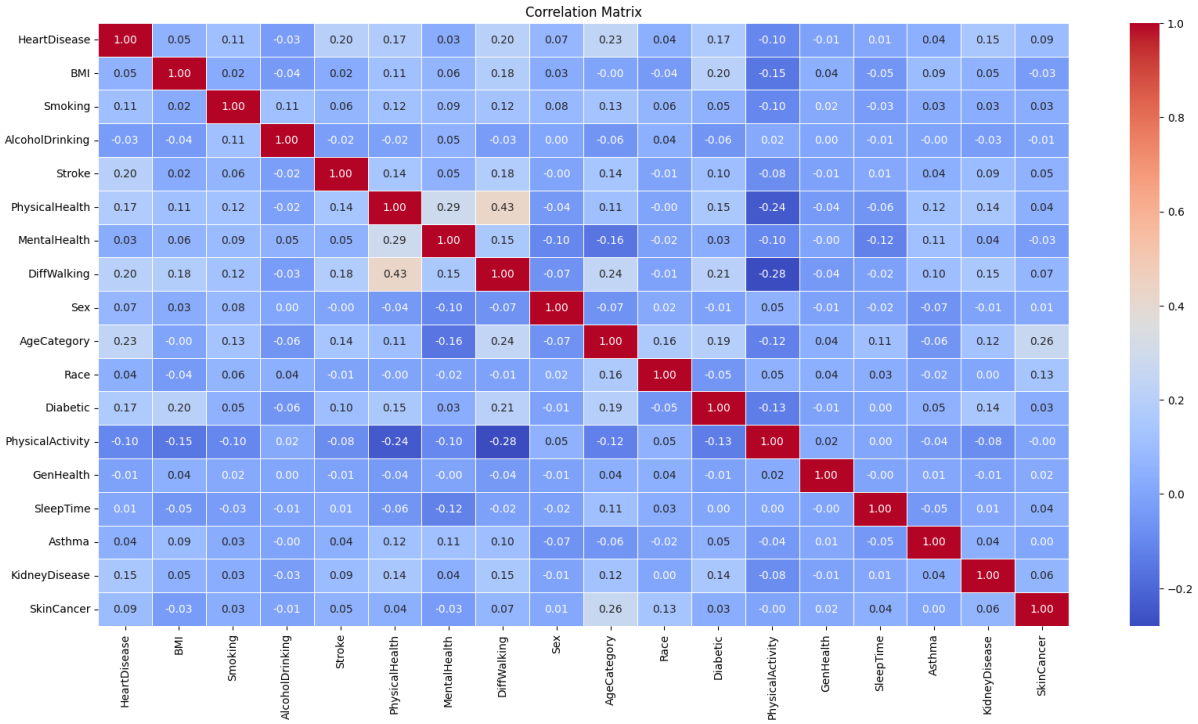
  Ref.[19]

- **Model performance:** evaluation the effectiveness of the predictive models in accurately predicting heart disease risk, was done considering factors such as sensitivity, and specificity. The model chosen was done according to the performance of the ROC AUC score, which provided a comprehensive evaluation of the model's ability to distinguish between the positive and negative classes. The metric considers both sensitivity and specificity, making it suitable for imbalance datasets like the one in this project.

  See below **correlation matrix** generated from our Jupyter notebook code:

  Ref.[8]

  First let's start with why is a correlation matrix important? Because it summarises a large dataset to identify patterns and make decisions according to it. In this case, creating a correlation matrix, mainly shows the attributes related to heart disease risk factors, serves to identify relationships between variables.



The observation of such we can reveal correlation between factors such as smoking, alcohol consumption, physical activity, and heart disease incidence. Understanding those relationships aids in identifying key predictors and informing feature selection for predictive modelling.

Classification Report for Logistic Regression:

	precision	***recall***	***f1-score***	***support***
No	0.92	0.99	0.96	58530
Yes	0.53	0.10	0.17	5429
accuracy			0.92	63959
macro avg	0.73	0.55	0.57	63959
weighted avg	0.89	0.92	0.89	63959

Source: output of jupyter notebook code visualisation. See notebook attached for more details about coding output.

- **Interpretation:** Analyse the implications of the model predictions for heart disease prevention and early detection strategies, identifying actionable insights for healthcare decision-making, such as early detection or disease prevention.
- **Project alignment with its goals:** To ensure that the project meets the initial objectives of developing a prototype for personalised health tracking and heart disease prediction.
- **Feedback:** gathering feedback from stakeholders and end-users to refine the application and improve its usability and effectiveness, would be essential for the project growth.
- **Legal and ethical issues:** When handling sensitive data, it is important to gather the user content from the data collection point. It is important to mitigate

bias on the predictive algorithms that might happen from that point of view, as well as ensure the model fairness in predicting. Addressing such issues is crucial to build customer trust with users and stakeholders and ensure a responsible deployment of the application.

5. Deployment

- **Prototype implementation:** Develop a prototype of the health tracking application, integrating the predictive models for heart disease risk prediction. In further deployment focused on the application improvement, it is possible to consider adding a personalised output for user lifestyle and health improvement, in order to change such prediction rates.

  In anticipation of further deployment and application refinement, consideration was given to incorporating personalised output, aimed at fostering user lifestyle and health improvements. This interactive approach underscores the project commitment to continuous enhancement, and user-centric design, aiming to empower individuals with actionable insights to mitigate heart disease risks and promote overall well-being.

- **Testing:** Was conducted extensive testing for the application to ensure its functionality, and reliability.

- **User interaction:** Should be done through user feedback through simulated usage scenarios or user testing sessions to identify areas of improvement, but the user interaction should be first done through the form upload, and then after through the output PDF with lifestyle and health improvement recommendations, along with General Practice (GP) doctor appointment to explain further details and guide the user to better predictions.

- **Iteration:** based on the user feedback and performance evaluation, defining features and some visual contents and recommendations could be another improvement done over the deployment phase. User interaction and feedback sessions will play a crucial role in identifying areas for enhancement, such as optimising user interface elements, data input processes, and providing more intuitive health insights.

  By interactively incorporating user feedback and evaluating performance metrics, we can ensure that the application evolves to meet the changing needs and preferences of both users and healthcare professionals, ultimately enhancing its effectiveness and user satisfaction.

- **Documentation:** deployment process was documented, including technical considerations, challenges that we might have faced, and solutions implemented, were both highlighted over this report, and over the Jupyter code file.

  The details of this project were comprehensively documented within the project report and integrated into the Jupyter code file, ensuring transparency and accessibility for further reference. By thoroughly documenting the deployment process, including both successes and obstacles, valuable insights are provided for stakeholders and collaborators, which facilitate knowledge transfer and enable informed decision-making for future interactions or similar projects within the healthcare domain.

- **Legal and ethical issues**: as we will be handling user personal and health gathered data, it is important to make the application compliant with the laws and regulations of the region implemented, such as GDPR (General Data Protection Regulation). It is also our responsibility to store the data in a secure way and make the user aware of the processing of such data for better, personalised information about their health outcome. (“Ethical Issues Related to Data Privacy and Security: Why We Must Balance Ethical and Legal Requirements in the Connected World”)

Resource and risk Analysis (Cost and Benefits)

Ref.[20]

Performing a resource and risk analysis for the project, involves estimating the resources and conducting risk analysis. We considered the costs and benefits associated with each phase. Resource allocation encompassed personnel time, software tools, computational resources. We weighed the benefits of developing a personalised health tracking and disease prediction application against potential risks such as data security breaches and regulatory compliance challenges.

Furthermore, a thorough risk analysis identified potential obstacles such as data privacy concerns, model performance limitations, and user adoption challenges. To mitigate those, strategies might be further developed to address these risks, including encryption protocols for data security, continuous model refinement, and user interface optimization for improved usability. Proactively addressing those recognized risks, the project might be better at achieving its objectives.

In addition to the risk mitigation plan, careful resource management would be advisable to ensure a better utilisation of the available resources, maximising efficiency through the project lifecycle. By monitoring closely, and adjusting as necessary, we can maintain a cost-effective application, while delivering high-quality outcomes. This balanced approach to resource and risk management positions the project for a better achievement and overarching its goals.

Despite the investment required, the anticipated benefits include improved health outcomes, cost savings in healthcare, potential governamental partnership, and enhanced patient engagement, understating the project's potential value in transforming healthcare delivery.

Modelling performance evaluation - Diving deeper in the results

Now diving deeper into the analysis of each model’s performance:

1. Logistic Regression:
    - Accuracy of 91.61%, representing the percentage of correctly classified instances out of the total instances.
    - Cross validation score of 83.99%, the average accuracy obtained in this step, indicates how well the model generalised to unseen data.
    - ROC AUC score of 84.29%, the area under the receiver operating characteristics curve, measured the model ability to distinguish between positive and negative classes.
    - Classification Report, provided precision, recall, and F1-score for both classes (yes and no), see below the output table.

Classes	Precision	Recall	F1-score	Support
No	0.92	0.99	0.96	58530
Yes	0.53	0.10	0.17	5429

2. Decision Tree:
    - With accuracy of 91.54%, similar to logistic regression, it presents the percentage of correctly classified instances.
    - Cross validation Score of 72.86%, lower compared to logistic regression, which might indicate potential overfitting or poor generalisation.
    - ROC AUC score of 73.67%, lower than logistic regression, suggesting less ability to discriminate between classes.
    - See below the classification report for this model.

Classes	Precision	Recall	F1-score	Support
No	0.92	1.00	0.96	58530
Yes	0.53	0.03	0.06	5429

3. Random Forest:
    - Accuracy of 91.52%, close to logistic regression, indicating a good overall performance.
    - Cross Validation Score of 80.28%, higher than decision tree but lower than logistic regression.
    - ROC AUC score of 80.19%, also higher than decision tree, but still lower than logistic regression.
    - See below classification report for this model.

Classes	Precision	Recall	F1-score	Support
No	0.92	1.00	0.96	58530
Yes	0.59	0.00	0.01	5429

Logistic regression demonstrated better performance at correctly identifying individuals with heart disease, minimising both false positives and false negatives. Additionally, Logistic regression offered several advantages suitable for our application, which includes simplicity, interpretability, and efficiency in handling binary classification tasks.

Its probabilistic nature allows for straightforward interpretation of results, providing clinicians and users with actionable insights for preventive care and

early intervention. Furthermore, logistic regression's computational efficiency enables real-time prediction, which is essential for seamless integration into our health tracking application.

Overall, logistic regression was the optimal choice for our deployment, aligning with our goal of developing an accurate, interpretable, and user-friendly model to empower individuals in managing their heart health effectively.

To summarise, logistic regression outperforms the other models in terms of precision, recall, and F1-score, for the positive class (heart diagnosed disease), therefore the model chosen for the deployment of the application was such.

It demonstrates a better balance between correctly identifying positive cases, while minimising false positives, making logistic regression the most suitable model in this case of predicting disease risk in the context of the provided data.

Diving deeper into prototype deployment phase

The development phase, explained deeper. The prototype involved integrating the selected logistic regression model into the health tracking and disease prediction application. This model was chosen due to its good performance predicting the likelihood of developing heart disease, as demonstrated by its high accuracy and ROC AUC score during validation.

The application allows users like 'Laura' to input their daily habits and health metrics, such as smoking habits, alcohol consumption, physical activity, and other features. Based on this input, the logistic regression model generates predictions about the likelihood of the user/patient of developing heart disease.

The main idea is to allow the user/patient over the first application interaction to submit their own personal data and health data to be analysed and therefore a personalised output will be done for her/him. But as explained before, to reach this step of the deployment, the application might face some legal and ethical steps to make it fully compliant with the law and regulations of the country deployed, as the given example of the GDPR (General Data Protection Regulation). As the application will be handling gathering, storing, and processing both personal data, and health data, which requires extra security features.

Therefore, for testing purposes on the deployment phase, we will be using an fictional dataset to be making predictions, as if the user had input those, so then there will be the integration of a separated dataset, where the fictional data will be stored, so each time that the predictions will be made, we will be able to gather a new output, due to the random selection of the user over the

‘user_data’ dataset, which is where the information will be gathered, rather than the actual submission of a form, as said over the final prototype project.

The insights extracted from this analysis can be invaluable for both users and healthcare professionals. For users, the application provides personalised insights into their health status and identifies potential risk factors for heart disease, empowering them to make informed decisions about lifestyle choices.

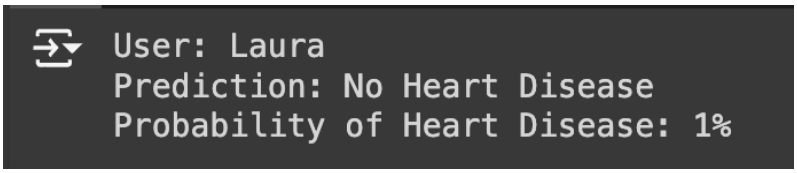
For healthcare professionals, the model generates actionable insights into their health status that can aid in early detection and intervention strategies for heart disease.

There will be also the need to direct the user for a healthcare professional, to make a guided analysis of the results for the user, such guidance is essential and in further steps must be integrated to the application, for questions of health safety, the input of a healthcare professional is essential.

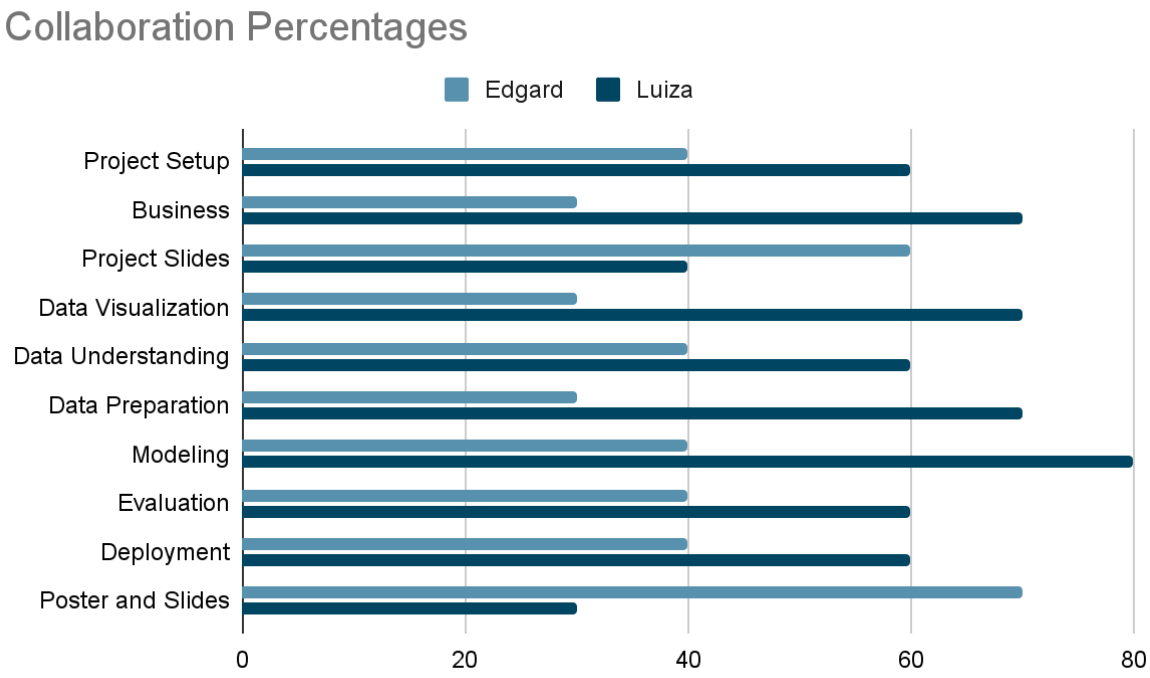
Overall, the development of this prototype, represents a significant step towards leveraging machine learning and data analytics for proactive health management and disease prevention.

As it is possible to see below, the result of the model that ran, we can see the prediction done to ‘Laura’, one of the random selected users, which indicates a low probability of heart disease, which was tested several times, and checked with the original dataset ‘user_data’ provided.

Where we can assure that the model was running as it should, Laura in our original dataset, had no heart disease, did not smoke, or drank alcohol, in general, lives a healthy lifestyle. Exemplifies how the model can provide reassurance and guidance to individuals seeking to maintain their cardiovascular health.



Roles and Responsibilities - division by team members



Luiza

In this project, I played several roles, starting from the initial stages of collaboration and integration. I established the main communication tool and selected the working tools and programming languages essential for the development of the project and collaboration.

During the business understanding phase, I had the main idea of the project approach, gathered requirements, defined the objectives, and facilitated the stakeholder needs with Edgard's assistance. Transitioning into the data understanding phase, I led efforts to explore potential datasets, checked for data quality, and selected suitable sources of information, as well as performed data visualisation steps and data engineering, before starting the modelling phase.

Additionally I took charge of the data preparation process, optimising efficiency by performing data cleaning within the data gathering code snippet, as well as performing a well detailed EDA (Exploratory Data Analysis), see details over the attached jupyter notebook file.

Utilising a predictive classification model, was also my initiative, as also detailed over the Jupyter notebook snippet, I opted to perform 3 different modelling, so we could choose the best outcome for our prototype. The idea to generate the running prototype prediction was also done by me, as a more tangible output for the project. Adaptations from the original project had to be done along the way, for legal, and ethical reasons, explained in previous steps.

Edgard

In this project, I worked on a key role in various stages. I collaborated with Luiza to establish communication channels and select the appropriate tools and programming languages. During the business understanding phase, I assisted in gathering requirements and defining objectives.

In the data understanding phase, I explored potential datasets, checked data quality, and contributed to data visualisation and engineering. I was actively involved in data preparation, performing data cleaning, transformation, and Exploratory Data Analysis (EDA).

I collaborated with Luiza to implement and evaluate three predictive classification models, ensuring we selected the best outcome for our prototype. Additionally, I checked the code to fix errors and optimise performance.

I was also responsible for creating the poster and slides for our presentation. Finally, I helped adapt the project to meet legal and ethical requirements, ensuring data integrity and compliance with relevant regulations.

Conclusion

The project goal was to develop a health tracking and disease prevention application, taking into consideration the integration with machine learning for personalised insights and predictive analytics. Through collaboration and usage of the CRISP-DM framework, we organised and addressed data understanding, preparation, modelling, evaluation and deployment phases. The challenges we faced included data quality assurance and feature selection, but despite that we demonstrated effective team work and delivered a prototype with potential for health management improvement.

In conclusion, this project delivered an innovative health tracking and disease prediction application, leveraging machine learning and data analytics. Through collaboration, we made a comprehensive planning, and model evaluation to develop a prototype aligned with healthcare industry needs. With further engagement, this solution promises to empower users and revolutionise preventative healthcare practices.

References

1. Six Steps in CRISP DM - The Standard Data Mining Process. Pro Global Business Solutions (PGBS). Available from: https://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/ [accessed 01 May 2024].

2. Sharma, N. (2023). Understanding and Applying F1 Score: AI Evaluation Essentials with Hands-On Coding Example. Arize. Available from: https://arize.com/blog-course/f1-score/ [accessed 11 May 2024].

3. Appendices. Oxford Brookes University. Available from: https://www.brookes.ac.uk/students/academic-development/online-resources/appendices#:~:text=An%20appendix**%20comes%20at,main%20body%20of%20your%20work. [accessed 12 May 2024].

4. Bhor, Y. (2024). Guide for building an End-toEnd logistic regression model. analyticsvidhya.com. Available from: https://www.analyticsvidhya.com/blog/2021/09/guide-for-building-an-end-to-end-logistic-regression-model/ [accessed 13 May 2024].

5. Navlani, A. (2023). Decision tree classification in python tutorial. datecamp.com. Available from: https://www.datacamp.com/tutorial/decision-tree-classification-python [accessed 13 May 2024].

6. Shafi, A. (2023). Random Forest Classification with Scikit-Learn. datacamp.com. Available from: https://www.datacamp.com/tutorial/random-forests-classifier-python [accessed 14 May 2024].

7. Hadhrami, T., Abdullah, S., Eid, H. (2023). Utilising random forest algorithms for early detection of academic underperformance in open learning environments. ncbi.nlm.nih.gov. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10703007/ [accessed 15 May 2024].

8. Deshpande, T. (2022). Heart Failure Prediction: CV Score(90%+) | 5 Models. Kaggle.com. Available from:https://www.kaggle.com/code/tanmay111999/heart-failure-prediction-cv-score-90-5-models [accessed 15 May 2024].

9. Rehkopf, M. Scrum sprints. atlassian.com. Available from: https://www.atlassian.com/agile/scrum/sprints [accessed 15 May 2024].

10. Pan, A. (2022). A Gentle Introduction to Machine Learning Models. wandb.ai. Available from: https://wandb.ai/wandb_fc/gentle-intros/reports/A-Gentle-Introduction-to-Machine-Learning-Models--VmlldzoyOTUxNjQw [accessed 15 May 2024].

11. What is Machine Learning in Healthcare? coursera.org. Available from: https://www.coursera.org/articles/machine-learning-in-health-care [accessed 15 May 2024].

12. Investigating with databases: Verifying data quality. datajournalism.com. Available from: https://datajournalism.com/read/handbook/verification-2/5-investigating-with-databases-verifying-data-quality [accessed 15 May 2024].

13. OZair, F. F., Jamshed, N., Sharma, A., Aggarwal, P. (2015). Ethical issues in electronic health records: a general overview. ncbi.nlm.nih.gov. Available

from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4394583/ [accessed 15 May 2024].

14. Ethical Issues Related to Data Privacy and Security: Why We Must Balance Ethical and Legal Requirements in the Connected World. digitalprovacy.ieee.org. Available from: https://digitalprivacy.ieee.org/publications/topics/ethical-issues-related-to-data-privacy-and-security-why-we-must-balance-ethical-and-legal-requirements-in-the-connected-world [accessed 15 May 2024].

15. Riddell, C. (2024). Data Security Explained: Challenges and Solutions. blog.netwrix.com. Available from: https://blog.netwrix.com/data-security/ [accessed 15 May 2024].

16. What is Logistic Regression? statisticssolutions.com. Available from: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/ [accessed 16 May 2024].

17. (2023). Decision Tree. geeksforgeeks.org. Available from: https://www.geeksforgeeks.org/decision-tree/ [accessed 16 May 2024].

18. Donges, N. (2024). Random Forest: A Complete Guide for Machine Learning. builtin.com. Available from: https://builtin.com/data-science/random-forest-algorithm [accessed 16 May 2024].

19. Jordan, J. (2017). Evaluating a machine learning model. jeremyjordan.com. Available from: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/ [accessed 16 May 2024].

20. Assessing resources and risks. thensmc.com. Available from: https://www.thensmc.com/content/assessing-resources-and-risks-1 [accessed 16 May 2024].

21. Heart disease. mayoclinic.org. Available from: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118#:~:text=Heart%20disease%20describes%20a%20range,born%20with%20(congenital%20heart%20defects) [accessed 15 May 2024].

Appendix:

Ref.[3]

Group Reflection

During our group reflection, we acknowledged the dataset's strengths in capturing variable health factors but identified gaps in family medical history data. We emphasised the importance of ensuring data consistency and accuracy, addressing outliers was prioritised for improved analysis and model performance. In this comprehensive project reflection, we can highlight the successful collaboration across phases, emphasising effective communication and task delegation.

We recognize the significance of adhering to the project plan's timeline, ensuring each phase received appropriate attention. Feedback loops were mainly for improving and delivering a better prototype despite constraints. We also added legal and ethical concerns as there will be gathering and processing of health and personal data, and how there will be a need to make the application compliant with such regulations.

Evidence of group work

The first proposal of evidence of group work would be our collaboration link, which is the following GitHub repository: https://github.com/luizaalbuquerque/Strategic-Analysis.git. Second, we have the google docs, collaboration link:https://docs.google.com/document/d/1SyoBxgqPGcTNsxcVqMHrGYCPiW-skzTKQfs1BsvtNGk/edit?usp=sharing. Where both collaborated on the written project, see above table for the breakdown of the tasks assigned. As well as regular meetings over google meet were performed, and face-to-face meetings to proceed with collaboration were also done.

The project was executed through collaborative efforts, with team members' contribution in various phases, showcasing their own expertise. During the collaboration process, we opted to change the collaborative tool, from anaconda platform to google colab platform, where we could both edit in real time the same file, and keep updating it without further challenges, since we were having trouble in keeping track of the versions that needed to be uploaded over the previous collaboration tool. The link for such file was:https://colab.research.google.com/drive/1QswzXC4OyWmo8hf444cmgF1IN2hcpBiP?usp=sharing.

Evidence of group work included above explained share responsibilities, in data exploration, model development, and validation. Regular meetings, task delegation, and shared documentation also proves effective collaboration, which ensures a comprehensive and cohesive outcome of the project.

Individual Contribution Report

Luiza

Through the project interaction, I have covered several aspects of the application, which was focused on health tracking and disease risk prediction. Starting from the project plan. I have outlined the key phases of work, such as business understanding, data exploration (focused on data visualisation and preparation, the exploratory data analysis (EDA) performance).

With my pair, I discussed the roles and requirements, which involved gathering information, data exploration, model development and others. By making use of the CRISP-DM framework, I focused over the business objectives, stakeholder needs, and user requirements, data understanding aspect, involved exploration, quality checks, and documentation. I have addressed some concerts such as data consistency and accuracy.

In modelling, I have discussed with my pair some technique selections, research, model development, optimization, and validation. Over the Evaluation phase, which included addressing the model performance, we had a few troubles with coding the correct sets, and also updating libraries to be used, but also implications regarding healthcare decision-making, such as which feature would be important for analysis.

Deployment, I focused over the prototype implementation, and also highlighted my pair reflection, over the dataset strengths, and effectiveness of the group collaboration to adhere to the project timelines. To conclude, we added relevant references, citations and acknowledgements of shared efforts and contributions. The idea to generate the running prototype prediction was also done by me, as a more tangible output for the project. Adaptations from the original project had to be done along the way, for legal, and ethical reasons, explained in previous steps.

Edgard

Throughout the project, I actively contributed to various aspects of the application focused on health tracking and heart disease risk prediction. Starting with the project plan, I collaborated closely with Luiza to outline the key phases of work, including business understanding, data exploration, data visualisation and preparation, and the performance of exploratory data analysis (EDA).

Together, we discussed roles and requirements, involving the gathering of information, data exploration, model development, and other tasks. By utilising the CRISP-DM framework, I concentrated on business objectives, stakeholder needs, user requirements, and data understanding, which involved exploration, quality checks, and documentation. I addressed concerns such as data consistency and accuracy.

In the modelling phase, I worked with Luiza to select appropriate techniques, conduct research, develop models, optimise, and validate them. During the evaluation phase,

we encountered some challenges with coding the correct sets and updating libraries. We also considered the implications for healthcare decision-making, identifying important features for analysis.

For deployment, I focused on implementing the prototype and reflected on the strengths of the dataset and the effectiveness of our collaboration to adhere to project timelines. Additionally, I was responsible for creating the poster and slides for our presentation. I ensured that our references, citations, and acknowledgements of shared efforts and contributions were properly included.

The idea to generate a running prototype prediction was a joint effort, providing a tangible output for the project. Along the way, I made necessary adaptations to the original project to meet legal and ethical requirements, ensuring compliance with relevant regulations.

Lastly, I took the lead in checking and fixing code errors to optimise performance, ensuring the integrity and reliability of our application. Through our effective collaboration and comprehensive approach, we successfully developed a prototype with significant potential for health management improvement.