School of Computer Science and Informatics

CM3203 – Individual Project Initial Plan

Machine Learning for Malware Detection

Author: Kriti Shewaramani - C1845909

Project Supervision: Yuhua Li

Project Description:

Malware is intrusive software designed to damage and destroys computers and computer systems. Malware can crack weak passwords, bore into systems, and spread through networks. Malware detection is the process of scanning the computer and files to detect malware. It effectively mitigates a possible security breach because it involves multiple tools and approaches.

This project will predict a Windows machine's probability of getting infected by various families of malware based on the different properties of that machine. This project is important because malware can lock up essential files, spam you with ads, or redirect you to malicious websites. Malware attacks can result in anything from data theft to the destruction of entire systems or devices.

In recent years, the emergence of Machine Learning algorithms has become an essential tool in the field of Cyber Security. With the ability of machine learning to be leveraged to improve malware detection, triage events, recognise breaches and alert organisations of security issues, the cyber security industry can benefit, and its landscape can be altered, benefiting individuals and significant corporations around the world.

The telemetry data containing these properties and the machine infections were generated by combining heartbeat and threat reports collected by Microsoft's endpoint protection solution, Windows Defender. This dataset is split into training and testing data that will be used in a supervised learning approach.

Aligning to the School guidance on research ethics, this project will not require ethical review and approval as there is no research that includes human participants, human material or human data.

Project Aims and Objectives

The aim of this project is to implement a machine learning model that predicts a computers probability of getting infected by a virus using the properties of a Windows computer. This model will be evaluated, and the hyperparameters will be tuned to achieve the highest possible accuracy.

Objectives

Objective	**Relevancy**
Research and construct a project plan	Initial research is essential for the foundations of the project, but research throughout the project is required.
Background reading and literature review	Research about machine learning models classification within supervised learning and review previous research and projects within this topic.
Data Understanding and pre-processing	Download, clean, and understand the data before running machine learning models. Data must be prepared to understand which machine learning algorithm should be used in the implementation phase.

Machine Learning model implementation	The model should be able to provide consistent high accuracy and low error.
Evaluation and result analysis	The results section is important to evaluate the results from the project and discuss findings. The conclusion section highlights the final findings and any future work which can be carried out.

Work Plan

To make my work plan, I have used the SMART methodology to generate smart, measurable, realistic, and time-bound objectives. These objectives are specific as they are linked to sub-tasks that are measurable and can be measured by key milestones (highlighted in red and set out in the Gantt chart). The objectives are time-bound as they are time-sensitive, with the number of days shown on the Gantt chart.

The work plan below documents the project timeline and defines the main objectives for each week.

Supervisor meetings: Weekly - Full project review meeting dates have been documented below.

Week 1 31/01/22 – 07/01/22 – Deadline for Initial Plan

Main Objective: Create an initial plan for the project

- Create a document establishing aims and objectives, project description, project timeline, risks and considerations, and work plan.

Week 2 07/01/22 – 14/01/22

Main Objective: Research and planning

- Research literature to find work conducted in this field.
- Research Machine Learning algorithms that can be used for this problem.
- Research tools that can be used for this project.

Week 3-4 21/01/22 – 07/03/22

Main Objective: Define a list of tasks for the main report

- Complete a MoSCoW analysis for requirements and task prioritisation.
- Create an in-depth risk plan.
- Create Flow charts and depict Machine Learning design considerations.

Week 5 14/01/22 – 21/01/22

Main Objective: Define the dataset

- The dataset needs to be:
    - Understood
    - Cleaned
    - Normalised
    - Pre-processed for the implementation
    - Explanatory Data Analysis

Week 6 07/03/22 – 14/03/22

Main Objective: Continue working on the model and the final report

- Continue modelling and document the justification of using a specific model
- Write the initial sections on the report.

Week 7-8 14/03/22 – 28/03/22

Main Objective: Continue working on the implementation phase

- Have a review meeting with a supervisor to check the project is meeting deadlines and discuss any barriers, additions or alterations that need to be made to the overall project.

Week 9 28/03/22 – 04/04/22

Main Objective: Interim Review

- Have a final meeting before Easter break to ensure the project remains on track.
- Continue work on the implementation and determine if the model is ready to move on to the evaluation phase.
- Test project and document challenges faced, and changes made from the initial plan.

---Easter Break---

Week 10-11 25/04/22 – 09/04/22

Main Objective: Complete final Evaluation and Report

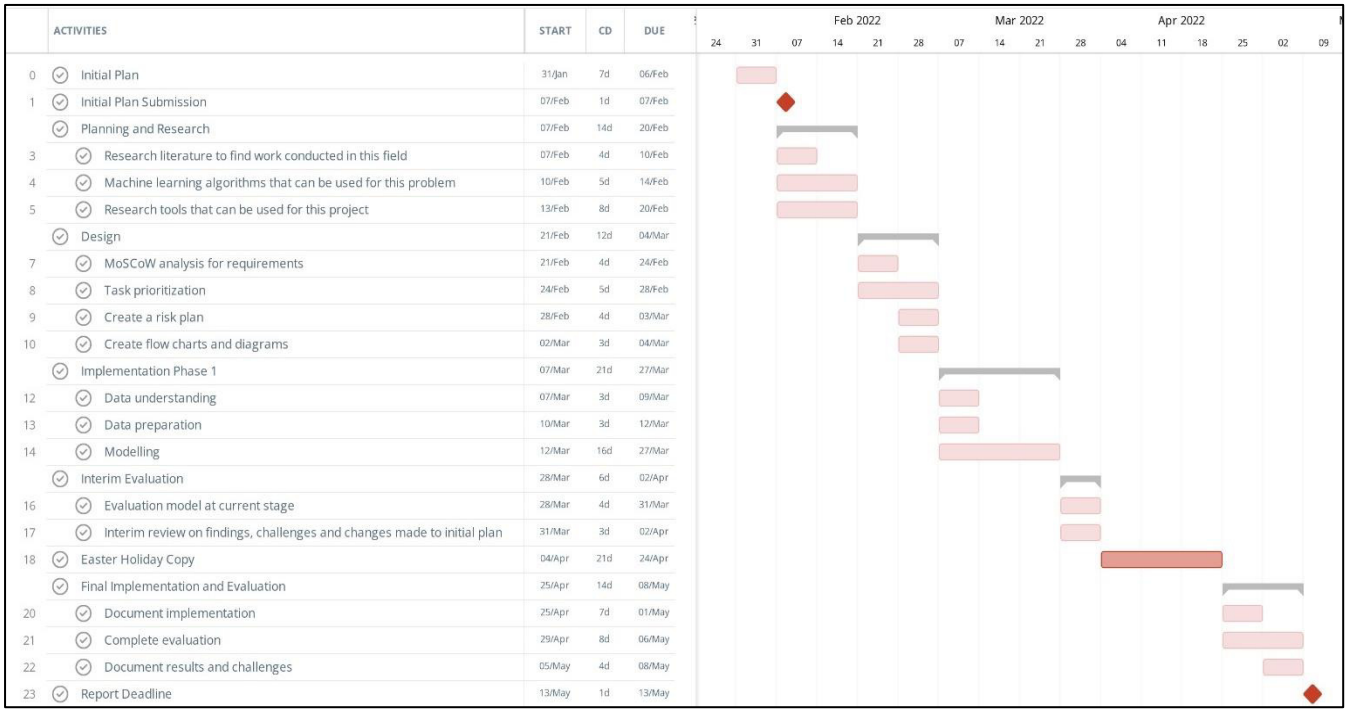
- Continue to evaluate the machine learning model and make sure all changes and results are documented.
- Test model against predefined criteria and user/test cases.

- Complete the report and leave time to proofread and make last-minute changes.

Week 12 – 09/04/22 – 16/04/22 Deadline for Final Dissertation Report

Main Objective: Submit the final report and the machine learning model

Gannt Chart



	ACTIVITIES	START	CD	DUE
0	Initial Plan	31/Jan	7d	06/Feb
1	Initial Plan Submission	07/Feb	1d	07/Feb
	Planning and Research	07/Feb	14d	20/Feb
3	Research literature to find work conducted in this field	07/Feb	4d	10/Feb
4	Machine learning algorithms that can be used for this problem	10/Feb	5d	14/Feb
5	Research tools that can be used for this project	13/Feb	8d	20/Feb
	Design	21/Feb	12d	04/Mar
7	MoSCoW analysis for requirements	21/Feb	4d	24/Feb
8	Task prioritization	24/Feb	5d	28/Feb
9	Create a risk plan	28/Feb	4d	03/Mar
10	Create flow charts and diagrams	02/Mar	3d	04/Mar
	Implementation Phase 1	07/Mar	21d	27/Mar
12	Data understanding	07/Mar	3d	09/Mar
13	Data preparation	10/Mar	3d	12/Mar
14	Modelling	12/Mar	16d	27/Mar
	Interim Evaluation	28/Mar	6d	02/Apr
16	Evaluation model at current stage	28/Mar	4d	31/Mar
17	Interim review on findings, challenges and changes made to initial plan	31/Mar	3d	02/Apr
18	Easter Holiday Copy	04/Apr	21d	24/Apr
	Final Implementation and Evaluation	25/Apr	14d	08/May
20	Document implementation	25/Apr	7d	01/May
21	Complete evaluation	29/Apr	8d	06/May
22	Document results and challenges	05/May	4d	08/May
23	Report Deadline	13/May	1d	13/May

Risks

Risk Description	Status	Probability	Impact	Mitigation Plan	Recommended actions
Loss of source code.	Unacceptable	5	9	Source code must be pushed to the git repository.	Regularly save source code to a personal machine. Backup source code to external hard drive for emergencies.
Failure to build an optimum classification model.	Reasonably Acceptable	5	9	Basic knowledge should be gained before making the model. Sources should be used to understand algorithms.	Use sources such as Udemy, Coursera, and Google to build a model. Consult with the supervisor as soon as possible and ask for advice.
Loss of accuracy of the model.	Reasonably Acceptable	7	6	Parameters must be considered if there is a loss in accuracy.	If there is a loss in accuracy, then the parameters must be considered in the case of overfitting.
Problems with programming tools being used	Reasonably Acceptable	6	6	Keep backup tools that can be used to do similar operations.	Research similar tools that could have been used and use them as a substitute.
Supervisor's absence for a period.	Reasonably Acceptable	3	6	Regular updates to the supervisor must be sent for clear communication.	Small updates every week should be sent to the supervisor. A plan should be made before the supervisor does go on absence for a period.
Other deadlines require attention.	Reasonably Acceptable	4	7	A thorough project schedule is required to organise key dates.	Deadlines must be given priority. A clear plan should be made with timeframes allocated to the other module.
Job applications.	Acceptable	6	3	Applications must be considered when making the project plan. Priority must be given to university work.	University work (project) must be given priority to job applications. The project plan must be updated regularly so that time can be made for other activities.
Poor time management.	Unacceptable	3	8	The project scheduling tool must be continuously updated weekly.	Use other tools like Trello or Asana to help in managing tasks.
Poor project design.	Unacceptable	3	9	A clear project plan must be created in the design section.	Any problems met in the implementation phase regarding the design must be documented, and the design of the project must be updated.
Project too complex to meet final aims and requirements.	Reasonably Acceptable	5	7	Strong communication with the supervisor is required to avoid problems of complexity.	Seek continuous communication with the supervisor to explain problems that are being faced.
Loss of planning documentation.	Reasonably Acceptable	2	8	Regular back-ups must be taken during the project.	Regularly save source code to the personal machine. Backup source code to external hard drive for emergencies.

References

Ieeexplore.ieee.org. 2022. *The promise of machine learning in cybersecurity*. [online] Available at: <https://ieeexplore.ieee.org/abstract/document/7925283?casa_token=kbE5yNdzL7sAAAAA:FmG09zZgZQI70Z8pX9xH7zTVpwL8LUYoKpfmcPQfeakecPXbpUz_tfSlQiMa-hkvmWn4_U6AZkhd> [Accessed 6 February 2022].

2022. [ebook] Available at: <https://www.researchgate.net/profile/Metin-Turan-2/publication/336414245_Sentiment_Analysis_of_Tweets_Using_Machine_Learning_2019_Turkey_Van_pages_85-87/links/5da066e0a6fdcc8fc347436a/Sentiment-Analysis-of-Tweets-Using-Machine-Learning-2019-Turkey-Van-pages-85-87.pdf#page=101> [Accessed 6 February 2022].

Services, P. and (AMP), A., 2022. *What is Malware? - Definition and Examples*. [online] Cisco. Available at: <https://www.cisco.com/c/en_au/products/security/advanced-malware-protection/what-is-malware.html> [Accessed 6 February 2022].

Comodo Enterprise. 2022. *What is Malware Detection? | Importance of Malware Tool*. [online] Available at: <https://enterprise.comodo.com/what-is-malware-detection.php#:~:text=Malware%20detection%20is%20the%20process,less%20than%2050%20seconds%20only> [Accessed 6 February 2022].