Initial thoughts:

\* IT MAY BE BEST TO DOWNLOAD THE COLUMNS ONTO MY COMPUTER THAT ARE USED TO DEFINE WHO HAS PREDIABETES AND WHO PROGRESSES TO DIABETES. IT SHOULD ONLY BE A FEW COLUMNS SO IT WILL NOT TAKE UP A LOT OF ROOM ON MY COMPUTER AND IT SHOULD BE QUICK COMPUTATIONALLY. THEN I CAN LOOK AT THE DATA ON A JUPYTER NOTEBOOK AND IT WILL BE MUCH EASIER TO DO WHAT I WANT.

To begin, it would be most efficient to focus on finding the certain patients that will be involved in this study so that we do not have to waste time waiting for computations on many pateints to run initially. We need to keep only the patients with prediabetic symptoms to start off AND THOSE THAT CAME BACK FOR A CHECK UP BECAUSE WE NEED TO CLASSIFY THESE PATIENTS BASED ON IF THEY DEVELOPED DIABETES OR NOT OVER TIME! If there were three times when patients were checked up on, we will need to figure out the first date when a patient's blood glucose and HbA1c were taken. Then we can keep only the patients who have results in a prediabetic range. Then we look at the next time a patient is checked up on and see if they have diabetes at that point or not. This will give us our targets. We have to take into consideration if a patient was checked up on more than once. Even if diabetes was not diagnosed the first check up, but it is for the second check up, this patient will be targeted as having developed diabetes.

Another thing we need to consider is gestational diabetes. We will not include this in our analysis because the cause of development of diabetes is due to pregnancy mostly and not other factors we are looking for.

Columns Needed - UDI:

In HTML: ukb42385.csv.gz

Glucose - 30740-0.0

Glucose - 30740-1.0

Glucose assay date - 30741-0.0

Glucose assay date - 30741.1.0

Glycated haemoglobin (HbA1c) - 30750-0.0

Glycated haemoglobin (HbA1c) - 30750-1.0

Glycated haemoglobin (HbA1c) assay date - 30751.0.0

Glycated haemoglobin (HbA1c) assay date - 30751.1.0

In HTML: ukb26867.csv.gz

Diabetes diagnosed by doctor - 2443-0.0

Diabetes diagnosed by doctor - 2443-1.0

Diabetes diagnosed by doctor - 2443-2.0

Age diabetes diagnosed - 2976-0.0

Age diabetes diagnosed - 2976-1.0

Age diabetes diagnosed - 2976-2.0

Gestational diabetes only - 4041-0.0

Gestational diabetes only - 4041-1.0

Gestational diabetes only - 4041-2.0

The next step after defining patients with prediabetic symptoms is to figure out which of them are diagnosed with diabetes and when. If we diagnose a patient with prediabetes, but they are already diagnosed with diabetes, then we must throw them away. If the patient is not diagnosed with diabetes, then we have to look at the later dates and see if diabetes was diagnosed after the initial prediabetes classification using the blood glucose and HbA1c. We then mark these patients as progressing to diabetes, and all other patients will be marked as not progressing to diabetes.

We then need to include all the columns using the HTML files in the phenotypes folder and merge all the columns together by eid for the patients with prediabetic symptoms which we have decided to keep in the analysis. AT THIS POINT, DEPENDING ON HOW MANY PATIENTS WE HAVE LEFT, IT MAY BE POSSIBLE TO SCP THE DATAFRAME WE CREATE TO THE COMPUTER AGAIN. THIS WOULD ALLOW US TO DO THE DATA MANIPULATION ON A BETTER VISUALIZED PLACE AND MAKE COMMENTS EASILY AS WE GO ALONG THE PRE-PROCESSING AND MACHINE LEARNING STEPS. THERE SHOULD BE A LOT OF FEATURES TO BEGIN WITH, SO IT MAY NOT BE POSSIBLE TO DO THIS, BUT FOR THE MODEL WE CREATE USING ONLY FEATURES WE KNOW ARE IMPORTANT TO DIABETES DIAGNOSIS IT SHOULD BE POSSIBLE.