# CMPT 318
# PROJECT REPORT

MEMBERS:

**Sicong Liu 301269558**
**Bowen Yang 301240295**
**Xin Tian 301201034**

# Track This Project

**Git repository:** https://csil-git1.cs.surrey.sfu.ca/boweny/cmpt318_project.git

# Project Abstract

Weather-related data is useful for finding out possible conditions on Earth such as weather descriptions. Recording these data to perform the prediction manually is expensive and time consuming consider the resources needed. If the task can be done automatically, then an efficient way to record the Earth conditions is produced. The problem is to find an automated data collecting process and data analysis strategy to obtain and make the data useful. The final data collecting and analyzing automation will provide valuable results that tell people what is going on in the world.

One of the possible solution to automate the process is to locate a webcam out of a window and record images of the weather outside. The images can be used to predict the weather-related data automatically by implementing a machine learning model. To execute the conceptual implementation, classification machine model can be used. One can combine the known images obtained by the webcam and their corresponding weather observation recordings to train the machine learning model. This is based on the assumption that similar images will result in similar weather observations. With the successfully trained machine learning model, unrecorded data inputs can have their corresponding classification predicted by the model. The overall process is that the image weather data will be inputed to the machine learning model to predict correlative result data such as weather description or temperature of the input data.

Several different candidate machine learning models and techniques are tested in this project to approach the best data analysis automation process. One image can produce several different correlative data, therefore the desirable machine learning model must have accuracy in producing the different types of correlative data such as weather description or temperature. The project reviews the feasibility of this automation process for each candidate machine learning model and their accuracy on the predicted data.

# Data in the project

Two data sets are required for this project as mentioned. The project needs the webcam images data and the corresponding weather-related data. The two data sets will be combined for the purpose to train the machine learning models.

The webcam images are pictures of English Bay, Vancouver from June 05th, 2016 to June 21th, 2017 gathered by Kat Kam. These images are recorded hourly from 6AM to 9PM every day within the stated time period. There are in total 5046 images. Each image has 256 x 192 pixels and each pixel is represented by its RGB value.

The corresponding weather observation for the weather-related data are historical weather data at the YVP airport, Richmond posted by the Canadian government and gathered by the Vancouver Airport weather station. There are 9480 observations in total, each recording 26 different weather conditions hourly from June 01st, 2016 to June 30th, 2017.

Both data sets are loaded into the Pandas dataframe from for model training purpose. Prior to using these data to train the models, appropriate transformation and cleaning are performed to combine the data. Both dataframe requires an appropriate form to been merged and some conditions are needed to match the dataframe for the two data sets.

Since each image is represented by 256 x 192 pixels, there are 49152 pixels for each image, the pixels each has an RGB value of 3 integers. Therefore, there are 14756 integers to represent each image. The original shape of the webcam images dataframe is 5046 x 192 x 256 x 3. After reshape of its dataframe, the new shape of it is 5046 x 147456. The taken time of each image is provided in the image file name. They are added to the dataframe by extracting the date time from the images' file names and adding them as new coloums to the dataframe. Four new columns are added which are year, month, day and time. The resulting dataframe represent 5046 images and each image has 147456 features.

Useless features in weather observation data are dropped. Columns dropped include the "Index" column which is redundant to the dataframe index when load the data from file, "Data Quality" column which has the same value for all observations, and all the "Flag" columns which contain NULL values. "Hmdx" and "Wind Chill" columns are also eliminated because the columns have very few data records in the columns. After all the unnecessary columns are dropped, the group found out that the vital column "Weather" has many entries without a value. So these entries are

cleaned out. The valid data result include 4176 observations with 13 features for each one. To carefully examine the weather descriptions in the “Weather” column, the weather descriptions categories have been made to look like human-generated and are not suitable for this project. There are 19 categories made. The data cleaning process categorized these 19 categories into 5 categories instead. Categories “Clear” and “Mainly Clear” are combined into one category “Clear”. “Cloudy” and “Mostly Cloudy” are combine into one as “Cloudy”; similar for “Fog” and “Freezing Fog”, they are now just “Fog” . “Drizzle, Freezing Rain”, “Heavy Rain”, “Moderate Rain”, “Moderate Rain Shower”, “Rain”, “Rain Showers”, and “Thunderstorms” are combined into “Rain”. “Moderate Snow”, “Snow Pellets”, “Ice Pellets”, “Snow Showers” are combined into “Snow”.

After both data sets are cleaned appropriately to combine, the two dataframes have merged together based on the condition that they have the same recorded time. The final data set has 2244 valid data rows and 147469 features for each row.

# Techniques Used in the Project

To analyze the feasibility of the automation, there are different approaches. The first approach has been tested is to use solely the image data to predict a weather description of an image input.To execute and test the first problem approach, the merged dataframe has been extracted to form input data X and target data y. All the image points are set to be X, and “weather” column alone is set to be y. Since X and y are present, through applying the train_test_split function, both train and test data sets have been split out.

The first problem approach used four different machine learning models; each model has been tested and draw an accuracy score.

1. Bayesian classifier model which uses probability to predict the input data’s category is tested and its accuracy score is generated by using the test data set. A normality test is also performed on the model to ensure its feasibility.

2. Support Vector Classifier (SVC) model which differentiates the categories through a data boundary line is tested. The kernels are the “linear” kernel and the “rbf” kernel. To increase SVC’s margin width, the C parameter has first been set to 1e-1. The model took a considerable time to run; therefore a pipeline is used and the Principal Component Analysis

(PCA) techniques are applied to reduce the dimensions of the input dataset for time efficiency. Different combinations of PCA parameters and SVC parameters have been tested and an accuracy score is generated.

3. Nearest Neighbours model which uses relative data to decide the category with a parameter n with value 5 and 10 is tested to generate two model and their corresponding accuracy scores.

4. Neural Network method which weights the features and uses the weigh values to analyze data is the forth tested model. A perceptron is first tested and the parameter solver is set to "lbfgs". Then a neural network use the MLPClassifier model within inside layer size of 10 x 30 has been tested. Two model results and corresponding accuracy scores are generated.

The second approach to analyze the feasibility of the automation is to combine the image data with stated weather conditions to predict the weather description of the input image. A different input data set X has been extracted from the merged dataframe. All the image points and the weather condition data together are set to be X. Y remains the same as the data in the "weather" column. Since weather conditions and the image data points have different range, scaling techniques have been applied to scale them into the same range. The standardScaler function is used for the data scaling. Then all four of the mentioned machine learning models have been tested with different corresponding parameters. The accuracy scores for each model result are recorded.

In the "weather" column, some entries have multiple description for a given time. To solve this multiple label problem, one approach is to use only the first weather description in each data entry. Another approach is to use a different model to handle the multiple labels. MultiLabelBinarizer has been used to reshape the y data set. OneVsRestClassifier model is then used for testing the data and an accuracy score is draw from the model.

# Result /findings /conclusions of the project

#### PREDICT WEATHER DESCRIPTION FROM IMAGE ONLY

The first analysis approach in this project is to predict weather description from only input images. By comparing the accuracy scores of each model tested, a conclusion is made to determine the most suitable model for this approach. The SVC model and PCA techniques

combination did the best job. The combination has an accuracy score of 0.6684. This is the highest score retrieved, but it is still not fully satisfactory.

The Bayesian Classifier model in the first approach has an accuracy score of 0.48484, which is bad. The reason for this unacceptable score is because the data set used does not meet the normal distributed requirements for this model. The p-value of normality test for data set X is very small, so the normality test is not significant.

The K-Nearest Neighbours (KNN) model did a better job than Bayesian model. KNN has an accuracy score of 0.6096 for set n equals to 5 and an accuracy score of 0.61675 for set n equals to 10.

The Neural Network model scored the lowest accuracy. By applying the MLPClassfier model, both the perceptron and the Neutral Network with layer size 10 x 30 have similar accuracy scores around 0.3992.

By solely using the SVC model, an accuracy score of 0.6684 is generated which concludes the SVC model to be the best model for the first approach. However, this model is very time consuming. The process to train the model requires around 40 minutes. Therefore, PCA techniques have been used to reduce the dimensions of the input data set so model training time would decrease. The result is good. By trying to use PCA to reduce X into 100, 1000, 5000 components respectively, the model trained by 5000 dimensions data has the best accuracy score of 0.65418 and the time for the training process took only 2 minutes. The accuracy of SVC model combined with PCA techniques is almost as good as using only SVC model, but the combination is much more time efficient.

One observation to note is that when using PCA to reduce the dimensions to 100, it took a long time. The time of the reduction took more than 40 minutes. The reason here is that to reduce the dimension from 147469 to 100 but also keep important components in data requires significant amount of operations. Less dimensions imply that more time is needed to do the reduction work when using PCA.

**PREDICT WEATHER DESCRIPTION FROM IMAGE AND WEATHER CONDITIONS**

The second approach in this project is to add additional weather conditions to the input data X and use both weather conditions and image data to predict the weather description. The best

model for this approach is still the SVC model combined with PCA techniques. The combination has an accuracy score of 0.704. A higher accuracy score is obtained than the first approach which only uses the image data to predict weather description. The reason is because there is more features to consider; more data are used to train the model thus the accuracy score increases.

The Bayesian Classifier model's accuracy score also increased in the second approach. An accuracy score of 0.549109 is obtained and the process time is less than 1 minute.

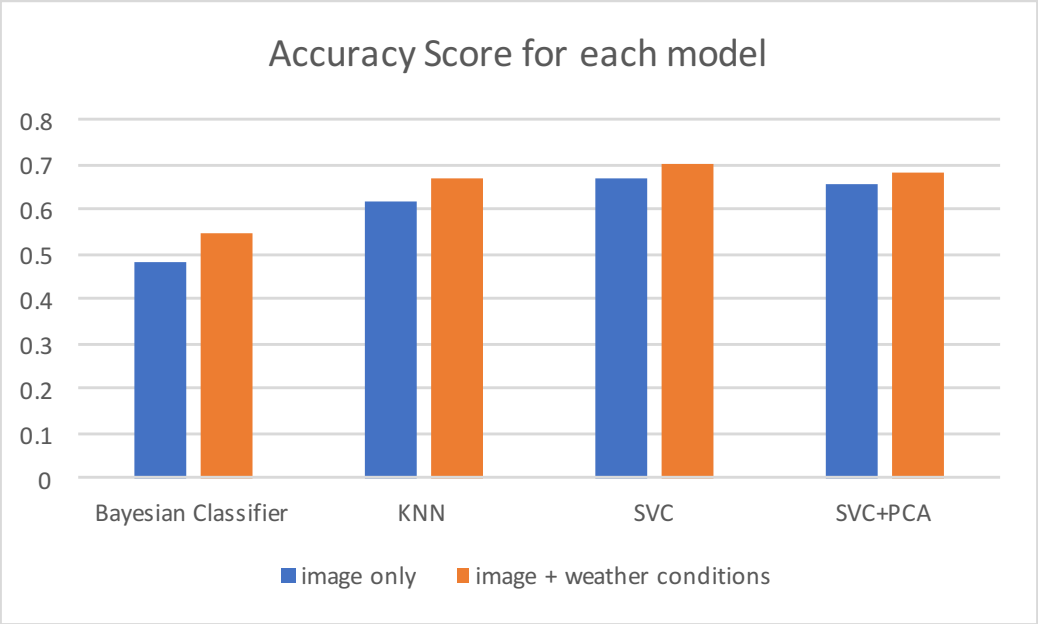
The KNN model's accuracy has increased as well. An accuracy score of 0.6684 is obtained with n is 10.

Scaling techniques are used to uniform the different ranges of the weather condition and image data points in the second approach as well. The SVC and PCA combination produced an accuracy score of 0.704 after the scaling. When leaving the scale function out, the accuracy score is still similar to the score obtained with using the function. An accuracy score of 0.693 has been obtained use SVC and PCA combination without the scaler.

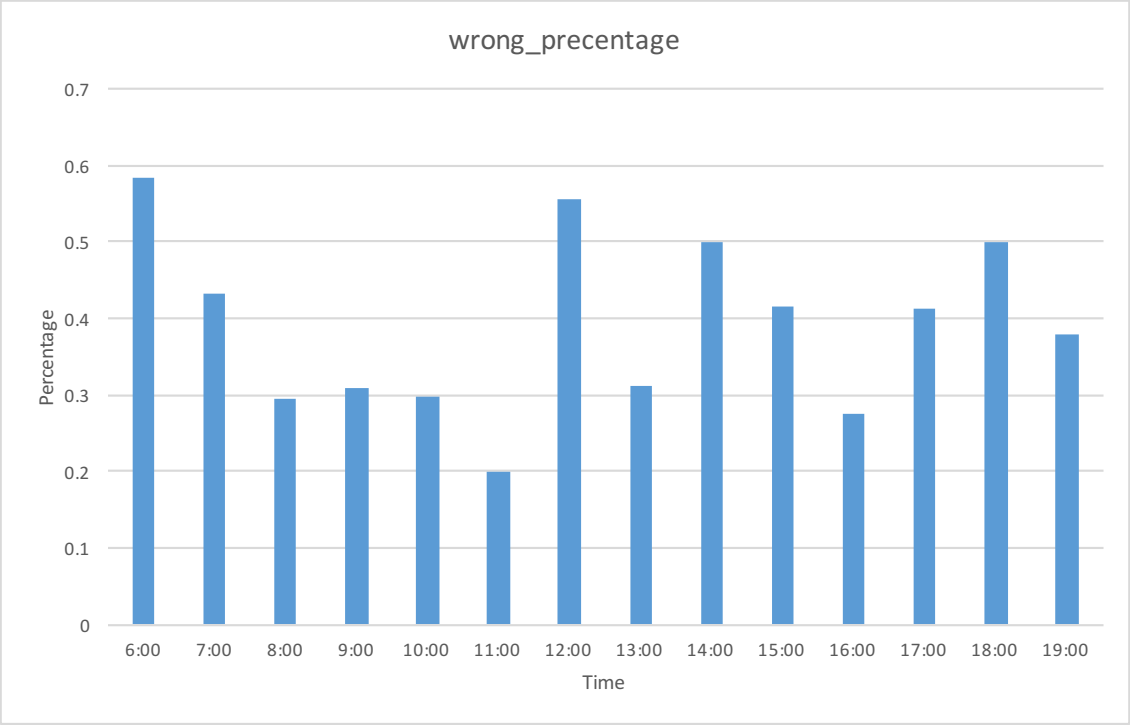
One thing that may cause problems is the multiple labels in the "weather" column. There may be multiple descriptions for the same image. Two approaches are used to fix the problem. First approach eliminates the extra weather descriptions in the "weather" column, so each column entry contains one label. Then use the SVC and PCA combination, an accuracy score of 0.701 is obtained, which the accuracy did not improve. Also, to arbitrarily drop the extra weather descriptions may not be valid and the action may lead to wrong prediction.
The second approach is use another model to solve the problem. Use MultiLabelBinarize to reshape the input data and use OneVsRestClassifier help train the model. However, an accuracy score of 0.57 is generated which is not acceptable. The reason here is perhaps most "weather" column entry has one single label but this does not apply to this model well. The mixture of single and multiple labels with majority of single labels is the possible reason for the low accuracy since we believe the nature of this method works well on all entries with multiple labels.

To conclude the two approaches for the data analyzing automation. The tests performed on the four machine learning models in the two approaches are similar in fashion, but the second approach produced higher accuracy scores overall. This is the result of more data being analyzed. The second approach included extra data, the weather conditions. Intuitively, more data will lead to more accurate analysis because the machine learning model is better trained. Although the

second approach's result model is better, but it requires more data resources. Only using images for weather predictions may be a better methodology for the automation process but the result is not satisfactory in the first approach.




The above chart illustrates the accuracy scores comparison for each trained machine learning model in the two different approaches. From the visualization, an conclusion can be draw that the SVC model of using image data with additional weather conditions produces the best accuracy score in the analysis.




The chart illustrates the wrong percentage of images with incorrect predictions in each hour of the day of the testing period. The visualization shows that 6AM, 12PM, and 6PM have the highest wrong percentage. This reason behind is probably due to the high darkness or brightness of the image taken; the sunrise, sunset, and mid-noon sunlight have large impact on the image's clarity.

#### PREDICT TIME OF THE DAY FROM IMAGE

Time of the day can also be predicted from the images. An accuracy score of 0.46 is obtained by using only image data points to train the SVC model. A better approach is to combine SVC and PCA together. The result accuracy score is 0.5048. The results are not satisfactory. The reason here is perhaps due to the difficulty to distinguish consecutive time like 10AM and 11 AM. Consecutive time has images with very similar looks.

# Limitations

While doing the project, one major problem encountered is that some models take a long to be trained. In additional, tuning the parameters of each model is difficult. These obstacles caused less efficiency in completing the project. If more time is permitted, enhancements can be added. To be able to predict the visibility of the image would be the next stage of the project. Through the project, one thing that should be done in retrospective is to scale the data before training the KNN model in the approaches. Maybe by doing so, a better result can be achieved.

# Members' Summaries

## Summary from Sicong Liu:

- Loaded and stored images as data using Python's Numpy and Pandas libraries to complete the preparation for data analysis on machine learning model training.

- Cleaned the weather observation data using regular expression to recategorize the target data set, in which made the input data more accurate to produce better data analysis results.

- Created a Support Vector Classifier machine learning model through Python's SKLearn library to predict weather descriptions from the weather image inputs.

- Produced a report to conclude the findings of the data analysis which included reasons for inaccuracy and accuracy and best trained machine learning model.

## Summary from Bowen Yang:

- Loaded weather observations data from several files by for loop, and concatenate them together to be a Pandas dataframe.

- Cleaned weather observation data by removing useless column, such that some column have no data or too little data. Also dropped rows with 'NaN' weather description, since they cannot be used for training.

- Developed the weather description category part with team members and classified varied weather description to 5 main categories.

- Using "Bayesian Classifier", "SVC", " Nearest Neighbors" model training data and compare their accuracy by score function. The model "SVC" is found as best model for predict weather.

- Applied "Scaler" and "PCA" to reduce the time consuming problem on "SCV" model. The processing time are reduced to 2 min.

# Summary Xin Tian 301201034

- Taught the computer how to predict weather based on images.
- Loaded image data using skimage.io library and weather data by glob library.
- Added ‘year’, ‘month’, ’day’, and ‘hour’ column for image data.
- Joined image data set with Cleaned weather data set
- Separated X and y set. X Set is the combination of image and weather conditions such as wind speed, and y set is the weather descriptions.
- Applied different machine learning models and compared them.
- The best model is PCA+SVC with about 70% accuracy.
- Predict the time of the day with approximately 50% accuracy by image data.