INFO411 ASSIGNMENT 2 REPORT

Authors: Isaiah Grant, Chris Yang, Jason Lu

Reflection and future improvements:

Balancing our diverse schedules was our main challenge for this assignment. Chris works full time and has family commitments, while Isaiah and Jason study full-time but on different courses, thus making communication and schedule coordination vital to our assignment. Furthermore, Chris is proficient with tools like GitHub, whereas Isaiah has limited exposure, so he could not run notebooks from Chris's GitHub on his Julia localhost.

Despite these challenges, we established a routine, convening for team meetings at 9 p.m. twice weekly and sharing screens to ensure we were all on the same page.

Our assignment benefitted from role delegation:

- Jason focused on EDA and dashboard creation.
- Chris handled EDA, data imputation, modelling, and the dashboard.
- Isaiah took on EDA, data imputation, modelling, and report writing.

For future projects, having individual GitHub accounts with identical notebook versions and increasing communication frequency would be beneficial.

This collaborative assignment has served as a catalyst for a deeper understanding of effective teamwork dynamics, highlighting the significance of clear communication, adaptable strategies, and a cohesive approach to achieving shared goals. Our collective experience has underscored the importance of leveraging individual strengths and fostering an environment that encourages open dialogue and innovative problem-solving. Through this process, we have gained practical insights into managing timelines, optimizing resource allocation, and navigating through unexpected challenges, ultimately reinforcing the value of synergy and a unified vision within a team. Furthermore, we have gained a deeper appreciation and understanding of machine learning techniques:

- Machine learning is a complex and multifaceted field that demands a blend of practical expertise and a sound theoretical understanding. The completion of this research paper marks the initiation of our journey into this expansive domain. To further enrich our comprehension and proficiency, dedicated practice and continuous learning from upcoming studies and practical experiences are imperative.
- Effective communication of the model outcomes to diverse audiences, irrespective of their technical acumen, is an essential skill for any machine learning engineer. Additionally, the development of a robust logical framework for machine learning, facilitating the identification of areas for troubleshooting and model enhancement, holds paramount significance.
- Embracing the concept of lifelong learning is fundamental in this trajectory. It serves as a guiding principle, ensuring the evolution and refinement of skills, insights, and methodologies throughout the course of our careers. This journal marks merely the inception of what promises to be an enriching and enlightening professional voyage.

Task 1 – Exploratory data analysis (EDA):

During our exploratory data analysis (EDA) of the DS1 heart disease dataset, we analysed a set of attributes to predict the absence (1) or presence (2) of heart disease.
The Statlog (Heart) dataset by the UC Irvine Machine Learning Repository (DS1) includes the following key attributes:

1. *Age (Real)*
2. *Sex (Binary: Male/Female)*
3. *Chest pain type (cp) (Nominal: 4 types)*
4. *Resting blood pressure (trestbps) (Real)*
5. *Serum cholesterol in mg/dl (chol) (Real)*
6. *Fasting blood sugar (fbs) > 120 mg/dl (Binary)*
7. *Resting electrocardiographic results (restecg) (Nominal: values 0,1,2)*
8. *Maximum heart rate achieved (thalach) (Real)*
9. *Exercise-induced angina (exang) (Binary)*
10. *Oldpeak: ST depression induced by exercise relative to rest (Real)*
11. *Slope of the peak exercise ST segment (Ordered)*
12. *Number of major vessels (0-3) colored by fluoroscopy (ca) (Real)*
13. *Thalassemia types (thal) (Nominal: 3 = normal; 6 = fixed defect; 7 = reversible defect).*

Despite forming our group later than desired, with members juggling work and study commitments, we each started individual EDAs. Once we formed our group and a cohesive routine, we combined our EDAs to derive insights.

Isaiah's pairwise histogram accentuated the distribution of heart disease across the attributes. Age displayed a typical bell-curve distribution, while resting blood pressure, serum cholesterol, and oldpeak showed a rightward skew. In contrast, the maximum heart rate achieved leaned leftward.

Chris's pair plot illustrated inter-attribute relationships, revealing patterns like increasing age correlating with higher resting blood pressure and serum cholesterol. The upper triangle of the pair plot provides a binned representation of these attributes' relationships in DS1. In contrast, the lower triangle of the plot shows correlation line plots. To summarise the lower triangle correlations, the age distribution is roughly bell-shaped, peaking around 55-60 years; resting blood pressure, 'trestbps', largely falls between 120-140 mmHg. Additionally, cholesterol levels for most individuals rest between 200 and 300, the maximum heart rate achieved, 'thalach', predominantly ranges from 150 to 175 beats per minute, and 'Oldpeak', which represents ST depression induced by exercise relative to rest, is mainly near 0.

One insight is that, as age increases, resting blood pressure and cholesterol tend to rise, whereas the maximum heart rate achieved decreases. Moreover, a clear negative correlation between 'thalach' and 'oldpeak' indicates that as the heart rate rises, the oldpeak value generally drops. The other correlations are either weak or not distinctly observable.

Our final visualisation was a correlation matrix by Isaiah that concentrated on the non-categorical attributes. Insights include the intricate connections between attributes, emphasising the importance of a multi-dimensional heart health analysis. The most pronounced positive correlations were between age and trestbps (0.273), followed by trestbps and oldpeak (0.223), age and chol (0.22), chol and trestbps (0.173), and age and oldpeak (0.194). On the other hand, notable negative correlations were observed between age and thalach (-0.402) and between thalach and oldpeak (-

0.349). These correlations revealed that individuals tended to exhibit higher resting blood pressure and cholesterol levels as they age, and their maximum heart rate capability diminishes. The heatmap provides insight into the interrelationships between various cardiovascular metrics, conveying the need for a comprehensive perspective when analysing heart health indicators.

Task 2 – Data Imputation

We chose to impute the following three processed location datasets: processed.hungarian, processed.switzerland and processed.va. As our portfolio shows, when we open the tables for each of these datasets in Julia, we can see the attributes but also 'missing' in the datasets, representing missing data. This is what we aimed to impute for this task. Isaiah attempted to impute these missing data to be replaced with the mode of each column, while Jason and Chris each found KNN to be better as k replaced the missing values better As K-Nearest Neighbours (KNN) can be effective for data imputation on missing values, particularly when dealing with categorical data, due to its inherent ability to handle similarity measurements and pattern recognition. Moreover, each of us agreed that dropping any record with missing values across the columns is easier too. This works as "missing values are imputed based on the values of the attributes of the k most similar instances" (Sim et al., 2015, p. 3). Moreover, each of us agreed that dropping any record with missing values across the columns is easier too.

However, it is not as simple as simply removing data with missing values. Upon deeper inspection into the processed Hungarian and VA datasets, we see they have only one valid value. Further analysis into the missing data for all three datasets, we found that the ca column would be good to drop because there is a very high missing ratio in each dataset. For example, ca in the processed Hungarian dataset has a 98.98% missing rate with only 3 valid data, ca in the processed via dataset has a 99.0% missing rate with only 2 valid data, and ca in the processed Switzerland dataset has a 95.3% missing rate with only 5 valid data. So the ca column in each dataset can be dropped due to such high rates of missing data, but to resolve this high ratio, Chris and Jason came to the conclusion that using K-Nearest Neighbours (KNN) is appropriate during the imputation process due to this method's ability to replace the missing values with a k value, as opposed to Isaiah's attempt at replacing with a mode value. We all agreed the KNN method was the best imputation method.

Furthermore, Isaiah discussed with Jeremiah about how subsequent modelling produced low testing accuracies from the KNN-imputed datasets. Jeremiah recommended our group try the BetaML random forest imputation package. The imputation on each dataset was performed seamlessly as the missing values in each dataset was replaced with a new integer value or sometimes also left blank.

Isaiah shared his discussion within the team, and subsequent modelling performed by Chris using the BetaML random forest-imputed datasets revealed consistently low accuracy tests on the plots, similar to our KNN-imputed test accuracies, but just below. However, the team still agreed to use BetaML's Random Forest imputation on missing values for the following reasons:

- *Random Forests can capture complex relationships between variables, making them suitable for datasets with intricate and nonlinear patterns. This ability allows them to impute missing values more accurately, especially when the missing values are part of a complex pattern.*
- *Random Forests can handle both numerical and categorical data effectively, making them well-suited for datasets with various data types. This capability allows them to perform imputations on mixed data without requiring extensive preprocessing or data transformation.*
- *Random Forests are relatively robust to outliers, as they consider subsets of features and data points during the imputation process. This robustness helps prevent outliers from significantly influencing the imputed values, leading to more stable and reliable imputations.*

Task 3 – Modelling

Isaiah and Chris tackled the modeling phase of our assignment. Initially, we experimented with the random forest and decision tree regression algorithms. Isaiah employed the random forest regressor from the DecisionTrees package in Lab 8 of INFO411 (2023). Chris used the Decision Tree Classifier.

We examined three processed location datasets from DS2 (Janosi et al., 1988):

- Switzerland: 5 outcomes (0-4)
- Hungarian: Binary outcomes (0 or 1)
- VA: Similar to Switzerland with 5 outcomes (0-4).

We split the datasets (80%/20% for training/testing) and used the TunedModel package for optimal configurations. Chris then visualised the RMS values from different subfeatures and max depth as a heatmap to show the Switzerland dataset; however, he found the heatmap messy with no discernable patterns. Our findings for the Switzerland dataset indicated an optimal RMS of 0.91932, but testing accuracy stagnated around 0.3199, revealing that the random forest regressor was not optimal for multi-category predictions. Chris rounded floating-point predictions to the nearest integer to circumvent this, rendering them pseudo-categorical.

Isaiah's random forest regression achieved a better accuracy of 0.7159 for the Hungarian dataset due to its binary nature. With the VA dataset, our accuracy dropped to 0.25, supporting the notion that random forest regression struggled with multi-category datasets.

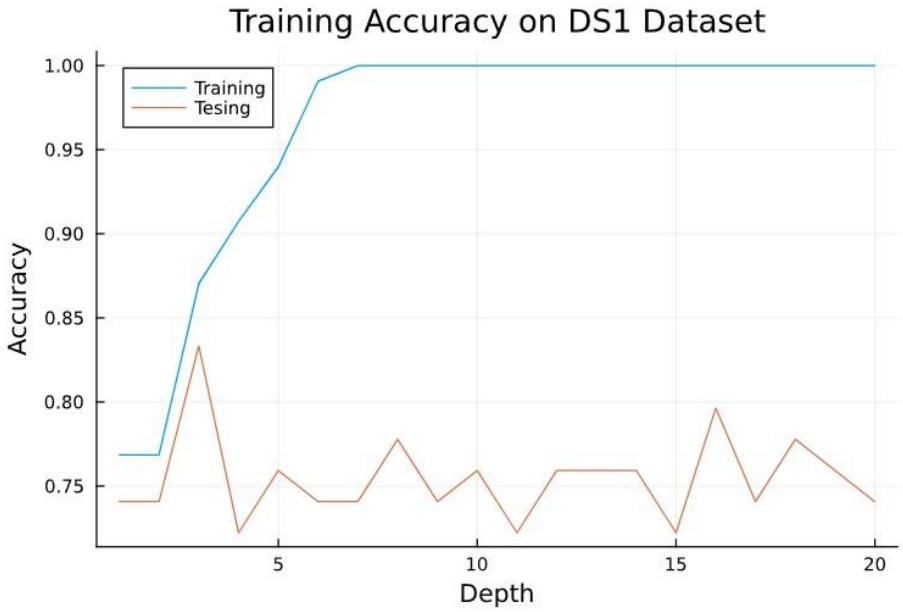
Chris tried use decision tree classification modelling, the training accuracy consistently outperformed testing accuracy across datasets. However, the decision tree classifier model demonstrated more promising results than the random forest regressor. Optimal training accuracies achieved were 86.441% (Hungarian, max_depth=6), 44.0% (Switzerland, max_depth=7), and 42.5% (VA, max_depth=8).

Given these improved accuracies, we suspect that overwriting missing values introduced biases. To address these biases, we dropped the following columns with over 40% missing data: ca, thal, slope, and fbs. However, retraining did not improve prediction accuracy. Chris the used the KNN method, with known K defined based on unique outcomes find in EDA. AS KNN is also a classification method, which shows have better performance than Regression mode, and also showed the challenges in predicting binary outcomes but had better results for categorical outputs.

Overall, our classification model outperformed our regression model. The Decision Tree and KNN Classification models notably outperformed the Random Forest Regression, especially for binary outcomes. The Decision Tree Classification model performed best for binary datasets, while KNN classification model performed best with multi-category datasets if K was known beforehand.

We deduce that the drop in accuracy for the Switzerland and Hungarian datasets is due to missing values; Tang and Ishwaran (2017) noted that “heavy missingness”, particularly over 75% missingness in datasets can result in low accuracy for random forest regressors. They also note that attributes with stronger correlation tend to have better imputation performance (2017). Given that our correlation plot in our EDA showed mostly medium correlations and a couple strong and weak correlations, Tang and Ishwaran’s observations that low to medium correlation enable RF algorithms to perform “noticeably better than the popular KNN imputation method” did not resonate with our own findings (2017, p. 16). Our KNN imputation and subsequent finding of the decision tree classifier

performing better than the decision tree random forest regressor contrasts with Tang and Ishwaran. Chris' experience with the BetaML package that Isaiah previously used to impute these datasets corroborates this insight. Using BetaML-imputed datasets, Chris produced consistently low testing accuracies, slightly lower than the KNN-imputed datasets' visualisations. Therefore, such consistently low testing accuracies between imputation methods and models indicate an inherent problem in the datasets – heavy missingness. Moreover, the consistency in low testing accuracy between the BetaML random forest imputation and the KNN imputation could be attributed to the relatedness of random forests and KNN methods, as Tang and Ishwaran observe that both are nearest neighbour methods (2017). On the other hand, the DS1 dataset had an impressive accuracy rate of about 83% under the same decision tree classification model.




```
begin
    summary_text_ds1,train_ds1, test_ds1, X_ds1, y_ds1, plot_ds1=
    find_best_DT_model(df_ds1,1:20, "DS1")
    plot_ds1;
end
```

```
The best training accuracy rate for DS1 data set is 83.333% where the max_de
pth is 3
```

Task 4 – Dashboards & business insights:

In dashboard 1, Chris and Jason have added a drop-down menu showing the Switzerland, Hungarian and VA datasets. By selecting each location, you can see how the statistics and variables from the Hungarian, Switzerland, and VA datasets compare. Furthermore, we can see that businesses have various opportunities to tailor their products and services. Considering age, the Hungarian dataset, with an average age of 47.8 years, presents an ideal demographic for businesses offering preventative healthcare measures, fitness equipment, or health tech wearables. This group might be more inclined towards proactive health measures given their relative youth. On the other end of the spectrum, the VA dataset, having a more senior-leaning average age of 59.35 years, could be a suitable demographic for businesses specialising in senior healthcare products, wellness items or tailored medications.

Furthermore, the gender distribution in these datasets also holds valuable insights. With the VA and Switzerland datasets dominated by males at 97% and 92%, respectively, they become primary targets for male-centric products or campaigns. Meanwhile, while male-skewed at 72%, the Hungarian dataset offers a more balanced demographic for products intended for a broader audience.

The datasets provide insights into cardiovascular health, which could benefit businesses in related sectors. For example, as the Hungarian and VA datasets indicate elevated blood pressure readings, products related to blood pressure management, whether devices, medications or health tech applications, could find a substantial market. Regarding cholesterol management, the Hungarian dataset stands out with its average cholesterol levels, indicating a potential market for dietary supplements or related products. Notably, the VA dataset, with about 34.5% of participants potentially having elevated fasting blood sugar levels, might be a focal audience for businesses in diabetes management tools, sugar alternatives or specific medications.

Fitness and wellness brands have a clear roadmap too. With the VA dataset indicating a broad heart rate range, heart rate monitoring devices or fitness trackers could be attractive. Moreover, given the 74% of VA participants experiencing exercise-induced angina, there is a niche for low-impact exercise equipment or programmes tailored to such needs.

Healthcare services have a robust foundation for crafting strategies. The pronounced concerns of heart disease in the Switzerland and VA datasets, averaging around 1.80488 and 1.52, convey the need for cardiac screening services or heart health campaigns in these regions. The predominance of asymptomatic chest pain types in the VA and Hungarian datasets also hints at a gap in awareness. Even without overt symptoms, campaigns emphasising the importance of routine cardiac check-ups could resonate well.

For those in the domain of medical research, these datasets could drive novel research initiatives. They offer a chance to develop innovative medical devices or champion personalised treatment methodologies. However, amidst these opportunities, there are limitations to consider. The male dominance in all three datasets means businesses should tailor their strategies accordingly or communicate the gender-specific nature of their research. Likewise, missing metrics, like the 'chol' in the Switzerland dataset, must be considered to ensure a comprehensive approach.

Overall, the dashboards for the Hungarian, Switzerland, and VA datasets are more than just numbers; they offer significant insights for businesses to understand, innovate, and cater to diverse health needs.

Dashboard 2:

In dashboard 2, Chris and Jason show a drop-down menu for each dataset’s testing accuracy visualisations from the Decision Tree Classifier model. This dashboard allows users to dynamically change the performance graph for each of the four datasets by adjusting the testing and training ratio and the maximum depth value.

Firstly, dashboard 2 illustrates the balance between model simplicity and complexity. If the model is too simplistic, then nuances are not captured, and the model is biased and underfitting. If it is too complex, it has high variance and fluctuates with any new data, risking overfitting. This balance is vital for businesses where computational time and resources are precious. Adjusting tree depth lets decision-makers balance computational costs and performance optimally. Also, the slider indicating the data's training percentage highlights the trade-off between training and testing data, offering insights into the model's robustness.

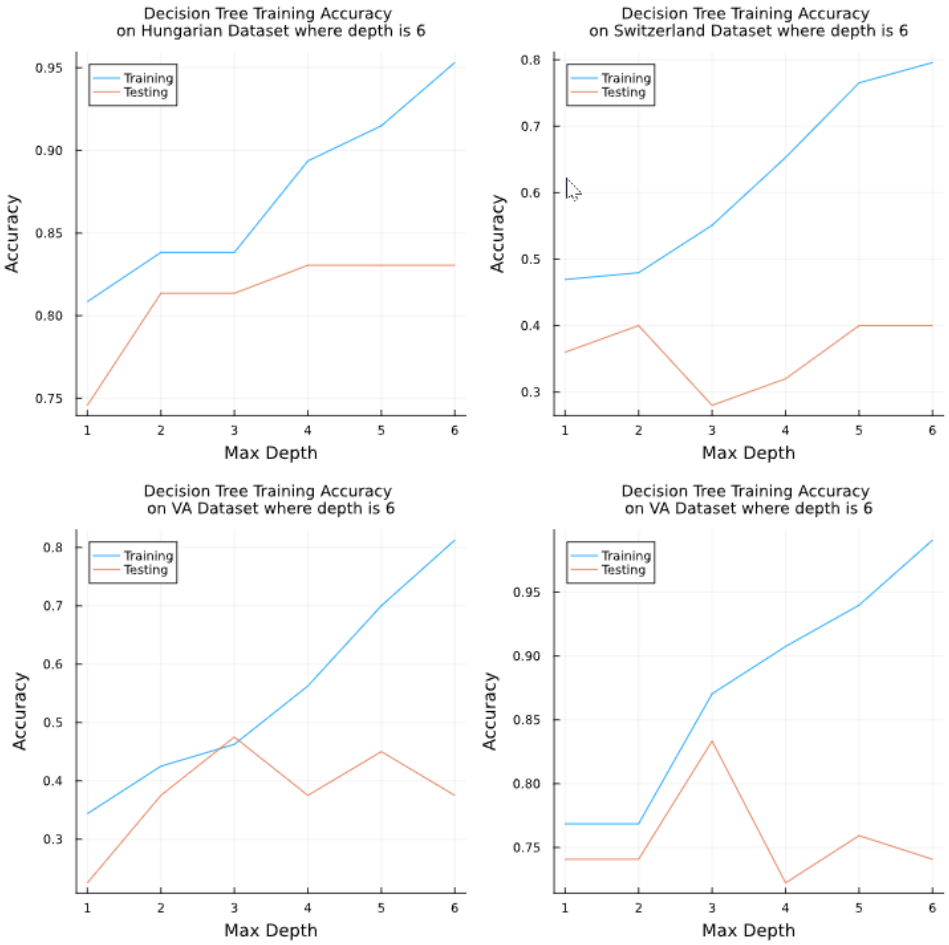
Moreover, with its interactive features, this dashboard translates abstract concepts into tangible insights, enabling stakeholders from varied backgrounds to comprehend the implications of model decisions. It paves the way for scenario-based strategic planning, answering crucial "what if" questions in real-time.

Finally, presenting how different parameters influence outcomes fosters a culture of iterative model improvement; this is especially significant due to the heavy missingness in our datasets. As new data is collected, the parameter adjustment features in the dashboard will improve the model and enable more insights. This dashboard bridges the gap between intricate machine learning concepts and tangible business objectives, empowering informed and strategic decision-making.

Of particular note is the significantly higher accuracy rate of the DS1 clean data compared to the other datasets. These datasets were trained using the same machine learning model and parameters, highlighting the significant impact of input data quality on the prediction accuracy rate.

Decision Tree Training with depth is 6 and 80% of data is used for training

```
RefValue(UUID("1a19f2c6-ce22-40f7-b53f-becfee24f04b"))
```



References:

INFO411 Lab 8. (2023).

Janosi, A., Steinbrunn, W. , Pfisterer, M., & Detrano, R. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Sim, J., Lee, J. S., & Kwon, O. (2015). Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications. *Mathematical Problems in Engineering, 2015*(538613), 1-15. https://doi.org/10.1155/2015/538613

Statlog (Heart). (n.d). UCI Machine Learning Repository. https://doi.org/10.24432/C57303.

Sylvaticus. (n.d). The BetaML.Imputation Module. GitHub repository. https://sylvaticus.github.io/BetaML.jl/stable/Imputation.html#BetaML.Imputation.RandomForestImputer

Tang, F., & Ishwaran, H. (2017). Random Forest Missing Data Algorithms. *Statistical analysis and data mining*, *10*(6), 363–377. https://doi.org/10.1002/sam.11348