Uppsala University

Decentralized Air Quality Monitoring and Prediction

(1DT054) - Autumn 2020

Course Report

Group 2:

Dhanush Kumar Akunuri
Md Tahseen Anam
Jude Felix
Lokesh Kumar
Venkata Sai Teja Mogillapalle
Seema Negi
Ankit Pandey
Anil Poudel
Vandita Singh
Praveen Swamy

January 17, 2021

Contents

Abstract

The course Project CS (1DT054) is a full-time course held during Autumn 2020. The aim of the project was to develop a product for a customer who was the industry partner, Ericsson for this year. The product DAMP was developed to predict the quality of air for the next few hours by using Federated Learning. This report comprises of the course-specific information and focuses on the experiences of the team during the project. There is insight about how the project was completed with the challenges posed to the team during the Coronavirus pandemic and the available resources. The take-away for all the team-members from this project have been presented in this report. The technical details of the product have been described in detail in the adjoining Product Report for DAMP Autumn 2020.

.

1 Introduction

1.1 Course - Project CS

The course 'Projekt DV/Project CS (1DT054) required us to develop a software for distributed system as a team of 10-15 students. The main aim was to get an insight into how a real-time industry based project is run with modern software development principles and practices.

The project De-centralised Air Quality Monitoring and Prediction targets to predict the air quality using Federated Learning.

1.2 Ericsson

The customer was Ericsson and Konstantinos Vandikasas from Ericsson as the Industry partner. The specifications for the project were presented to our team and this marked the start of the project to develop the product DAMP.

1.3 Objective

The objective was to develop a product for monitoring and prediction of Air Quality in Stockholm by using Federated Learning. The system takes as input air concentration data along with selective meteorological parameters to predict the Air Quality at different stations in Stockholm. The major pollutants of concern were Nitrogen dioxide (NO_2) and Particulate Matter (PM_{10}).

The data was provided by Sweden's Meteorological and Hydrological Institute (SMHI). The product is targeted to provide a forecast stating the level of pollution in Stockholm which would be helping people belonging to certain sensitive groups in deciding when they should/should not go out and thus reduce their risk of being exposed to Air Pollution.

This report comprises of the course related details and provides an insight into the experiences of the team while developing the product DAMP. A technical description of the product DAMP has been presented in the

Product report submitted alongside this Course Report for 1DT054 (Autumn 2020).[1].

2 Resources

2.1 Hardware Resources

Usually, a project lab is assigned which is said to be well-equipped with state-of-the-art hardware resources. However, this time due to the restrictions imposed by Coronavirus pandemic, it was required for us to work remotely. Systems with adequate computing facilities were available with all the team-members and no difficulty was faced in terms of lack of appropriate hardware resources.

2.2 Software Resources

This year due to the COVID-19 pandemic, the course was conducted online. The following resources were used during the development of the product.

2.2.1 Cloud Resources

Each team member was provided an id for accessing the cloud resources at the Department of ITC, Uppsala University.

2.2.2 Clockify

Clockify [2] is a project management tool, used to monitor and manage the activities efficiently within a project. It helps mainly in tracking the amount of work invested in different tasks and thus we can rearrange the resources based on the task complexity to improve the team productivity. Since, each team member periodically updates the time-sheet so everyone knows the task on which the group members are working on and the time taken by them.

2.2.3 Trello Board

Trello [3] is a visual collaboration platform which helps in organizing the projects and the tasks in the form of boards. It was used in this project for creating different tasks within each sprint and assigning the members for each board. The drag and drop feature is very handy in organizing the boards. In one glance we get to know what's being worked on and who is working on what. We can imagine it as a white board, filled with a list of sticky notes, with each note as a task for the team members.

2.2.4 Zoom

Zoom was used as the main channel for all the scrum meetings, discussions within the subgroups and for the group activities.

2.2.5 Slack Channel

Slack [4] channel was used for all the public and private communications within the group and for sharing important resources related to the project. Several channels were created based on different tasks to organize the conversations. The add-on feature Simple-poll was used for creating subgroups and for other decisions.

2.3 People

Selim Ickin, Jalil Taghia and Konstantinos Vandikas were the industry representatives from Ericsson Research [5] this year. However, most of the interaction was done with Konstantinos Vandikas who introduced the team to the specifications of the product. The main person responsible for supervision of the course was Vera Von Zoest and assistance was provided by Tobias Mages and Shenghui Li.

The development of the Decentralized Air quality Monitoring and Prediction was done by us as a team of ten members assisted by Shenghui Li. All the team members, being from different backgrounds had varied expertise areas which proved to be very useful for the project to be a success. All the decisions were made jointly after discussing in the daily meetings that we had over zoom.

2.4 Tools

The development of the project was carried out using the following tools.

2.4.1 Google Colaboratory

The Google Colaboratory [6] facilitated us to write, run and share code within Google drive using Python. This platform supports versioning of code - with a single click the notebook gets saved to the github repository. It also has dedicated GPUs and TPUs which is a great addon for intensive Machine learning tasks.

2.4.2 Jupyter Notebook

Jupyter notebook is an open source web application used to write, run and share the python jupyter notebook codes. It was formerly called as Ipython Notebooks. A Jupyter Notebook document is a JSON document, following a versioned schema, containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media, usually ending with the ".ipynb" extension.

2.4.3 Git and GitHub

Git is the most commonly used for version control system[7]. Git tracks the changes made to files, so there is a record of what has been done, and it can revert to specific versions that is needed at times. Git also makes collaboration easier, allowing changes by multiple people to all be merged into one source.

GitHub is a subsidiary of Microsoft which provides hosting for software development and version control using Git. It offers the distributed version control and source code management functionality of Git, plus its own features.

3 Project methodology and organization

The project started with understanding the specifications provided by the industry partner- Ericsson. At various stages, we were given an opportunity to have a Q/A session with the representative from the Industry. This activity greatly helped us to develop a clearer understanding of the project goals. The different roles were assigned to certain members but the opportunity to be the scrum member was given to each team member on a rotation basis. There was however, no dedicated planning team for the project though everyone was free to propose any idea/ suggestion, which was discussed over and decision used to be made using a poll.

There was an interesting competition organised for the development and optimization of the centralized model. Both the teams (our team Group 2 and the other team Group 1) working on the project DAMP, were required to make submission as a Notebook along with the Comma-separated values (csv) files for the stations - Stockholm E4/E20 Lilla Essingen (18644), Stockholm Sveavagen 59 Gata (8779), Stockholm Hornsgatan 108 Gata (8780) and Stockholm Torkel Knutssonsgatan (8781). This was an amazing opportunity for the team to make improvements to the centralized model. This helped us to achieve good scores by the last Sprint.

Virtual Activity was organised by Ericsson for all the course participants which helped to establish better capabilities for working in a team.

3.1 Scrum

Scrum is a project management methodology where team members work and develop to tackle the issues in the cleverest way[8].

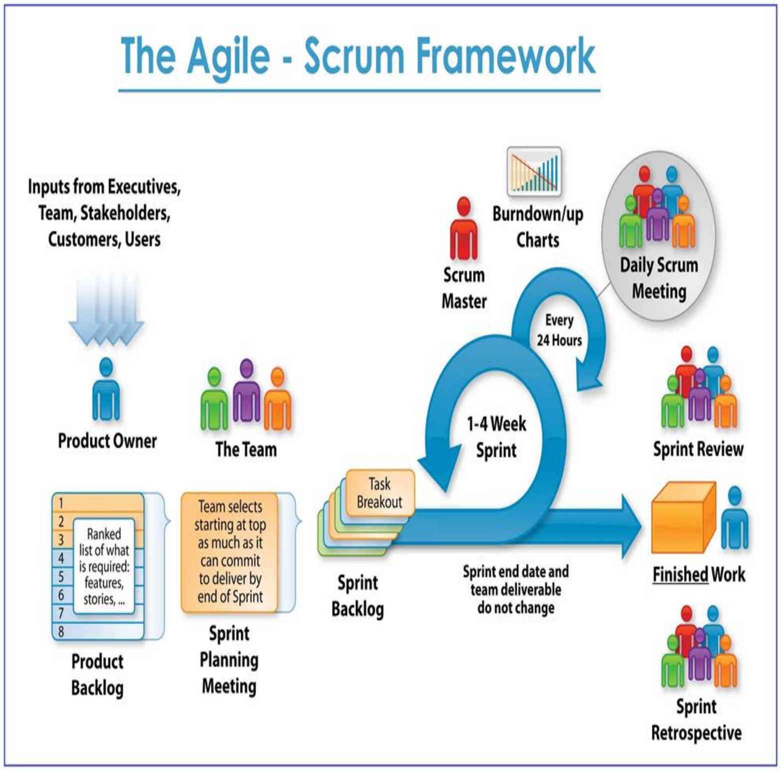
The Agile - Scrum Framework
Inputs from Executives, Team, Stakeholders, Customers, Users
Product Owner
The Team
Scrum Master
Burndown/up Charts
Daily Scrum Meeting
Every 24 Hours
1-4 Week Sprint
Sprint Review
Task Breakout
1
2
3
4
5
6
7
8
Ranked list of what is required: features, stories, ...
Team selects starting at top as much as it can commit to deliver by end of Sprint
Sprint Backlog
Finished Work
Sprint end date and team deliverable do not change
Product Backlog
Sprint Planning Meeting
Sprint Retrospective

Figure 1: Scrum Framework[8]

It was planned to follow scrum framework[9] in order to organize and carry out the project. The tasks were divided into sprints. Every morning from Monday to Friday, members would meet online to discuss what was done by everyone up to that day and plan about the tasks needed to be done in the upcoming days. It was decided that each member in the group would be provided with the opportunity to be the scrum master for one week. Therefore, a 10 weeks window was created and each member was assigned one week within that window when that member would perform as a scrum master. This process would continue in a cycle. So at the end of the last week of the window, same process would start again from the beginning.

3.1.1 Responsibilities of scrum master

It was agreed unanimously that the following task would be performed by a scrum master:

1. Writing the meeting's short overview.
2. Taking status updates of each member in the meeting.
3. Recording the ongoing tasks.
4. Updating the Trello board accordingly.

3.1.2 Advantages

Scrum methodology worked really well for us. It kept everyone updated about the ins and outs of the project. All of us knew what others in the group were doing with the project. Daily morning meeting helped everyone to get involve in the project and maintained the team environment required to carry out the project successfully. It kept everyone motivated. Because everyone got the opportunity to perform as a scrum master, all members took the ownership of the project and perform their tasks sincerely. Everyone was well informed and aware all the time. The morning meetings were held as open discussions where anybody could seek any kind of help or discuss about anything even if it is not within the context of the project. We believe that the project was a wonderful demonstration of team work.

3.1.3 Disadvantages

It was discovered later on that not everyone was able to attend the morning meetings everyday because of various issues. Therefore, instead of arranging a morning meeting everyday, it was decided to arrange the morning meetings three days in a week. Those three days of the week were unanimously selected by discussing among the members of the group.

3.2 Sprint Breakdowns

The project development was organised into several sprints with each Sprint lasting for a period of two-three weeks. Each Sprint was set to achieve a defined target and had some set of tasks to be completed by sub teams having high bandwidth communication among them.

The following sub sections provide some description about the goals and tasks of the team. An instance from the Trello Board can be seen in Figure 2.

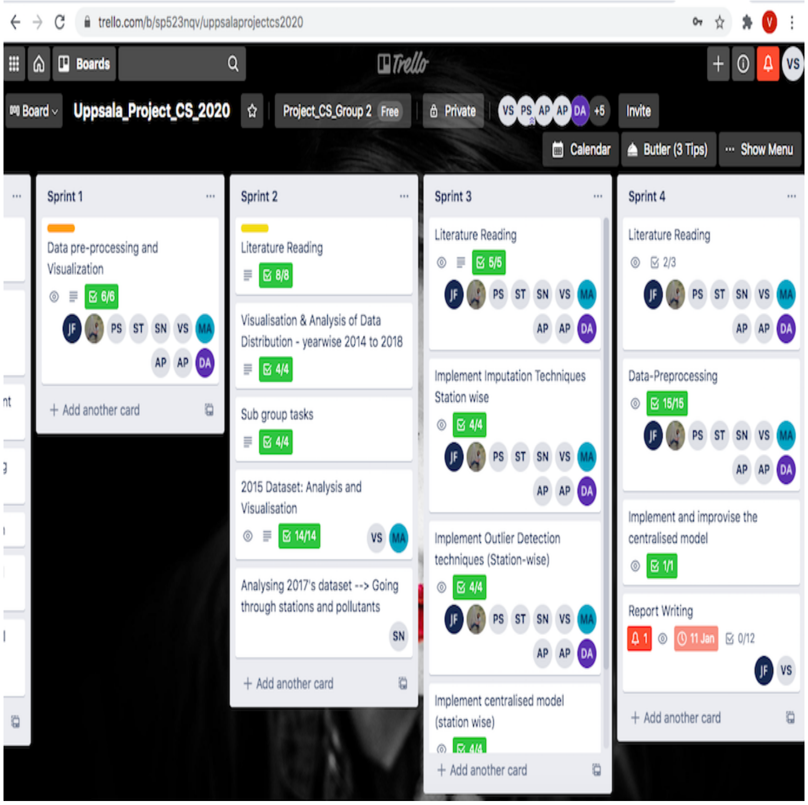
trello.com/b/sp523nqv/uppsalaprojectcs2020
Boards
Trello
Board
Uppsala_Project_CS_2020
Project_CS_Group 2 Free
Private
Invite
Calendar
Butler (3 Tips)
Show Menu
Sprint 1
Data pre-processing and Visualization
6/6
Add another card
Sprint 2
Literature Reading
8/8
Visualisation & Analysis of Data Distribution - yearwise 2014 to 2018
4/4
Sub group tasks
4/4
2015 Dataset: Analysis and Visualisation
14/14
Analysing 2017's dataset --> Going through stations and pollutants
Add another card
Sprint 3
Literature Reading
5/5
Implement Imputation Techniques Station wise
4/4
Implement Outlier Detection techniques (Station-wise)
4/4
Implement centralised model (station wise)
4/4
Add another card
Sprint 4
Literature Reading
2/3
Data-Preprocessing
15/15
Implement and improvise the centralised model
1/1
Report Writing
11 Jan
0/12
Add another card

Figure 2: Trello Borad : An Instance

Sprint 1

During the first sprint, it was essential to understand the basic concepts which form the basis of the project development. This sprint lasted during the first three weeks from the start of the project. Each team member was involved in conducting a thorough Literature Survey with the aim of reflecting on what was read by them. The articles were mainly related to Air Pollutants, Development of Air Quality Prediction Systems, Possible Pre-processing Methods for Air Quality data, Feature Engineering.

This task was followed by performing Visualization of the data either using Tableau or 'matplotlib' package in Python. The data was downloaded from Visualization and analysis of the data distribution was done for the years 2017 and 2018. Imputation Methods were also tried. A naive centralized model(from tutorial was also implemented).

Sprint 2

The second sprint lasted for two weeks with the end goal of visualizing and analysing data distribution for the years 2014 , 2015, 2016, 2017 and 2018. The team was divided into five sub-groups of two members each, where each sub-group was assigned the task of visualising and analysing data for the years 2014, 2015, 2016 , 2017 and 22018 respectively. The hourly distribution of air concentration data was combined with weather components - Imputation Methods were also tried.

Sprint 3

During this Sprint, there were two groups of 5 members each working on the selected stations for the years 2014 - 2018. Visualization and Analysis of the Data Distribution for the years 2014 , 2015, 2016, 2017 and 2018 was done Station-wise. Imputation Methods were tried and evaluated. Methods for Outlier Detection were also being tried. Centralized Model based on Tutorial was a good resource to be followed but could not be taken further for building the time-series prediction model.

Sprint 4

During this Sprint, the division into sub-teams was the same as previous Sprint, each sub-team working on the selected stations for the years 2014 - 2018. The main target was to build the time-series based centralised model and perform the evaluation. The various tasks accomplished were (i) combining more meteorological features with the air concentration features, (ii) Implementation of various methods for Imputation, Handling Negative Values, Outlier Detection, (iii) Implementation and Improvisation of Time-series prediction model (DNN and LSTM) for the next 24 hours, (iv) Applying the "SMAPE" metric to evaluate the performance of model, (v) Addition of

Time-Based features (Weekends, Weekdays, Morning, Afternoon, Evening) to the model. The Literature Reading was a continuous process throughout the different stages of the project.

Sprint 5

During this Sprint, the target was to build the best performing centralized model. Optimization was being done to make the centralised model perform better. The work was presented in Mid-term presentation. The tasks for optimisation were similar to the last sprint with focus on optimising the features for the

Sprint 6

This was one of the most crucial Sprints during the course with the target to build the best performing centralized model, implement Attention Mechanism, Federated Learning Model, Build Front End for the product- DAMP, Plot Cross Correlation for the input features and finishing the first draft of the reports. There were four sub-teams- one working on implementing the Cross correlation method, another working on the Attention Model, third working over and Front End development and fourth over Federated Learning Model implementation. Articles related to Cross Correlation, Implementation and Optimisation of Centralized Model,Front-end Implementation, Federated Model Implementation, Attention Mechanism were studied. Cross Correlation Method was implemented and plots were generated . The final feature selection and analysis was performed. The 24 hour model and 1 hour model for Time Series Prediction from 01/10/2019 to 31/12/2019 was implemented, improvised and optimised. The Attention Mechanism was applied to the Centralized Model. The implementation of the Federated Learning model was started. The product Report and Course Reports were started to be written by each person as per the assigned section.

Sprint 7

This Sprint was all about finalizing the model, implementing federated learning and attention model being tried and writing reports. It was also decided to create a web user interface where user can interact with the product. The end goal of this sprint was to come up with a working product that showcases our works and meets the project specifications successfully.

With the completion of each Sprint, considerable progress was made in the project and the activities completed in the last Sprint were reviewed with each sub-team presenting/sharing/reporting their work.[10]

3.3 Rules

In order to get the best out of the project, certain rules were set. The Base hours were set from 08 to 17, however the core hours remained to be 09-16.The Daily standup was held at 9.00 everyday. Other meetings were decided to be held during the core hours at times decided unanimously by the team. The Lunch time was usually from 12 to 13 or according to individual preference. Break hours were flexible as per the individual preference. Clockify was used to do Individual tracking of 7 hours worth of work per day. Doing work on weekends was not much encouraged. The team members were expected to be on time for meetings and presentations and in case they were not able to attend the next day's meeting, it was required to be communicated through Slack well in advance (i.e. one day before). The language used for the communication was English only, even during breaks or casual activities. In case of any problem with the team dynamics or the behaviour of a group member, it was required that the matter be brought up forward as early as possible.

4 Conclusion

Despite the challenges posed by the Covid-19 pandemic and limited physical engagement in the course work, the project turned out to be successful. The course was well coordinated and co-operation was received by all the resource persons associated with the project. There were at times felt some lack of synchronisation among the team-members which was a matter of concern for timely completion of certain tasks however, steps were taken to deal with such issues. During the start of the project there was observed, a lack of planning and organisational issues, but was tried to be regulated by discussion among the team members to organise and plan sprint tasks and goals more effectively. Certain issues with communication and planning were faced by the team in the early stages but were dealt with ease with the willingness of each member to collaborate for achieving the overall goal of making the project a success. Relatively less work was done on

The kick-start meetings helped us all keep track of the on-going work where each member used to share what task(s) were and would be done by them. This was a commendable way to keep up the motivation, share knowledge and make everyone feel involved. However, during later stages, lesser meetings lead to certain issues related to lesser communication and synchronization.

The Retrospective held at the end of each Sprint helped the team realise the direction of work and whether it would or would not lead to our final goal. Likewise, the tasks completed in previous Sprint was either taken forward or sometimes completely or partially discarded.

The tasks for different stages of the project were shuffled among the team members which greatly helped everyone to learn and gain experience in every aspect- technical and other skills.

There were some activities which helped the team members to know each other well. During some of the meetings, Fika was organised and a potluck lunch was planned among the team-members but could not become a success due to stricter guidelines issued due to the Coronavirus pandemic. There were some fun activities organised by Ericsson and were enjoyed by all of us. The work was tried to be completed taking into consideration the best practices and made us learn the real essence of working together as a team.

Acronyms

DAMP Decentralized Air Quality Monitoring and Prediction. 1, 3, 6

References

[1] U. U. Dept of Information Technology. (). De-centralized air quality monitoring and prediction () product report, project cs ht 2020.

[2] Clockify, *Clockify*. [Online]. Available: https://clockify.me/.

[3] Trello, *Trello*. [Online]. Available: https://trello.com/en.

[4] Slack, *Slack*. [Online]. Available: https://slack.com/intl/en-se/.

[5] Ericsson, *Ericsson research*. [Online]. Available: https://labs.ericsson.com.

[6] Google, *Google co-laboratory*. [Online]. Available: https://colab.research.google.com/notebooks/intro.ipynb.

[7] S. Chacon and B. Straub, *Everything you need to know about git*. Apress, 2020.

[8] Edelon, *Scrum*. [Online]. Available: https://edelalon.com/blog/2017/09/scrum/.

[9] K. Schwaber, *Agile Project Management with Scrum*, ISBN: 10 073561993X, 13 9780735619937.

[10] J. Shore. (). On the dangers of doing scrum the wrong way, [Online]. Available: https://www.jamesshore.com/v2/blog/2008/the-decline-and-fall-of-agile.

A Appendix A: Individual Contribution by Team Members of group 2.

A.1 Dhanush Kumar Akunuri

Being part of project DAMP was an excellent opportunity for me to have an industrial experience and made me a better team player. During this COVID situation, working from home was new for me. We used to have daily meetings which helped me to stay motivated and keep track of my work.

At the initial stages of the project, my teammates and I devoted our time to Literature readings. We started by collecting all the articles and research papers related to Air Pollution, Time Series Forecasting, and Federated Learning. We divided all articles among ourselves, and at the end of this sprint, we updated a short description of these articles into a word document. In the next sprint, we were divided into five sub-groups with two people in each group. I, along with Praveen, was assigned with the year 2014 data. I started following different techniques to pre-process and visualize the data. I tried a few imputation techniques to fill missing values in the data. I've learned using Tableau for data visualization. By the end of sprint 3, MICE, DNN and K-NN were three imputation techniques I worked on. After this, I downloaded Meteorological data and performed some data pre-processing steps. I built a final dataset by combining Meteorological data and Air Quality data.
After Pre-processing the data, Feature Selection and building a centralized model was our sprint 4,5 goals. I used correlation plots to finalize meteorological features like air pressure, wind speed, and air temperature. While building a centralized model, I followed the TensorFlow tutorial for time series prediction. Data windowing, converting data into time series data format are a few techniques I was not familiar with and had to work on them. Then trained data on the LSTM model and also tried different architectures of LSTM's. Our TA hosted a competition to speed up the development process where our models are ranked based on loss values. Out of all the imputation techniques mentioned above, KNN imputation was performing better and decided to use it. The goal for the next sprint was to optimize our model performance. Auto-correlation, time-based features like day sin, day cos were tried to improve model performance.

In the final phase of the project, I, along with Anil and Ankit, worked on building a visualization dashboard. I created an outline of the dashboard using Dash, and plotly, later we agreed on continuing with Anil's approach. When it comes to Federate Learning, I tried implementing using Praveen's method. Finally, I contributed to the report writing and presentation review part.

A.2 Md Tahseen Anam

The project offered a great opportunity to learn different methods used in data pre-processing, analysis, visualization and different sequence models used in time-series analysis and data forecasting. The project work was organized in multiple sprints. During the first sprint, I and my group members mostly focused on literature reviewing, studying data analysis methods and were trying to understand our dataset. I started the project by reading research articles, preparing a summary of each article and sharing the concepts and methodologies mentioned in those articles with my other group members.

After reading some research papers and articles, me and my group members started analysing our dataset. We decided to distribute the stations among us so that each of us can analyse the data of one station and later combine our findings. I at the beginning used an interactive data visualization software called Tableau to visualize and analyse our dataset. The software provides tools to generate plots and graphs, moreover it provides tools to generate trend lines and provides equations explaining those trend lines. I shared my findings with my group. Me and my group members generated .csv files from the dataset in order to start processing the data.

I along with my group members then started implementing different preprocessing techniques on our dataset. I combined air quality and meteorological data, generated co-relation matrices between features, observed the percentage of missing values in each feature, worked with different imputation techniques like KNN and outlier detection techniques such as Isolation forest. I also generated auto-correlation plots and studied the correlation at different lags. My group members also worked with different imputation and outlier detection methods.

Then we started designing models to predict the air quality. I used RNN and LSTM models and my group members also worked with different models to predict the air quality. I also participated in the competition for the centralized model and therefore, predicted PM2.5, PM10, NO2, NO2 as NOX features. I experimented with different window sizes and found that predicting one hour worked better than predicting the next 24 hours. From the results of all models we tried, the best model was chosen unanimously. Then I kept working on improving the centralized model and simultaneously contributed to the report writing.

A.3 Jude Felix

The project decentralised air quality prediction was a project which had scope for data analysis, preprocessing and machine learning with federated learning. This is was great opportunity for me to work with a bigger team with an industry partner and good chance to put my theoretical knowledge into practical use. Working on the project made me learn new technical concepts and more importantly evolved me into a better team player and better professional.

All the phases of the project was implemented with sprint breakdowns. There were seven sprints in total and I took part in all the sprints with great enthusiasm and brought the best results. To begin with, the team decide to read articles on the data analysis of time series data for forecasting, different preprocessing techniques of time series for forecasting and different machine learning techniques for forecasting weather. Articles on federated machine learning was referred for the understanding it. Each of us read different articles.

After having a understanding of the basic concepts for the project. Each of us began to work to analyse the data from different stations between the years 2014-2018. I analysed the data for the year 2018 and the data from the station Stockholm Torkel Knutssonsgatan. The data was analysed for the different features and the finding the correlations other visualisations. After having a detailed understanding of the different techniques used for preprocessing of the time series data, I and my other teammates began to implement the techniques we read about and visualize the data to prepare for the models. After visualizing the data we were able to see some missing values and outliers. I and my other teammates worked on different imputation techniques and outlier techniques.

After finishing with the imputations and the outlier detection techniques, I worked on implementing deep neural networks for forecasting the PM10. I worked with different architectures after referring to articles. I tried with different versions of deep networks like convolution neural networks and adding and removing different layers. I also participated in the competition for the centralized model.After finalising the centralised model, I helped in implementing the federated models. Later on contributed in the product and course reports and helped in the teams' presentations. I was constantly helping my teammates whenever they needed some assistance.

A.4 Lokesh Kumar

Though I had previous Industry experience of working on a non-machine learning project in an agile work environment. But, This project gave me a unique opportunity to work on a machine learning problem in an Industry like setup with a big team. Under this project in such a big team I learned how to approach a machine learning problem and divide it into multiple essential small phases involved in any machine learning problem. Like Data Prepossessing, Data Visualization to understand underlying data-set, How to Impute available data if the available data is not perfect or missing and finally build ML models, predicting and comparing score to finalize best model and its parameters. Through out the course Apart from the Technical aspect of the project I took part in the daily meetings over zoom most of the times but missed sometimes due to serious health issues.

The team work kick-off with the Literature reading phase to make understanding of different articles related to Air Quality, pollutants monitoring and relevant ML techniques preferred in this kind of problem. I went though many articles and shared the knowledge with team during the meetings. The team then conceived some essential sub phases of the project and picked Data processing and visualisation task to start with. I contributed in Data prepossessing work by working on many Imputation techniques like mean/median, Most Frequent, K-nearest neighbour on station no. 8781 for 2014 till 2018. Other Team members also shared the results of various Imputation techniques and most of team went with KNN imputation technique for centralized model later on. I also tried to work on some outlier detection techniques but the results didn't improve so I discarded this part from my code.

After Imputing the data and finalizing the stations based on common(NO2, NOX as NO2, PM10, P2.5) pollutants, I repeated the process with other finalized remaining stations as well. I merged few meteorological features like

Air pressure, Air pressure, Rel Relative Humidity etc after discussion with the team. After this Our team Decided to work on Centralized model. I went through tensor flow tutorial to build a LSTM model(RNN model) to predict PM10 pollutant using a Bidirectional windowing system as input. I tried to improve my model score by normalizing the the meteorological data as suggested by Shenghui Li but my results didn't improve much so I didn't submit the model to the competition. Later I shared the code with the team rather. After working on centralized model, I did some work on Federated learning model as well but due to less time and deteriorated health I couldn't able to complete it. Later on I contributed in the report writing and presentation review work.

A.5 Venkata Sai Teja Mogillapalle

This project has helped me gain some understanding and hands-on experience with some of the state-of-the-art technologies that were used for solving the problem statement. I have some previous experience in Machine Learning, but I haven't worked with federated learning, attention mechanisms and using LSTMs for time series forecasting and I got some good knowledge of the stated topics. Working online was a bit tough in the beginning, but we had daily meetings and weekly sprints that helped us to be on track and know how other people in the group are approaching the problem.

Our work started with a literature survey and I have read multiple research papers on federated learning, air quality prediction, and time series forecasting and prepared a summary of the research papers that I have read. Next, we have divided the tasks as Data Visualization, Data pre-processing, building a centralized model, and building a federated model. For the data visualization part, I have proposed the team for using Tableau software and people accepted it and they showed enthusiasm for learning a new tool for visualization. I have made some visualization plots by combining the data from different years and different stations. Next task was dividing the entire data into 5 parts and we were divided into 5 groups with 2 people in each group. I worked on a station with id - 8779. I have worked on trying different pre-processing steps like finding correlation plots, some imputation methods - linear imputation, imputation after removing seasonality and trends in the data, statistical imputation methods; time-based attributes etc.,

Next task was to build a centralized model. I have followed TensorFlow tutorial for - Time series forecasting using LSTMs for this task. I have built an LSTM model following the tutorial. Shenghui Li has brought in the discussion for using Attention models for the time series forecasting. I found the topic new and pretty interesting so I started to read some research papers on attention models and how they are used in time series forecasting(attention models were mostly used in tasks like text translation, so using them in a different domain has caught my eye). I developed a centralized model based on attention mechanism, the score was decent but not better than the usual centralized model so I didn't upload my solution to the competition. The model could have been improved but I couldn't work more on the model due to time constraints. Though I haven't implemented the federated model, I understood in detail how it worked from the solution that was already submitted. I have played with some parameters and observed how the result was changing for the federated model. I have also actively participated in the mid-term and final presentation sessions and have made my contributions to the product report as well.

A.6 Seema Negi

The DAMP project was an excellent and unique experience for me. Working on real-world data allowed me to understand how to deal with non-uniform and missing data. Working remotely in a big group due to COVID situation was a new and unique experience. It allowed me to understand how to work in a team which works remotely. We used to have daily standup meetings which made us keep track of the work and helped us learn from each other. The initial work of the project consisted of reading articles and research papers. The research papers were divided among the group members, and the task was to read the paper and write a short summary. I wrote the summary for an article and did some initial study by reading articles on time series prediction and other related technologies.

The next task was to know more about the dataset. The group was divided into five subgroups containing 2 group members each. I worked on year 2017 data. It included data analysis which consists of finding the no. of missing values of an air pollutant, finding the stations with common attributes, analyzing the meteorological features etc. The next task was Data Preprocessing, where groups were divided into two subgroups containing five members each. My team and I worked on data imputation techniques. We read about the imputation techniques, implemented and compared to see the results. I worked on the MICE and KNN imputation technique. A fair

amount of time was given for data preprocessing
. After the data preprocessing phase, it was decided to work on the centralized model. I implemented the DNN, Multistep dense model and LSTM models with the help of Tensorflow's official tutorials and documentation for time series forecasting. I submitted two different models(DNN and LSTM) for the competition. To improve the performance of the model, I tried to add different time-based features like time of the day(morning, afternoon, evening, night), Weekdays, Weekends etc. However, it did not improve the performance too much.
Later, my team members and I worked on implementing the attention mechanism. As this was a new concept, I understood it by reading articles and then moving to the implementation part. Due to time limitation, the visualisation part of the attention weights was not completed but results were achieved for PM10, PM2.5, NO2 and NOX as No2. Apart from this, I also contributed to the report writing and actively participated in Midterm and final presentation.

A.7 Ankit Pandey

This project provided me a great platform to gain some hands-on experience on real time data and aggravate my knowledge base. Though I had some prior knowledge of Machine Learning, I had never worked on federated learning or time series forecasting. Through, the course of the project, I gained good amount of knowledge to work on these topics. I have some work experience of working with diverse teams and with agile teams, but this was a different and unique experience especially with the COVID situation. It was a little difficult initially, but since we had daily morning meetings, we quickly synced with each other. This also helped us keep track of each other's work. We started our work with literature study. I read research papers and articles on air quality monitoring, air pollutants, time series forecasting, federated learning, data imputation and a few more topics. I made summaries and discussed them with the team in our meetings. This helped us cover a wide range of topics since everyone discussed their summaries, which in turn helped us learn from each other. After the literature study, I started working on the practical aspect of the project. After downloading the data, my team and I planned to analyse it. I started learning Tableau to analyse the data. I found some trends and seasonalities which would help us with knowing our data better. Along with that, I also worked on finding how to fix the missing values issue. I worked on a few imputation techniques like Multiple Imputation Using Chained Equations (MICE), KNN Imputer, Simple Imputer. I shared the results with my team, and we decided on which technique would work best for our model. Since data pre-processing is an integral part of the DAMP project, I spent a good amount of time to get that part right. I also tried to implement the Generative Adversarial Imputation Network to impute data. But due to time constraints, I could not completely get the desired results. I experimented with different types of time features as well. Along with that, I researched into different meteorological features after which we check correlations to check if a particular feature is important. After the data preprocessing part, we decided to start working on the centralized model. I implemented the baseline mode, DNN model and a single step LSTM model. I used tensorflow's time series forecasting tutorial to implement the above-mentioned models. I tried different prediction windows and compared the result. I also submitted 8 solutions to the centralized model competition. One of my models, where I predicted 1 hour into the future using DNN, got the best average SMAPE score of 0.28. I tried to predict 24 hours, 6 hours and 1 hours using different Input windows as well. I also participated in the mid-term and final presentations as well. To conclude, I feel this project was a great learning curve for me. I got some hands-on experience of working with real time data. I also learnt more about federated learning and how it could be useful during this time when the topic of data privacy is highlighted to a great deal.

A.8 Anil Poudel

This course was exciting and equally challenging for me, with not having much experience in Machine Learning techniques. As soon as the course started, I started learning the ML techniques side by side, working with the project. Before starting with the project, We assigned some roles to make the project more structured and managed. The roles were Team leader for managing the overall group activities, Time manager for organizing the zoom meetings, and Secretary for keeping contact with the teacher and the external partner representing the group. The group gave me time management responsibility, and I managed the group meetings without missing a single meeting throughout the course.

Intending to make this project flexible and smooth, we divided the tasks into different sprints. Before jumping directly into the models, we needed to understand various air pollutants and Federated learning concepts, so we decided to start the first sprint with a literature study focusing on air pollution, Federated Learning, and prediction models. I started reading different articles and research papers, and some suggested by the course teacher, course teaching assistant, and group members. After this sprint, we started working with the data preprocessing, where we divided into subgroups and worked together by downloading the air pollution and meteorological data. Then went through data analysis, visualization, and tried different imputation techniques and other outlier detection techniques. When completed the data preprocessing, I started working on the centralized models asking my teammates' help when required. I tried both DNN and RNN based LSTM model but wished to stick with LSTM model. After successfully building the centralized LSTM model, I worked on optimizing the model. During this process, I tried improving the model by adding and removing different meteorological features, handling negative values, and imputing the values with varying imputation techniques.

After this, the next step was to build a Federated Model. But during the task division, I was assigned to work on building the product, a WebApp where we could display our results to the users. With my previous learnings, tutorials, and help from my teammates, I successfully created a Python Flask WebApp called DAMP. After completing this, I wanted to understand the Federated model and learned that too from my teammates working on a federated model. Then the next steps were to work on the final report and the final presentation. We worked together for preparing report and presentation slides.

The intention for taking this course was to learn new skills and in the same way, get an experience of working on the project with the team. Concluding it was an excellent experience for me working with such amazing teammates, mostly learning and sharing knowledge among us.

A.9 Vandita Singh

The project Decentralized Air Quality Monitoring and Prediction(DAMP) was a wonderful opportunity for me to work as a part of a big team as in a real industry-like setup. The experience gained not only helped me to learn some new technical concepts but grow and develop as a better work professional by being able to communicate, resolve conflicts and work towards the achievement of team goals.

The first phase of the project demanded an intensive literature study to understand the concepts that were relatively new to most of us such as Air Pollution and Federated Learning. I was assigned three Research Papers based on Federated Learning, Outlier Detection and Air Quality Prediction using Deep Learning, for which the summary was prepared and shared by me. I delved into reading and tried to develop an understanding about different Data Visualisation and Pre-processing methods which I could later put to implementation and generate useful plots for the part of data assigned to me.

The task of Data Collection, Visualization and Pre-processing of the data for the year 2015 was successfully completed by me and sub-team where we worked using both Tableau and Python libraries. The next sub-team I worked with, was assigned the task of Pre-processing and Implementation of the centralized model, where I worked on the station id 8780 and shared my code on the GitHub repository for our team. Several different techniques at each step were tried by me and I discussed those with my team and seek for help from them wherever necessary. I gained experience in applying relevant Imputation Techniques such as K-NN and try certain Outlier Detection Methods. I also implemented the time-series prediction model for forecasting the air quality for the next 1 hour using tensor flow based Time-Series Prediction tutorial, using Air quality and meteorological features such as Temperature, Relative Humidity , Wind Speed and Direction. The next task was to optimise the model and different combination of features were tried and tested. I worked on the centralised models implemented using Deep Neural Network (using tensor flow) and LSTM (using keras).

I evaluated the performance using Mean Absolute Error(MAE) initially. At later stages, the features were required to be optimised and some time-based features were extracted to get the best possible SMAPE scores. It was decided that the Attention Mechanism be incorporated into the centralised model to better understand the relative importance of the input features, I enjoyed helping my sub-team to implement the Attention Mechanism. I worked on three approaches for implementation of the Attention Mechanism. The first approach was using Seq-Self-Attention Package based on Bi-LSTM, the second approach was based on DNN using iNNvestigate and the third using Bahdanau's approach. Alongside the implementation work, I was also trying to keep up with the literature study part. Due to the need to fast track our work, we got divided into sub-teams, where I could not have the first hand experience of implementing the Federated Model but learned about it through my team members.

I was actively involved with the project management and planning activities which not only helped me as an individual to be better able to organise but the entire team as a whole to focus on achieving the well-defined short term goals. Another task was the preparation of presentations and reports to which significant amount of contribution was made by me.

The project certainly helped me to accomplish my personal learning objectives while also ensuring my team succeeds in achieving the targets set.

A.10 Praveen Swamy

The problem definition in this project had a full scope of end-to-end implementation involving an immense data pre-processing, analysis, visualization, and machine learning model implementation for the forecast of the PM10 pollutant. Every phase involved a detailed work to be done, and thus we organized the project work in different sprints. During each sprint, we had subgroups within us so that each subgroup would be part of a different task to achieve the result. Each sprint had a scrum master who was responsible for taking the daily status updates and update the work log on shared drive. During the first sprint, I along with my other teammates focused on reading the research articles related to Predicting air pollution using different ML techniques. Understood different data pre-processing techniques by referring to Kaggle examples, open source git hub projects and several blogs on Medium.com and towardsDataScience.com. Also, studied some of the outlier detection techniques from multiple resources.

Once, I had an idea of different methods and the right techniques to be followed for this problem, I along with Dhanush Akunuri started working on using different techniques to visualize and understand the variance of the data set 2014-2018 first, using different plots such as scatter plots, missing number etc. Also, checked on the missing values and understood the data variations and trends according to hours, days, weeks, months, and years. Then, we tried different data imputation techniques to impute the missing values.

Once, I had a clean dataset then I and my other teammates switched to work on building a centralized model using LSTM networks. Before working on the Machine learning model the dataset needed a conversion into time series data in the format of samples,time-steps, and features for the LSTM models. We then worked on the windowing technique referring to the tensor-flow tutorials. I tried out different architectures and types of LSTM like Gated recurrent neural network and Bi-directional LSTMs. Then, we trained the model on the normalized dataset with proper splits for training, validation, and test sets. Then, we started evaluating our data preprocessing and imputation techniques used earlier based on the test results. To Fastrack the work and to achieve an optimized model we then participated in the competition hosted by our TA - Shengui Li. This helped us in optimizing the model with a good accuracy.

Later, I worked on optimizing the model by trying different techniques like changing the imputation technique, trying outlier detection methods, changing the network architectures, adding and removing the weather features and other air-pollutants that had higher missing values, added some time features like a weekday, week number and added some regularization techniques into the model. All these techniques helped us in improving the model accuracy. Then, I worked on implementing the Federated model with the help of tensor-flow tutorials which was the final phase. Finally, I contributed in writing the sections within the product and course reports to map the implementation results with the documentation.

Though, the project was held completely digital i.e. work from home we were successfully able to cross all the hurdles and reach all the milestones which we had aimed for. At times, we felt bit difficult to just sit in home and work alone. So we decided to meet once in two weeks to share our knowledge on different technologies and talk for a while with coffee and chocolates sponsored by Anil Poudel most of the time. Due to the bigger team size we got opportunity to work with each other during different sprints.