APPLIED DATA SCIENCE CAPSTONE PROJECT

ON

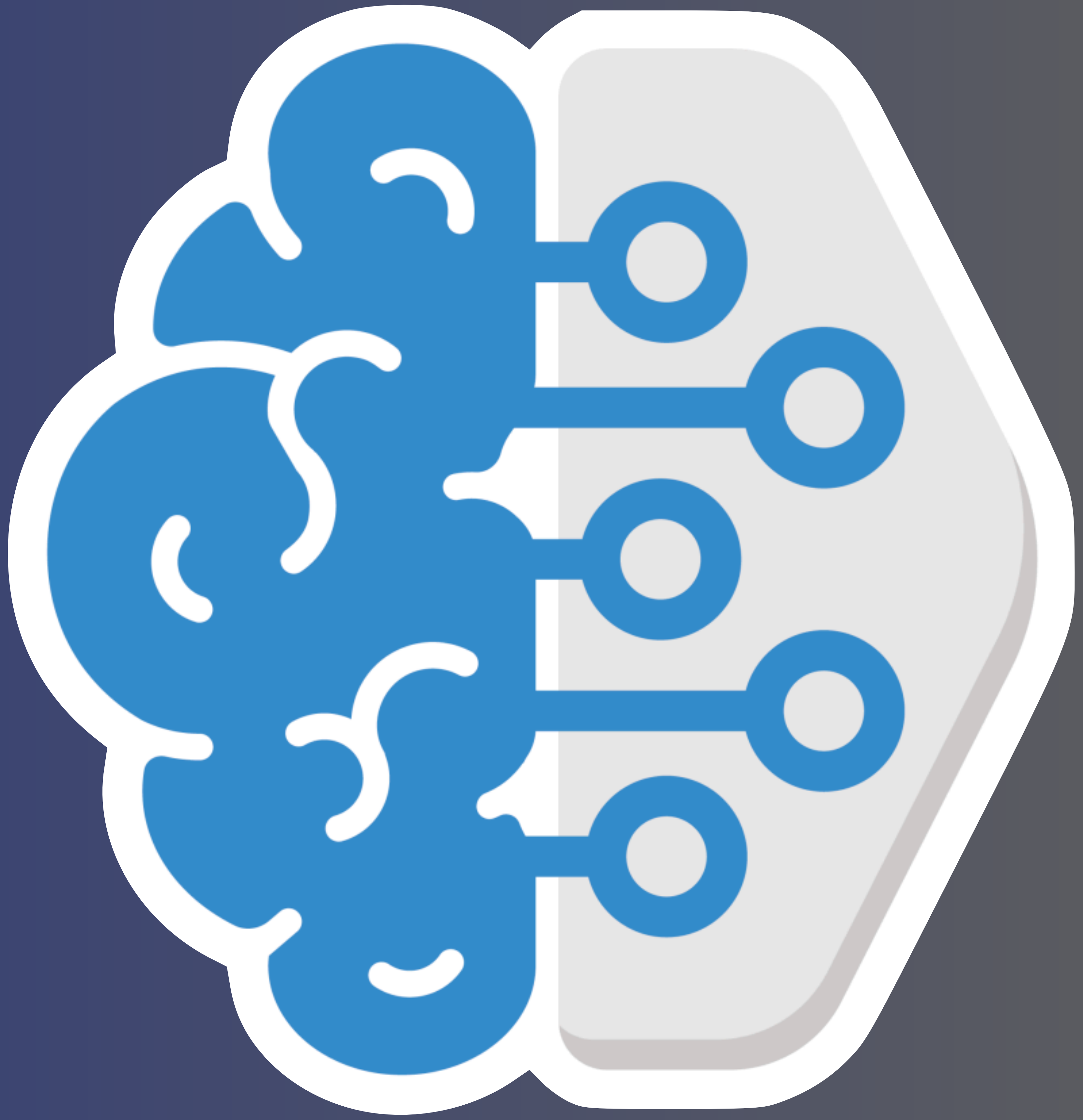
CAR ACCIDENT SEVERITY - PREDICTION USING MACHINE LEARNING ALGORITHMS

SUBMITTED BY

SANKET AMDALLI

CONTENTS

- Introduction
- Data Understanding & Preparation
- Methodology
- Results
- Discussion & Recommendations
- Conclusion



INTRODUCTION

INTRODUCTION

BACKGROUND

- Over 1.35M deaths and 20-50M injuries caused annually due to road accidents - World Health Organization.
- Primary Causes include speeding, driving under influence, non-use of helmets and seat-belts, distracted driving, unsafe vehicles, unsafe road infrastructure and inadequate post-crash care

PROBLEM

To analyse Accident data and determine the extent to which attributes like road conditions, weather conditions, light conditions, etc. impact road accidents, and consequently develop a model using the key factors to predict the severity of potential accidents.

INTRODUCTION

INTEREST

The primary consumers of our model (target audience) will be the government - road/transport authorities, who can:

- Build a mechanism to alert drivers about adverse driving conditions.
- Based on expected severity, inform police and medical authorities to prepare for quick response.
- Use insights from the model to better plan construction of new roads in the future.

For the scope of this project, the municipal authorities of Seattle, Washington are the target audience.

DATA UNDERSTANDING & PREPARATION

DATA UNDERSTANDING & PREPARATION

DATA SOURCE

Records of vehicle collisions in the city of Seattle

- Collected by Seattle Police Department, maintained by Seattle Department of Transport (SDoT).
- Has attributes such as the accident severity, location, number of people/vehicles involved, number of injuries/fatalities, whether a driver was speeding/driving under influence, road conditions, light/visibility conditions and weather conditions among others
- Data since 2004 (about 220k), updated weekly.

Data Home: https://dataseattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

DATA UNDERSTANDING & PREPARATION

TARGET VARIABLE

‘SEVERITYCODE’ is chosen as the target variable. The variable has coded values for different levels of severity of accidents.

- 0 — Unknown Severity
- 1 — No Injury, Only Property Damage
- 2 — Injury Collision
- 2b — Serious Injury Collision
- 3 — Fatality Collision

Removed the rows with code 0 (Unknown), as they are of no use to the model. Also combined codes 2, 2b and 3 to make target variable binary and have 2 categories - Non-Injury vs Injury type

DATA UNDERSTANDING & PREPARATION

EXPLORATORY DATA ANALYSIS

Descriptive statistics to understand the relationship between different features and with the target variable. The following relationships were explored:

- Severity of accident vs No. of people/vehicles involved
- Road, Weather and Light conditions vs No. of vehicles involved
- No. of accidents involving parked cars vs Severity
- Correlation Matrix between Severity and no. of people/cyclists/pedestrians/injuries

In addition, some relation-plots and box-plots were mapped to understand variable relationships, data distributions and outliers. Refer the Jupyter Notebook or the Report for charts and results.

DATA UNDERSTANDING & PREPARATION

FEATURE SELECTION

Since the problem statement is to predict the severity of accidents based on external causes, the three features chosen as input variables are:

- ‘ROADCOND’ – Road Conditions
- ‘WEATHER’ – Weather Conditions
- ‘LIGHTCOND’ – Light Conditions

Variables such as driving under influence, speeding, inattention, etc. have been excluded as these factors, albeit causes for accidents, are in control of the drivers, and municipal authorities cannot predict potential accidents based on these factors.

DATA UNDERSTANDING & PREPARATION

DATA CLEANING & PREPARATION

Necessary cleaning processes to prepare the dataset for analysis

- Handling Missing/Null values
- Converting the variables into the appropriate data type
- Removing outliers (if any)
- Balancing the dataset.

Rows having NaN values have been removed, or substituted with the mode values of the respective features. We see that the no. of records with SEVERITYCODE = 1 are twice that of SEVERITYCODE = 2, so we have made them equal to ensure there's no bias in our model.

DATA UNDERSTANDING & PREPARATION

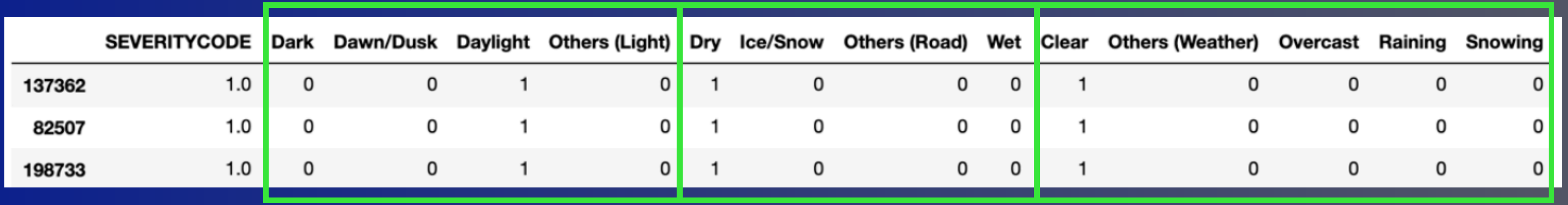
FEATURE ENCODING

Convert categorical input variables into binary variables.

There are plenty of categories for the three variables, these have to be reduced by grouping some categories based on similarity of conditions.

- LIGHTCOND has 9 categories, reduced to 4 - Daylight, Dark, Dawn/Dusk, and Others (Light)
- ROADCOND has 9 categories, reduced to 4 - Dry, Wet, Ice/Snow, and Others (Road)
- WEATHER has 12 categories, reduced to 5 - Clear, Overcast, Raining, Snowing, and Others (Weather)

The 13 total categories have then been converted to binary variables using One-Hot encoding technique. The final dataset looks something like this:



LIGHTCOND ROADCOND WEATHER

	SEVERITYCODE	Dark	Dawn/Dusk	Daylight	Others (Light)	Dry	Ice/Snow	Others (Road)	Wet	Clear	Others (Weather)	Overcast	Raining	Snowing
137362	1.0	0	0	1	0	1	0	0	0	1	0	0	0	0
82507	1.0	0	0	1	0	1	0	0	0	1	0	0	0	0
198733	1.0	0	0	1	0	1	0	0	0	1	0	0	0	0

METHODOLOGY

METHODOLOGY

The 3 models chosen for our analysis are:

- K-Nearest Neighbours (KNN) Classification
- Decision Tree Classification
- Logistic Regression

All three models are suitable for predicting binary variables.

Before deploying the models, the dataset needs to be split into a Training set (on which the models will learn) and a Test set (used to validate the accuracy). All the 3 models will use the same samples for Train and Test. I have used a 70-30 Train-Test split

METHODOLOGY

K-NEAREST NEIGHBOURS (KNN) CLASSIFICATION

The algorithm works on the principle that the closer two data points are, the more similar they'll be in nature.

Based on the current weather, road and light conditions, a data point will be generated which will have 13-dimensional coordinates and exist somewhere in that 13-dimensional space.

The kNN algorithm will compare it with 'k' closest data points (from our Train set) in the 13-dimensional space, and whatever is the most common severity code among those k data points, the same will be assigned to this data point.

METHODOLOGY

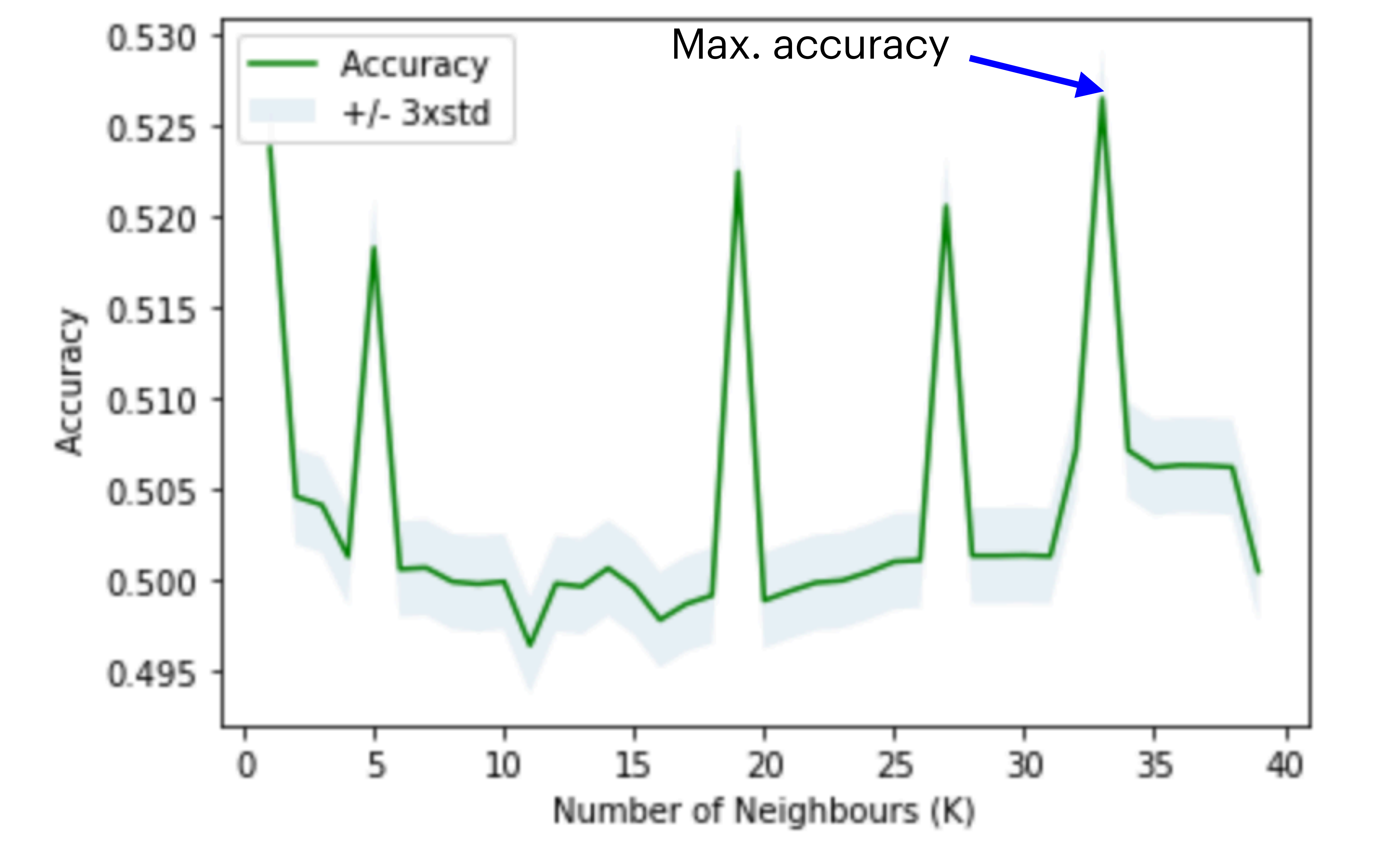
K-NEAREST NEIGHBOURS (KNN) CLASSIFICATION

The algorithm is initially run with k=6, then the value of k is varied to find the maximum accuracy that can be achieved.

For k=6, we get a Test accuracy value of 50.1%.

As k varies, the accuracy fluctuates marginally between 49.8% and 52.6%.

Maximum accuracy is 52.65% at k=33



METHODOLOGY

DECISION TREE CLASSIFICATION

The algorithm uses a tree-like model to show each decision and its expected outcome.

Every unique combination of the 13 variables is considered to be a unique decision, and assigned an expected value based on the frequency of its occurrences in the training set.

e.g. If a particular combination of road, weather and light conditions has 100 occurrences in our training set, of which 65 are severity 1, and 35 are severity 2, then the expected values of 1 and 2 will be 0.65 and 0.35 respectively.

The model then assigns the severity to the samples in the test set based on these expected values.

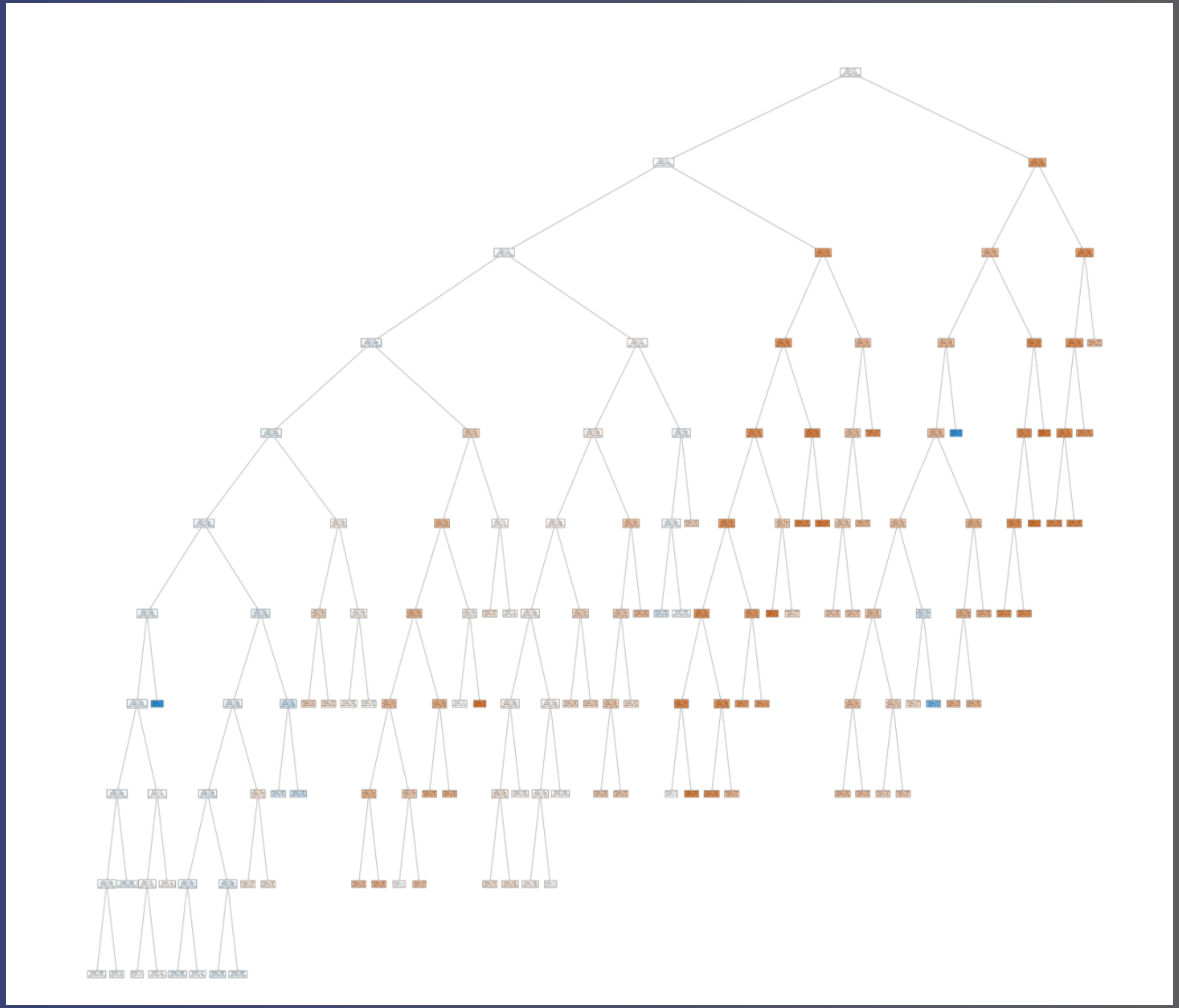
METHODOLOGY

DECISION TREE CLASSIFICATION

The model was implemented with a maximum depth of 13 because we have 13 variables.

The tree obtained is displayed on the right (click here for full image). It has 76 leaf nodes, indicating there are 76 unique combinations of road, weather and light conditions.

The model Test accuracy is 53.3%



METHODOLOGY

LOGISTIC REGRESSION

The algorithm uses logistic function to model binary dependent variables, and predict the probability of events having 2 outcomes like win/loss, pass/fail, alive/dead, yes/no.

Can be used as a classifier by deciding a threshold value, and assigning classes to data points by comparing their probability values with the threshold.

The model has been implemented keeping the regularization parameter c as 1, since all input variables are binary, and there’s no need for regularization.

METHODOLOGY

LOGISTIC REGRESSION

The model Test accuracy is 53.3%, and we get a decent logarithmic loss score of 0.684

The Confusion matrix is shown on the right.

The model is correctly able to predict 5,158 values for Severity 1 (28%), and 14,339 values for Severity 2 (78%) correctly.

		PREDICTED VALUES	
		1	**2**
ACTUAL VALUES	**1**	5,158	13,156
	2	3,916	14,339

METHODOLOGY

MODEL EVALUATION

In addition to the Test accuracy, I have also compared the models on two other metrics:

- Jaccard Similarity Index: measures the similarity between two sets of data. It can range from 0 to 1. The higher the number, the more similar the two sets of data.
- F1- Score: denoted as Harmonic mean of the precision and recall values calculated for the model. It can range from 0 to 1. The higher the number, higher is the precision and recall.

RESULTS

RESULTS

All three models show fairly similar accuracy – KNN (Max accuracy of 52.65% at K=33), Decision Tree (53.28%) and Logistic Regression (53.32%).

The highest Jaccard was observed for KNN, while F1 Scores were again fairly similar.

The results of the three models are summarized below:

	Test Accuracy	Jaccard	F1-Score	Log Loss
KNN	52.65%	0.40	0.49	NA
Decision Tree	53.28%	0.23	0.50	NA
Logistic Regression	53.32%	0.23	0.50	0.68

DISCUSSION & RECOMMENDATIONS

DISCUSSION & RECOMMENDATIONS

From the exploratory data analysis, it was clear that the target variable was well distributed and that there was little correlation between the input variables (Road, Weather & Light conditions) and the severity of accidents. Hence the accuracy values, Jaccard and F1 scores obtained for the three models are not surprising.

All three models are able to predict the severity of accidents with a >50% accuracy, meaning all of them can be used to classify accidents.

However, the Jaccard score for KNN Classification is significantly higher than the other two, and hence, I would recommend this model to be used for predicting the severity of accidents.

DISCUSSION & RECOMMENDATIONS

FUTURE SCOPE

More effective insights can be generated if additional data – The number of cars that are on the road during different weather, road and light conditions – is added to the analysis.

Currently from our dataset, since the proportion of accidents is the same in different conditions, we cannot predict whether the conditions are causing more severe accidents. Having data of the number of vehicles on the road in different conditions will give us an idea of what percentage of vehicles on the road get involved into an accident.

Rather than directly looking at the severity of accidents, we will be able to answer a more fundamental question about the likelihood of an accident happening if the conditions are adverse.

CONCLUSION

CONCLUSION

Our problem statement was to predict the severity of potential accidents based on the weather, road and light conditions.

The approach was to:

> *extract data from the mentioned source, identify target and input variables, and perform exploratory data analysis to understand relationships. Once this was done, perform the data cleaning processes and prepare the dataset to be fed into the models.*

I selected 3 models – KNN, Decision Tree and Logistic Regression to perform the classification.

> *The data was split into training and test sets, and the models were first trained on the training set, then evaluated on the test set. The metrics used for evaluation were Test accuracy, Jaccard Similarity Score, F1 Score and Logarithmic Loss Coefficient (for Logistic Regression).*

CONCLUSION

The results show that all three models have a fairly similar accuracy rate (52-53%), with KNN having a significantly higher Jaccard Score than the other two.

All three models can be used as classifiers for now, but KNN is recommended because of better Jaccard Score.

More effective classification can be performed in the future if additional data is added to the analysis.

The Jupyter Notebook, with the complete code, can be accessed here.

The full report can be accessed here.

THE END