**VISVESVARAYA TECHNOLOGICAL UNIVERSITY**

Jnana Sangama, Machhe, Belagavi, Karnataka 590018



A

Phase II Report

on

# IMPACT OF TEMPERATURE VARIATIONS OVER THE BAY OF BENGAL ON THE CLIMATE OF EASTERN COAST OF INDIA.

*Submitted in partial fulfillment of the requirement*

*for the award of the degree of*

Bachelor of Engineering

in

# Information Science & Engineering

by

**Bhavana (1BG14IS008)**

**H Sudhanva (1BG14IS017)**

**Sai Navaneeth V (1BG14IS041)**

**Satish Kumar M S (1BG14IS045)**

Under the Guidance of

**Dr. Saritha Chakrasali**

Professor

Dept. of ISE

B.N.M.I.T



*Vidyaya Amrutham Ashnuthe*

**B.N.M. Institute of Technology**

12th Main, 27th Cross, Banashankari II Stage, Bangalore 560 070.

Department of Information Science and Engineering

2017-2018

**B.N.M. Institute of Technology**

12th Main, 27th Cross Banashankari II Stage, Bangalore - 560070

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING



*Vidyaya Amrutham Ashnuthe*

##### CERTIFICATE

# Certified that the phase II of project work entitled, “Impact of temperature variations over the Bay of Bengal on the climate of eastern coast of India” is carried out by Mr./Ms. Bhavana (1BG14IS008) , H Sudhanva (1BG14IS017), Sai Navaneeth V (1BG14IS041), Satish Kumar M S (1BG14IS045) , the bonafide students of B.N.M Institute of Technology in partial fulfillment for the award of Bachelor of Engineering in Information Science & Engineering of the Visvesvaraya Technological University, Belagavi during the year 2017-2018. It is certified that all corrections / suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The project synopsis has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

**Dr. Saritha Chakrasali**

**Professor** , Dept. of ISE

BNMIT

**CONTENTS**

**1 Introduction 1**

**1.1 Problem Statement 1**

**1.2 Motivation 2**

**2 System Design 3**

**2.1 Proposed System 3**

**2.1.1 Module-1: Data Collection 6**

**2.1.2 Module-2: Data Cleaning and Filtering 9**

**2.1.3 Module-3: Feature Extraction 13**

**2.1.4 Module-4: Implementing Regression model 14**

**2.1.5 Module-5: Final model 15**

**3 References 16**

Chapter 1

**INTRODUCTION**

Artificial intelligence (AI) traditionally refers to an artificial creation of human-like intelligence that can learn, reason, plan, perceive, or process natural language. Artificial intelligence is further defined as “narrow AI” or “general AI”. Narrow AI, which we interact with today, is designed to perform specific tasks within a domain (e.g. language translation). General AI is hypothetical and not domain specific, but can learn and perform tasks anywhere. This project focuses on advances in narrow AI, particularly on the development of new algorithms and models referred to as machine learning. Computer algorithms organize enormous amounts of data into information and services, based on certain instructions and rules. It’s an important concept to understand, because in machine learning, learning algorithms – not computer programmers – create the rules. Instead of programming the computer every step of the way, this approach gives the computer instructions that allow it to learn from data without new step-by-step instructions by the programmer.

Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements

Weather simply refers to the condition of air on earth at a given place and time. The application of science and technology are to predict the state of the atmosphere in future time for a given location which is important due to its effectiveness in human life. Today, weather forecasts are made by collecting quantitative data about the current state of the atmosphere and using scientific understanding of atmospheric processes to project how the atmosphere will evolve. The chaotic nature of the atmosphere implies the need of massive computational power required to solve the equations that describe the atmospheric conditions.

Various papers related to weather forecasting using Machine Learning algorithms and other data mining concepts were studied. The relevant papers have been discussed belo

**1.1 PROBLEM STATEMENT**

The project aims to achieve the following objectives:

* To analyze related weather data using data mining techniques.
* To obtain correlation between sea surface temperature to corresponding coastal temperature.
* Implementing supervised machine learning algorithm on the available dataset.
* To forecast the weather conditions of the respective city.

**1.2 MOTIVATION**

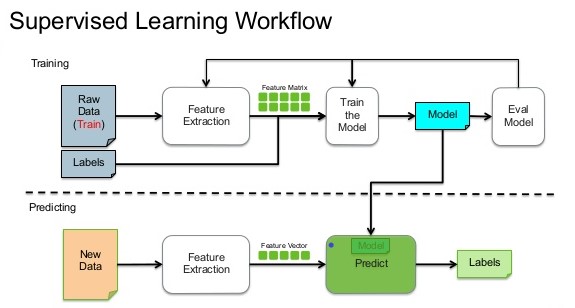
The motivation behind this project is to be able to recognize and establish a pattern between sea and land climate conditions upon considering a large data set for a specific location. Thus, being able to forecast weather more efficiently while improving the process of doing so.

**Chapter 2**

**SYSTEM DESIGN**

**2.1 PROPOSED SYSTEM**

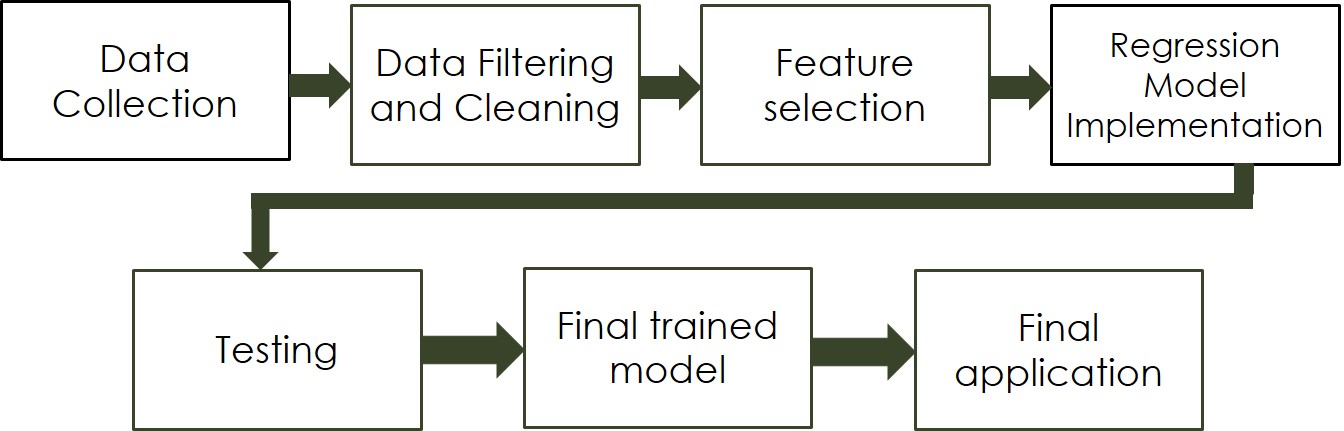
Data collection allows us to gather records of measurements that are already conducted by data mining techniques. The data mining stage is divided into three phases. At each phase the algorithms are used to analyze the available meteorological datasets. The testing method adopted for this research is percentage split, that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge will be identified. With the identified patterns, we apply the supervised machine learning techniques such as linear regression and functional regression to build a Machine Learning model as shown in Fig 1. This Machine Learning model will be able to predict climatic changes which affect the temperature of Chennai city in coastal region due to corresponding change in Bay of Bengal temperature.



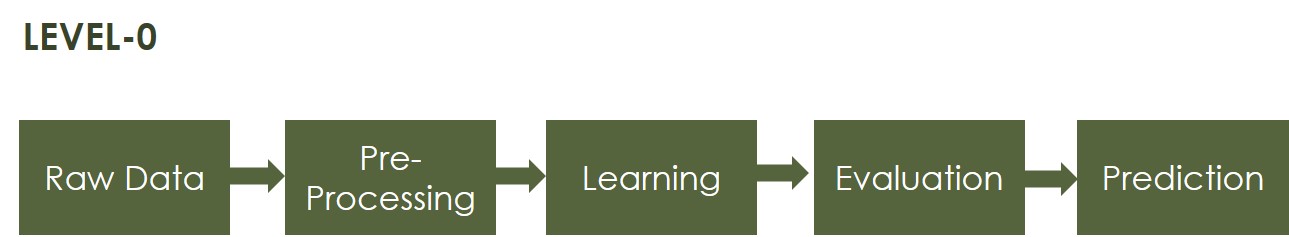
**Fig 1: Supervised Learning Workflow [11]**

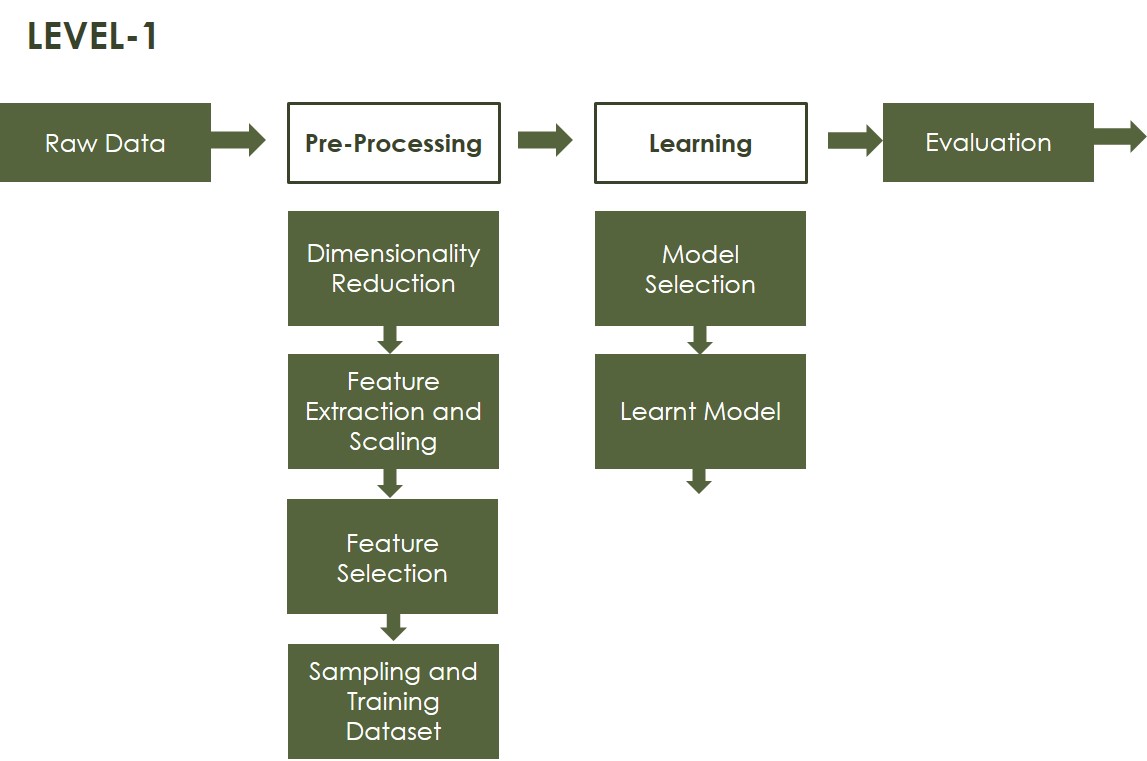
The dataset that will be used for this work is obtained from the following sources:

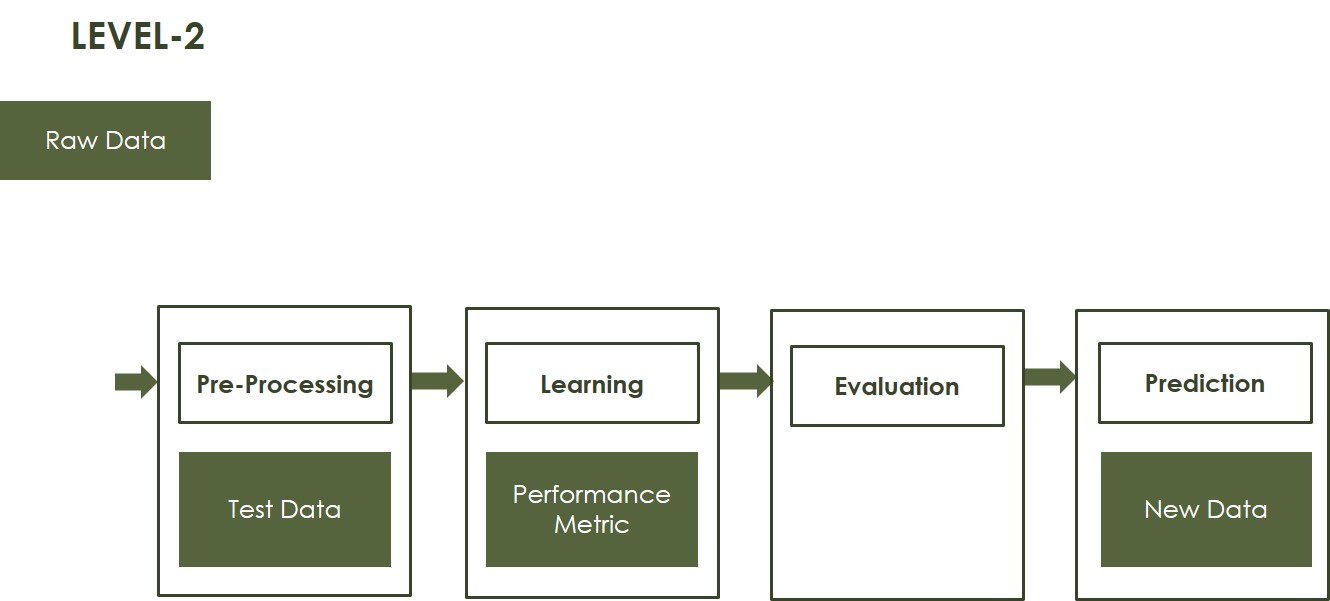
* Land temperature [12]
* Sea surface temperature [13]

****

**DATAFLOW DIAGRAM**

****

****

****

This project has been divided into the following modules:

Module-1: Data Collection

Module-2: Data Cleaning and Filtering

Module-3: Feature Extraction

Module-4: Implementing Regression model

Module-5: Final model

**2.1.1 MODULE 1- DATA COLLECTION**

Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a component of research in all fields of study including [physical](https://en.wikipedia.org/wiki/Physical_science) and [social sciences](https://en.wikipedia.org/wiki/Social_science), [humanities](https://en.wikipedia.org/wiki/Humanities), and [business](https://en.wikipedia.org/wiki/Business). While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed. [14]

The data used for this project is collected from Weather Underground's free tier API web service. Weather Underground provides many different web service API's to access data from but, the one we will be concerned with is their history API. The history API provides a summary of various weather measurements for a city and state on a specific day.

Data collection process takes the following steps:

Step 1: Initialize base url, api key, target date, feature name list.

Step 2: open a csv file in write mode

set the fieldnames to feature list values

Step 3:for each in range(days)

send a request to the base url defined

convert the api response to json format

write the json data to the opened csv file

increment the target date to next day

end for

Step 4: end

For the purpose of this project, the following code has been implemented for the data collection process:

from datetime import datetime, timedelta

import time

from collections import namedtuple

import pandas as pd

import requests

import matplotlib.pyplot as plt

import urllib.request, json

import csv

API\_KEY = "dbe60ccafe99d108"

BASE\_URL = "http://api.wunderground.com/api/{}/history\_{}/q/IN/chennai.json"

features = ["date", "meantempm", "meandewptm", "meanpressurem", "maxhumidity", "minhumidity", "maxtempm", "mintempm", "maxdewptm", "mindewptm", "maxpressurem", "minpressurem", "precipm"]

target\_date = datetime(2015, 1, 1)

def extract\_weather\_data(url, api\_key, target\_date, days):

with open('f2016.csv', 'w', newline='') as f:

thewriter = csv.DictWriter(f, fieldnames=features)

thewriter.writeheader()

for \_ in range(days):

url = BASE\_URL.format(API\_KEY, target\_date.strftime('%Y%m%d'))

response = urllib.request.urlopen(url)

data = json.loads(response.read())

thewriter.writerow({

'date' : target\_date,

'meantempm' : data['history']['dailysummary'][0]['meantempm'],

'meandewptm' : data['history']['dailysummary'][0]['meandewptm'],

'meanpressurem' : data['history']['dailysummary'][0]['meanpressurem'],

'maxhumidity' : data['history']['dailysummary'][0]['maxhumidity'],

'minhumidity' : data['history']['dailysummary'][0]['minhumidity'],

'maxtempm' : data['history']['dailysummary'][0]['maxtempm'],

'mintempm' : data['history']['dailysummary'][0]['mintempm'],

'maxdewptm' : data['history']['dailysummary'][0]['maxdewptm'],

'mindewptm' : data['history']['dailysummary'][0]['mindewptm'],

'maxpressurem' : data['history']['dailysummary'][0]['maxpressurem'],

'minpressurem' : data['history']['dailysummary'][0]['minpressurem'],

'precipm' : data['history']['dailysummary'][0]['precipm']})

print (data['history']['dailysummary'][0]['meantempm'])

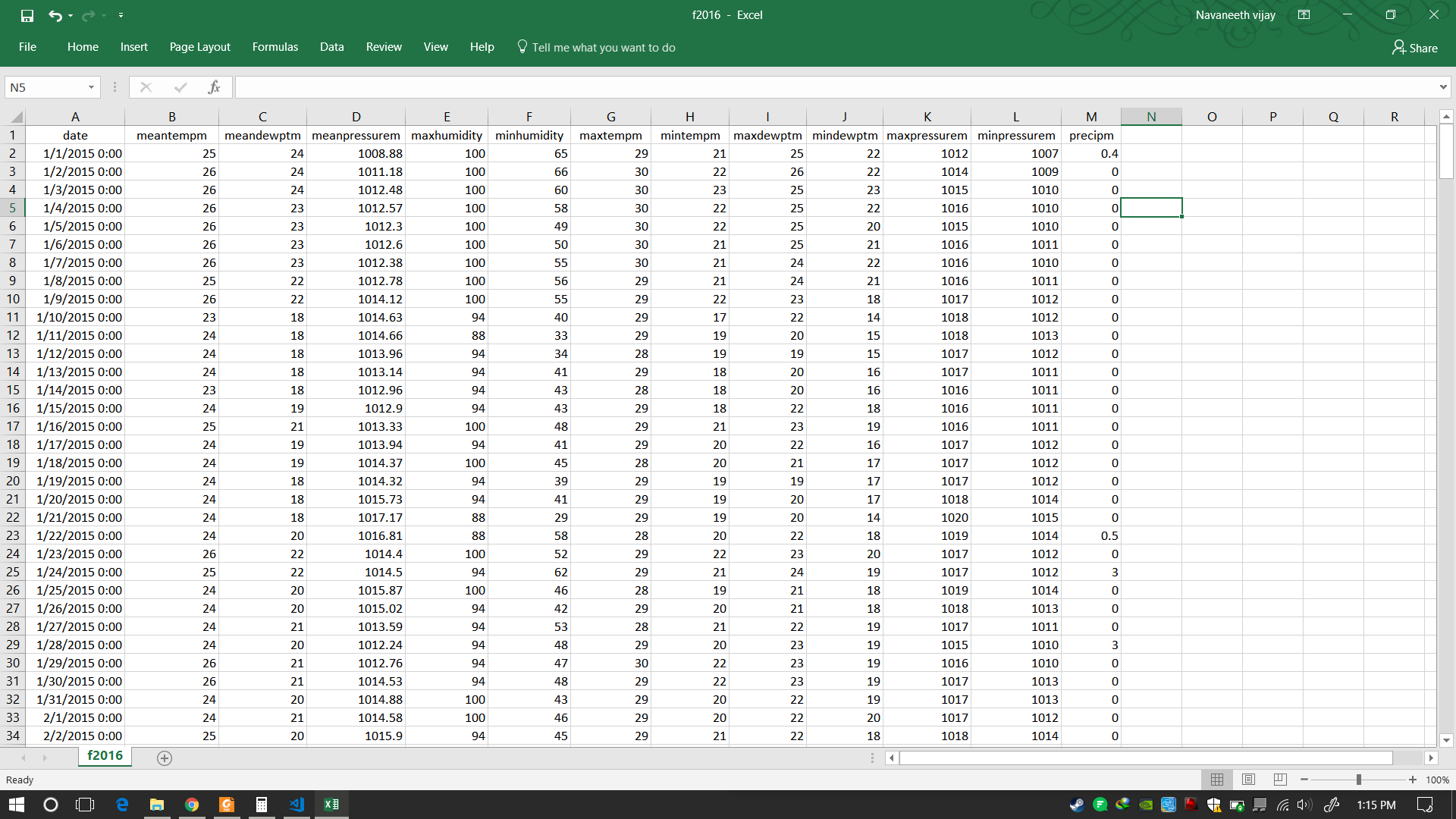
time.sleep(6)

target\_date += timedelta(days=1)

return

extract\_weather\_data(BASE\_URL, API\_KEY, target\_date,365)

Sample output of the above code being as follows:



**2.1.2 MODULE 2- DATA FILTERING AND CLEANING**

Data filtering in IT can refer to a wide range of strategies or solutions for refining data sets. This means the data sets are refined into simply what a user (or set of users) needs, without including other data that can be repetitive, irrelevant or even sensitive. Different types of data filters can be used to amend reports, query results, or other kinds of information results. [15]

Data cleaning, (or data cleansing, data scrubbing) is an aspect of [data processing](http://psychology.wikia.com/wiki/Data_processing) and is the process of detecting and correcting (or removing) corrupt or inaccurate [records](http://psychology.wikia.com/wiki/Storage_record?redlink=1&action=edit&flow=create-page-article-redlink) from a record set, [table](http://psychology.wikia.com/wiki/Table_(database)), or [database](http://psychology.wikia.com/wiki/Database). Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this [dirty data](http://psychology.wikia.com/wiki/Dirty_data?redlink=1&action=edit&flow=create-page-article-redlink). [16]

The steps used for the process of data filtering and cleaning are as follows:

Step 1: Make a list containing features to remove

Step 2: Make a list containing features to keep

Step 3: Convert the selected feature objects into numerical values

Set not available values to 'coerce’.

Step 4: Transpose the DataFrame Set.

Step 5: Calculate IQR and remove outliers

Step 6: Data visualization.

Step 7: End

The following code has been implemented for the data filtering and cleaning process:

import pandas as pd

import matplotlib

import numpy as np

import matplotlib.pyplot as plt

df = pd.read\_csv(r"C:\Users\LalithaShankar\Desktop\FinalYearProject\VtuProject\vtuproject\f2016.csv").set\_index('date')

#tmp = df[['meantempm', 'meandewptm']].head(10)

N = 1

features = ["date", "meantempm", "meandewptm", "meanpressurem", "maxhumidity", "minhumidity", "maxtempm",

"mintempm", "maxdewptm", "mindewptm", "maxpressurem", "minpressurem", "precipm"]

def derive\_nth\_day\_feature(df, feature, N):

rows = df.shape[0]

nth\_prior\_measurements = [None]\*N + [df[feature][i-N] for i in range(N, rows)]

col\_name = "{}\_{}".format(feature, N)

df[col\_name] = nth\_prior\_measurements

for feature in features:

if feature != 'date':

for N in range(1, 4):

derive\_nth\_day\_feature(df, feature, N)

#df.columns

#df.shape

#Data\_Cleaning

# make list of original features without meantempm, mintempm, and maxtempm

to\_remove = [feature

for feature in features

if feature not in ['meantempm', 'mintempm', 'maxtempm']]

# make a list of columns to keep

to\_keep = [col for col in df.columns if col not in to\_remove]

# select only the columns in to\_keep and assign to df

df = df[to\_keep]

#df.columns

#df.info()

df = df.apply(pd.to\_numeric, errors='coerce')

#df.info()

# Call describe on df and transpose it due to the large number of columns

spread = df.describe().T

# precalculate interquartile range for ease of use in next calculation

IQR = spread['75%'] - spread['25%']

# create an outliers column which is either 3 IQRs below the first quartile or

# 3 IQRs above the third quartile

spread['outliers'] = (spread['min']<(spread['25%']-(3\*IQR)))|(spread['max'] > (spread['75%']+3\*IQR))

# just display the features containing extreme outliers

spread.ix[spread.outliers,]

%matplotlib inline

plt.rcParams['figure.figsize'] = [14, 8]

df.maxhumidity\_1.hist()

plt.title('Distribution of maxhumidity\_1')

plt.xlabel('maxhumidity\_1')

plt.show()

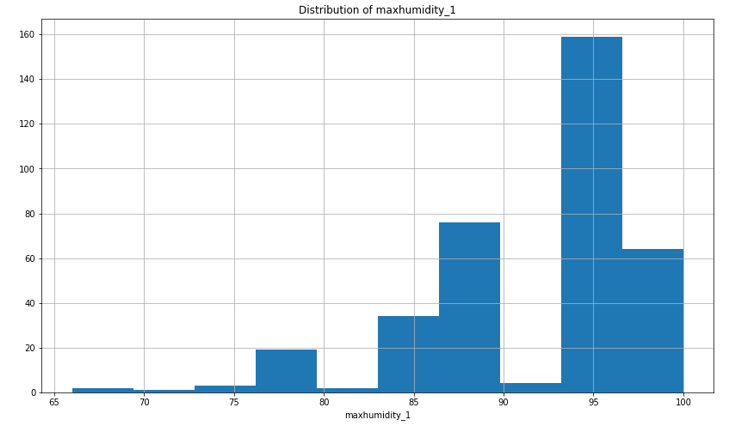
df.minpressurem\_1.hist()

plt.title('Distribution of minpressurem\_1')

plt.xlabel('minpressurem\_1')

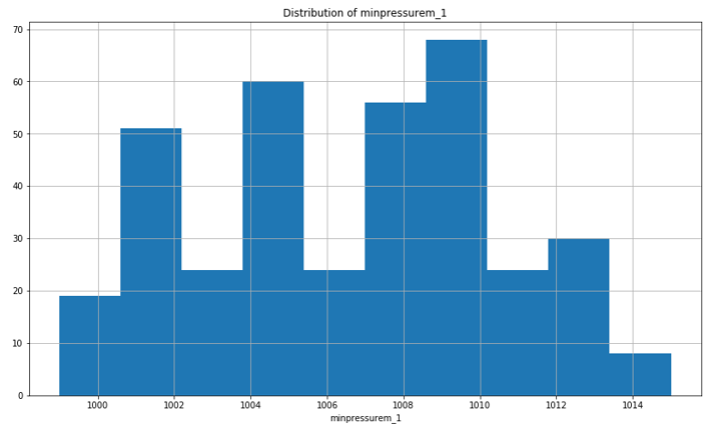
plt.show()

The first set of features all appear to be related to max humidity. Looking at the data one can tell that the outlier for this feature category is due to the apparently very low min value. To take a closer look at it, use a histogram.

****

Looking at the histogram of the values for maxhumidity the data exhibits quite a bit of negative skew. This must be kept in mind when selecting prediction models and evaluating the strength of impact of max humidities.

Now consider the minimum pressure feature distribution.

****

This plot exhibits another interesting feature. From this plot, the data is [multimodal](https://en.wikipedia.org/wiki/Multimodal_distribution), from which it can be considered that there are two very different sets of environmental circumstances apparent in this data.

**2.1.3 MODULE 3- FEATURE SELECTION**

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [statistics](https://en.wikipedia.org/wiki/Statistics), feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

* simplification of models to make them easier to interpret by researchers/users,
* shorter training times,
* to avoid the [curse of dimensionality](https://en.wikipedia.org/wiki/Curse_of_dimensionality),
* enhanced generalization by reducing [overfitting](https://en.wikipedia.org/wiki/Overfitting) (formally, reduction of [variance](https://en.wikipedia.org/wiki/Bias-variance_tradeoff)) [17]

**2.1.4 MODULE 4- IMPLEMENTING LINER REGRESSION MODEL**

Linear regression is a [linear](https://en.wikipedia.org/wiki/Linear) approach for modelling the relationship between a scalar [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) *y* and one or more [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or independent variables) denoted *X*.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

* If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.
* Given a variable *y* and a number of variables *X*1, ..., *Xp* that may be related to *y*, linear regression analysis can be applied to quantify the strength of the relationship between *y* and the *Xj*, to assess which *Xj* may have no relationship with *y* at all, and to identify which subsets of the *Xj* contain redundant information about *y*. [18]

**2.1.5 MODULE 5- TESTING**

Software testing is a method of assessing the functionality of a [software](http://searchsoa.techtarget.com/definition/software) [program](http://searchsoftwarequality.techtarget.com/definition/program). There are many different types of software testing but the two main categories are [dynamic testing](http://whatis.techtarget.com/definition/dynamic-testing) and [static testing](http://whatis.techtarget.com/definition/static-testing).

Dynamic testing is an assessment that is conducted while the program is [executed](http://searchcio-midmarket.techtarget.com/definition/executable); static testing, on the other hand, is an examination of the program's [code](http://whatis.techtarget.com/definition/code) and associated [documentation](http://searchsoftwarequality.techtarget.com/definition/documentation). Dynamic and static methods are often used together. [19]

**3 REFERENCES**

[1] Mark Holmstrom, Dylan Liu, Christopher Vo. Machine Learning Applied to Weather Forecasting. Stanford University. 2016.

[2] Siddharth S. Bhatkhande, Roopa G. Hubballi. Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques. *IJARCCE* (2016).

[3] Aditya Grover, Ashish Kapoor and Eric Horvitz. A Deep Hybrid Model for Weather Forecasting.

[4] John K. Williams and D. A. Ahijevych, C. J. Kessinger, T. R. Saxen, M. Steiner and S. Dettling. A machine learning approach to finding weather regimes and skillful predictor combinations for short-term storm forecasting. National Center for Atmospheric Research, Boulder, Colorado.

[5] Aastha Sharma, Setu Chaturvedi and Bhupesh Gour. A Semi- Supervised Technique for Weather Condition Prediction using DBSCAN and KNN.International Journal of Computer Applications (0975 – 8887) Volume 95– No. 10, June 2014.

[6] Mrs. C. Beulah Christalin Latha, Dr. (Mrs.) Sujni Paul, Dr.E.Kirubakaran and Mr. Sathianarayanan. A Service Oriented Architecture for Weather Forecasting Using Data Mining.Int. J. of Advanced Networking and Applications. Volume: 02, Issue:02, Pages:608-613. 2010.

[7] C. Johansson, M. Bergkvist, O. De Somer, D. Geysen , N. Lavesson and D. Vanhoudt. Operational demand forecasting in district heating systems using ensembles of online machine learning algorithms.The 15th International Symposium on District Heating and Cooling September 4-7, 2016, Seoul, Republic of Korea (South Korea).

[8] Daniel Bejarano and Adriano Quiroga. Wind Prediction: Physical model improvement through support vector regression.Stanford University. December 2013.

[9] Emilcy Hern ́andez Victor Sanchez-Anguix Vicente Julian Javier Palanca and N ́estor Duque.Rainfall prediction: A Deep Learning approach.

[10] Kiran Kumar. R 1, Usha Rani. R.Weather Prediction through Machine Learning.Dept of Computer science, Krishna University, Machilipatnam. AP, India.

[11] <http://en.proft.me/2015/12/24/types-machine-learning-algorithms/>

[12] <https://www.wunderground.com/>

[13] <https://seatemperature.info/>

[14] <https://en.wikipedia.org/wiki/Data_collection>

[15] <https://www.techopedia.com/definition/26202/data-filtering>

[16] <http://psychology.wikia.com/wiki/Data_cleaning>

[17] <https://en.wikipedia.org/wiki/Feature_selection>

[18] <https://en.wikipedia.org/wiki/Linear_regression>

[19] <http://whatis.techtarget.com/definition/software-testing>