Day 1 STEM Games Science Arena

Introduction

Welcome to the 2023 STEM Games Science Arena. Through this arena you will handle some sensitive and private patient data and discuss what are the benefits and challenges in regards to patient data.

Healthcare records, also known as medical records or health records, are documents that contain a patient's health information and medical history. The primary use of these records are for treating the patient but they have an important secondary use in research. Healthcare records contain a wealth of information about patients, including their medical history, treatments received, and health outcomes, which can be used to conduct research studies to improve healthcare practices, identify trends, and develop new treatments. However, it is important to ensure that patient privacy is protected through appropriate de-identification and data security measures.

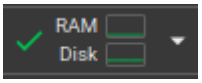
Day 1 description

For day 1 you will step into the shoes of a data scientist and machine learning specialist. You have gathered data from patients suffering cardiovascular disease and built several models for predicting if a person is at risk of having heart failure.

In order to answer tasks for Day 1 you will have to run the code found on this link:

https://colab.research.google.com/drive/1ykREBD0wQVJFzWbjvN1a1NuK4VvQ8nz6#scrollTo=208CUyD0ZYXh

The code is written in the R, a programming language and software environment commonly used in life sciences research. It is an open-source, free-to-use software that is widely used for statistical analysis, data visualization, and modeling in a variety of fields, including genetics, genomics, proteomics, and bioinformatics.

In order to run the analysis firstly you will need to make a copy of the R notebook into your drive. To do so, select File and Save a copy in Drive. Next, press connect on the top right corner of the screen. This may take some time and when this appears: 

you are ready to run the code.

Parts of the R notebook that contain code are called cells or chunks. You should run them one by one and in the order that they appear in the notebook. To run a cell select the  button on the left side of the cell which will appear when you hover your mouse over the cell.

After running a cell some output will appear underneath. Some of the outputs will be just log type messages which are not particularly important but some outputs will contain plots or other useful information. Additionally, some of the questions are general questions about statistics/data science which require you to do some research, and others just require you to look around in the notebook.

Please write short and concise answers. Some questions require one word answers. When questions require descriptions please use a maximum of 300 characters.

Questions from the R notebook.

?Q1

Q1.1. What are the two types of supervised machine learning tasks?
Q1.2. Give examples of these tasks in healthcare or life sciences.
Q1.3. What type of task are we facing in this analysis?

?Q2

Q2.1. Since the provided description of the variables is scarce, do some of your own research and write a more detailed description for these variables:

- ChestPainType,
- RestingPB,
- Cholesterol,
- FastingBS,
- RestingECG,
- MaxHR,
- ExerciseAngina,
- Oldpeak,
- ST_slope.

Q2.2. Out of **all** variables, determine which are categorical and which are continuous.
Q2.3 For each continuous variable write the interquartile range and average value.

?Q3

Q3.1. How can missing values be represented?
Q3.2. Does our dataset contain any missing values?
Q3.3. Analyze each plot to detect outliers. There are two variables that have unusual values - which variables are they?
Q3.4 What are the methods for dealing with missing values and outliers?

?Q4

Q4.1. Which variables are used for imputation?
Q4.2. What are the most correlated variables and what are the r values?
Q4.3. Which of the variables seem to implicate a higher risk of heart disease.

?Q5

Q5.1. Which variables are most significant in predicting the risk of Heart disease with logistic regression model?
Q5.2. What are the accuracy, precision, recall and the F1 score of the logistic regression model?
Q5.3. What are the top3 variables with the highest importance measures in the random forest model?
Q5.4. What are the accuracy, precision, recall and the F1 score of the random forest model?
Q5.5. Which are the best values for the hyperparameters "gamma" and "cost"? What is the error value for that parameter combination for the SVM model?
Q5.6. What are the accuracy, precision, recall and the F1 score of the SVM model?