

Smartbridge

Machine Learning with Python & IBM Watson

**An Internship Report submitted in partial fulfilment of the
requirements for the award of the degree of**

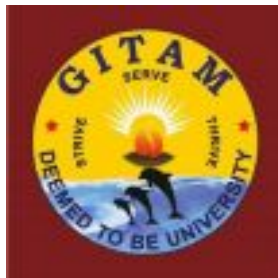
BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Phaneendra Allu, 121710304004



**DEPARTMENT OF COMPUTER SCIENCE &
ENGINEERING (Deemed to be University)**

Visakhapatnam

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GITAM INSTITUTE OF TECHNOLOGY**

GITAM
(Deemed to be University)



Declaration

We, hereby declare that the internship review entitled “**Machine Learning with Python & IBM Watson**” is an original work done in the Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University) submitted in partial fulfilment of the requirements for the award of the degree of B.Tech. in Computer Science and Engineering. The work has not been submitted to any other college or University for the award of any degree or diploma.

Date: 13th November 2020

Name: Phaneendra Allu

CERTIFICATE



Date: 13/05/2020

SB ID: SB20200001257

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Mr./Ms. Allu Phaneendra** pursuing B.Tech in Computer Science Engineering from Gitam Deemed to be University, Visakhapatnam has successfully completed his/her Summer Internship from **15/04/2020** to **13/05/2020**.

During this period he/she had learned the concepts of **Machine Learning with Python & IBM Watson** and successfully completed a project **"Food Requirement Analysis in an Area"**.

Refer the enclosed **Certificate of Merit** for his/her performance during the tenure of Internship.

We wish him/her all the best for his/her future endeavours.

For **SmartBridge Educational Services Pvt. Ltd.,**



Jayaprakash.ch,
Program Manager
RSIP-2020

Certificate of Merit

This is to certify that **Mr./Ms. Allu Phaneendra** has completed his/her internship program from **15/04/2020** to **13/05/2020** with **Machine Learning with Python & IBM Watson** as the specialization and secured a **SKILL INDEX 9** of 10

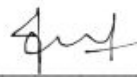
Career Readiness Factor (CRF)

Evaluation Metrics: (on a scale of 1 to 4)

- 1 - Rarely/poorly displays characteristic 2 - Occasionally displays characteristic
 3 - Frequently displays characteristic 4 - Always displays characteristic
 NA – Not Applicable

Motivation/Enthusiasm	4
Leadership Qualities	3
Flexibility towards work	4
Professionalism/Work Ethics	3
Self-Confidence	4
Ability to work independently	3
Oral/written communication	4
Problem solving skills	4
Over All Score	29

Date: 13/05/2020


 Jayaprakash.ch



CERTIFICATE

This is to certify that the internship report entitled “**Machine Learning with Python & IBM Watson**” is a bonafide record of work carried out by **Phaneendra Allu (121710304004)** student submitted in partial fulfilment of requirement for the award of degree of Bachelors of Technology in Computer Science and Engineering.

INTERNSHIP REVIEWER PROJECT GUIDE **Prof. Suresh Chittineni Kalki**

Seetharaman (Dir. Skill and Development) (Assistant Professor) Dr.

Subhadra Kompella

(Assistant Professor)

Vinod Babu

(Assistant Professor)

TABLE OF CONTENTS

Abstract.....	7 1)
----------------------	-------------

Introduction.....	8-9 2)
Literature Survey.....	10 3)
Theoretical Analysis.....	11-13
4) Methodology.....	14-21
4.1: Data Preprocessing	
4.2: Label Encoding	
4.3: One-Hot Encoding	
4.4: Modelling	
5)	
Outcomes.....	22
6)	
Conclusion.....	23
7) Bibliography.....	23

Abstract

Food analysis is a prerequisite for ascertaining product quality, implementing regulatory enforcements, checking compliance with national and international food standards, contracting specifications and nutrient

labeling requirements. One third of all food produced is lost or wasted – around 1.3 billion tonnes of food –costing the global economy close to \$940 billion each year. If one quarter of the food currently lost or wasted could be saved, it would be enough to feed 870 million hungry people. In this project, we will be using Machine Learning Algorithms to predict the number of orders a restaurant should produce given certain variables to minimise the food wastage.

1) Introduction:

1.1 Overview:

Food is central to human well-being: it provides the body with nourishment, offers livelihoods that lift people out of poverty, and brings communities together. Although food is a basic human need,

too many people are trapped in a cycle of hunger by forces beyond their immediate control, like poverty, disaster, conflict and inequality.

Despite decades of progress in reducing world hunger, 2017 saw increases in the number of people who are hungry. More than 820 million people still go to bed hungry every night — that's one in every nine people who don't have the food they need to live a healthy, productive life.

The World Health Organization considers this to be the single greatest threat to global health. Hunger is cyclical and generational: it inhibits people's ability to work and learn to their fullest potential, which can curb their future and trap them and their families in more poverty — and more hunger.

People in poverty generally spend between 60 and 80 percent of their income on food, which can force them to prioritize feeding their families over meeting other basic needs or reaching long-term goals, like sending their children to school. If an emergency strikes, they may need to skip meals in order to cope financially — and the cycle of hunger continues.

According to the Food Security Information Network, conflict and insecurity were primary drivers of food insecurity in 2017, alone accountable for putting 74 million people in need of urgent assistance. Climate change is also eroding existing efforts to improve food security.

1.2 Purpose:

Recent years, the ML methods have become popular as they allow researchers to improve prediction accuracy and used in many applications.

In this study, regression algorithms are used to predict various food material requirement. Based on this predictions farmers and producers can produce them to meet the requirement. Here ML regression methods are used for prediction. The study aimed to determine the most successful regression methods by comparing the logistic regression, decision tree (DT), random forest (RF), linear regression, multiple regression, support vector regression, Naïve Bayes.

2) Literature Survey:

2.1 Existing Problem:

Our world population is expected to grow from 7.3 billion today to

9.7 billion by 2050. Finding solutions for feeding the growing world population has become a hot topic for food and agriculture organizations, entrepreneurs and philanthropists. These solutions range from changing the way we grow our food to changing the way we eat. To make things harder, the world's climate is agriculture. Hence, it is necessary that we analyze the food production and act faster rather than repenting later.

2.2 Proposed Solution :

Using a machine learning model, we can predict the number of orders on a food item per day for any kind of cuisine located in various regions and cities which is directly proportional to the food required. We can take various inputs as a base price, region code, city code, food item id, type of cuisine etc to the algorithm. We then use regression algorithms to predict which food item gets how many orders. A food item is considered to be most interesting when the number of orders has more than 500 per day. The model will be trained based on the algorithm features which will give the higher accuracy. Finally, it will be integrated into a web based application. The final system allows the user to give the details required as input and the system will display the output number of orders.

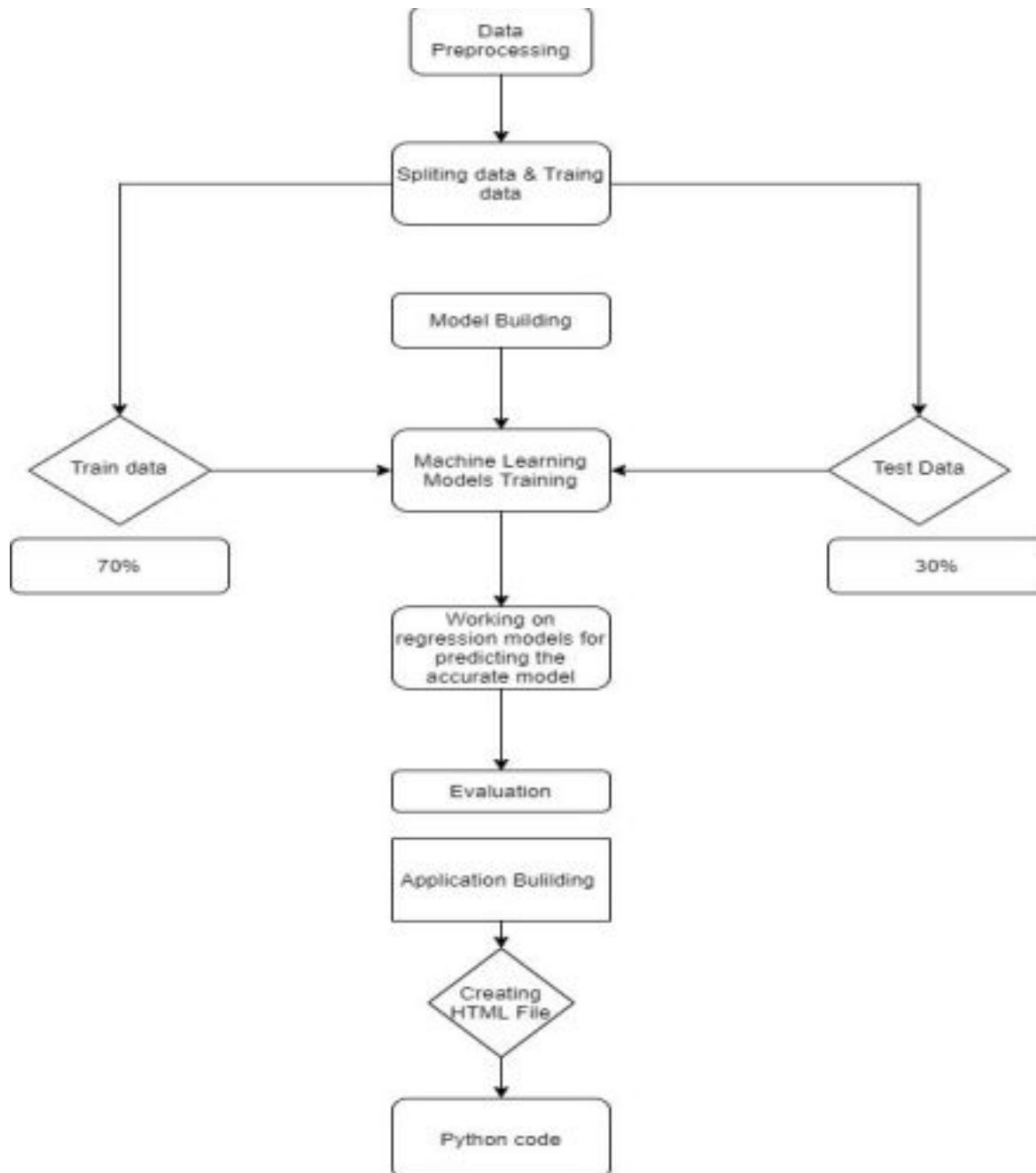
3) Theoretical Analysis

3.1 Hardware / Software designing:

The software used to implement this project is mainly Jupyter Notebook and Spyder. The other programming skills required are Python, Flask, HTML and CSS. The project works in 2 parts. The first part is building the model and the other part is building the frontend. The model building is done by using Machine Learning libraries by the help of Python. The frontend development is done by using HTML and the designs are added by using CSS. Both the frontend and backend are connected by using Flask. The Flask generates localhost for the application to run.

	values Label
	Encoding
	OneHot Encoding
Food requirement analysis	Feature Scaling
Data Collection	Splitting data
Data Preprocessing	Model Building
Import libraries	Application Building
Import data set	Create html file
Data Visualization	Build python code
Handling null	

3.3 Flowchart:



13

4.METHODOLOGY:

4.1 DataPreprocessing

Data Merging

- Data merging is the process of combining two or more data sets into a single data set. Most often, this process is necessary when you have raw

data stored in multiple files, worksheets, or data tables, that you want to analyze all in one go.

- We have three datasets. We merge them to form a single dataset.

Merge datasets

```
data1=pd.read_csv("Food-demand.csv")
```

```
data2=pd.read_csv("fulfilment_center_info.csv")
```

```
data3=pd.read_csv("meal_info.csv")
```

```
dataset = data3.merge(data1)
```

```
dataset=data2.merge(dataset)
```

```
dataset.head()
```

	center_id	city_code	region_code	center_type	sp_area	meal_id	category	cuisine	id	week	checkout_price	base_price	emailer_for_promotion	homeq
0	11	679	58	0	3.7	1885	0	3	1103215	1	138.83	138.83	0	
1	11	679	58	0	3.7	1885	0	3	1091358	2	133.88	135.88	0	
2	11	679	58	0	3.7	1885	0	3	1195933	3	135.88	133.88	0	
3	11	679	58	0	3.7	1885	0	3	1425802	4	134.88	135.88	0	
4	11	679	58	0	3.7	1885	0	3	1248127	5	148.53	148.53	0	

Data Cleaning

- Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. Data Cleaning is one of those things that everyone does but no one really talks about. It surely isn't the fanciest part of machine learning and at the same time, there aren't any

hidden tricks or secrets to uncover. However, proper data cleaning can make or break your project. Professional data scientists usually spend a very large portion of their time on this step.

- Before we go for modeling the data, we have to check whether the data is cleaned or not. And after cleaning, we have to structure the Data. For the cleaning part, First I have to check whether there exists any missing values. For that I am using the code snippet isnull()

Finding missing values

```
: dataset.isnull().any()
center_id           False
city_code           False
region_code         False
center_type         False
op_area             False
meal_id             False
category            False
cuisine             False
id                  False
week                False
checkout_price      False
base_price          False
emailer_for_promotion False
homepage_featured   False
num_orders          False
dtype: bool
```

There are no missing values in our dataset. So we continued with other operations.

4.2 LABEL ENCODING:

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning.

Label Encoding

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
dataset['category'] = le.fit_transform(dataset['category'])
dataset['center_type'] = le.fit_transform(dataset['center_type'])
dataset['cuisine'] = le.fit_transform(dataset['cuisine'])
```

```
dataset.head()
```

	center_id	city_code	region_code	center_type	sp_area	meal_id	category	cuisine	id	week	checkout_price	base_price	enabler_for_promotion	home
0	11	679	56	0	3.7	1895	0	3	1100215	1	136.83	136.83	0	
1	11	679	56	0	3.7	1895	0	3	1061356	2	133.86	135.86	0	
2	11	679	56	0	3.7	1895	0	3	1185833	3	135.86	133.86	0	
3	11	679	56	0	3.7	1895	0	3	1425602	4	134.86	135.86	0	
4	11	679	56	0	3.7	1895	0	3	1248127	5	146.53	146.53	0	

16

4.3 ONE HOT ENCODING:

A one hot encoding allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers.

One Hot Encoding

```
from sklearn.preprocessing import OneHotEncoder
oh=OneHotEncoder()
z=oh.fit_transform(x[:,3:4]).toarray()
p=oh.fit_transform(x[:,6:7]).toarray()
q=oh.fit_transform(x[:,7:8]).toarray()
x=np.delete(x,3,axis=1)
x=np.delete(x,6,axis=1)
x=np.delete(x,7,axis=1)
x=np.concatenate((q,p,z,x),axis=1)
```

```
x.shape
```

```
(456548, 32)
```

17

4.4 MODELLING

IMPORTANT PYTHON LIBRARIES

- Numpy
- Pandas
- Scikit Learn

In modelling the data is fed into machine learning algorithms.

So we split the data into two parts training and testing.

The training data is used to train the algorithm and testing data is used to test the model .

Generally the model having the highest accuracy is chosen as the good model.

The problem here we are dealing comes under the classification as we have to predict whether the applicant can get loan sanctioned or not?

There are many algorithms for Regression in the scikit learn library. The algorithms here we are going to use are

- ☐ Logistic Regression
- ☐ Decision Trees Regression
- ☐ Polynomial Regression
- ☐ Random Forest Regression

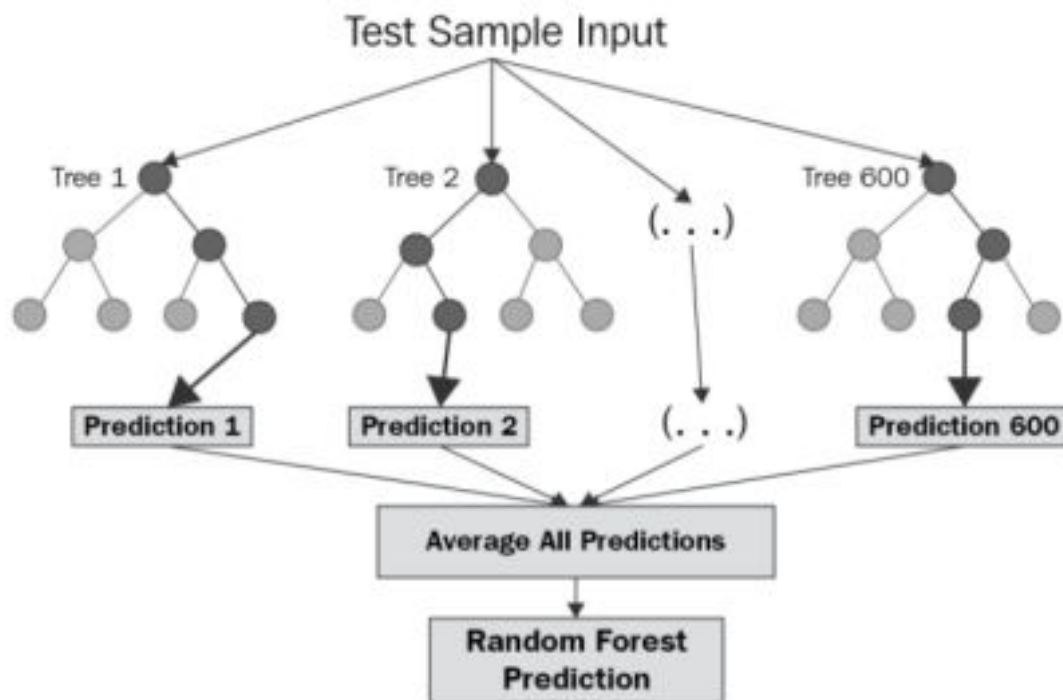
Random Forest Regression

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

Feature and Advantages of Random Forest :

1. It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
2. It runs efficiently on large databases.
3. It can handle thousands of input variables without variable deletion.
4. It gives estimates of what variables that are important in the classification.



19

Random Forest Regressor

```
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 100, random_state = 0)
regressor.fit(x_train, y_train)
```

```
<ipython-input-44-3073cddef914>:3: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
regressor.fit(x_train, y_train)
```

```
RandomForestRegressor(random_state=0)
```

```
randforest = regressor.predict(x_test)
randforest
```

```
array([ 581.39,  46.64, 1042.93, ..., 144.91, 110.1 , 104.31])
```

```
from sklearn.metrics import r2_score
acc3 = r2_score(y_test, randforest)
acc3*100
```

```
83.17683267299506
```

Saving the model :

After the successful execution of the model, we will save it in the pickle (.pkl) format using dump() method.

Exporting pickle file

```
import pickle
pickle.dump(regressor, open('num_orders.pkl', 'wb'))
```

20

Flask code :-

Flask is often referred to as a micro framework. It aims to keep the core of an application simple yet extensible. Flask does not have built-in abstraction layer for database handling, nor does it have form a validation support. Instead, Flask supports the extensions to add such functionality to the application.

```
from flask import Flask, render_template, request

import pickle
import numpy as np

model=pickle.load(open('num_orders.pkl','rb'))

app=Flask(__name__)

@app.route('/')
def home():
    return render_template('index.html')

@app.route('/login', methods=['POST'])
def login():
    base_price=request.form['base_price']
    Dishes=request.form['Dishes']
    if(Dishes=="Beverages"):
        s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14=1,0,0,0,0,0,0,0,0,0,0,0,0,0
    if(Dishes=="Extras"):
        s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14=0,0,0,1,0,0,0,0,0,0,0,0,0,0
    if(Dishes=="Soup"):
        s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14=0,0,0,0,0,0,0,0,0,0,0,0,1,0
    if(Dishes=="Other Snacks"):
        s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14=0,0,0,0,1,0,0,0,0,0,0,0,0,0
    if(Dishes=="Salad"):
        s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14=0,0,0,0,0,0,0,0,1,0,0,0,0,0
    if(Dishes=="Rice Bowl"):
        s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14=0,0,0,0,0,0,0,0,1,0,0,0,0,0
    if(Dishes=="Starters"):
        s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14=0,0,0,0,0,0,0,0,0,0,0,0,0,1
```

5. Outcomes :

Food Analysis

base_price

Select Dishes

Select restaurant

Select center

checkout_price

city_code

Select cuisine

Select meal

Select homepage_featured

id

Select meal_id

Select op area

Select region

Select week

Predict

num of orders per week: 142.41

When given values for variables such as base_price, Dishes, checkout_price, city_code, Cuisine, id, meal_id, the model predicts the number of orders to be produced as 142.41.

6. Conclusion:

In this study, a prediction model of Food Requirements in an area was established by various Machine Learning Algorithms. A dataset meticulously gathered and used to develop the Machine Learning Regression algorithms for prediction of number of orders per day for any kind of cuisine located in various regions and cities(directly proportional to the food required). Conclusions can be drawn as follows:

- Compared to all other Machine Learning Models Random Forest Regression was best suitable for this data.
- Random Forest Regressor gave the maximum accuracy when tested using accuracy_score confusion matrix.
- Maximum accuracy received is 83.17 %.

7. Bibliography:

Books

- 1) Hastie, Friedman, and Tibshirani, the Elements of Statistical Learning, 2001
- 2) Bishop, Pattern Recognition and Machine Learning, 2006 abs/1809.02862, 2018
- 3) C. Anderson, "A survey of food recommenders," *ArXiv*, vol.

abs/1809.02862, 2018

- 4)C. M. Lăcătușu, E. D. Grigorescu, M. Floria, A. Onofriescu and B. M. Mihai, “The Mediterranean Diet: From an Environment-Driven Food Culture to an Emerging Medical Prescription,” *International Journal of environmental research and public health*, vol. 6, no. 16, 2019