

School of Computing Final Year Research Project Project Initiation Document

Can machine learning improve geoscience prediction models

Kane Swartz

May 4, 2023

Contents

1	Basic Details														
2	Degree Suitability														
3	Outline of the project environment and problem to be solved 3.1 Who is the client														
4	Project aim and objectives														
5	Project constraints														
6	Facilities and resources														
7	Log of risks 7.1 Risk 1	5 5 5 5 5 6 6 6 6 6 6 6													
8	Project Deliverables														
9	Project Plan Approach														
10	10 Supervisor Meetings														
11	11 Legal, ethical, professional, social issues (mandatory)														
12	Appendix A: Ethics Certificate	8													
13	Appendix B: Gannt Chart	10													
14	References	12													

1 Basic Details

Student Name: Kane Swartz

Draft Project Title: Can machine learning improve geoscience prediction

models

Course Year: Software Engineering year 3 Project Supervisor: Dr Rich Boakes

Client Organisation: University of Portsmouth School of the Environment,

Geography Geosciences (SEGG)

Client Contact Name: Dr. Mohammad A Hoque

2 Degree Suitability

When I enrolled onto the Portsmouth Software Engineering course it was to develop the skills that would empower me to make impactful software. This project demonstrates a range of both software development and academic skills including problem identification, appropriate choice of software tools and the ability to do appropriate research in order to understand a specific real world problem sufficiently to specify and develop a solution. This project shows my ability to develop impactful software demonstrating that I have realised the quintessential purpose of this degree by having garnered skills to this effect.

3 Outline of the project environment and problem to be solved

3.1 Who is the client

Dr. Mohammad A Hoque is a senior lecturer in the SEGG whose ambition lies in reducing human consumption of contaminants. After procuring a BSc and MSc in Dhaka University he then went on to complete a PhD in University College London.

3.2 What is their problem

Dr. Mohammad A Hoque is a contributor to an interdisciplinary project between the SEGG and the School of Computing (SoC), iArsenic. iArsenic is a tool which predicts a data point's classification based on its features.

3.3 Why does it need to be solved

At its core the purpose of iArsenic is to contribute to the scientific effort to improve access to safe drinking water by aiding the identification of unsafe drinking water.

The need for integration between the disciplines of environmental science and machine learning has been detailed (Fleming, Watson, Ellenson, Cannon, Vesselinov, 2021, pp. 878-880).

This dissertation will work toward the goal of integrating these fields by introducing machine learning techniques to an already existing environmental science project in order to compare the accuracy of predictive models developed from an environmental science perspective to models developed from a software engineering / machine learning perspective.

4 Project aim and objectives

This is an exploratory project with the aim of comparing existing predictive models developed within the University of Portsmouth SEGG with a geoscience discipline with models developed from within a software engineering / data science discipline using machine learning techniques, which are not leveraged in the existing models.

The objective of the project will be to develop a model or multiple models, leveraging machine learning. These models will then be evaluated and their error will be compared to the existing models.

5 Project constraints

The models developed will be constrained by the objective of comparing their error to the existing models. This means the new models will need the same source data-set and produce the same format of outputs.

Time constraints and the need to produce a deliverable also come together as another fundamental constraint, meaning the research, development and review areas of the project will likely benefit from more time, almost indefinitely, requiring time to be provisioned accordingly.

6 Facilities and resources

As this project will be leveraging supervised machine learning techniques on a data-set approximately 1.3 million rows in size, a regular computer without sophisticated graphics capabilities will be suitable.

If the data-set was significantly larger or deep learning techniques were being considered, cloud computing technologies would likely be considered though this is not anticipated.

- Personal laptop
- Meetings with the client
- Occasional meeting with machine learning experts will be beneficial

7 Log of risks

7.1 Risk 1

7.1.1 Description

Time constraint

7.1.2 Impact

Almost all aspects of this project, from problem identification to algorithm research and design to execution could take more time on their own than is available for the entire project

7.1.3 Mitigation / Avoidance

The project will be split into stages and each stage will have a deadline after which the project will move on to the next stage. It is anticipated there may be temptation not to move on when the time comes, so planning ahead of time will help keep time allocation realistic.

7.2 Risk 2

7.2.1 Description

Burnout / stress related health deterioration

7.2.2 Impact

A final year dissertation is well associated with student stress. This could negatively impact not only the quality of output from myself with regards to the dissertation but could lead to more profound negative personal impact.

7.2.3 Mitigation / Avoidance

At all times I aim to prioritise my well-being and if the strain appears significant enough my strategy is to remain aware that abandoning the project for the sake of my well-being is always an option and perhaps a sensible choice as my well-being is foundational to the dissertation itself.

While I don't expect to abandon the project, maintaining this frank and honest perspective will hopefully allow me to make sensible decisions with regards to my well-being and thus my productivity too.

7.3 Risk 3

7.3.1 Description

Data loss or corruption

7.3.2 Impact

This would require work to be done again which will be time inefficient. As a time sensitive project the ramifications of this would range from minor to so severe the the project cannot be delivered.

7.3.3 Mitigation / Avoidance

Work will be stored on GitHub and google drive and backed up frequently.

8 Project Deliverables

The primary deliverable will be a written comparison and analysis of the error of all of the existing models and the models to be developed.

To produce this deliverable a machine learning model will need to be developed and will also be delivered as an artefact.

9 Project Plan Approach

The primary strategy in executing this project is to break it into distinct stages. Broadly these will be:

- 1. Problem research / problem identification (what should the deliverable look like)
- 2. Solution development (developing the solution outlined by problem research)
- 3. Solution evaluation (reflecting on the suitability of the solution)

Stage one is already underway and formed the basis of this PID, this stage will be concluded by the end of November 2022.

Stage two will involve developing a machine learning model and a suitable evaluation method, common to the new model and the previous models, to be concluded by the middle of February 2023 with a soft deadline to have a demonstrable machine learning based model for the demo on the 3rd February 2023. Stage three will be the write up of the project and should be concluded one week before the submission deadline on 5th May 2023 allowing for at least 1 week to review and make amendments.

10 Supervisor Meetings

Regular communication with the project supervisor will be over Discord and in person meetings subject to availability.

Having worked with Dr Rich Boakes for a number of years we have developed effective though ad-hoc communication methods such as attending a first or second year lecture to borrow his time in the last 15 minutes of the class. I have confidence that the supervisor will be available for support when required in a timely fashion on a suitably regular basis.

11 Legal, ethical, professional, social issues (mandatory)

Any interpretation of the output or results of these models has to be taken in the context of experimental or research tools. This project is not intended to provide advice or guidance on whether any source of drinking water is safe and accepts no liability for the use of these models output for this purpose or any other purpose outside of academia. Drinking water pollution is a sensitive and current topic and this project will be conducted with consideration to the human impact associated with this ongoing phenomena.

12 Appendix A: Ethics Certificate



Certificate of Ethics Review

Project title: Can machine learning improve geoscience prediction models

You must download your referral certificate, print a copy and keep it as a record of this review.

The FEC representative(s) for the School of Computing is/are Haythem Nakkas, Dalin Zhou

It is your responsibility to follow the University Code of Practice on Ethical Standards and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers including the following:

- University Policy
- Safety on Geological Fieldwork

It is also your responsibility to follow University quidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

Which school/department do you belong to?: School of Computing

What is your primary role at the University?: Undergraduate Student

What is the name of the member of staff who is responsible for supervising your project?: Dr Rich Boakes

Is the study likely to involve human subjects (observation) or participants?: No

Will financial inducements (other than reasonable expenses and compensation for time) be offered to participants?: No

Are there risks of significant damage to physical and/or ecological environmental features?: No

Are there risks of significant damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: No

Does the project involve animals in any way?: No

Could the research outputs potentially be harmful to third parties?: No

Could your research/artefact be adapted and be misused?: No

Will your project or project deliverables be relevant to defence, the military, police or other security organisations and/or in addition, could it be used by others to threaten UK security?: No

Please read and confirm that you agree with the following statements: I confirm that I have considered the implications for data collection and use, taking into consideration legal requirements (UK GDPR, Data Protection Act 2018 etc.), I confirm that I have considered the impact of this work and and taken any reasonable action to mitigate potential misuse of the project outputs, I confirm that I will act ethically and honestly throughout this project

Supervisor Review

As supervisor, I will ensure that this work will be conducted in an ethical manner in line with the University Ethics Policy.

Supervisor comments:

Supervisor's Digital Signature: rich.boakes@port.ac.uk Date: 21/10/2022

13 Appendix B: Gannt Chart

primary focus secondary focus tertiary focus

deadline

Date	0 2 / 0 9 / 2 2	0 7 / 0 9 / 2	0 9 / 0 9 / 2	1 6 / 0 9 / 2	2 3 / 0 9 / 2 2	0 / 0 9 / 2	0 7 / 1 0 / 2	1 4 / 1 0 / 2	2 1 / 1 0 / 2	8 / 1 0 / 2	0 4 / 1 1 / 2	1 / 1 / 2	1 8 / 1 1 / 2	5 / 1 1 / 2	0 2 / 1 2 / 2	1 5 / 1 2 / 2	/ 2	3 / 1 2 / 2	3 0 / 1 2 / 2	0 6 / 0 1 / 2 3	3 / 0 1 / 2	2 0 / 0 1 / 2 3	7 / 0 1 /	0 3 / 0 2 / 2 3	1 0 / 0 2 / 2 3	1 7 / 0 2 / 2 3	2 4 / 0 2 / 2 3	0 3 / 0 3 / 2 3	1 0 / 0 3 / 2 3	1 7 / 0 3 / 2 3	3 / 2	1 / 0 3 /	0 7 / 0 4 / 2 3	1 4 / 0 4 / 2 3	2 1 / 0 4 / 2 3	2 8 / 0 4 / 2 3	0 5 / 0 5 / 2 3	1 2 / 0 5 / 2 3	1 9 / 0 5 / 2 3
research potential projects																																							
Research to identify exactly what the problem is to be solved																																							
Prototype and evaluate potential models																																							
Evaluate potential models																																							
Select viable models for evaluation																																							
Evaluate developed models and compare to existing models																																							
Do write up																																							
Prepare final demo																																							

14 References

Fleming, S. W., Watson, J. R., Ellenson, A., Cannon, A. J., Vesselinov, V. C. (2021). Machine learning in Earth and environmental science requires education and research policy reforms. Nature Geoscience, 14(12), 878–880. https://doi.org/10.1038/s41561-021-00865-3.