Slide Number	Slide Name	Speaker	Script	
1	Title Page	Lishani	Hi everyone, our group looked into creating a cardiovascular disease predictor. In our group there was Ayroza, Ash, Myself, and Savina.	
2	Introduction	Lishani	Cardiovascular disease is a disease of the heart or blood vessels which limits blood flow to the brain, heart or body. It is the leading cause of death for men, women, and most people of minority ethnic groups in the United States. (Centers for Disease Control and Prevention 2023). During this presentation 18 people will lose their battle to cardiovascular disease.	
3	Case study: Sybil Jones	Lishani	This is Sybil Jones who is a mother to 3. Her husband, Marcus works for the Navy and therefore the family have had to relocate a lot. At one of their farewell party’s Sybil had been drinking and enjoying her friends company when the mood suddenly changed, she thought she had just drank too much but she later found out that she had experienced a stroke caused by a cardiovascular disease she didn’t even know was lurking within her. ♂♂An MRI showed a clot deep in the right side of Sybil's brain which needed to be plucked out. However, these sorts of procedures can either leave the person feeling like they never had a stroke or having to accept a new norm of reality. Thankfully, after a lot of emotional months, Sybil did eventually recover but had there’d been wide stream resource which could have detected her underlying cardiovascular disease sooner, her stroke could have been avoided.	
4	Tableau visual showing ethnic minority risk	Ayroza	We created a tableau workbook to shows mortality rate throughtout the US by state, gender and race. The image on screen shows CVD mortality by race. Ethnic groups account for 43.6% of the US population whilst the white population makes up the remaining 56.4% (Census 2022). Knowing this, it is evident to see, that cardiovascular mortality is increasingly determined by geographic location, wealth, and education. This could imply, not everyone has the luxury of modern technology, such as their smartwatch, detecting anomalies in the data it collects and reporting back to the user. More preventative measures are required to help reduce the 696 thousand deaths from heart disease in 2021 (CDC 2023). Therefore, we decided to explore creating a model which will predict cardiovascular disease diagnosis, based on risk prediction indicators.	
5	Dataset	Ayroza	The dataset used to train our CVD predictor, was sourced from the 2021, Behavioral Risk Factor Surveillance System Dataset from the centres for Disease Control and Prevention. This dataset originally contained over 300,000 records including 18 risk factors for heart disease and whether the individual had a heart disease or not.	2min 50
6	Data exploration	Ash	To familiarise ourselves with the cleaned dataset, 3 visuals were created per risk factor. 1) The overall distribution of the risk factor demographic within the full dataset - in this case how many people exercised regularly and how many did not. 2) The proportion of each risk factor present in heart disease cases - in this case, of the group of people who did have heart disease, what % exercised regularly and what % did not. 3) The prevalence of heart disease in risk factor cases to remove biases of group ratios - in this case for each exercise category, how many people had heart disease. This was done for all 18 risk factors to uncover trends in the data.	
7	Data pre-processing	Ash	Then, when it came to buiding our model, we realised the classes in our dataset were heavily imbalanced. For every 1 person who had heart disease, there were 12 who did not. This would lead to a model that may be able to recognise if a person is not at risk of heart disease, but would not be able to accurately predict if a person was at risk. To combat this, our majority class - in this case the ‘No’ category in the heart disease column - was cut down by removing all datapoints outside of 1 standard deviation from the mean. This, along with random oversampling allowed us to balance the classes in our dataset.	4min
8	Supervised machine learning models	Savina	Five machine learning models were utilised: 1) Logistic Regression is a statistical model that quantifies the relationship between multiple predictor variables (e.g., age, BMI, general health) and the binary outcome of whether an individual has CVD or not. It calculates probabilities and employs a sigmoid function to classify individuals into one of the two categories. 2)Support Vector Machines find a hyperplane in a multi-dimensional space that classifies two classes(CVD and non-CVD), by maximizing the margin between data points of different classes. 3) Decision Trees are hierarchical structures that recursively split the dataset based on the most informative features using if-then-else conditions to reach a decision. 4)Random Forests are ensemble methods that combine multiple Decision Trees to improve predictive accuracy and reduce overfitting and providing a robust and accurate model. They also allow feature importance analysis, aiding in understanding which factors contribute most to CVD prediction. 5) Neural Networks consist of interconnected artificial neurons organized in layers and carry out feature extraction and nonlinear modeling, making them suitable for CVD prediction when ample data is available. These models find applications in various CVD-related tasks, aiding Healthcare professionals to analyze patient data, predict outcomes, and identify potential interventions.	5min 25

9	Accuracy and Precision Table (Unoptimised Models)	Savina	Here are the accuracy and precision scores we obtained after testing our model on unseen data. The random forest model performed the best with an accuracy and precision of 93%, whilst the support vector machines model performed the poorest and wouldn't be as reliable as in comparison, it's acuracy dropped by 14% and it's precision dropped by 12%. However before, any conclusions are reached...	
10	Optimisations	Savina	Several optimisations were conducted. Hyperparameter tuning was explored to enhance the logistic regression model's performance. Initially, the 'sag' solver was considered for its effectiveness with large datasets, however, because it's an older version of 'saga' it was not used.For the optimization of the SVM model, we adjusted its kernel hyperparameters to potentially better suit our dataset's structure. The optimization of the Decision Tree model focused on manually fine-tuning hyperparameters such as max_depth, min_samples_split, and min_samples_leaf to specific values. By limiting the tree's depth and controlling the granularity of splits, we aimed to prevent overfitting and ensure the model's generalization to new data. The optimization of the Random Forest model involved defining a parameter distribution dictionary for key hyperparameters with a RandomizedSearchCV object named random_search, employing cross-validation and parallel processing to find the best hyperparameter combination for improving model accuracy. In the first optimization of the neural network, a flexible Sequential neural network was configured with Keras Tuner, allowing it to dynamically select activation functions and layer architectures. Secondly, specific hyperparameters were defined based on the best-performing configuration from Keras Tuner, featuring hidden layers to simplify the model.	
11	Accuracy and Precision Table (Optimised Models)	Savina	Intrestingly, the decision tree outperformed random forests by 1% in precision.	7min 10
12	Limitations	Ash	As for limitations, the primary limitation in our models is overfitting. Overfitting occurs when the model excessively fits the training data, hampering its ability to generalize to new data. Though we implemented early stopping and hyperparameter tuning, overfitting persisted, which was evident in the higher training accuracy compared to validation accuracy. As mentioned previously, we removed outliers from the majority class of our data. While this balanced our training classes, the significant reduction in the size of the dataset may have led to loss of valuable information. Furthermore, our dataset is geographically limited to the United States which restricts its global applicability.	7min 50
13	Conclusion	Lishani	To conclude, the initial random forest model performed the best in terms of accuracy and precision. However, after optimisation, our decision tree model outperformed the random forest model in precision. From a business perspective the consideration of the sensitivity of this topic carries great weight. We wouldn't want to have told Sibyl that she did not have heart disease when she did actually have one or vice versa for another individual's case. Therefore, we would look at increasing the dataset and optimising further before deploying. Ideally there should be an accuracy and precision of 100% or as close to this value as possible. This is why we have decided to recommend the random forests model to be explored further due to its nature of ensemble learning we believe this model has the best potential of reaching this goal.	
10	Github page	Lishani	The best means of prevention is through education, we created a web page which contains links to resources related to each cardiovascular disease risk indicator. For example, for someone in the US who doesn't exercise and wanted to find a class, they could visit our page and find a link which will help them to find a local class. https://ashejaz.github.io/project-4-predicting-CVD/	9min
	Conclusion	Ayroza	In the future, we plan to explore advanced regularisation techniques to further combat overfitting, seek methods to balance class distribution without sacrificing dataset richness, expand our data sources beyond the United States, and place increased emphasis on feature engineering and ongoing model evaluation to enhance the predictive capabilities. Thank you, we will open up to any questions you may have.	9min 20