ACCIDENT SEVERITY

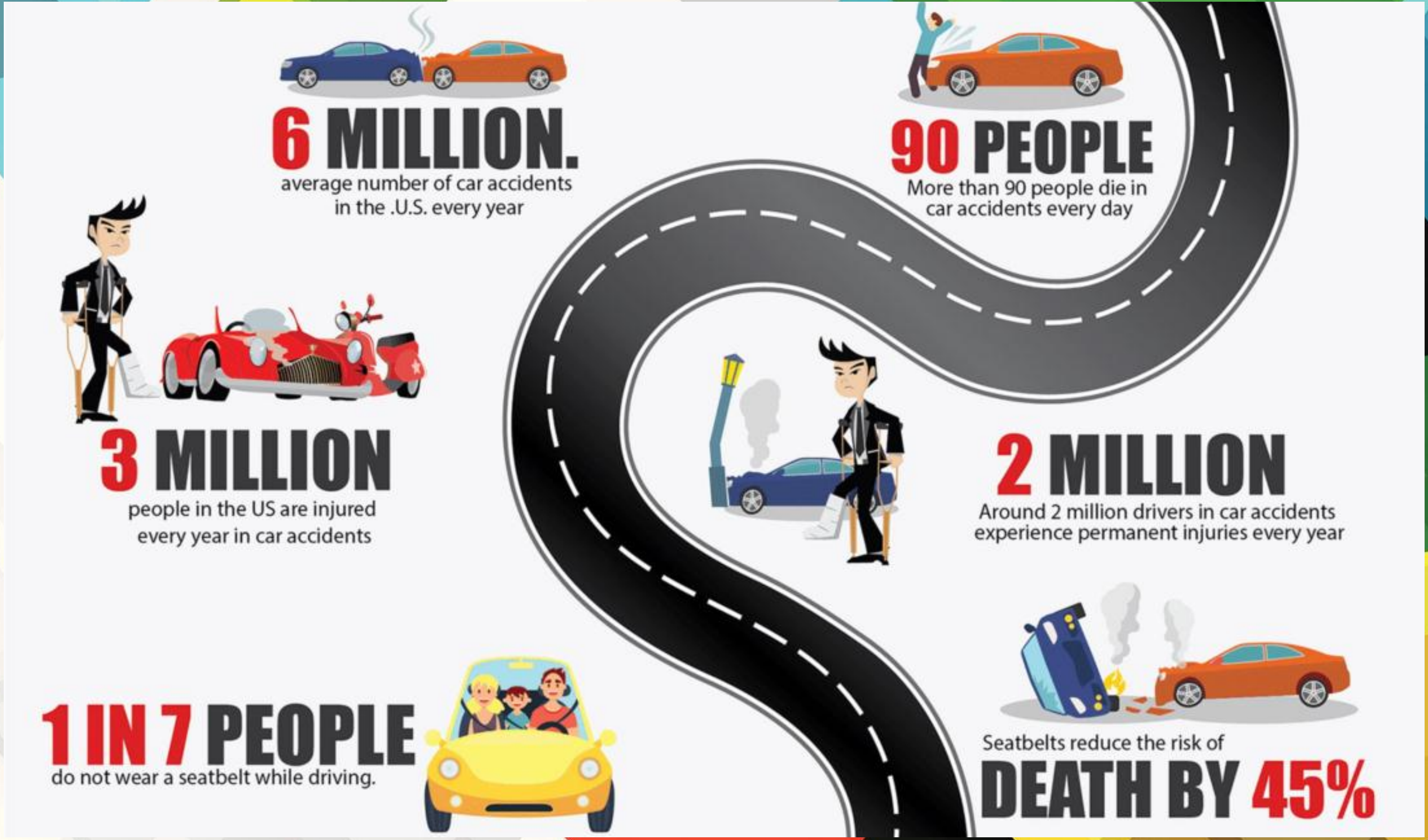
CAPSTONE FINAL PROJECT

CONTENT

1. INTRODUCTION
2. DATA
3. METHODOLOGY
4. RESULT AND DISSCUSSION

INTRODUCTION

1. INTRODUCTION



6 MILLION.
average number of car accidents in the .U.S. every year

90 PEOPLE
More than 90 people die in car accidents every day

3 MILLION
people in the US are injured every year in car accidents

2 MILLION
Around 2 million drivers in car accidents experience permanent injuries every year

1 IN 7 PEOPLE
do not wear a seatbelt while driving.

Seatbelts reduce the risk of
DEATH BY 45%

1. INTRODUCTION

How to Improve Road Safety?

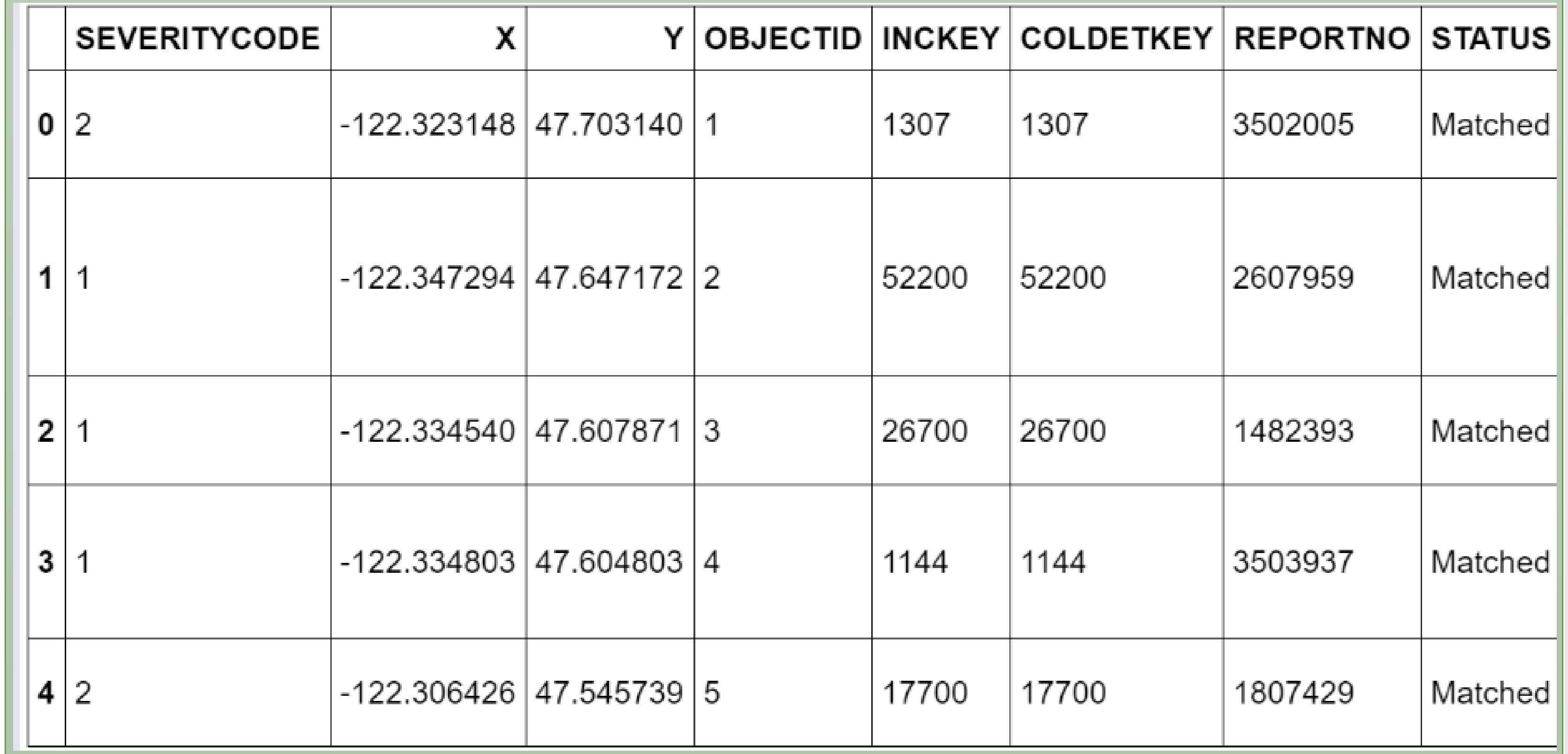
- MACHINE LEARNING TO PREDICT THE LIKELIHOOD OF SEVERE TRAFFIC ACCIDENTS, THEREBY WARNING DRIVERS OF THE DANGERS
- PREDICT SEVERITY OF ACCIDENTS COULD HELP MEDICAL FACILITIES PREPARE IN ADVANCE SO AS TO DECREASE FATALITIES

BETTER AWARENESS

FEWER FATALITIES

LESS WORK FOR POLICE

DATA



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched

DATA

1. Data Source

   Seattle Department of Transportation (SDOT). Updated weekly, from 2004 to present.

   Email: DOT_IT_GIS@seattle.gov

2. Metadata

   The raw dataset contains 38 columns and 194673 row. Except the first column being the label, all other 37 columns are features.

   Complete metadata: click here.

METHODOLOGY

Data Processing

Eliminating Bias

- Raw data contains far more instances of SEVERITYCODE 1 than of 2 (around 2.34:1)
- Uses dataframe.sample() method to sample from SEVERITYCODE==1 instances an amount equal to the number of SEVERITYCODE==2 instances

BALANCED DATA

BIAS ELIMINATED

BETTER TRAINING

EXPLORATORY DATA ANALYSIS

Which features affect the SEVERITYCODE?

- Dataframe.groupby(feature_name)[].value_counts() is used on each column to determine the ones correlated with accident severity
- Converts INCDATE to data objects and then to day of the week, but finds no correlation with SEVERITYCODE

WHICH FEATURES

WHY THESE FEATURES

TO BE ONE-HOT ENCODED

ONE HOT ENCODING

How could categorical features be used to train the model?

- Dataframe(feature_name).replace() was used on each feature to convert categorical variables into numerical ones
- Test the post-processing dataset with dataframe.dtypes to double check

NUMERICAL VALUES

READY FOR TRAINING

DECREASED COMPLEXICTY

Feature Selection And Normalization

How could features on different scales be used without bias?

- ★ Selects 14 features from dataset, including weather, road condition, lighting, etc.
- ★ Uses dataframe.dropna() to drop rows of the feature set with NaN values and preprocessing.StandardScalar().fit().transform() to normalize the feature set.

NO EMPTY CELLS

WITHOUT BIAS

READY FOR TRAINING

Model Training and Testing

How to train the ML models with existing data and test them?

- Uses the train_test_split() method to split the datasets into X_train, y_train, X_test, y_test.
- Imports four ML classification models (KNN, Decision Tree, SVM, and Logistics Regressioin, trains them with X_train and Y_train, and tests them with X_test and y_test to obtain their performance.

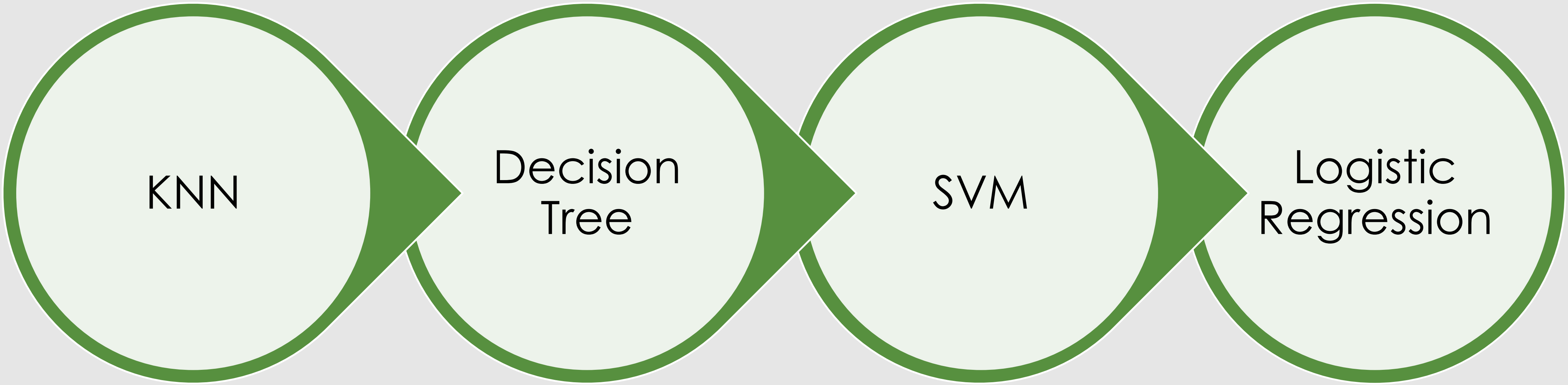
MODELS TRAINED

MODELS TESTED

PERFORMANCE OBTAINED

RESULT AND DISSCUSSION

Results and Discussion



KNN

Decision Tree

SVM

Logistic Regression

Accuracy Score: 0.694

K=25, consumes the most computing time

0.702

Max_depth=15, consumes little computing time

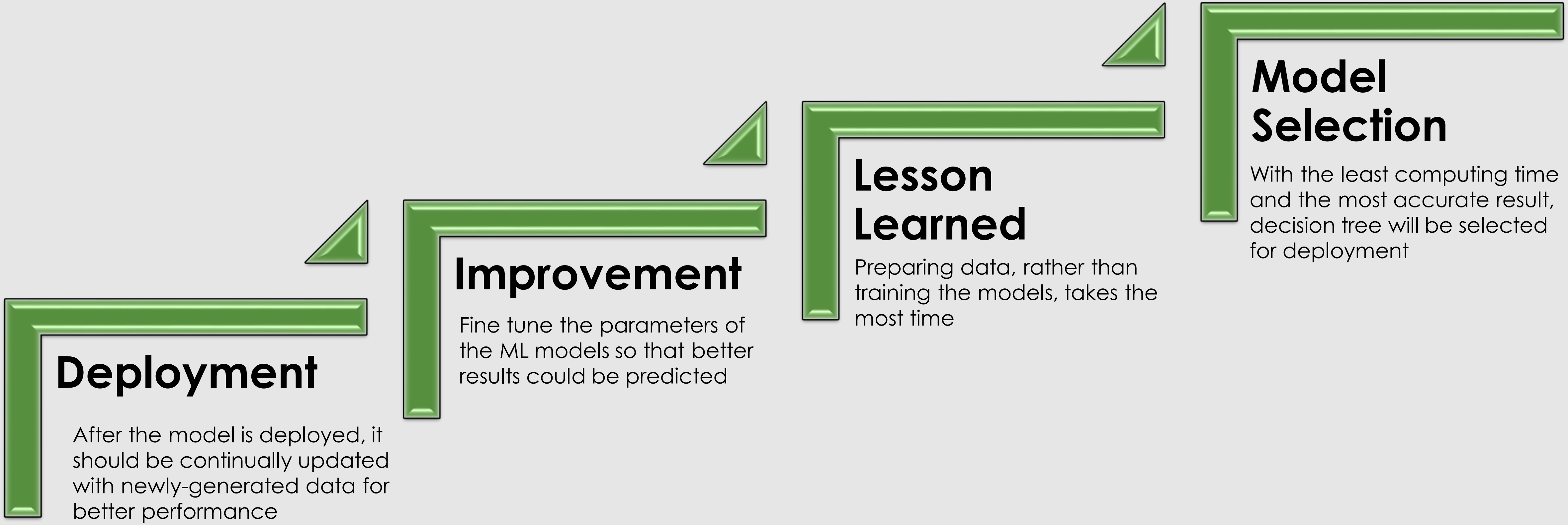
0.692

Kernel = ‘rbt’ , took much computing time

0.659

C = 0,1, took moderate computing time

Results and Discussion



Deployment

After the model is deployed, it should be continually updated with newly-generated data for better performance

Improvement

Fine tune the parameters of the ML models so that better results could be predicted

Lesson Learned

Preparing data, rather than training the models, takes the most time

Model Selection

With the least computing time and the most accurate result, decision tree will be selected for deployment