GLOBAL STEM ALLIANCE
The New York Academy of Sciences

ANALYTIC CHRONICLES

BIG DATA SPECIALTS

Team Members:-
Ms Linda Witte (Mentor)
Vikram M. (Team Leader)
Gaurang B.
Johan J.
Miu S.
Catherine A.
Erol B.

THE JUNIOR ACADEMY

INTRODUCTION

The analysis of big data has been a struggle for years. Numerous different projects have been implemented to ensure the complete use of large datasets in real life applications, in a way that is beneficial. Today, we have great opportunities when compared to any other time, to utilize the information that is in our hands and perform wonders. That's what we did together as a team. Using Machine Learning, our team hopes to utilize large data-sets pertaining to focused information with respect to weather, climate and numerous other factors, to successfully predict natural disasters. Have a look, at the possibilities achievable through big data.

WHY BUILD A SOFTWARE?

- The analysis of numerous natural disasters require years of observation and research, which consumes a lot of time and money. Our experience with analytical data suggests that there are few other places from which we can directly analyse information and use it in supplying weather updates to the world.
- A software that can accurately analyze the changes in weather patterns, and detect any significant threats of natural disaster's that may occur in different places long, before the disaster actually strikes.
- Software's are convenient and available 24/7, so anyone can access the information that might be vital for the preservation of the environment and human life.

BACKGROUND

- According to Karlsruhe Institute of Technology, Germany, natural disasters have been responsible for over $7 trillion in damages and 8 million deaths.
- Places can be evacuated if sufficient warning is given to the public.
- Research could be made, based on the recent disasters.
- Injury and death toll during such events could decrease significantly.
- The Government can make the necessary arrangements for providing displaced homes, food and other basic needs for people during the evacuation process.
- The financial operations of the region can be shifted, moreover cash can be convereted to online credit, so that there is no the disaster does not cause the economy to disintegrate or fall.

R E S E A R C H

Our research was, the premise of work and idea. From looking for a viable and doable solution to creating it. The research revolved around the common problems that existed in the world today even though some problems inevitably don't have a solution, there is a chance of preventing or perhaps decreasing the detrimental impact of such natural disasters that occur, when compared to their impact in the past. Along with that, we were looking for a viable solution that could be implemented on a large scale moreover something that could be created within a limited time span. This required a lot of research and frequent brainstorms nevertheless, we came up with the idea of integrating our tech skills into this challenge by constructing a software that could provide accurate information about the climate, weather patterns, and the percentage of different atmospheric conditions .

LOOKING FOR TRENDS IN EGYPTIAN HEATWAVES

Month of August, 2014

« Previous Month

Daily | Weekly | Monthly | Custom

	Max	Avg	Min
Temperature			
Max Temperature	38 °C	35 °C	32 °C
Mean Temperature	31 °C	30 °C	28 °C
Min Temperature	25 °C	24 °C	23 °C

Month of August, 2015

« Previous Month

Daily | Weekly | Monthly | Custom

	Max	Avg	Min
Temperature			
Max Temperature	42 °C	36 °C	33 °C
Mean Temperature	37 °C	31 °C	29 °C
Min Temperature	31 °C	26 °C	24 °C

Sunday, August 16, 2015

« Previous Day

Daily | Weekly | Monthly | Custom

	Actual	Average	Record
Temperature			
Mean Temperature	36 °C	-	
Max Temperature	42 °C	32 °C	41 °C (2015
Min Temperature	29 °C	22 °C	22 °C (1996
Cooling Degree Days	30		
Growing Degree Days	46 (Base 50)		

Week of August 16, 2015 through August 22, 2015

« Previous Week

Daily | Weekly | Monthly | Custom

	Max	Avg	Min
Temperature			
Max Temperature	42 °C	36 °C	34 °C
Mean Temperature	36 °C	31 °C	29 °C
Min Temperature	29 °C	26 °C	24 °C

The heatwave experienced in Egypt during the month of August 2015, claimed the lives of many. The heatwave had suddenly made a rampage throughout the country without a prior warning or a noticeable threat.

HYPOTHESIS

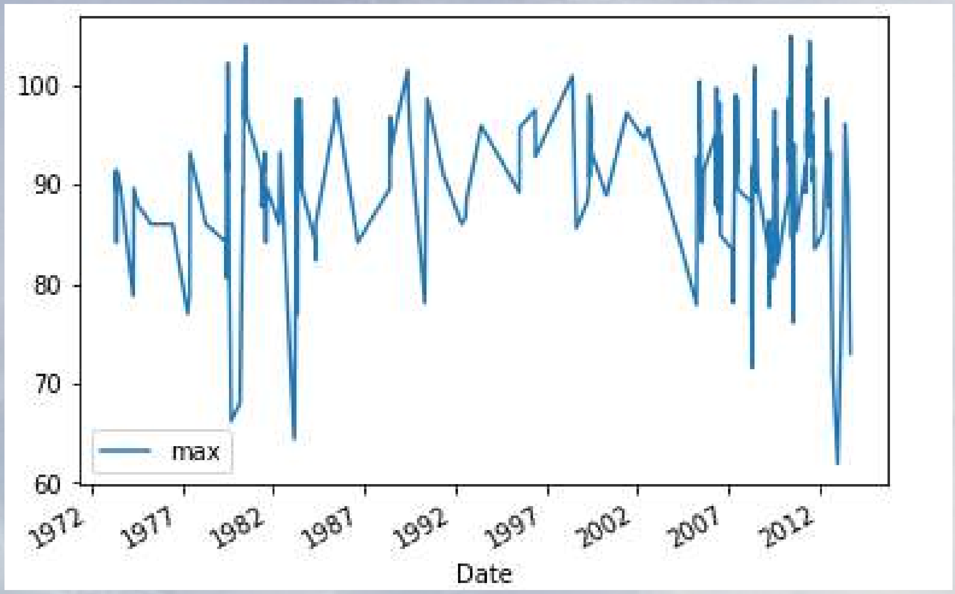
- Through the trends in weather and climate data, our solution will be to predict the natural disasters that would happen worldwide and notify users about them. We plan to present the raw data that we receive from the software into a graphical form (Info-graphics, Graphs, Statistics etc), in order to provide valuable information about the climatic conditions to people in simple terms so that they can understand, and be prepared for any natural disaster that might occur.
- Using a decision tree model and previous data of maximum temperature, mean temperature, precipitation, due point and the wind speed, the algorithm has been able to successfully understand when a heatwave should occur based on previous incidences.

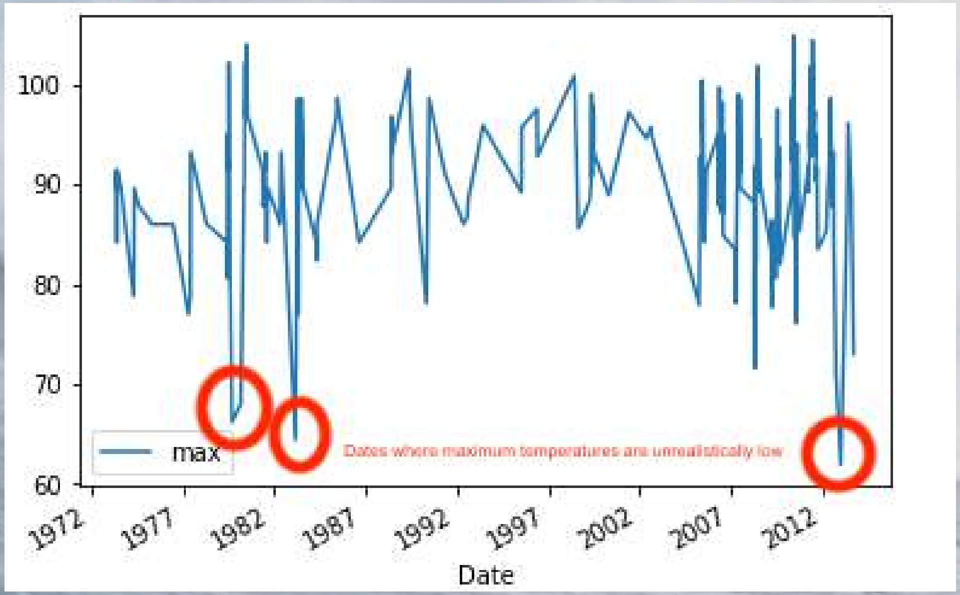
METHODS

- To ensure that our software/algorithm works seamlessly, we did not rely on basic constraints that restrict the algorithm in terms of data analysis. That includes basic if-else statements to differentiate between the different numerals in the data. Instead, we relied on machine learning to teach our algorithm the trends that revolve around the natural disasters that we were analyzing. We did tests with previously measured disasters to see if our algorithm worked accurately, and debugged accordingly.
- The machine learning algorithm draws a best fit line and determines whether the disaster has a possibility of occurring. The algorithm is generic so that it can be used for any natural disasters.

REFINING

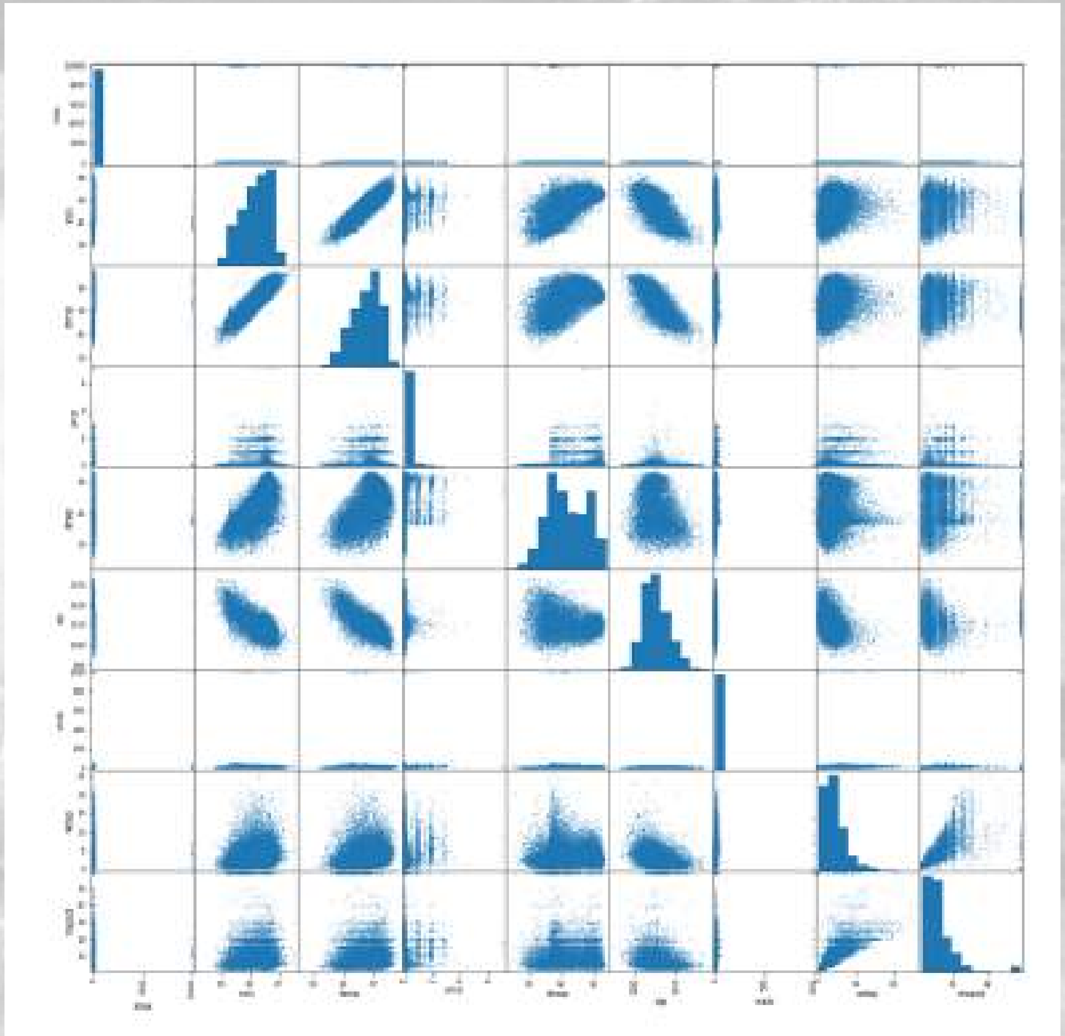
The graph obtained was not linear as expected, so this meant that either heatwaves were recorded incorrectly or there might have been some exceptions in data wherein heatwaves occurred even at low temperatures. For the ease of the application, we chose to take an average value of the maximum temperature for the whole week and eliminated data which contained low temperatures.





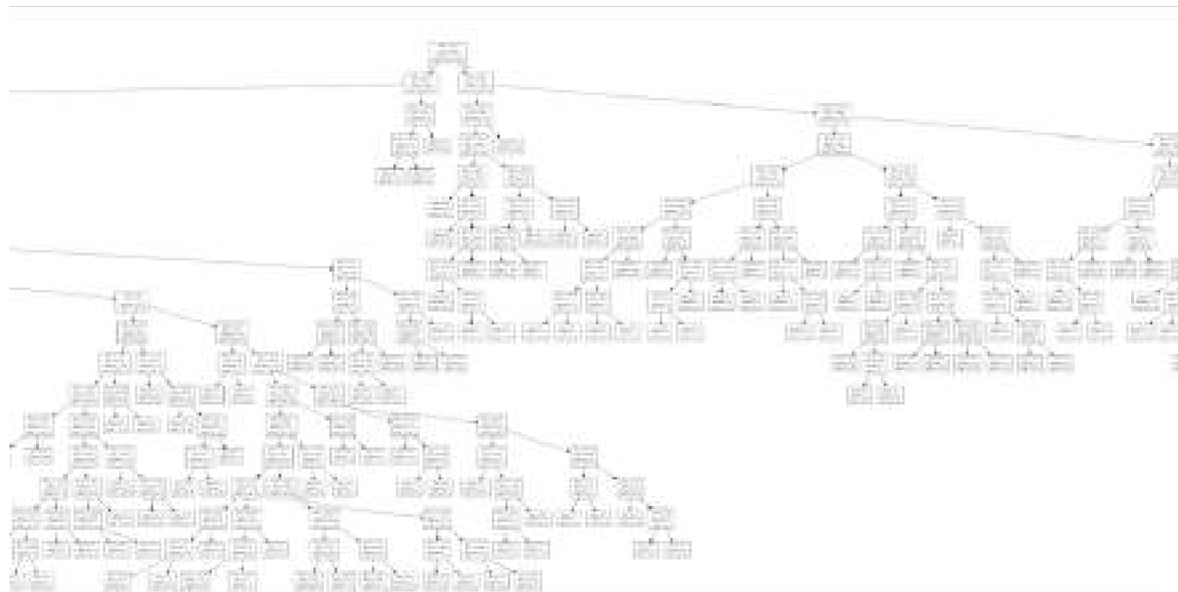
REFINING

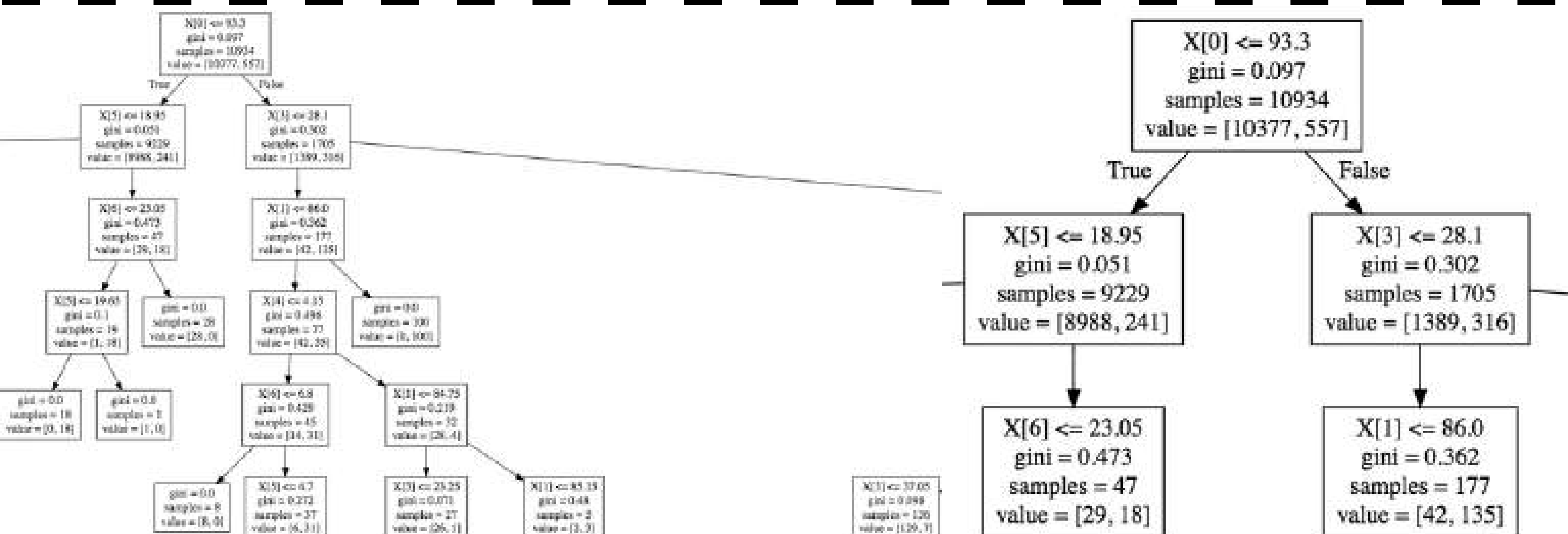
- To handle missing values, we imputed data from other days in the week.
- Features which had too many values missing or were not relevant to our prediction were dropped
- When two features showed high correlation one was dropped as it became irrelevant.
- Correlation graph (right)



DECISION TREE SNIPPETS

The decision tree uses both classification and regression to make decisions based on factors and plot them.





CONCLUSIONS

- Data balancing was an important step in the algorithm because the number of heatwaves not recorded was exponentially higher than the number of heatwaves recorded which caused more predictions of a heatwave not occuring which increased the accuracy rate but didn't effectively predict heatwaves.
- With a test sample of 20% and learning sample of 80%, the algorithm was able to produce successful predictions with high accuracy allowing people to identify when heatwaves will occur.
- The raw data is converted to graphs for ease of understanding for the public using libraries such as plotly, figure factory, graphviz, ggplot and matplotlib.

RESULTS

Accuracy of upto 83% after balancing data.

Accuracy of 97% before balancing data.

Using a decision tree it allowed the algorithm to take multiple factors into consideration for a heatwave to occur.

```
In [197]: df_heatwaves = df_downsampled
          # Column min, slp (sea level pressure) was removed since its highly correlatd with temp
          feature_list = [x for x in ['max','temp', 'prcp','dewp','visib', 'wdsp', 'mxpsd']]
          final_heatwave_set = df_heatwaves[feature_list]
          X = final_heatwave_set
          Y = df_heatwaves['heatwave']
          X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=.2, random_state=82)

In [198]: y_test.shape
          type(y_test)
          y_inter = y_test.to_frame()
          y_inter
          y_inter.heatwave.value_counts()[1.0]
          y_inter_train = y_train.to_frame()
          y_inter_train
          y_inter_train.heatwave.value_counts()[1.0]

Out[198]: 555

In [199]: clf = tree.DecisionTreeClassifier()
          clf = clf.fit(X_train, y_train)
          clf

Out[199]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                      splitter='best')

In [200]: with open('heatwave_classifier.txt", "w") as f:
              f = tree.export_graphviz(clf, out_file=f)

In [201]: predicted = clf.predict(X_test)
          predicted
          print("Accuracy is ", accuracy_score(y_test,predicted)*100)

          Accuracy is  81.72043010752688
```

FINAL SOLUTION

- A machine learning algorithm which uses data collected from weather stations for many years. For this project, we took Chihuahua, Mexico as an example because the data-set we found contained the most number of heatwaves recorded in Mexico.
- The algorithm works by testing on parts of the data-set provided or through manual entry of data by the user. The user enters data such as the maximum temperature and other factors which the model uses to predict weather to see if a heatwave will occur or not. The algorithm learns from heatwaves that occur constantly and will get even better over time.

FUTURE SCOPE

- In the future, this can become a strong algorithm which will be able to precisely identify many natural disasters well before they strike and preventive measures can be taken to stop the natural disaster.
- The algorithm can be used by authorities and organizations that are responsible for the prediction of natural disasters. This can be one of the most accurate and easiest ways of analyzing data and intepreting it to get the information.
- The algorithm can be used to maximize crop yield and economic benefits, as robots and machines will be informed of natural disasters.
- The machine can work much cheaper than humans and professional analysts will be replaced by machine learning algorithms.

REFERENCES

- RED, UNISDR - LA. "United Nations DesInventar Open Source Initiative - Official Website."www.desinventar.net/
- https://www.reference.com/science/heat-waves-occur-world-d71783e16176a1#
- http://www.redcross.org/get-help/how-to-prepare-for-emergencies/types-of-emergencies/heat-wave-safety#About
- https://edition.cnn.com/2017/06/19/world/killer-heat-waves-rising/index.html
- https://www.sciencenews.org/article/are-we-ready-deadly-heat-waves-future?utm_source=editorspicks040818&utm_medium=email&utm_campaign=Editors_Picks
- https://en.wikipedia.org/wiki/Heat_wave
- https://cosmosmagazine.com/climate/all-you-need-to-know-about-heat-waves
- https://www.wunderground.com
- https://www.wmo.int/pages/prog/wcp/ccl/opace/opace4/meetings/documents/fiji2015/D3-5-Alexander_heatwaves.pdf
- https://www.thoughtco.com/heat-waves-weathers-number-one-killer-3443910
- Books, Magazines, Newspaper Articles etc.

ACKNOWLEDGMENTS

- Special thanks to:
- Our mentor: Linda Witte for her expertise as a STEM teacher in motivating and supporting us all the way.
- The Junior Academy for providing us a viable platform to connect and address such great problems that can revolutionize the way we think about this world at large.
- Launchpad for giving us a great platform to interact on and work with our teams seamlessly through video chats, and other means of social communication online.
- Our relatives and our parents for their continuous support, throughout this challenge.