Weather Forecasting using Machine Learning Report

Project Overview

With the aim to guarantee the security and effectiveness of aircraft operations, weather forecasting is crucial in airports. Precise weather forecasts assist in arranging landings, takeoffs, and aircraft routes, hence minimizing the probability of meteorological accidents. By providing early warning of unfavorable weather conditions, it helps to reduce delays and cancellations. Through early preparation for weather impacts, forecasting also helps ground operations, including de-icing, refueling, and luggage handling. It also improves communication about possible weather disruptions, which improves the passenger experience.

This project of Weather Forecasting using Machine Learning project aims to develop an accurate weather prediction model utilizing machine learning models and algorithms. The project seeks to leverage historical weather data from Toronto Pearson Airport to train the model, which will then be used to forecast weather conditions accurately.

Project Objectives

1) Develop a comprehensive prediction model based on historical weather data, including temperature, humidity, wind speed, and precipitation.
2) Explore and pre-process the collected data to ensure its quality and suitability for machine learning modelling.
3) Implement a machine learning algorithm, such as regression, to build a predictive model.
4) Evaluate the performance of the model using appropriate metrics and optimize them for accuracy and efficiency.
5) Deploy the final weather forecasting model for real-time predictions.

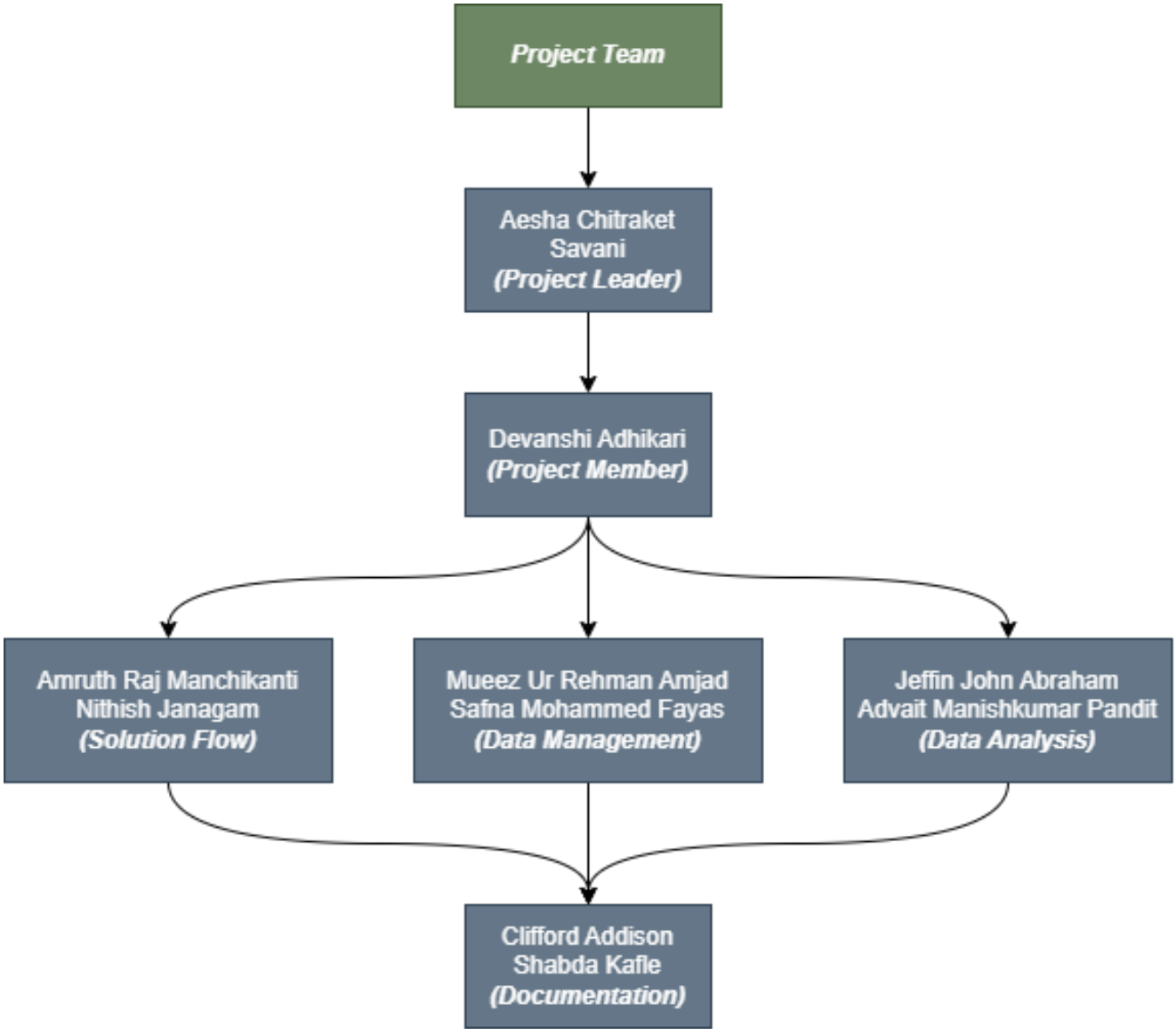
Project Deliverables

- Cleaned and pre-processed weather dataset.
- Trained machine learning models for weather forecasting.
- Final deployment for real-time predictions.
- Project presentation slides and documentation.

Overall Tasks

1. Collect historical weather data from reliable sources.
2. Clean and pre-process the data to remove outliers and missing values.
3. Select and implement machine learning algorithms suitable for weather forecasting.
4. Train initial models using the pre-processed data.
5. Evaluate model performance and iterate for improvement.
6. Conduct thorough evaluation of the trained models using appropriate metrics.
7. Identify areas for optimization and fine-tune model parameters.
8. Deploy the final model for real-time forecasting.
9. Prepare for the presentation by organizing project findings and results.

Project Scope and Timeline

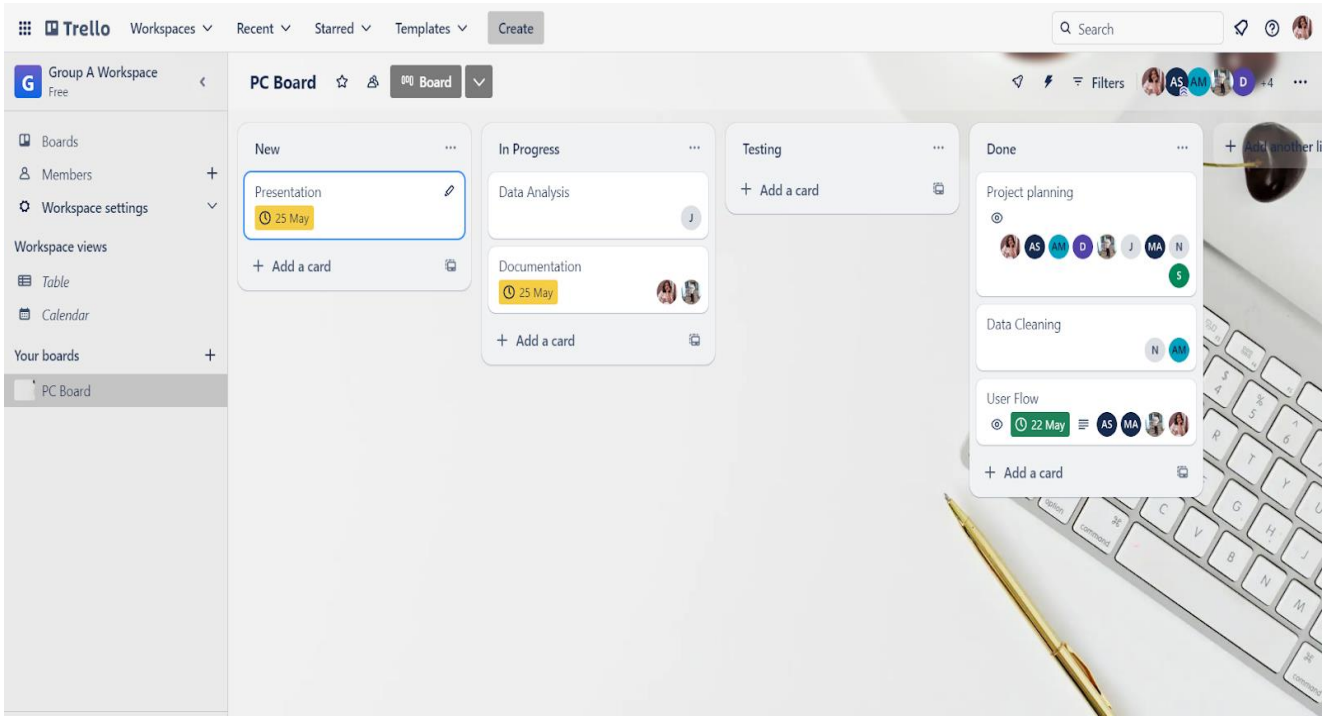


Phase1: Project Management

Project Team

Project Leader:	Aesha Chitraket Savani
Project Team:	Devanshi Adhikari Mueez Ur Rehman Amjad Safna Mohammed Fayas Shabda Kafle **Clifford Addison** Amruth Raj Manchikanti Jeffin John Abraham Advait Manishkumar Pandit Nithish Janagam

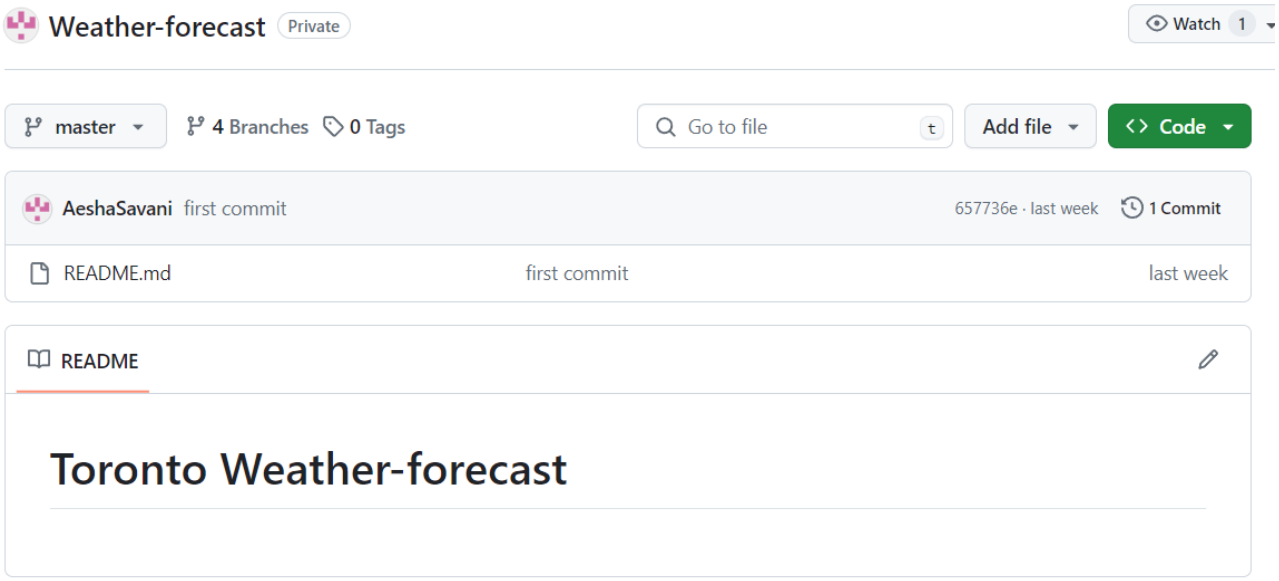
Trello



We are using Trello for project planning; this is an easy-to-use, versatile project management tool. With boards, lists, and cards, Trello easily allows us to organize tasks and keep track of where our project is. Each task can be assigned to a specific card, which could be transferred across different lists representing various stages of completion, for instance, \"To Do, \" \"In Progress, \" and \"Done.\" This visual workflow will keep the team in line and remind everyone of their responsibilities. Trello also allows for team collaboration through the option of commenting on cards, attaching documents, and setting due dates to keep the project on course. This tool's flexibility and usability make it invaluable in project planning.

GitHub

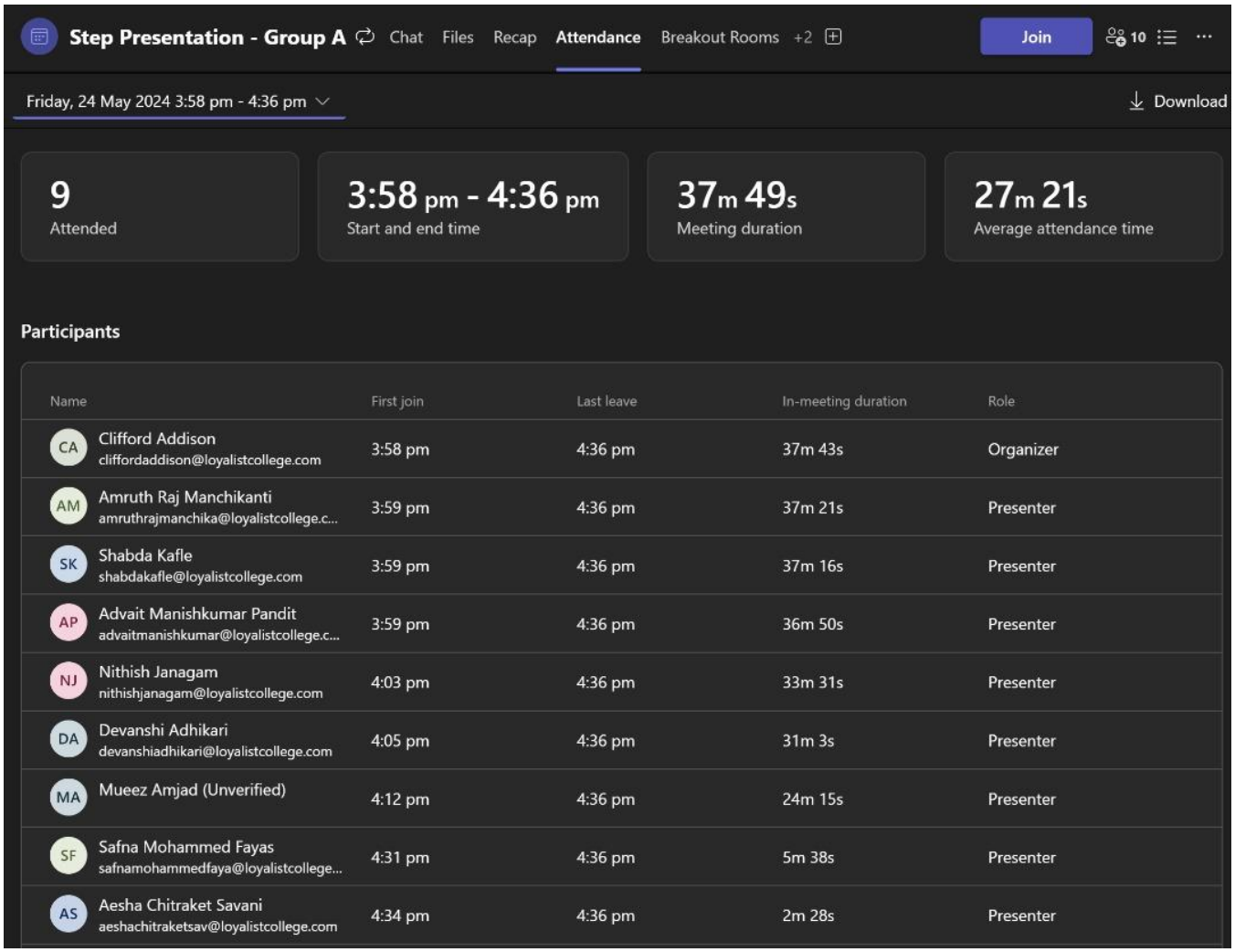
GitHub is our core platform for version control and collaborative development. It provides a centralized repository where the code and all the associated documentation can be stored to ensure that all people are accessing the latest versions. Branches allow us to work on different features and bug fixes at the same time without disturbing the main code. The pull request and the code review allow maintaining high-quality code, besides sharing knowledge amongst team members. For the identification, assignment, and management of tasks related to software development, the GitHub issue tracking system will be most helpful in a way that no bug or feature request went unnoticed. Using GitHub means that we can perform our development work in a streamlined and structured way, making it more effective and collaborative.



Microsoft Teams

Microsoft Teams will be the central communication and collaboration tool throughout the project. It offers a single platform where team members can engage in real-time chat, video meetings, and file sharing. Teams allow us to have dedicated channels for different aspects of the project, such as planning,

development, and testing, which keeps the conversation organized and focused. The collaboration will be easy with the integration of other Microsoft Office applications, such as Word and Excel, to collaborate on documents and share instantly within the team. Routine check-ins and video conferences on Teams ensured that everyone was aligned, and any issues could be addressed promptly. Its features supported effective remote collaboration, keeping the project on track and the team connected.



Meeting Minutes Summary

Date:	Monday, 13 May 2024
Time:	3:55 pm - 4:28 pm
Duration:	33 minutes 28 seconds

Participants: 7

Clifford Addison, Amruth Raj Manchikanti, Safna Mohammed Fayas, Aesha Chitraket Savani, Shabda Kafle, Mueez Ur Rehman Amjad, Nithish Janagam

Key Points Discussed:

1) Team Collaboration:
- Emphasis on working as a team.
- Importance of each team member playing their part to ensure timely and high-quality completion of tasks.

2) Task Assignment:
- Plan to discuss individual strengths to appropriately assign tasks.
- Task assignments will be done the following day to commence work.
- Selection of data

3) Next Meeting:
- Scheduled for the following day.

4) Conclusion:
- The meeting adjourned with thanks and reminders about the next day's activities and assessments.

Meeting End: 4:28 pm

Meeting Minutes Summary

Date:	Wednesday, 15 May 2024
Time:	3:58 pm - 4:29 pm
Duration:	30 minutes 37 seconds

Participants: 10

Clifford Addison, Shabda Kafle, Amruth Raj Manchikanti, Safna Mohammed Fayas, Jeffin John Abraham, Devanshi Adhikari, Mueez Ur Rehman Amjad, Aesha Chitraket Savani, Advait Manishkumar Pandit, Nithish Janagam

Key Points Discussed:

1) Collaboration and Task Management:
- Emphasis on teamwork and the importance of collective effort for success.
- Coordination needed for tasks and ensuring all work aligns.

2) Google Sheet for Task Sequence:
- Mueez will create a Google Sheet for the project flowchart and sequence of tasks.
- Sheet to be shared with the group via Google Drive, Moodle, and percept.

3) Finalization of Flowchart and Data:
- The flowchart and necessary data need to be finalized by the end of the day.
- Discussion about task allocation and deadlines to occur after class the following day.
- The goal is to start working on the project tasks properly once data and flowchart are finalized.

4) Task Identification and Steps:
- Identify and select relevant data and steps required for the project.
- Establish a clear base to ensure smooth project execution.

5) Next Steps:
- Address any final questions or comments before closing the meeting.
- The meeting adjourned with a reminder to prepare for the next steps starting the following day.

6) Conclusion:
- Meeting adjourned at 4:29 pm with thanks to all participants.

Meeting End: 4:29 pm

Meeting Minutes Summary

Date:	Wednesday, 15 May 2024
Time:	4:01 pm - 4:41 pm
Duration:	40 minutes 3 seconds

Participants: 7

Clifford Addison, Safna Mohammed Fayas, Shabda Kafle, Amruth Raj Manchikanti, Aesha Chitraket Savani, Jeffin John Abraham, Mueez Ur Rehman Amjad

Key Points Discussed:

1) Punctuality and Attendance:
- Emphasis on everyone arriving on time for meetings.
- If unable to attend, members should post updates.

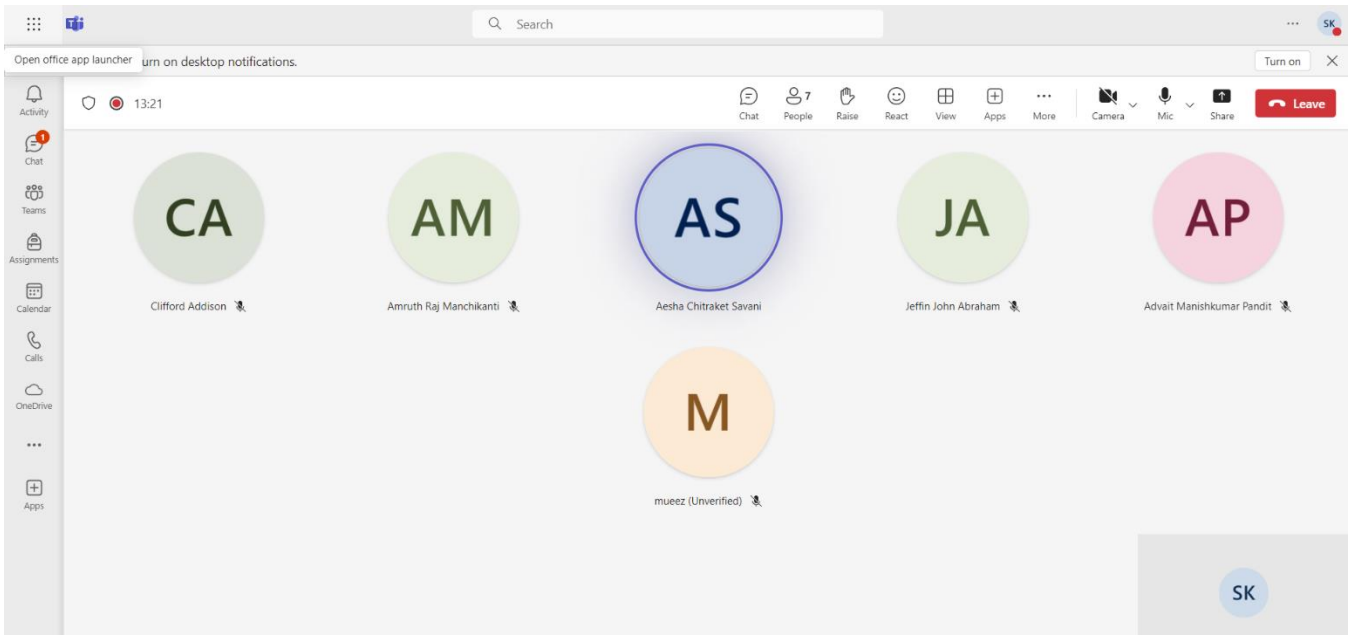
2) Use of Git and Communication Tools:
- Encouragement to use Git for project updates and share via Teams or WhatsApp.
- Aesha plans to Microsoft Teams for better coordination of updates in case of absence and create a Git Repo.

3) Project Updates and Expectations:
- Members should keep each other updated and follow through on assigned tasks.

4) Meeting Conclusion:
- The meeting concluded with well wishes for upcoming exams and tests.

5) Conclusion:
- Meeting adjourned at 4:41 pm with thanks to all participants.

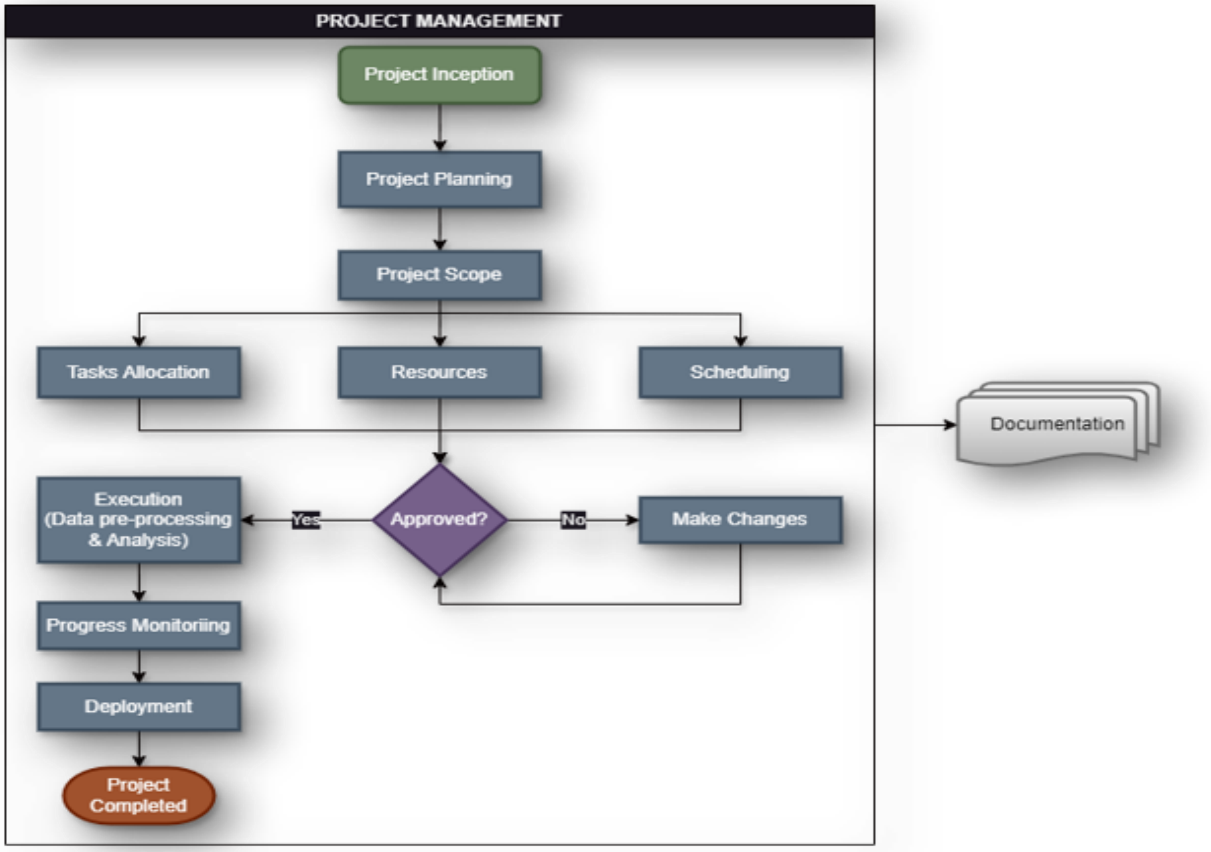
Meeting End: 4:41 pm



Phase 2: Solution Flow

We selected the National Oceanic and Atmospheric Administration (NOAA) weather dataset for Toronto Pearson Airport based on several significant factors. First off, NOAA is a trustworthy resource with a solid reputation for precise and thorough weather information. It was easy to get the data and incorporate it into our project because of NOAA's platform's accessibility and ease of use.

Given the importance of Toronto Pearson Airport as a key hub, the dataset was chosen to ensure that a broad range of people can benefit from the information. The dataset provides large quantities over a long temporal span of historical and present meteorological data required for in-depth research. Detailed hourly observations are also included in the data, which is essential for accurate weather monitoring and forecasting.



After gathering the Toronto International Airport weather dataset, the subsequent stage involves conducting exploratory data analysis (EDA) to thoroughly understand the data. In order to comprehend the dataset's patterns, trends, and any unusual patterns or outliers, this entails going deeply into it. We will then determine which features are most crucial for forecasting future weather conditions based on the knowledge gathered during EDA. These could include wind speed, precipitation, and temperature among other things.

We next use machine learning methods, such as regression and decision trees, to construct our predictive model using this revised collection of characteristics. We carefully assess the accuracy and performance of this model, which will be trained on a subset of the dataset, to make sure it can accurately predict the weather in the future. When we are satisfied with the model's performance, we will put it into use so it can begin generating forecasts using fresh inputs of meteorological data.

Phase 3: Data Management

In this dataset, we have 3960 observations. We changed the temperature unit from Fahrenheit to Celsius using preprocessing techniques such data transformation, missing data imputation, and outlier detection.

These methods make sure that the data is reliable, consistent, and of high quality before it is sent into the machine-learning algorithm or algorithms of our choice for testing, validation, and training. The columns that are in the dataset are:

Station: The station Id of the data

Name: Name of the city or place

Date: Date of the observation

PRCP: Precipitation

Snow: Snow in inches

SNWD: Snowfall Depth

TAVG: Average temperature of the day

TMAX: Maximum Temperature of the day

TMIN: Minimum Temperature of the day

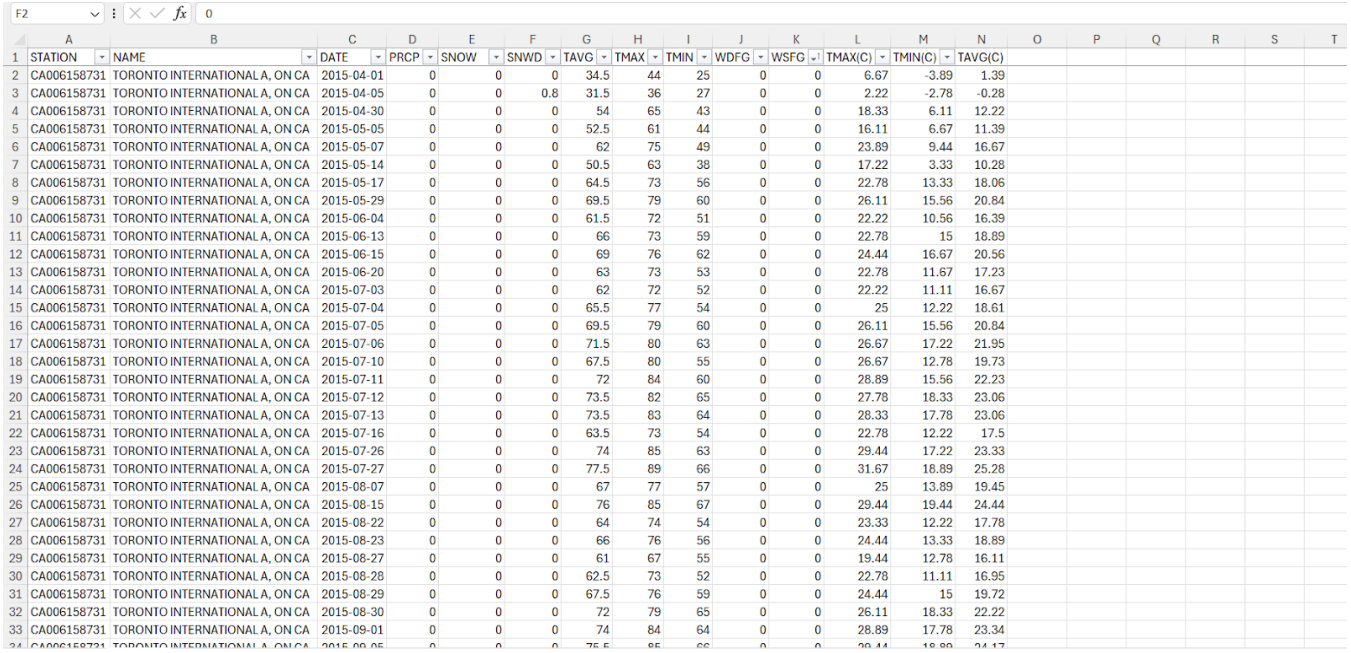
WDFG: Direction of Peak wind gust (degrees)

WSFG: Peak Gust wind speed (miles per hour)

Pre-Processing:

Data Columns and Observation:

- Since there is just one site for which the data is recorded—Toronto—the "Station" and "Name" columns are consistent throughout the data.
- Every observation has a different "Date" column, which indicates a different day for every observation.
- “PRCP” Precipitation is any form of water, liquid or solid, that falls from the atmosphere and reaches the ground. This includes rain, snow, sleet, and hail.
- “SNOW” column represents the snowfall in mm.
- “SNWD” column represents the snow depth.
- “TAVG” is the average temperature of the day
- “TMAX” is the maximum temperature of the day
- “TMIN” is the minimum temperature of the day
- “WDFG” Direction of Peak wind gust (degrees)
- “WSFG” Peak Gust wind Speed (miles per hour)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	STATION	NAME	DATE	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN	WDFG	WSFG	TMAX(C)	TMIN(C)	TAVG(C)
2	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-04-01	0	0	0	34.5	44	25	0	0	6.67	-3.89	1.39
3	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-04-05	0	0	0.8	31.5	36	27	0	0	2.22	-2.78	-0.28
4	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-04-30	0	0	0	54	65	43	0	0	18.33	6.11	12.22
5	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-05-05	0	0	0	52.5	61	44	0	0	16.11	6.67	11.39
6	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-05-07	0	0	0	62	75	49	0	0	23.89	9.44	16.67
7	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-05-14	0	0	0	50.5	63	38	0	0	17.22	3.33	10.28
8	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-05-17	0	0	0	64.5	73	56	0	0	22.78	13.33	18.06
9	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-05-29	0	0	0	69.5	79	60	0	0	26.11	15.56	20.84
10	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-06-04	0	0	0	61.5	72	51	0	0	22.22	10.56	16.39
11	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-06-13	0	0	0	66	73	59	0	0	22.78	15	18.89
12	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-06-15	0	0	0	69	76	62	0	0	24.44	16.67	20.56
13	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-06-20	0	0	0	63	73	53	0	0	22.78	11.67	17.23
14	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-03	0	0	0	62	72	52	0	0	22.22	11.11	16.67
15	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-04	0	0	0	65.5	77	54	0	0	25	12.22	18.61
16	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-05	0	0	0	69.5	79	60	0	0	26.11	15.56	20.84
17	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-06	0	0	0	71.5	80	63	0	0	26.67	17.22	21.95
18	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-10	0	0	0	67.5	80	55	0	0	26.67	12.78	19.73
19	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-11	0	0	0	72	84	60	0	0	28.89	15.56	22.23
20	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-12	0	0	0	73.5	82	65	0	0	27.78	18.33	23.06
21	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-13	0	0	0	73.5	83	64	0	0	28.33	17.78	23.06
22	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-16	0	0	0	63.5	73	54	0	0	22.78	12.22	17.5
23	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-26	0	0	0	74	85	63	0	0	29.44	17.22	23.33
24	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-07-27	0	0	0	77.5	89	66	0	0	31.67	18.89	25.28
25	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-07	0	0	0	67	77	57	0	0	25	13.89	19.45
26	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-15	0	0	0	76	85	67	0	0	29.44	19.44	24.44
27	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-22	0	0	0	64	74	54	0	0	23.33	12.22	17.78
28	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-23	0	0	0	66	76	56	0	0	24.44	13.33	18.89
29	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-27	0	0	0	61	67	55	0	0	19.44	12.78	16.11
30	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-28	0	0	0	62.5	73	52	0	0	22.78	11.11	16.95
31	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-29	0	0	0	67.5	76	59	0	0	24.44	15	19.72
32	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-08-30	0	0	0	72	79	65	0	0	26.11	18.33	22.22
33	CA006158731	TORONTO INTERNATIONAL A, ON CA	2015-09-01	0	0	0	74	84	64	0	0	28.89	17.78	23.34

Data Cleaning:

- There were some missing (blank) values in the “PRCP” column. Therefore, we considered that the value at that time was zero and replaced them with zero suggesting that there was no precipitation on that day.
- The “SNOW” column had missing values that indicated that there was no snowfall during that period. We made this conclusion because we observed that most of the dates fell in the summer season and snowfall was not possible at that time.

- The "SNWD" column had some missing values. However, upon cross-referencing with the Snowfall data, we found that there was no snowfall during those observations. Therefore, the missing values indicate there was no snow depth.
- The "TMAX" column had some missing values for the maximum temperature of the day. To complete the data, we retrieved the missing temperature values from the internet.
- Similarly, the "TMIN" column had some missing values for the minimum temperature of the day. We also used online sources to fill in these missing temperature values.

Phase 4: Data Analysis

In our project of weather forecasting there are some machine learning models that can be useful but according to our dataset, Multiple Linear Regression algorithm will be highly suitable.

The reasons are mentioned below:

- In our dataset we have multiple output features so multiple linear regression could be able to handle that data very well.
- It helps to provide coefficient for each predicting variable and that leads to identify strength of all variables to target variable.
- Multiple linear regression can perform complex calculations.
- We can also get an idea about the features that are non-linear by applying this algorithm so that we could be able to fix them.
- It helps to simplifying the model and reducing the overfitting
- Moreover, it is easy to implement and interpret.
- Uncertainty of data can also be measured through confidence intervals for forecasting values
- We could also be able to find out trends of data.

Failings

- Finding the appropriate dataset was one of the major setbacks we faced in our initial phase. We searched through Kaggle, cloud databases etc. But the dataset either was missing a lot of columns or was not enough for what we wanted to do with. Finally, we found our dataset on NOAA.
- The second hurdle we went through was establishing proper communication between the team members. But after proper planning, the idea of daily stand-up was proposed where every member would update the work done between us.

- Finding and working with suitable project management tools was also difficult, as not everyone was familiar with the way it worked. Monday board was proposed but since it is only available on paid version, it was dismissed, and Trello board was finally used.

Resources Used

- Weather Forecasting Data

Weather and climate resources. (n.d.). National Oceanic and Atmospheric Administration. https://www.noaa.gov/tools-and-resources/weather-and-climate-resources

- Integrated Development Environment (IDE)

Google Colab. (n.d.). https://colab.research.google.com/drive/1dWyOce5CaHdpKfGFr_aRp6KZk29rMIaM?usp=sharing

- Repository

AeshaSavani/Weather-forecast. (n.d.). GitHub. https://github.com/AeshaSavani/Weather-forecast

- Project Board

The basics of a board. (2024, May 9). TrelloBoard.

Appendix

Phase 1: Project Management

Duration: 05/04/2024 – 05/11/2024

Lead: Aesha Chitraket Savani & Devanshi Adhikari

Tasks

- Choosing roles
- Choosing a project
- Deciding on a timetable
- Choosing a platform
- Meeting documentation/minutes
- Proof of discussion

Phase 2: Solution Flow

Duration: 05/04/2024 – 05/11/2024

Lead: Amruth Raj Manchikanti & Nithish Janagam

Tasks

- Choosing a data source
- Deciding on the type of analysis
- Picking the output format

Phase 3: Data Management

Duration: 05/04/2024 – 05/11/2024

Lead: Mueez Ur Rehman Amjad & Safna Mohammed Fayas

Tasks

- Choosing the source dataset
- Data quality considerations
- Data cleaning
- Data warehousing
- Data ingestion

Phase 4: Data Analysis

Duration: 05/04/2024 – 05/11/2024

Lead: Jeffin John Abraham & Advait Manishkumar Pandit

Tasks

- Data processing
- Selection of Mathematics to use.
- Modelling
- Analysis

Phase 5: Documentation

Duration: 05/04/2024 – 05/11/2024

Lead: Clifford Addison & Shabda Kafle

Tasks

- Reports & Presentation
- Recording project errors, failings and debugging attempts
- Capturing all the resources used
- Formatting

Project Management

Tracking Roles + Accountability

Tracking of roles is managed by project leader Aesha and assistant Devanshi. Team members include: [list all members including Advait responsible for EDA and visualizations].

Conflict Resolution

We had to agree on changing the dataset for our weather forecasting model because the one we initially acquired had fewer features, making it difficult to do much with it even after feature engineering.

Project Divergence/Adaptations

Adaptations included modifying the dataset and tools based on initial setbacks. We shifted from the initial dataset to a more feature-rich dataset which allowed for better model training and accuracy.

Timetable Divergence

Initially planned timetable had to be adjusted due to data and communication issues. Updated timelines were communicated via Trello and Teams, ensuring all team members were aligned with the new schedule.

Platform Progress

We used Trello for task tracking and GitHub for version control. Regular updates and progress tracking were done to ensure the project was on track. This helped in managing the tasks efficiently and meeting the deadlines.

Executing the Plan of Action

Each task was assigned and tracked using Trello. Regular check-ins and meetings were conducted to ensure adherence to the plan. This approach ensured smooth execution of the project plan.

Meeting Documentation/Minutes

Meeting documentation and minutes were recorded and shared with all team members to ensure transparency and keep everyone updated on the project progress.

Proof of Discussion

Proof of discussion is available in the project documentation and meeting minutes. This includes decisions made, changes to the project plan, and any adaptations implemented.

Visualization

Choosing a Visual Type

For our data, we used bar charts, line graphs, and heatmaps. Bar charts were used for categorical data comparison, line graphs for trends over time, and heatmaps for correlation analysis.

Matching Type to Analysis

Each visualization type was carefully chosen to match the type of analysis being performed. Bar charts for comparing different categories, line graphs to show trends over time, and heatmaps to visualize correlations between features.

Addresses Multiple Audiences

Visuals were designed to address different audiences. Simple visualizations were used for general audiences, while detailed charts catered to technical audiences.

Ease-of-Use, Intuitiveness

Charts were labeled clearly, with distinguishable colors to ensure ease of use and intuitiveness. This made the data easy to understand for all viewers.

Balance of Detailedness/Complete

A balance was maintained between providing detailed information and ensuring completeness. This ensured that the visuals were informative without being overwhelming.

Interactivity

Interactive dashboards were created using tools like Plotly and Tableau. These allowed users to engage with the data and gain deeper insights.

Integration on Platform

Visuals were integrated into the final report and presentation. This ensured that the data insights were easily accessible and could be effectively communicated to stakeholders.

Coding Solution

Code Commit/Versioning

Code commit and versioning were managed using GitHub. This allowed for efficient collaboration and version control.

Unit Testing

Unit tests were implemented for all critical functions to ensure code reliability and correctness.

Extensive Edge/Corner Case Testing

Edge and corner cases were extensively tested to ensure robustness of the code.

Reliability

Reliability was ensured through peer reviews and thorough testing.

Use of Dummy Data

Initial tests were run on dummy data to verify logic before using real data.

Modular Structure

The code was structured in modules for better organization and readability. This made it easier to manage and understand.

Efficient Structure

Code was optimized for performance and speed, ensuring efficient execution.

Commenting

The code was well-commented to ensure understanding and maintainability.

Speed/Latency

Efforts were made to ensure low latency in predictions, making the model responsive.

Meta Testing (e.g., Latency)

Meta tests were conducted to measure and optimize performance.

Consistency (Authorship)

Consistency in coding style was maintained across team members to ensure uniformity.

Readable/Self-Explanatory Naming Conventions

Meaningful variable and function names were used to make the code readable and self-explanatory.

Easy-to-Adapt, Extendable, Scalable

The code was designed to be easily adapted, extended, and scaled for future needs.

Stress-Tested

Stress tests were performed to ensure the system could handle high loads.

Appropriately Parametrized

Parameters were used to make the code flexible and reusable.

Collaboration

Collaboration was fostered through regular code reviews and team discussions. This ensured that everyone was on the same page and could contribute effectively.

Task 2: Updating the Documentation

1. **Project Management Section**
    - Add details about the roles and accountability.
    - Describe the conflict resolution process.
    - Include adaptations and changes to the project plan.
    - Detail the updates to the timetable.
    - Outline the platform progress and how the plan was executed.
    - Summarize meeting documentation and proof of discussions.
2. **Visualization Section**
    - Expand on the types of visualizations used.
    - Match visual types to specific analysis needs.
    - Ensure visuals address different audience types.
    - Focus on ease of use and intuitiveness.
    - Maintain a balance between detail and completeness.
    - Incorporate interactivity where possible.
    - Discuss integration of visuals into the platform.
3. **Coding Solution Section**
    - Provide the coding practices followed.
    - Include details on unit testing, edge case testing, and reliability measures.
    - Describe the modular and efficient structure of the code.
    - Highlight commenting and naming conventions.
    - Discuss scalability, stress-testing, and parametrization.

- Emphasize collaboration and consistency in coding.