THE UNIVERSITY of ADELAIDE

Retrospective Sprint 2 of Group RTMST1

Real-Time Wine Sensing Tool

by

a1724222 Andrew Grace
a1783911 Harnoor Bandesh
a1716640 Harrison Greven
a1670268 Jay Hansen
a1720561 Joshua Lennon
a1671245 Ross Pickett
a1785307 Ruonan Fu
a1608613 Ryan Swiggs
a1779153 Vandit Gajjar

Snapshots (Group):

First Snapshot:

Product Backlog and Task Board:

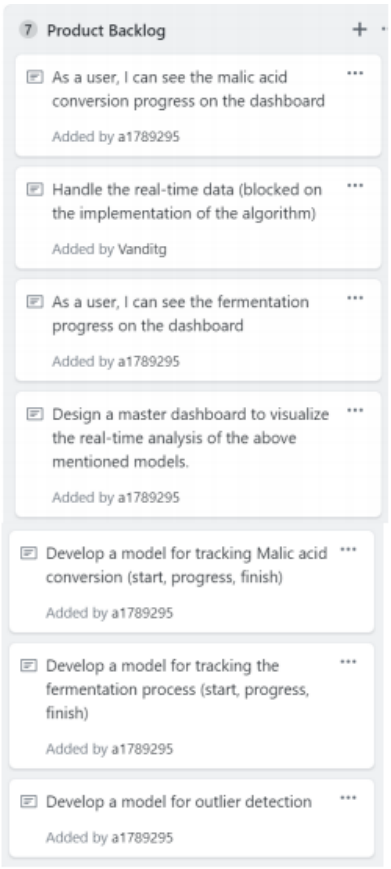

Figure 1. Product Backlog

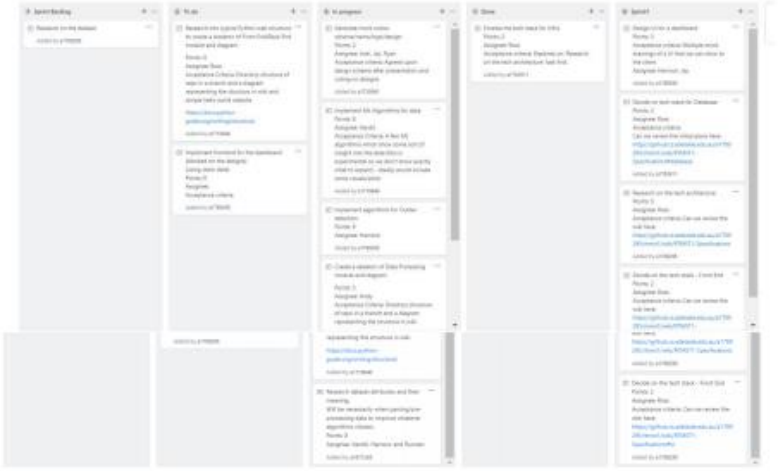

Figure 2. Task Board

Sprint Backlog and User Stories:

- **Decide on tech stack for Database:**

We need to decide on a tech stack, such as a programming language (SQL or panda for example) to be used as a database. Alternatively, we need to decide if a database is the best option for this project.

Acceptance Criteria:

We have decided on a tech stack for the database. We have researched and tentatively decided upon architectures that we will use.

- **Design UI for a dashboard:**

We want to have designed and reviewed a UI/GUI for a user dashboard, where the user will interact with the tool.

Acceptance Criteria:

We have designed and reviewed a UI that the group collectively accept.

- **Finalize the tech stack for infra:**

We also want to finalize our decisions for tech stack for the infra part of the design.

Acceptance Criteria:

We have decided and finalized our infra tech.

- **Decide on Tech stack for Frontend:**

We want to decide on a tech stack, such as a programming language and which IDE to use.

Acceptance Criteria:

We as a group have decided on a tech stack to use for front end development.

- **Research on the tech architecture:**

We want to research possible architectures for accessing and interfacing with the dataset, as well as processing the real time data, and architecture for how the user will access and interact with the tool.

Acceptance Criteria:

We have researched and tentatively decided upon architectures that we will use.

- **Implement the algorithm for outlier detection:**

There may be outliers for any machine learning algorithm we implement. We want to create another algorithm to detect and remove these outliers so we can get more precise results.

Acceptance Criteria:

We have an algorithm that successfully identifies and removes outliers.

- **Research Dataset Attributes and their meaning:**

The dataset contains several attributes that determine the value of the label. Understanding these labels will give the project greater significance and will give various results more meaning. We want to determine what each attribute means.

Acceptance Criteria:

We have a description and an understanding of what each dataset attribute means

- **Generate Mock of Color scheme/name/logo/design:**

Design of the front-end UI aesthetics that will be visible to the users. This will be the first thing users see so a pleasing design is important.

Acceptance Criteria:

Agreed upon by the entire group

Definition of Done:

The definition for each user story for this week:

- Generate mock color scheme/name/logo/design will be considered done when we have had a presentation to our group and agreed upon a design scheme.
- Implement ML Algorithms for data will be considered done when we have agreed upon and began implementing some experimental Machine Learning algorithms based on the dataset we have found.
- Implement algorithms for Outlier detection will be considered done when we have begun implementing some method of detecting outliers from a dataset.

- Create a skeleton of Data Processing module and diagram will be considered done when we have a file tree and skeleton files for the data processing module.
- Research dataset attributes and their meaning will be considered done when as a group we have agreed upon a basic understanding for what each of the dataset attributes means, as well as their importance for the dataset.

Summary of changes:

In this snapshot, we have formed separate teams: “web team”, “data team” and “product and subject team” to be able to separate tasks to each group member as their skills/preferences fit. This allowed us to tackle some of our product backlog tickets and get started on writing some actual code with a dummy wine fermentation dataset, as we are still waiting on our official data. We have begun research into algorithms for use in our data processing, as well as research into the attributes of the dataset and their importance, as well as some initial sketches of a UI and dashboard, experimenting with static vs dynamic graphs for data reporting.

Second Snapshot:

Product Backlog and Task Board:

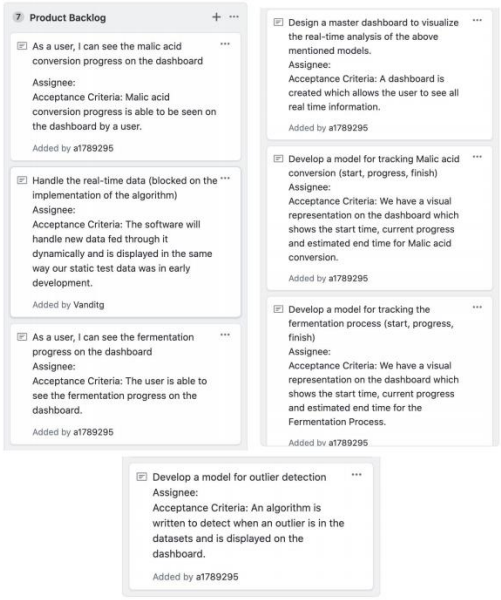

Figure 3. Product Backlog

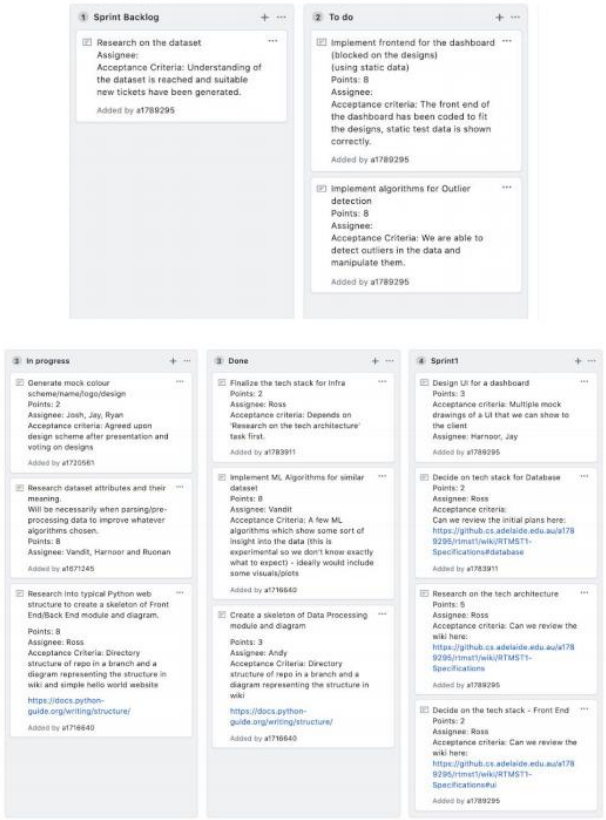

Figure 4. Task Board

Sprint Backlog and User Stories

Figure 5. Sprint Backlog

- **Research on the dataset:**

The dataset will be in a particular format and of a certain size. There may be other factors that influence what machine learning algorithm is most suitable to use. Research these factors.

Acceptance Criteria:

We have a description of the structure of the dataset so we can communicate with the user about the data on a technical level.

- **Implement algorithms for Outlier detection:**

The data will have outliers that will skew the results of our machine learning programs, these outliers need to be removed to provide more precise analysis of the data.

Acceptance Criteria:

We have a definition of an outlier. We have an algorithm that detects outliers which can reliably inform the user when a wine is at risk.

- **Generate mock color scheme/name/logo/design:**

Create some drafts of a User Interface design. Create a name and a logo for the project.

Acceptance Criteria:

As a group we have decided on a name, logo, and design for our project User Interface which would best engage and interest the user.

- **Research dataset attributes and their meanings:**

The dataset contains several attributes that determine the value of the label. Understanding these labels will give the project greater significance and will give various results more meaning. We want to determine what each attribute means.

Acceptance Criteria:

We have a description and an understanding of what each dataset attribute means so that we are able to explain any technical terms to the user.

- **Research into typical Python web structure to create a skeleton of Front End/Back End module and diagram:**

Create a Python project for the Front and Back End of the website. Create a diagram of the interaction between modules.

Acceptance Criteria:

We have a Python project directory setup, including files and subdirectories such that the product is reliable for the user. We have a diagram of module interaction so that any team member is able to explain our products structure to a user.

Definition of Done:

The definition for each ticket for this week:

- Generate mock color scheme/name/logo/design will be considered done when:
  - We have multiple logos/brandings to choose from (**Complete**).
  - The majority of the team agrees on the branding for the product.
- Implement ML Algorithms for data will be considered done when we have (**Complete**):
  - Have an in-depth analysis of the relation between the different features of the dataset.
  - Understand which features are most important.
  - Have a set of machine learning algorithms that can extract meaningful information from the dataset

- Implement algorithms for Outlier detection will be considered done when:
    - We have a series of algorithms which can detect when one of the features is outside of its reasonable bounds.
- Create a skeleton of Data Processing module and diagram will be considered done when (**Complete**):
    - We have a file tree and skeleton files for the data processing module.
    - Every team member is happy with the repository structure.
- Research and develop the code for a simple web application using Python/Django will be considered done when:
    - The skeleton directory structure of a web application is in the repository (**Complete**).
    - We are able to load a simple "Hello World" for the main page.
    - There is suitable documentation on the structure of the code.
- Research dataset attributes and their meaning will be considered done when:
    - As a group we have agreed upon a basic understanding for what each of the dataset attributes means, as well as their importance for the dataset.

Summary of changes:

The previous week of work has been very technically focused and has seen us make good progress on the product's development. To begin, we found a set of the most effective machine learning algorithms for classifying the quality of wine using a surrogate dataset with a similar problem space and attributes. The development of these algorithms also involved a detailed analysis of the features, many of which we hope will be similar to the actual dataset which we have not yet received. It is out hope that the algorithms developed will transfer effectively to the real dataset. Beyond this, we have finally created a skeleton repository structure for the code, for both the data processing side of the project and the web-application user interface. Beyond the simple skeleton structure we have also almost finished creating the foundations of a simple running website which we will build the dashboard on top of. In the design team we generated a digital mock-up of what

we hope the dashboard may look like based on the previous drawing design. Some members of the product team also generated a series of possible logos/product branding which we will vote on in our next meeting.

Declaration: I have attended sprint 2 planning meeting on 26th August 2020 and sprint 2 retrospective meeting/sprint 3 planning on 9th September 2020 with the tutor Mr. Navpreet Singh Ahuja.

What went well in the sprint (Individual):

Our second sprint was about the implementation of sub-modules, defining and making a rough outline for tech stack infrastructure, and researching how to create a tech stack skeleton of Front-End and Back-End communication. All of the above implementations were based on the user stories that our group members selected for the second sprint backlog. One of the most important things to note here is that we also updated the Wiki section of our GitHub repo which was and will be helpful currently and in the future for documentation purposes, respectively. Our project mostly focuses on the Machine Learning and Web Development aspects, in our initial sprint we were exploring three languages - C++, JavaScript, and Python. After rigorous research, we have finalized Python (Back-End) and JavaScript (Front-End) for the above aspects as both the languages have a number of open-source libraries and frameworks available to work with. In order to understand the Machine Learning concepts in a broad manner, I have presented a short demo on Google Collab in one of the scrum meetings, where the major focus was on how the different algorithms will work for our data [1].

What could be improved (Individual):

As we progressed to the third sprint, there were certain limitations I figured out during the second sprint retrospective.

- There was also one of the issues we found relating to documentation as everyone is working on different modules/tech infrastructures. Thus we came with a conclusion to update the Wiki section. This solution will help everyone on the team immensely as we can watch the progress as well as the contributions of individual team members.
- When we entered in our second sprint, the first week was good as we distributed most of the implementation tasks to each team members, did regular scrum meetings, but the second week was worst, as we some of the team members were not able to cope up with the scrum meetings which created communication issues. Thanks to our scrum master's regular updates

on our communication medium, we could manage to align with each members' task, but it was less organized. We could all develop an efficient communication channel that would allow all team members to collaborate and give feedback.

- Considering the Machine Learning aspects, I'm focusing on the task of outlier detection and as the dataset is still under the legal process, we added one more task of Wine Quality prediction in which I have considered five Machine Learning models listed below in a low-to-high accurate manner. Also, I have made a Google Collab notebook [2] to mitigate the computation cost issue and team members can plug and play with it to visualize the data, distributions, and prediction of Wine Quality using Machine Learning models. The notebook can be found at shorturl.at/fAEH6.
  - Naïve Bayes [3]
  - Decision Tree Classifiers [4]
  - Support Vector Machines (SVM) [5]
  - Random Forest Classifiers [6]
  - Logistic Regression [7]

What will be the group commit to improve in the next sprint(Individual)?

In our last scrum meeting of the second sprint, we discussed several issues and decided to follow certain rules and strategies in the research as well as the implementation of the successful development of our product.

- More informative Wiki section update for different submodule implementation which can be helpful to team members who do not have background/knowledge in those fields.
- For showcasing the visualization of data, distribution, and prediction of Wine Quality, we have decided to either extract raw data or in the CSV file to input in

our Front-End stack that would provide better insights about the project progress to the client.

- One of the important points is to complete the research for the integration of Front-End and Back-End and come up with the solution on how to develop communication between these two, which would be a baseline for our sprint 3.

References:

[1]Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems. 2009 Nov 1;47(4):547-53.

[2]Bisong, Ekaba. "Google Colaboratory." Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA, 2019. 59-64.

[3]Ng, Andrew Y., and Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems. 2002.

[4]Freund, Yoav, and Llew Mason. "The alternating decision tree learning algorithm." icml. Vol. 99. 1999.

[5]Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[6]Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[7]Menard, Scott. Applied logistic regression analysis. Vol. 106. Sage, 2002.