FIT3164 – Data Science Project 2 – Assignment 1 - Journal Entries #1 – Weeks 1 to 6
Name: Vionnie Tan (ID: 30092809) Group 4
Workshop: Thursday, 2pm – 4pm (Online)

Week 1
As the break spanned over a month, I found myself forgetting most of the contents written in our project proposal. This was critical as I felt like I was behind my teammates and needed to quickly get back on track. We've also realized that we've been lacking in the project management aspects and is still finalizing what approach to take with regards to our project, despite mentioning it on the project proposal already. We were also stuck and didn't know where to begin with in the project and needed some inspiration/help from our tutor for us to start our project as soon as possible. On a personal scale, I didn't do that much self-learning associated with knowing how to build the predictive model due to personal reasons hindered because of COVID-19. As this was a personal issue mainly towards myself, I needed to bounce back soon or else, the trajectory of our team's progress would just be a standstill. I'll try my best to put in some more contributions to the team wherever needed.

Week 2
During this week, although we have set up most of the fundamentals needed for the project, our team has yet to start coding yet (including myself) and I'm slightly worried that we won't meet the deadline for the project. To add to my concerns, all of us has no specific specialization that we excel at so it's a matter of working together to meet the project requirements. Hopefully, by the next week or so I can start with the coding part and prepare for our first interim presentation in this semester. The research that I've been doing seems to not be in the right direction for our project as I'm not implementing it practically and to add to this, I have had no experience with image processing and building the predictive model, so it's a bit overwhelming to me. Also, this week, I assigned myself with setting up the meeting environment for my teammates and asking the tutor for clarification on our doubts regarding on where to start with the project. I also contributed briefly into adding some of the required checkpoints on Trello. I also found some similar articles regarding cancer predictive modelling that might be useful for our project.

Week 3
Reflecting on this week, it was clear to all of us as a group that we are very behind in terms of progress as opposed to other teams. Our supervisor got back us up on our feet and asked each of us to be clearer in task delegation between ourselves. It was also apparent to all of us that if we do not start with the coding part of the project, we wouldn't be able to manage the workload and there's a high risk that we would have an incomplete project. It was also clear that, as we all forgot what we had initially pitched in the project proposal, wasn't followed through in this semester, so we were inconsistent with the answers we gave to our tutor. I immediately held up an emergency meeting to allow for clearer task delegation between us and increase our motivation towards finishing the project successfully. I also added some details to our Trello Boards and divided it into some minor checkpoints so that I have a clear idea of what I must complete during this sprint and so on. We also had our weekly meetings, and I delegated the tasks clearly to my teammates, instead of just collectively doing it together. I assigned myself to do the predictive model part. Also, we decided to code our project in Python, and I created the Jupyter Notebook environment for us to code our project and also finalized our GitHub repository.

Week 4
When coding the predictive model, it dawned upon us that the given dataset was too large (10GB) and it slowed the processes of the code pretty significantly as we didn't have high end specification PCs. So, I suggested that we train and test on a smaller part of the dataset, just to make sure that our code is working, and the desired output is obtained. During my code of the data augmentation and normalization,

I was unfamiliar with most of the techniques and struggled to get it right for the first time, so I had to delegate some extra personal time to understand what each line of code is doing, and which parts of code (if any) can be optimized. By the end of the week, our code still had some errors in them – on the label part of the visualization so we needed to fix that ASAP before our next class. Communication within teammates have also resulted in a positive outcome. The Messenger chats are highly active where we ask each other for feedback (and help, if necessary) and just an overall check-up on how everyone's keeping up with their task delegation. The Trello Board had also been a massive help to the increased productivity within our team – as we take the time at the end of the week to checklist our sprint's to-do-list and make necessary changes (if needed). My roles for this week included assigning myself with doing the data normalization and augmentation of the split training and testing dataset. I also organized the weekly meeting for this week, and we just kept up to date with the tasks that were delegated to each of us.

Week 5

As the labels part of our visualization is still not optimized – still showing the wrong labels, my teammate and I decided to implement a custom dataset in hopes to fix the labels of the graph, but to no avail. As this is becoming a bottleneck in terms of our progress, it is vital that we seek for help and guidance as to how to fix this bug. The trained model also posed a very low accuracy (only 40-50%), so this is an ongoing concern for us as it is below our stated MVP and some solutions to this would be to alter the batch size and epoch of our model. Progress was looking good overall, but I felt that it was hard to keep up with all the workload – which could be due to myself slacking off earlier in the semester, and less sprint time to make do of a bigger productivity. The worsening COVID-19 situation at my country added to these pains and it became increasingly difficult to put in a constant amount of effort for the project each week due to my other units as well. This week I tasked myself with filling up the presentation slides and organizing it in the order that we would be presenting, whilst also writing the script for our presentation. I also organized the meeting that we'd all use for recording of the presentation. I tried implementing the custom dataset as well this week and the initial creation of the resnet18 model.

Week 6

With the mid-semester mark underway, I was personally worried with the risks associated with our model. It took a very long time to train the large dataset, and it was extremely difficult to make changes to our code once training started. On one of our models, we posted a 78% peak accuracy which is a huge improvement compared to what we have started with. There were some minor bugs that needed to be changed/fixed on my parts of the code that I've written, and it's taking much longer than what I had expected due to my unfamiliarity with how pytorch works. I was able to show more meaningful plots with regards to training loss and test loss, as well as the model's AUC score and confusion matrix. What's needed to be done next is the integration with the website, and if that's done and dusted, the MVP of our product should be achieved, and I can focus on implementing further improvements on the model. This week, we had our Interim Presentation, and I answered the Q&As asked by our tutors present during the presentation. I also fixed some minor codes, created an inference notebook, looked at more resources that might be beneficial for our project. I delegated the integration part to my teammate, and I will solely focus on improving the accuracy of our model.