Student Name: Aleena Kotadiya

ADTA 5340: Discovery and Learning with Big Data

**Final Project Report**

# PART I: A Strategy to Employ Machine Learning in a Firm

Uber Eats : Uber Eats is the easy way to get the food you love delivered.

**Size**

Uber Eats is one of the largest food delivery platforms globally, operating in **45+ countries** and partnering with **600,000+ restaurants** as of recent estimates.

**Sector**

Uber Eats operates in the **online food delivery and logistics sector**

**Competitors**

1. DoorDash
2. Grubhub
3. Just Eat Takeaway
4. Deliveroo
5. Meituan
6. Zomato/Swiggy

How does it work?

Uber Eats has hundreds of restaurants to choose from. When you open the app, you can scroll through When you find something you like, tap to add it to your order. When you’re ready to check out, you’ll see your address, an estimated delivery time, and the price of the order including tax and delivery fee. When everything looks right, just tap Place order—and that’s it. Follow your order in the app. First you’ll see the restaurant accept and start prepping. Then, when the order’s almost ready, a nearby delivery person—in a car, on a bike, or on a scooter—will go to the restaurant to pick it up. Next, they’ll drive or ride to you. You’ll be able to see their name and photo and track their progress on the map.

Strategy :

Improving Personalized customization

GOAL:

We are to tune our recommendations such that the user tries new food and likes it - and gives business to restaurants that are not performing well.

**Enhance user engagement, satisfaction, and revenue**

**Plan of Action**

We are going to partner with restaurants that are not performing well with a deal. The deal is to boost the

restaurant on our app and in turn we are going to take a percentage(20 – 30%) of their revenue generated

from our app. We are just not recommending slow restaurants. We are to analyze customers’ food

preferences and a restaurant’s menu and find similarities and then recommend restaurants to customers that

are likely to order from these restaurants.

**Implementation** :

Resources Required :

1. HUMAN CAPITAL :

We require domain experts such as

1. Marketers
2. Content makers
3. Data scientist
4. ML engineers
5. Food Bloggers
6. Human Resource
7. Financial Manager
8. TECHNOLOGY:
9. Big data platform : like Hadoop to store and process large amounts of data produced by people. Analyze the data to provide recommendations
10. Customer Support Integration
11. Real-Time Monitoring
12. Machine Learning
13. Cloud services

We will use both cutting-edge technology and human skills to enhance underperforming restaurants on Uber Eats. While data scientists and machine learning experts provide tailored recommendations for clients based on their interests, marketers and content producers will spearhead focused campaigns. Machine learning models will customize meal recommendations, and big data platforms like Hadoop will evaluate customer and restaurant data to improve underperforming eateries. Cloud services and real-time monitoring will track performance and refine recommendations, guaranteeing ongoing enhancements to consumer engagement and restaurant profitability

REFERENCES:

[How Does Uber Eats Work? | About Uber Eats](https://about.ubereats.com/)

# PART II: Data Preprocessing

**DATASET LINK** : [Life Expectancy (WHO)](https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who)

**Data Set Introduction**: This data set has 2938 rows and 22 columns which are as follows:

1. Country
2. Year
3. Status
4. Life expectancy
5. Adult Mortality
6. infant deaths
7. Alcohol
8. percentage expenditure
9. Hepatitis B
10. Measles
11. BMI
12. under-five deaths
13. Polio
14. Total expenditure
15. Diphtheria
16. HIV/AIDS
17. GDP
18. Population
19. Thinness 1-19 years
20. Thinness 5-9 years
21. Income composition of resources
22. Schooling

**Data types** : The columns “Country” and “ Status” are string and everything else is of numeric data type (int and float)

**Data Collection** : The data was collected from WHO and United Nations website with the help of Deeksha Russell and Duan Wang

**Why was this data collected?**

This data gives information about various parameters like : Adult Mortality, infant deaths, Alcohol percentage, expenditure , Hepatitis B, Measles , BMI ,under-five deaths ,Polio, Total expenditure , Diphtheria, HIV/AIDS ,GDP ,Population , thinness 1-19 years , thinness 5-9 years ,Income composition of resources ,Schooling

And impact of these parameters on life expectancy of a country through the years from 2000 – 2015

**DATA PREPOCESSING:**

1. **Columns that cannot have “0” as a value**

1. Country : Countries are identified by names or codes, not numerical values like 0

2. Year : Years represent time, and a value of 0 is not valid in modern datasets

3. Status : This categorical variable (e.g., "Developed" or "Developing") cannot logically be 0.

4. Life Expectancy : It reflects human lifespan and cannot reasonably be 0

5. BMI : A value of 0 would indicate a complete absence of body mass, which is biologically impossible.

6. GDP : Gross Domestic Product measures economic output and cannot be 0 unless a country has no economy which is not in our case

7. Population : A country cannot have a population of 0 unless it is uninhabited.

8. Expenditure : Indicates spending or investment and cannot be 0 in any operational system.

If any of these columns have “0” as a value we replace them with “NAN”

1. **Finding out Nan values in the dataset**
2. Country :0
3. Year :0
4. Status :0
5. Life expectancy:10
6. Adult Mortality :10
7. infant deaths : 0
8. Alcohol : 194
9. percentage expenditure :0
10. Hepatitis B : 553
11. Measles : 0
12. BMI : 34
13. under-five deaths : 0
14. Polio : 19
15. Total expenditure : 226
16. Diphtheria : 19
17. HIV/AIDS : 0
18. GDP : 448
19. Population :652
20. thinness 1-19 years : 34
21. thinness 5-9 years : 34
22. Income composition of resources : 167
23. Schooling :163
24. **Finding the percentage of Nan values**

A screenshot of a computer screen

Description automatically generated

1. **Removing rows where columns have less percentage of missing data**

Here columns like 'Life Expectancy', 'Adult Mortality' ,'BMI', 'thinness 1-19 years' ,'thinness 5-19 years', 'Polio' and 'Diphtheria' have a very small percentage of missing data for some equivalent to zero.

For these we are going to drop the rows with NAN Values

As the percentage of missing data is very low dropping these records will not affect the variability in the data

1. **DATA Imputation**

So, the Columns with missing values

1. Alcohol - float64

2. Hepatitis B - float 64

3. Total Expenditure - float64

4. GDP- float65

5. Population -Floadt64

6. Income composition of resources - float64

7. Schooling - float64

For these columns a large proportion of data is missing, dropping these rows can loss to loss of critical data.

Since all these columns have numeric data type – we impute missing values with mean of the columns.

1. **Machine learning project that can be done with the dataset**.

The dataset can be used to predict **Life Expectancy** for a country based on the following predictors:

Status, Adult Mortality, infant deaths, Alcohol percentage, expenditure , Hepatitis B, Measles , BMI ,under-five deaths ,Polio, Total expenditure , Diphtheria, HIV/AIDS ,GDP ,Population , thinness 1-19 years , thinness 5-9 years ,Income composition of resources ,Schooling

We can even identify which predictors impacts life expectancy the most and which doesn’t have any effect.

# PART III: …

## EDA: Data Discovery

The dataset has 4177 rows and 9 columns. It has “Sex” column as string rest other columns are numeric type

1. Histogram

Length and Diameter are close to a normal distribution with a slight skew towards the left. Whole weight, Shucked weight .viscera weight, shell weight and age are positively skewed. Height has values between 0 and 0.25 it is negatively skewed.

1. Density Plot:

As mentioned above the height is between a small range therefore the spread is very tight. Whole weight and shell weight appear to be bi modal

1. Box Plot:

As we know from above, we not have a perfect normal distribution in our dataset

Therefore, there are outliers in each column: Length and diameter have outliers below their lower whisker. Height and age have outliers on either side of the whiskers. Whole weight, Shucked weight, viscera weight and shell weight have outliers above the upper whisker.

1. Scatter Matrix:

We see many perfect linear relationships and almost perfect linear relationships.

Perfect linear relationships :Length and Height, Length and Diameter, Diameter and height, Whole weight and shell weight, whole weight and shucked weight, whole weight and viscera weight, whole weight and height, shucked weight and viscera weight, shucked weight and height, height and viscera weight

We also have semi-linear relationship and Height has linear relationship with all variables

## EDA: Pre-Processing Data

1. First, we look at the null values of the dataset – we don’t have any null values
2. Then we look at the data types of each column : Which were all correct
3. Since we want to find the ‘Age’ : which we can get by adding 1.5 to Rings column so we add ‘1.5’ to rings column and create a new column name Age and drop Rings column

NOTE: In further analysis, I decided to go with linear regression, therefore I have to map Categorical data of ‘Sex’ column to integers which I did using cat.codes and converted the column to integer data type.

## Model Planning

I decided to go with linear regression because of the following reasons:

1. We want to predict ‘Age’ which is a continuous numeric variable
2. After looking at the data types of all the columns all the columns except SEX were of numeric datatype which is ideal for linear regression
3. After looking at the scatter plot matrix we can see there are linear relationship between the variables. Therefore, linear regression is the best model to work with in this case

## Model Building

Before Splitting the Data set into Input and Output we need to convert the values of ‘Sex’ column from categorical to integer

I used cat.codes to do that.

## Reporting Results

Now split the data into X and Y. X contains all columns except ‘age’ and Y contains ‘age’.

We do the train/test split where test size is 0.33 and build the model

We will use LinearRegression() to build our model and fit on model to X\_train and Y\_train.

R SQUARE:

We get R square value as 0.5223858580716234 which means 52.238 % variability in data is explained by this model

We created 2 new records to predict the age :

1. Sex :Male(2)

2. Length: mm: 0.44

3. Diameter: mm : 0.3

4. Height : mm : 0.15

5. Whole weight : grams : 0.406

6. Shucked weight : grams : 0.194

7. Viscera weight : grams : 0.151

8. Shell weight : grams : 0.12

Age predicted was 8.881

1. Sex :Female(0)

2. Length: mm: 0.45

3. Diameter: mm : 0.38

4. Height : mm : 0.11

5. Whole weight : grams : 0.768

6. Shucked weight : grams : 0.08

7. Viscera weight : grams : 0.21

8. Shell weight : grams : 0.1

Age predicted was 14.49

Based on the R square which is 52.22% the predicted values of 2 new records are moderately accurate but not precise because the model explains 52.22% variance in the target variable based on the predictors.

**K-FOLD CROSS VALIDATION**

By using negative mean square, we get -5.0644.

The average error between the predicted and actual values (in terms of squared units of the target variable) is **5.0644**.

The root of which is 2.25 that means the predicted value will deviate from the actual value by + or – 2.25. Which in our case is moderately good since abalone’s highest age is 31.5 and the range falls till there.

Since the range of age is not too big an error of + or – 2.25 might be significant.

# PART IV: …

## EDA: Data Discovery

The dataset has 48842 rows and 15 columns.

We drop column ‘Fnlwgt’ because it was used for the purpose of doing the survey.

1. Histogram

Age, capital\_gain and capital\_loss are positively skewed whereas education\_num and weekly\_hours are close to normal with a very high peak

1. Density

The spread or variation for capital\_gain,capital\_loss and weekly\_hours is very small and age has a good spread. In education\_num we see 3 peaks making it bimodal

1. Boxplot

Age, capital\_gain and capital\_loss have outliers above the upper whisker, weekly\_hours has outliers on either sides of the whiskers. Education\_num has very few outliers below the lower whisker.

1. Scatter plot matrix:

There is no clear or linear relationship among variables

## EDA: Pre-Processing Data

1. Frist thing to do is drop “Fnlwgt” because it is only used for the purpose of the survey
2. For columns ‘Emp\_type”, “Occupation” and “Country” we have ‘?’ which are basically ‘Nan’ values
3. I used excel to “Find and Replace” “?” with Nan value
4. For “Income” column we have values “<=50K” , “<=50K.” , “>50K”, “>50K.”

“<=50K” , “<=50K.” imply the same thing and “>50K”, “>50K.” imply the same value

1. We again make it uniform by using excel and replacing “<=50K.” with “<=50K” and replacing “>50K.” with “>50K”.
2. Now let’s look at the null values: (in percentage)

A screenshot of a computer

Description automatically generated

Since the percentage of data containing nan values is very less and considering the fact that we have 48842 row we can drop these rows with nan values

## Model Planning

I choose Random Forest

1. At first glance looking at the model we need to predict Income – which is categorical in this case and has 2 outcomes that is either >50k or <=50 k . I thought to use logistic regression
2. Now looking at the dataset more deeply :
3. The dataset is quite large it has around 48 thousand rows
4. It is a very imbalanced dataset (target classes are imbalanced ) and values for most columns are skewed,
5. Looking at the scatter matrix we can see there is **No strong linear relationships**
6. I decided to go with Random Forest for the following reason
7. Large or Complex Data : We have a very large data set with 48 thousand rows
8. Nonlinear Relationships : We don’t see any strong linear relationship among variables
9. Robustness to Noise and Outliers : There are lots of outliers and the data is imbalances (target classes are imbalanced .( <=50K is 77% whereas >50 k is 24%)
10. High Dimensionality : There are 15 columns
11. Ensemble Power: combines predictions from multiple decision trees for higher accuracy, random forest is ideal.

## Model Building

Before Splitting the Data set into Input and Output we need to convert the columns from categorical to integer

I used cat.codes to do that.

## Reporting Results

Now, split the data into X and Y. X contains all columns except ‘Income’ and Y contains ‘Income’.

We do the train/test split where test size is 0.33 and build the model

We will use RandomForestClassifier() to build our model and fit on model to X\_train and Y\_train.

Let’s Review the classification Report:

Class 0 (≤50K):

**Precision (0.88):**

88% of the instances predicted as class 0 were actually class 0. This shows that when the model predicts a person earning ≤50K, it’s mostly correct.

**Recall (0.90):**

90% of all the actual class 0 instances were correctly predicted by the model. This indicates that the model is good at identifying most instances of class 0.

**F1-Score (0.89):**

A balanced measure of precision and recall, showing overall good performance for class 0.

Class 1 (>50K):

**Precision (0.69):**

69% of the instances predicted as class 1 were actually class 1. This is relatively low compared to class 0, meaning the model has some difficulty in correctly predicting high-income earners.

**Recall (0.64):**

64% of all actual class 1 instances were identified by the model, which is even lower. This shows that the model is missing many true positives (high-income individuals), making more false negative errors.

**F1-Score (0.66):**

A lower F1-score for class 1, indicates the model's performance is weaker for this class.

**ACCURACY** is 84% which is a very good score

2 NEW RECORDS:

We are going to predict Income for 2 new records

1. AGE : 38

2. Emp\_type :State-gov(5)

3. Educatiom :Bachelors(9)

4. Eduction\_number :13

5. Marital:Divorced (0)

6. Occupation : Other-service (3)

7. Relationship: Not-in-family (1)

8. Race: white (4)

9. Sex:Male (1)

10. Capital\_gain:0

11. Capital\_loss: 0

12. weekly hours :40

13. Country: usa (38)

PREDECTION : 0

We are going to predict Income for 2 new records

1. AGE : 40

2. Emp\_type :State-gov(5)

3. Education: HS-grad(11)

4. Eduction\_number :9

5. Marital :Divorced (0)

6. Occupation : Other-service (3)

7. Relationship: Not-in-family (1)

8. Race: Black (2)

9. Sex:Male (1)

10. Capital\_gain:0

11. Capital\_loss: 0

12. weekly hours : 50

13. Country: cuba (4)

PREDICTION :0

With the accuracy of 84% the model performs well

K -Fold Cross Validation

Cross-validation scores for each fold: [0.84422333 0.83946932 0.84918178 0.84144184 0.83790358]

Average accuracy across all folds: 0.8424439696451156

Accuracy is similar to our classifiers accuracy that mean the model’s performs is good and the results are reliant.

# PART V: …

## EDA: Data Discovery

The dataset has 1728 rows and 7 columns.

1. Histogram

We see uniform charts for all columns except for ‘Evaluation’. It is skewed to the left. Moreover, we can see in ‘Passengers’ column there is no value for ‘3’.

1. Density : We can again see uniformity in density too except for passengers and Evaluation. In evaluation we see 2 major peaks, one more significant than other.
2. Box plot: There are no outliers present in the dataset
3. Scatter plot : We don’t see any significant relationships among variables

## EDA: Pre-Processing Data

1. There are no null values in the dataset
2. For numerical columns ‘Doors’ and ‘Passengers’ we see the data type listed as object – which is incorrect

‘Doors’ column has a value 5more which mean 5 or more doors with "5" which would mean that the car has 5 or more doors

We are replacing 5more with 5

‘Passengers’ column has "more" which mean 5 or more passengers with "5" which would mean that the car has 5 or more passengers

We are replacing more with 5

Finally convert the column data type to integer

## Model Planning

Since we have to use an unsupervised model for this part, I went with KMeans

## Model Building

We convert all Object data types to integer so that the model can process it.

We converted the Dataframe to an array and split the data into X and Y

X has all columns except ‘Evaluation’ and Y has only ‘Evaluation’

## Reporting Results

I went with 4 clusters since we want to predict if the car is classified as unacceptable, acceptable,  
good, or very good.

In the scatter plot we see the centroids of all 4 clusters and can see that 3 clusters have major points.

1. price: high

2. maintenance : high

3. Doors: 2

4. Passenger: 4

5. luggage :small

6. safety: low

For the first record model predicts [0] cluster

1. price: vhigh

2. maintenance : high

3. Doors: 4

4. Passenger: 4

5. luggage :small

6. safety: low

For second record it predicts [0] cluster

Both belong to ‘acceptable’ cluster

# PART VI: …

I Regression Models: Linear Regression vs. Decision Tree (CART) Regression

I selected abalone.csv for this part

The dataset has 4177 rows and 9 columns. It has “Sex” column as string rest other columns are numeric type

## EDA: Pre-Processing Data

1. First, we look at the null values of the dataset – we don’t have any null values
2. Then we look at the data types of each column : Which were all correct
3. Since we want to find the ‘Age’ : which we can get by adding 1.5 to Rings column so we add ‘1.5’ to rings column and create a new column name Age and drop Rings column

NOTE: In further analysis, I decided to go with linear regression, therefore I have to map Categorical data of ‘Sex’ column to integers which I did using cat.codes and converted the column to integer data type.

**Model Quality (R2)**

1. R2 of Linear Regression : 0.52238
2. R2 of Decision Tree (CART) Regression : 0.14985

R2 of Linear Regression implies that 52.238 % of variability in ‘Sex’ column or target column is explained by the model. R2 of Decision Tree (CART) Regression implies that 14.958% of variability in ‘Sex’ column or target column is explained by the model.

Here Liner Regression performs better than Decision Tree (CART) Regression because there is significant linear relationships among the variables and Decision Tree (CART) Regression is quite complex. Decision Tree (CART) Regression is useful when we have non-linear or complex relationships but, in our case, we have linear relationships among variables.

**Prediction**

1. 1. Sex :Male(2)

2. Length: mm: 0.44

3. Diameter: mm : 0.3

4. Height : mm : 0.15

5. Whole weight : grams : 0.406

6. Shucked weight : grams : 0.194

7. Viscera weight : grams : 0.151

8. Shell weight : grams : 0.12

Linear Regression : 8.88

Decision Tree (CART) Regression : 8.5

1. 1. Sex :Female(0)

2. Length: mm: 0.45

3. Diameter: mm : 0.38

4. Height : mm : 0.11

5. Whole weight : grams : 0.768

6. Shucked weight : grams : 0.08

7. Viscera weight : grams : 0.21

8. Shell weight : grams : 0.1

Linear Regression : 14.49

Decision Tree (CART) Regression : 10.5

For the first record the predicted value from both models are very close whereas for the second

you can see large discrepancy in the output this is because Decision Tree (CART) Regression is

performing very poorly compared to Linear Regression. Linear Regression predicting powers are also

moderate but still better than Decision Tree (CART) Regression because of their R2 values

**K-Fold Cross-Validation**

1. Linear Regression :

By using negative mean square, we get -5.0644.

The average error between the predicted and actual values (in terms of squared units of the target variable) is 5.0644.

The root of which is 2.25 that means the predicted value will deviate from the actual value by + or – 2.25. Which in our case is moderately good since abalone’s highest age is 31.5 and the range falls till there.

Since the range of age is not too big an error of + or – 2.25 might be significant

1. Decision Tree (CART) Regression

By using negative mean square, we get - 8.90668.

The average error between the predicted and actual values (in terms of squared units of the target variable) is 8.90668.

The root of which is 2.98that means the predicted value will deviate from the actual value by + or – 2.98. Which in our case is moderately good since abalone’s highest age is 31.5 and the range falls till there.

Since the range of age is not too big an error of + or – 2.98might be significant

**Comparison:**

Linear Regression’s RMSE (2.25) is smaller than Decision Tree’s RMSE (2.98), indicating that its predictions are closer to the actual values on average.

Linear Regression’s outperforms Decision Tree in both R sq and MSE.

The actual values for new records (+-2.5)

Record 1. Between range 6.63 and 11.13

Record 2 . Between range 12.24 to 16.74

Linear Regression is the better model as it has smaller errors and explains more variation in the target variable.

**II Classification Models: Logistic Regression vs. K-Nearest Neighbors**

## EDA: Pre-Processing Data

The dataset I selected was weather\_forecast\_data. It has 2500 rows and 6 columns.

All columns are of int data type except for our target variable ‘Rain’

1. First, we look at the null values of the dataset – we don’t have any null values

There is no Pre-processing required.

**Model Quality(Accuracy Level)**

Logistic Regression : 92.485%

K nearest neighbor : 94.78%

Logistic Regression classifies correctly 92.485% of the times whereas KNN does it 94.78%.

Both model performs very well, KNN slightly outperforms Logistic regression by 2.2995%. Therefore, predictions by KNN are more reliable.

**Prediction**

New Record 1:

:

1. Temperature: 22.5

2. Humidity: 78.3

3. Wind\_Speed: 6.4

4. Cloud\_Cover: 55.1

5. Pressure: 1015.2

Logistic Regression : No rain’

K nearest neighbor : No rain’

New Record 2:

1. Temperature: 30.2

2. Humidity: 50.1

3. Wind\_Speed: 10.7

4. Cloud\_Cover: 20.3

5. Pressure: 1012.7

Logistic Regression : No rain’

K nearest neighbor : No rain’

Since both model perform very well we can see that the predictions made by both models are same.

The predictions made by KNN are more reliable than Logistic regression based on accuracy level.

**K-Fold Cross-Validation**

Logistic Regression : 0.930 or 93%

K nearest neighbor : 0.958 or 95.8%

KNN performs better at capturing patterns in the dataset than logistic regression. The accuracy of the **Logistic Regression** model is **93%**, while the accuracy of the **K-Nearest Neighbor (KNN)** model is **95.8%**

Comparison:

Logistic regression misclassifies 7.515%( Error: 100%−92.485%=7.515%) and

KNN misclassifies 5.22% ( Error: 100%−94.78%=5.22%)

KNN has a lower error rate making it more accurate than logistic Regression. Both models make the same predictions for new data values.

KNN should be chosen to make predictions on the model since it has higher accuracy and lower error rate.