Department of Computer Science and Engineering (Data Science)

Subject: Machine Learning – I (DJ19DSC402)

AY: 2022-23

Experiment 10

(Mini Project)

Aim: Design a classifier to solve a specific problem in the given domain.

Tasks to be completed by the students:
Select a specific problem from any of the given domain areas, such as: Banking, Education, Insurance, Government, Media, Entertainment, Retail, Supply chain, Transportation, Logistics, Energy and Utility.
Task 1: Select appropriate dataset, describe the problem and justify the suitability of your dataset.
Task 2: Perform exploratory data analysis and pre-processing (if required).
Task 3: Apply appropriate machine learning algorithm to build a classify. Perform appropriate testing of your model.
Task 4: Submit a report in the given format.

- Introduction
- Data Description
- Data Analysis
- Reason to select machine learning model
- Algorithm
- Result Analysis
- Conclusion and Future Scope.
- Python notebook

Task5: Presentation

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)
Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2022-23

FOOD RELATED ILLNESS AND DISEASES

NAME: Sowmya Dadheech

SAP ID: 60009210163

Guided By

Dr Kriti Srivasta

CHAPTER 1: INTRODUCTION

A foodborne disease outbreak occurs when two or more people get the same illness from the same contaminated food or drink. While most foodborne illnesses are not part of a recognized outbreak, outbreaks provide important information on how germs spread, which foods cause illness, and how to prevent infection.
This project is to anticipate the analysis of illness with respect to various features such as State, Month, Year, Location and Food items.

I have solved the problem statement by following these steps:
1. Data Exploration
2. Exploratory Data Analysis (EDA)
3. Data Pre-processing
4. Predictive Modeling
5. Project Outcomes & Conclusion

CHAPTER 2: DATA DESCRIPTION

This dataset provides data on foodborne disease outbreaks reported to CDC from 1998 through 2015.

Data fields include:

- Year (the year of reported of outbreak)
- State (outbreaks occurring in more than one state are listed as "multistate")
- Location (where the food was prepared, reported food vehicle and contaminated ingredient)
- Etiology – Serotype/Genotype (the pathogen, toxin, or chemical that caused the illnesses)
- Status (whether the etiology was confirmed or suspected)
- Total illnesses
- Hospitalizations
- Fatalities

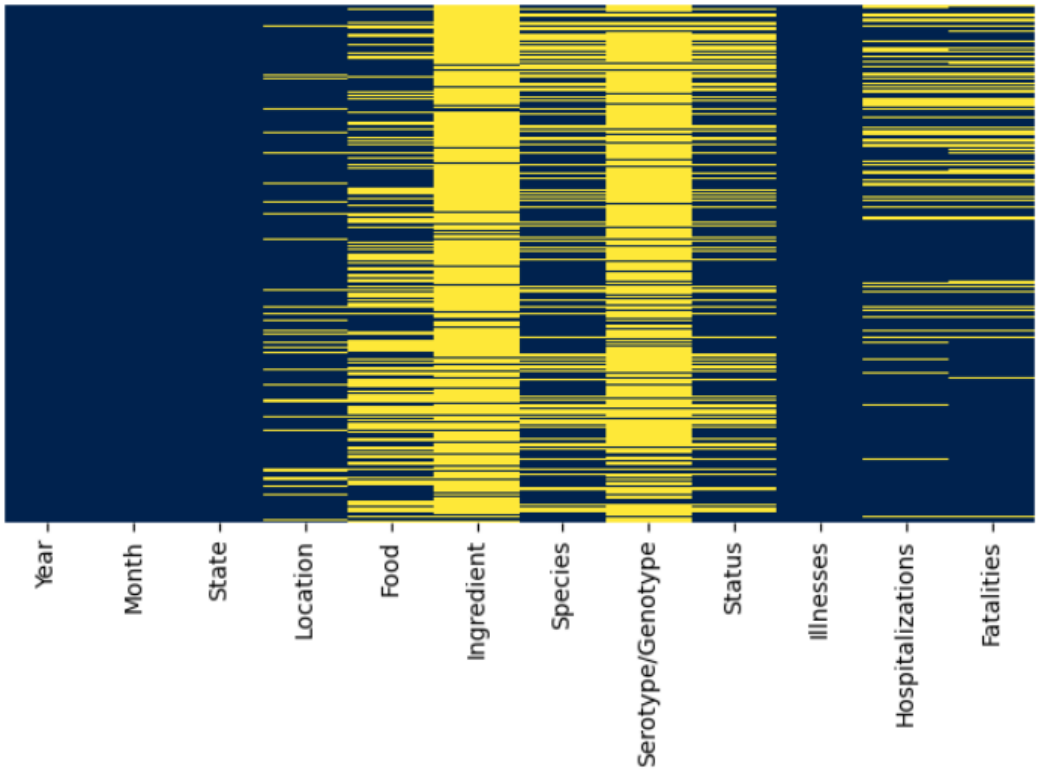
CHAPTER 3: DATA ANALYSIS

There were 19119 records in the dataset. The columns 'Ingredients' and 'Serotype' had almost the same number of null values. So, we did not consider these columns in our analysis.

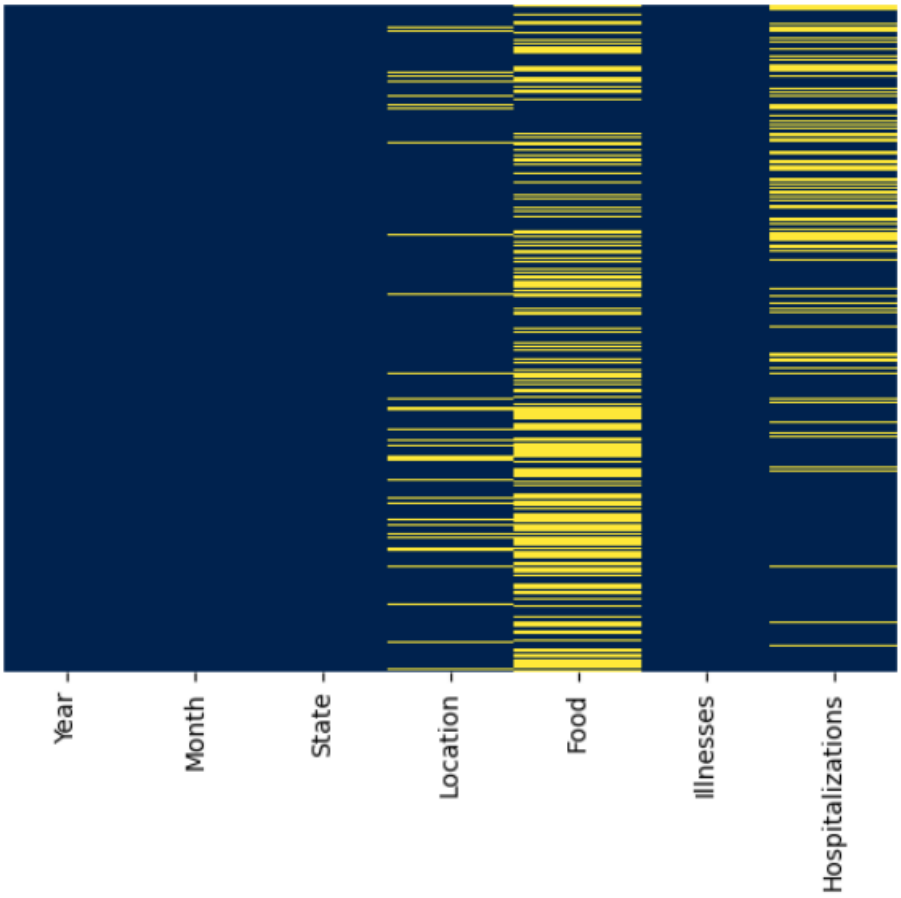
```
#Calculating no. of null values
df.isnull().sum()
```

```
Year                    0
Month                   0
State                   0
Location             2166
Food                 8963
Ingredient          17243
Species              6619
Serotype/Genotype   15212
Status               6619
Illnesses               0
Hospitalizations     3625
Fatalities           3601
dtype: int64
```

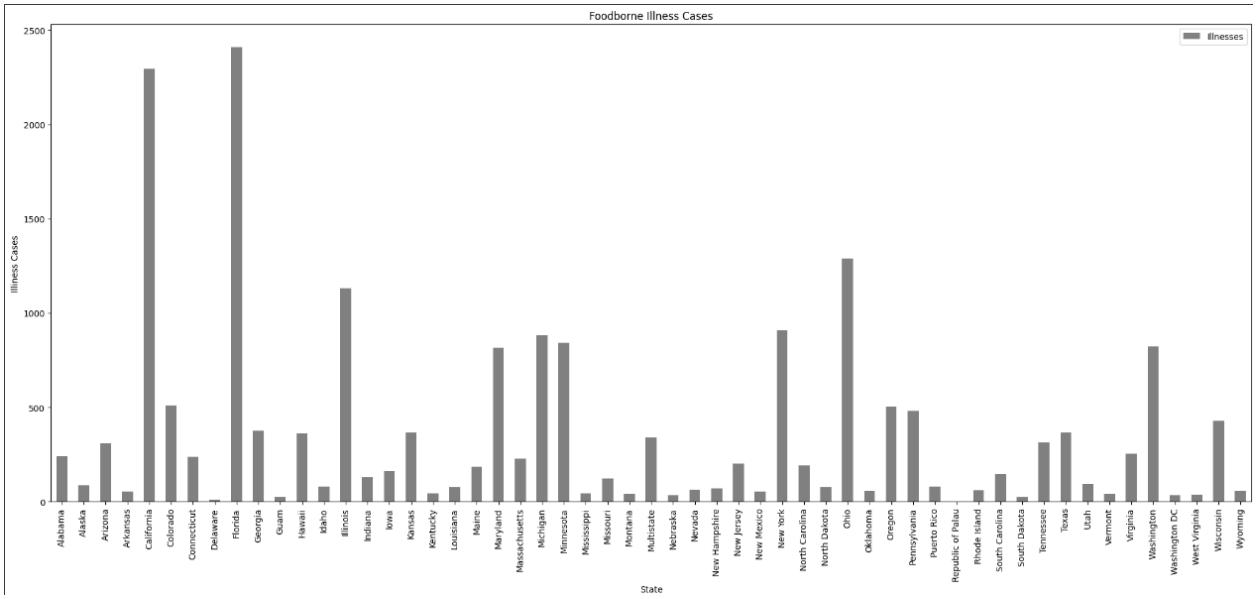
Following is a heatmap that shows the distribution of null values in the dataset. Yellow is represented as null values.



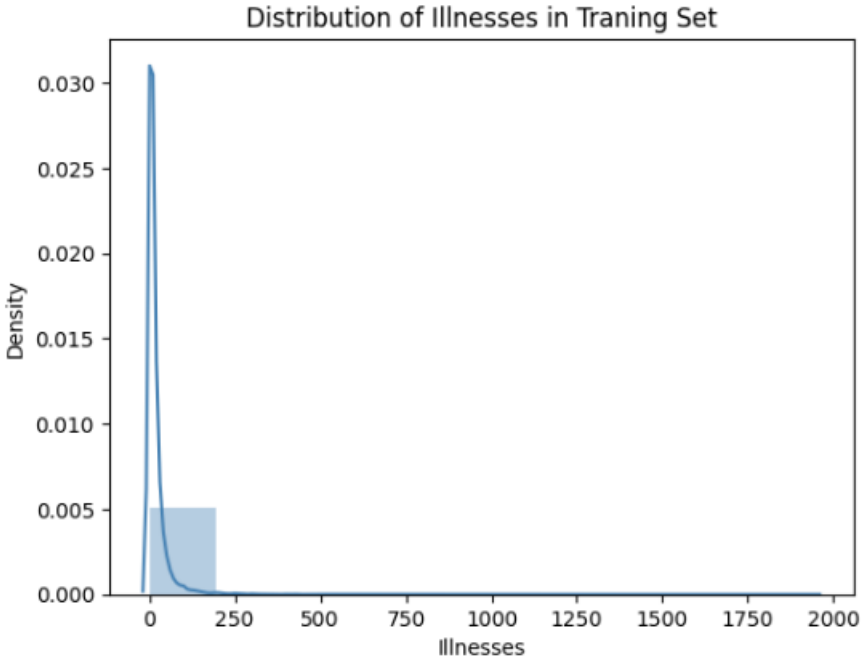
After dropping the columns we don't require we get a heatmap as follows.



Graph of illness per state was plotted to see their distribution in a particular state.

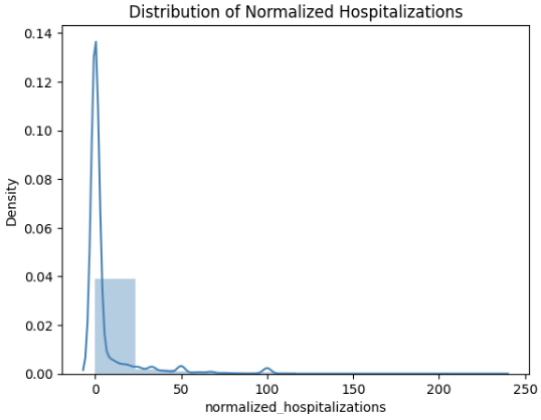


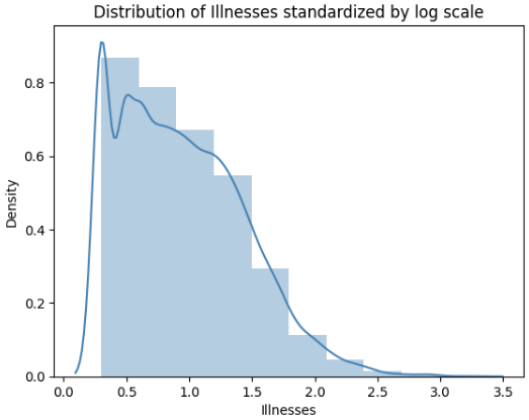
Plotting how Illnesses are distributed over the dataset.



Data preprocessing was applied to fill in missing values.

The distribution of illness was plotted and after analysis the distribution was skewed so it was transformed using log scale to get standardized results for our hypothesis.



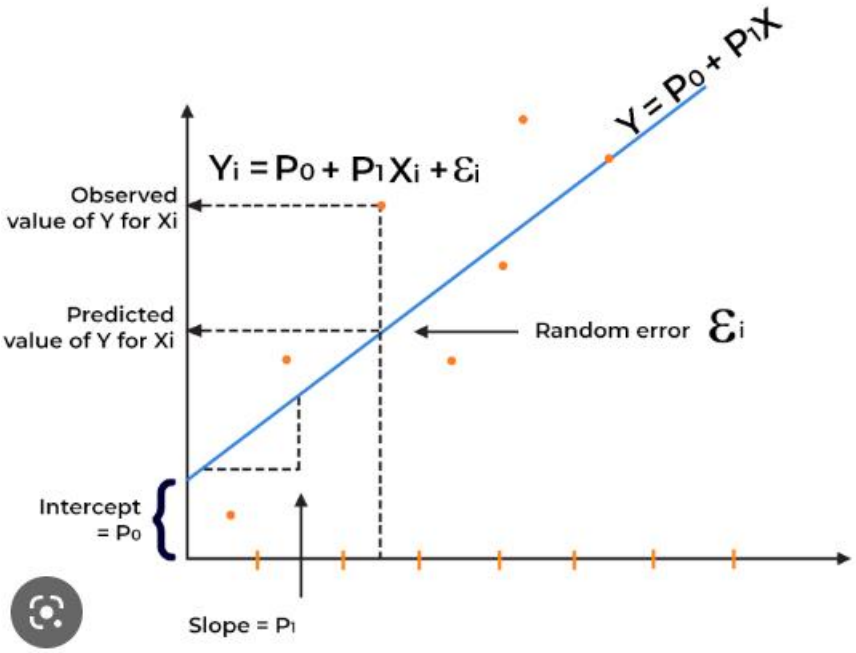


CHAPTER 4: DATA MODELLING

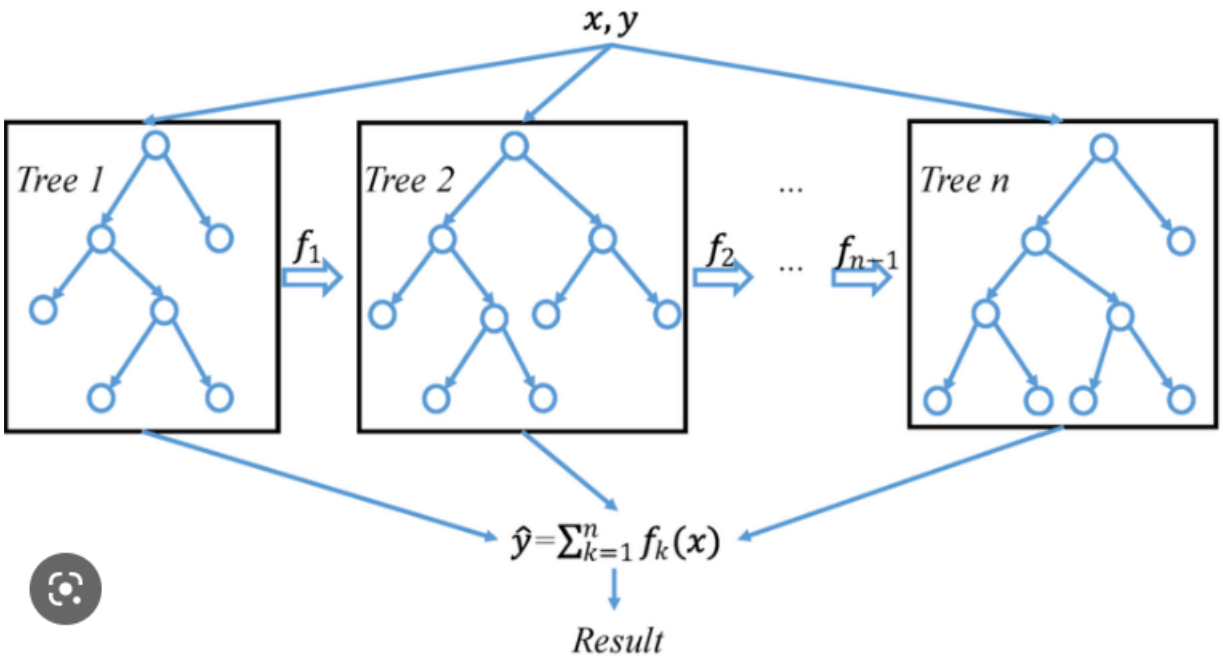
As the features are continuous in nature, therefore a regression type model is chosen to predict the value of target variable. Here we have taken different features into consideration and built two models – Linear Regressor and XGBoost Regressor on these attributes to test their performance.
Linear Regression is a simple and interpretable model that can provide insights into the relationships between the input features and the target variable.
XGBoost Regressor is a more complex and flexible model that can capture non-linear relationships between the input features and the target variable. It is based on gradient boosting, which combines multiple weak learners (decision trees) to improve the overall performance of the model. XGBoost Regressor is known for its high predictive accuracy.

ALGORITHM:

Linear Regression:



XGBoost Regressor:



We have taken different set of features for X (independent variable) to predict the value of y (dependent variable).

1. X variable consists of attribute Year. y consists of Illnesses_log as the dependent variable.

   The above set of data is fitted to a Linear Regression model.
   We get an **RMSE** value of **1.1389**

2. X variable consists of attributes State and Month. y consists of Illnesses_log as dependent variable.

   The above set of attributes are fitted to Linear Regression and XGBoost Regressor.
   RMSE for Linear Regression: **1.1312**
   RMSE for XGBoost Regressor : 1.0551

3. X variable consists of Location_modified, Food_Modified_new, State, Month. y variable consists of Illnesses_log.

   We have applied necessary preprocessing to Location and Food attributes.
   The Location and Food attributes contain many values for a single tuple.
   e.g : Child Daycare; Religious Facility

Applying necessary processing to split the multivalued tuples into single valued

The above set of attributes are fitted to Linear Regression and XGBoost Regressor.
RMSE value for Linear Regression : 1.1058
RMSE value for XGBoost Regressor : 0.9666

RESULT ANALYSIS:

Comparision between features taken with respect to model is shown below:

Features	Model	MAE	MSE	RMSE
X = Year Y = Illnesses_log	Linear Regression	0.9452	1.2972	1.1389
	XGBoost Regressor	0.9383	1.2938	1.1374
X = State, Month Y = Illnesses_log	Linear Regression	0.9344	1.2797	1.1312
	XGBoost Regressor	0.8463	1.1133	1.0551
X = State,Month, Location_modified,Food_modified_new Y = Illnesses_log	Linear Regression	0.9044	1.2227	1.1058
	XGBoost Regressor	0.7623	0.9344	0.9666

As we can see the RMSE values for XGBOOST Regressor model built on the third dataset is lowest of all. Therefore ,State,Month, Location_modified,Food_modified_new, attributes give better analysis for Illnesses.
We get a more generalized model with minimum error with the above set of attributes.

CHAPTER 5: CONCLUSION

Looking at the above Machine Learning model we can conclude that XGBoost Regressor gives a generalized model. The model has low Root Mean Squared Error.
This research findings could have several implications upon deployment. Consumers could benefit from being more informed about foods that are prone to causing foodborne illnesses and when they are most likely to occur. This knowledge would enable them to either avoid such foods or take extra precautions while consuming them.
However, we should note that our data only consists of reported instances of foodborne illnesses. In the case of fatalities resulting from foodborne illnesses, they would most likely have been detected during an autopsy.

Future scope is being able to take in all of the data about foodborne illnesses the model is likely to find some sort of correlation between factors that is directly leading to illnesses or hospitalizations and is only going to get stronger. This can be achieved by regularly gathering data from the CDC and using it to create test sets for the model. By feeding these test sets into the model's training set, the model can continue to learn and improve its accuracy over time, resulting in more robust and reliable predictions.

Dataset: outbreaks.csv

Colab Link :
https://colab.research.google.com/drive/1ANRS2H2zxa4mfjEYylOWl42vEmonHCgs#scrollTo=rgUI865XfGF5