Slide 1: Good day everyone, we are Wei Zhen, Pamela and Shi Ying from SC5 group 2. Today, we will be presenting our SC1015 mini project which is on hospital mortality prediction.

Slide 2: We will be covering our problem motivation, followed by data cleaning and preparation, exploratory data analysis, machine learning and lastly we will be sharing the outcomes of our research.

Slide 3: First off, let’s take a look at the motivation behind our problem.

Slide 4: With overall increasing life expectancies and low fertility rates, the proportion of Singapore’s population aged 65 years and above is rising rapidly. By 2030, 25% of the population is predicted to be aged 65 and above. This is almost double that of the 14.4% that was recorded in 2019.

Slide 5: As such, it is evident that healthcare quality is of paramount importance. Models that predict in-hospital mortality can assist with patients’ risk stratification, allowing for better level of care placement, closer follow-up, and improved health care quality.

Slide 6: The dataset that we have settled on is sourced from Kaggle and is called “In Hospital Mortality Prediction”, which outlines variables including demographic characteristics such as the age of the patient, their gender, BMI, as well as their vital signs such as heart rate, blood pressure. and the primary outcome of the study was in-hospital mortality, defined as the vital status at the time of hospital discharge.

Slide 7: We will be investigating this problem question: How can hospitals predict the survival outcome of their in-hospital patients?

Slide 8; We will start off by cleaning and preparing the data.

Slide 9: We dealt with NULL values in the numeric variables by filling them with the mean. As for the data column on outcome that consisted of ‘1’s and ‘0’s only, we have dropped all rows for which ‘outcome’ was NULL as there was only one NULL value for this variable.

As the columns ‘group’ and ‘ID’ are merely for identification purposes, we dropped these 2 irrelevant columns as they do not play a part in the outcome. However, this was not enough as we had more than 40 variables to work with. Thus, it was necessary to do feature selection to lessen the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

Slide 10: To begin, we started with pairwise correlation. Using a heatmap, we identified groups of highly correlated features and only kept one of them so that our model can have as much predictive power using as few features as possible.

Slide 11: We also used a low variance filter. Columns with low variance are likely to distract certain learning algorithms (in particular those which are distance based) and are therefore better removed.

Slide 12: After applying these 2 feature selection methods, we managed to reduce the number of independent variables to 15. These are the final variables that we will be using for our EDA.

Slide 13: We will now move on to our exploratory data analysis section.

Slide 14: We will be analysing the following variables that we have obtained from data cleaning.

Slide 15: To do so, we first did univariate exploration by combining a histogram and density plot.

Slide 16: These are our findings. It is important for hospitals to take note of the following information such that any extremities can be closely monitored.

Slide 17: We also did bi-variate exploration using a violin plot and a box plot.

Slide 18: Here are our major findings. For Creatine kinase, despite the presence of extreme outliers, we decided to keep them as high levels of creatine kinase is known to be dangerous and will affect the outcome. From the boxplot, we also inferred that EF does not seem to have a significant impact on outcome, and thus we dropped that variable.

Slide 19: Now, we will move on to building our models.

Slide 20: Since we are working on a classification problem, we will be using the following models: Logistic Regression, K-nearest Neighbours and random forest.

Slide 21: First, we explored the prediction of mortality outcomes using the K-Nearest neighbours machine learning model. K-nearest neighbours enables us to classify unclassified data points based on “closeness” to other obtainable data points. We have selected the number of nearest neighbours to be 15 in our model. We derived this optimal K value by taking the square root of the total number of samples we have. The classification accuracy of this model was decent at 86.39% with True Negative Rates at 86.03%. This signifies that we will be considerably confident with using this model to predict mortality outcomes.

Slide 22: However, we wanted to explore if there were even better models to carry out predictions. As a result, we also did machine learning with the Logistic Regression model.Typically, Logistic Regression is used to categorise data into discrete classes of True or False, or ‘1’s and ‘0’s. Since the goal of our project is aimed at predicting the mortality outcome of patients, this model will undoubtedly help us make a clear prediction using numerical predictors. The classification accuracy improved to around 88%, with a False Positive Rate of around 0.5% and a False Negative Rate of 11%. This shows that our machine learning model is confident in classifying outcomes into ‘0’s and ‘1’s.

Slide 23: This model provides a good prediction. Nonetheless, we wanted to see if we could further improve the accuracy of our predictions. Hence, we considered the Random Forest machine learning algorithm as it is known to provide high levels of classification accuracy due to cross validation. Indeed, the Random Forest Algorithm provided us with a promising accuracy score of 100% for both our test and train data.

Slide 24: Upon comparing the 3 machine learning models used, we observe that the Random Forest Algorithm provided us with the highest accuracy and F1 score. This is likely to be the most suitable model for predicting in-hospital mortality.

Slide 26: Finally, let’s take a look at our final outcomes from our project.

Slide 25: Now, we will analyse the data driven insights of our project and suggest how our project can help hospitals react better towards changes in patients’ medical data. Firstly, we acknowledge that the number of in-hospital patients increases with age. Therefore, hospitals should pay especially careful attention to older patients as they are at higher risk of mortality. Similarly, heart rate, systolic blood pressure, SPO2, Urine output etc. are notable determinants that can affect mortality outcomes. Therefore, hospitals should install more medical devices to monitor changes in patients’ medical statistics in these areas.

Slide 26: Next, we recommend that hospitals consider the use of the Random Forest machine learning algorithm to process collected data to predict the probable mortality outcome for patients. This would then enable hospitals to better allocate their resources and manpower to patients who either have better chances of survival; or need especially close attention from nurses and doctors as they may require emergency rescue any time.

Slide 27: Through this project, we were happy to learn more about how to do feature selection. Also, we better understood machine learning algorithms such as K-Nearest Neighbours, Logistic Regression and Random Forest to carry out data prediction.

Slide 28:

Slide 29: With that, we have come to the end of the presentation. We have enjoyed working on this project and we hope that this presentation has been insightful for you too. Thank you !!