Identify a problem from your own experience that you think would be amenable to data mining. For that problem describe:

1. What the data is.

The data is a person's health record which can be generated from a patient's health database. The relevant features are age, gender, blood group, and list of relevant diseases.

2. What type of benefit you might hope to get from data mining.

There are certain people who are not easily infected by covid-19 when they are exposed to the virus but some people get it immediately if they share the same space. For me this is something interesting to study about although there is plenty of information at our fingertips. We do know that our immunity strength is one of the main factors that decides how much vulnerable we are to various diseases. To my knowledge immunity can be defined by our underlying health conditions like pre-existing diseases, age, and more.

So through this data mining process we can group different people into different clusters and clearly understand the level of risk they have on the basis of their health conditions. The result can be shared by a credible agent to the public to make people more aware of the risk associated with one’s health. So after knowing the risk level the family member can decide who should be more protected from the virus and then take wise decisions accordingly. Such an initiative at a family level can help to save each other's lives and humanity as a whole.

In continuation with the immunity, the information that is there on the internet is not equally relevant to people across the globe because I feel depending on the diet, upbringing(hygiene, lifestyle), and environment, people may have different immunity or reaction to any unknown diseases. I do not mean the information available is useless, I feel the study is more biased to people in that country or continent but in general it could be true for everyone. For this purpose it would be more precise if we could study our people by using our own contextual data so that it doesn’t create confusion among people with contradicting facts.

3. What type of data mining (classification, clustering, etc.) you think would be relevant.

A descriptive task of data mining is relevant for this analysis problem because depending on the person's health record, we need to cluster people into low, medium, high risk groups. In this task, specifically a clustering algorithm can be used to segment people into different clusters.

There are three possible clusters:

1. Low risk: From my intuition, a low risk cluster can consist of people who are aged between 20 to 50 years without any underlying medical conditions. Young adults without any disease history are considered to be strong, energetic, and therefore, have the best immunity to guard diseases. People with blood group O are less susceptible as per studies.
2. High risk: This includes people who are newborn to 15 years old and more than 70 years old. Aslo people who have diseases like asthma, neumonia, TB, and other lung related diseases. Blood group A is more likely to get covid-19.
3. Medium risk: On the basis of age, people between 15 to 20 years old and 50 to 70 years old can fall in this risk level.

4. Name one type of data mining that you think would not be relevant, and describe briefly why not.

For this task particularly, I feel a regression analysis is not relevant because for regression problems we map every input to some continuous output function, in which the model learns based on the target variable and the learned model has to predict target variable for future unseen data. The aim of the problem is to clearly segregate people into different groups as high, medium, low risk based on the patterns in the data. Moreover, there is no predefined continuous target variable in a patient's database. If there exists a target variable which identifies each person in different groups, then a data mining task is not needed because by querying a database it can generate the result.

For each, illustrate with an example, e.g., if you think clustering is relevant, describe what you think a likely cluster might contain and what the real-world meaning would be.

Write one to two pages of 11 point single-spaced typeset text - you aren't writing a paper, but it isn't a short answer either.