 **Assignment Cover Sheet**

| **Subject Code: INFO411** |  |
| --- | --- |
| **Subject Name: Data Mining and Knowledge Discovery** | | |  |
| **Submission Type: Group report** | | |  |
| **Assignment Title: Report** | | |  |
| **Student Name: Mohamed Maarij Uddin Khan, Hassan Abdullah Ghaurii** | | |  |
| **Student Number: 6222560, 6882213** | | |  |
| **Student Phone : +971 50 6966480, +971 56 6346194** | | |  |
| **Student Email:** [**mmuk953@uowmail.edu.au**](mailto:mmuk953@uowmail.edu.au)**,** [**hag999@uowmail.edu.au**](mailto:hag999@uowmail.edu.au) | | |  |
| **Lecturer Name: Dr Soly Mathew.** | | |  |
| **Due Date: 5 / 12 / 2022** | | |  |
| **Date Submitted: 5/ 12 / 2022** | | |  |

| **PLAGIARISM:** The penalty for deliberate plagiarism is FAILURE in the subject. Plagiarism is cheating by using the written ideas or submitted work of someone else. UOWD has a strong policy against plagiarism.  The University of Wollongong in Dubai also endorses a policy of non-discriminatory language practice and presentation.  **PLEASE NOTE:**STUDENTS MUST RETAIN A COPY OF ANY WORK SUBMITTED | **DECLARATION:** I/We certify that this is entirely my/our own work, except where I/we have given fully-documented references to the work of others, and that the material contained in this document has not previously been submitted for assessment in any formal course of study. I/we understand the definition and consequences of plagiarism.  **Signature of Student: HASSAN, MOHAMED** |
| --- | --- |



| | **Optional Marks:** | | | --- | --- | | **Comments:** | |
| --- | --- | --- | --- |

| | **Lecturer Assignment Receipt**(To be filled in by student and retained by Lecturer upon return of assignment) | | | --- | --- | | **Subject:** | **Assignment Title:** | | | **Student Name:** | **Student Number:** | | | **Due Date:** | **Date Submitted:** | | | **Signature of Student:** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

| | **Student Assignment Receipt** (To be filled in and retained by Student upon submission of assignment) | | | | --- | --- | --- | | **Subject:** | | **Assignment Title:** | | | **Student Name:** | **Student Number:** | | | **Due Date:** | **Date Submitted:** | | | **Signature of Lecturer** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

**Accident Causes in UAE**

**Mohammed Maarij Uddin Khan, Hassan Abdullah Ghauri**

**UOWD,** [**mmuk953@uowmail.edu.au**](mailto:mmuk953@uowmail.edu.au)**,** [**hag999@uowmail.edu.au**](mailto:hag999@uowmail.edu.au)

1. **Abstract**

Our project aims to cover the amount of traffic accidents that happened in the United Arab Emirates.We have a dataset of UAE where accidents happened from 2013 to 2019 acquired from the Bayanat website as well as the Ministry of Interior. According to The National, the number of crashes increased in 2021 to 3,488 up from 2,931 in 2020. Accidents have been a big problem in the UAE as well other GCC countries with reasons ranging from overspeeding, talking on the cellphones and much more. According to World Gulf, Saudi Arabia is ranked 33rd worldwide for traffic accidents. Saudi Arabia has over 450,000 accidents per year and the average is about an accident every minute. Using the dataset provided by Bayanat and the Ministry of Interior, we will use the attributes included in the dataset to predict the accident size which is the target variable for which other attributes will relate to as well as the reasons of why an accident could happen. Moreover, our dataset includes the time, place, weather condition, road condition, the pedestrians, country, place and much more to help us solve our problem. The accident size consists of values such as Basic, Limited and Medium to express the magnitude of the accident. We will use a decision tree Machine Learning Algorithm by implementing python to see what is the cause of the accident and how it is influenced by different factors. Furthermore, we will visualize the different relationships between the variables which show us the factors are connected to each other. We will also go into a deep discussion into why the accidents happen and how we can prevent them to make sure this does not happen again in the future to promote safety in the public roads which vary from different conditions and circumstances.

**Keywords**

* Decision tree
* Number of traffic accidents
* Severity of traffic accidents
* Bayanat website
* Accident Size
* Python
* United Arab Emirates

**2. Introduction**

**Background**

Our project entails one of the recurring problems in the UAE which are Traffic Accidents. Traffic accidents have been prevalent in the UAE and according to the National, there has been an increase in the number of crashes in 2021 to 3,488 up from 2,931 in 2020. We tend to find out the reason why these things happen. We should not only know the reasons, but how the reason relates to other factors as well such as road condition, speed of the vehicle, pedestrians and much more. The UAE has been a victim of high car crashes from youngsters to people under influence of drugs and can make driving in the UAE less safe. The research on the topic of why the accidents take place are scarce and more attention needs to be done in this area.

**Objectives**

Finding out the factors that lead to road severity crashes in the UAE.

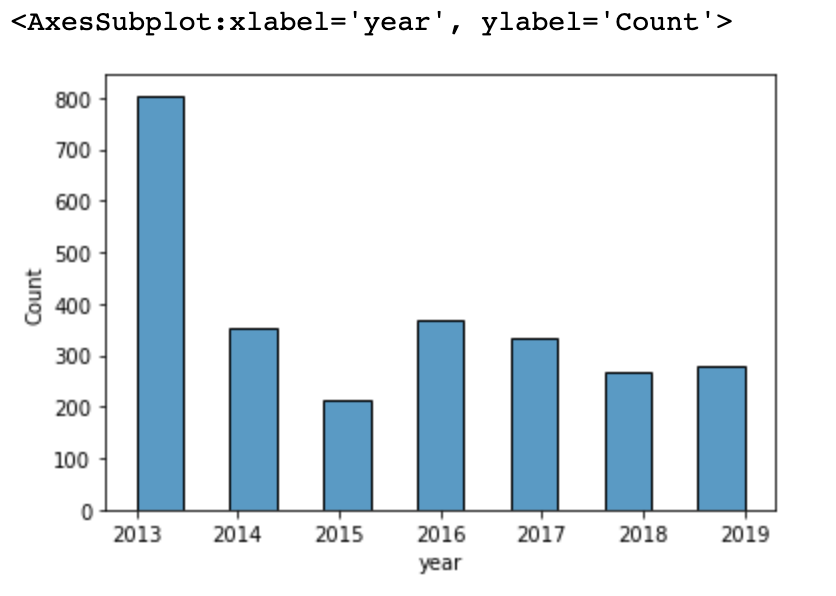
**Study Relevance**

This study won’t only tell the reason, but it can show what other factors lead to the accident. This is also not confined to the UAE but can be compared to other GCC because of the similar culture and road condition

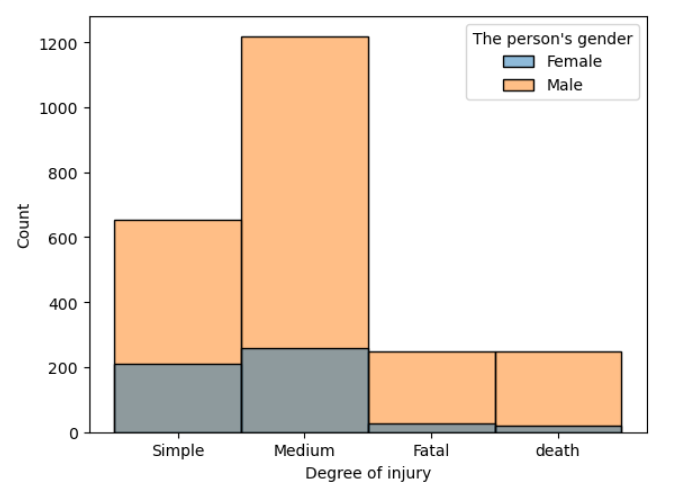
**3. Literature review**

We have used 2 datasets about the accidents that took place in UAE from 2013-2019 as well as in 2021. The decision tree algorithm was used to predict the cause of the accidents in UAE using the 2021 dataset and the same machine learning algorithm was used to predict the accident size in UAE from 2013-2019 while taking other factors into account like weather condition, road condition and so on. We found a similar study done by Francisco Daniel B and Dina M. Awadalla in May 2019. The study goes through a dataset with reports on crashes from 2012 to 2017. Similarly we have dataset ranging from 2013 to 2019 and even 2021 but unfortunately could not get our hands on 2020 accident reports. Our dataset was collected from the Ministry of Interior website while the other had collected about Abu Dhabi police. Our dataset for 2021 showed January, November and December were the highest rate where accidents happened while July, September and August were the lowest. In comparison to Daniel B findings, November and December accounted for **the lowest**crashes from 2013 to 2019 which contrary to what we found.

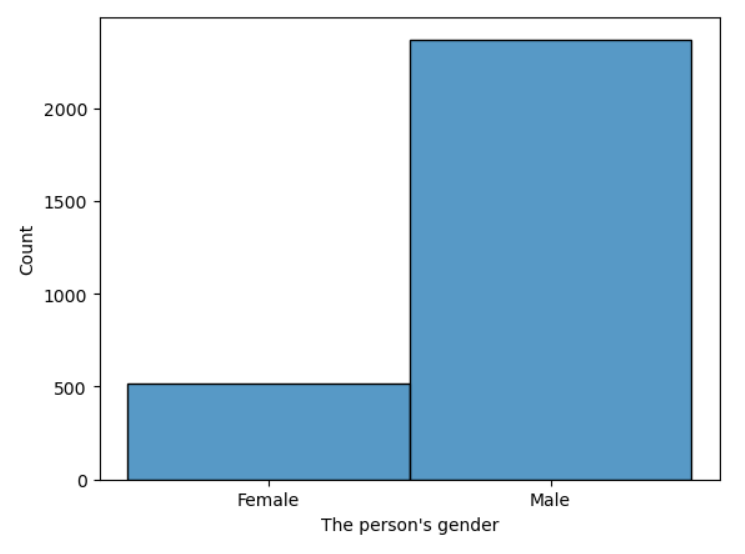
Moreover, we did find out that the accidents did decrease dramatically. From our findings, 2013 was recorded to have the most accidents from 2013-2021 (Excluding 2020). The accident rate dramatically decreased after 2013 and there was a constant rate of accidents which was half of the accidents that happened in 2013. Furthermore, comparing the genders which got into crashes, from our 2021 dataset we found out that Male have gotten into a lot of trouble lately with having more in 2021 as most crashes were done by male as compared to females (Figure 3). Also, since Males got into accident’s more the degree to injury was more for the Males. The degree of injuries ranged from Simple to death with Medium and Fatal in the middle. Francisco’s team found that there were more simple Injuries than medium injuries in comparison to what we did, in 2021 we found that there were more Medium type injuries but the data from 2013-2019 (Figure 5) matches well with that Francisco’s team found. Carrying on our data we also found that the amount of accidents that happened, most of the time the weather was clear which matches with what the researchers had found compared to what we found (Figure 6 and 7). The results have been matching so far with little variation of what the authors did. Though it is to be noted that the researchers have more datasets that we have as they are going reports which reaches 1 million while we are working with a dataset that reaches more than 2000.



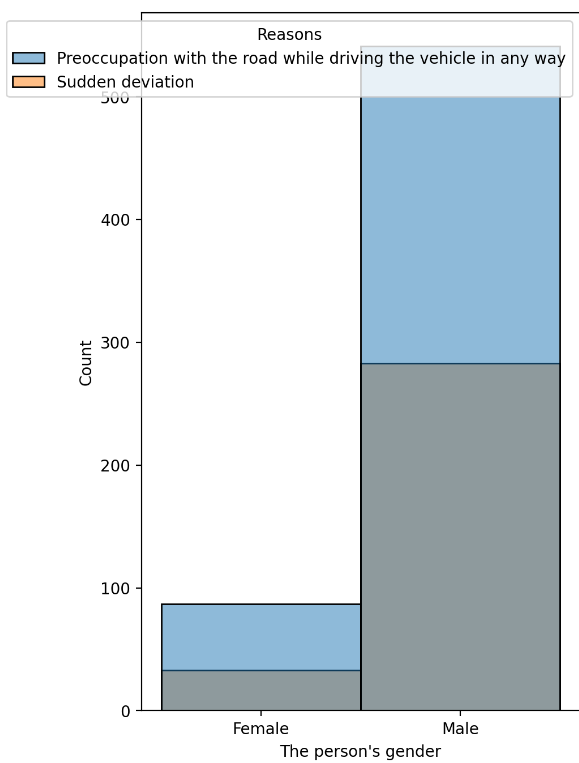
**Figure 1 (2013 - 2019)**



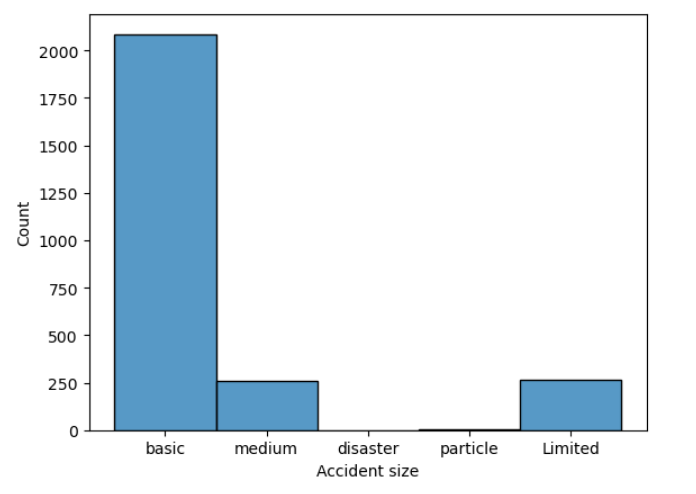
**Figure 2 (2021)**



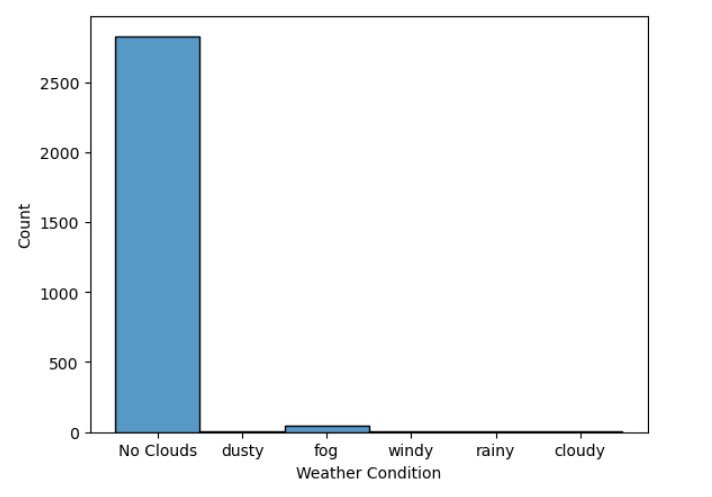
**Figure 3 (2021)**

****

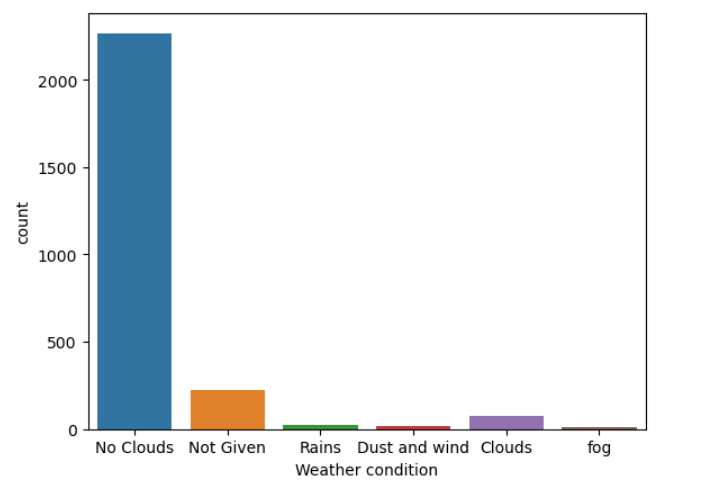
**Figure 4**

****

**Figure 5 (2013 - 2019)**

****

**Figure 6 (2021)**

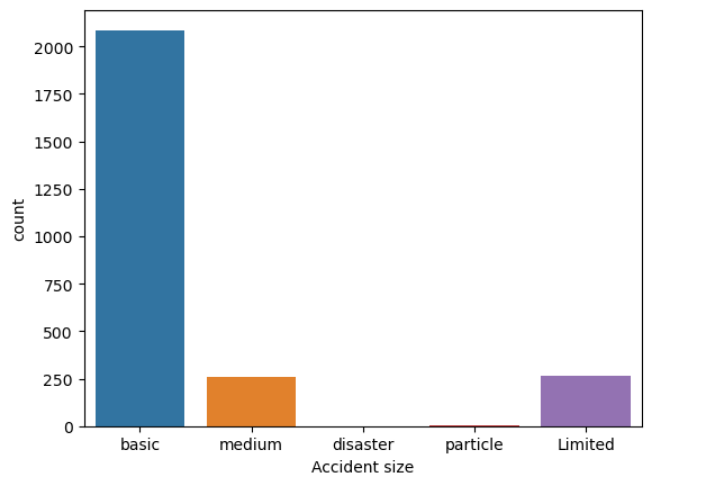
****

**Figure 7 (2013 - 2019)**

**4. Methodology**

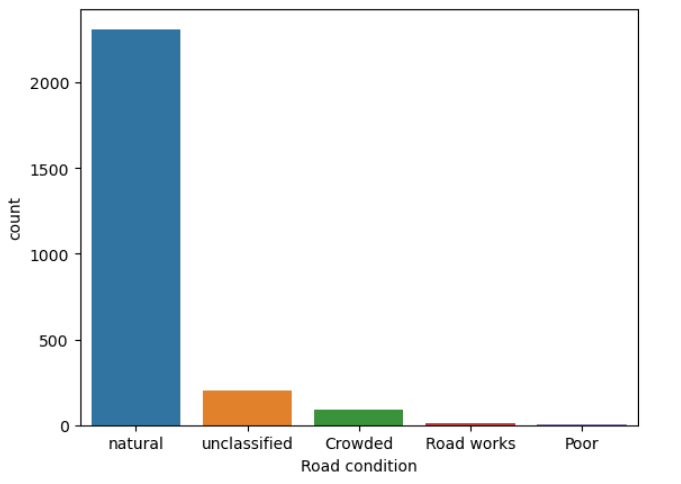
The method we used to predict the cause of the accidents was the Decision Tree Classifier. One of the strong Machine Learning models which classifies an accident based on parameters and in this case it were Weather, Road condition, Time, Area, Lighting, Location and so on. We also used data visualization such as hist, countplot, scatter. Since we had mostly text data, we tried our best to work with what we had. Classification report to evaluate the model properly. We used scikit learn to get the machine learning model, pandas for reading the file and other libraries which were used in the process. Moreover, searching for the dataset from UAE websites to get the best results from local sources for better work.

**5. Result and discussion**

****

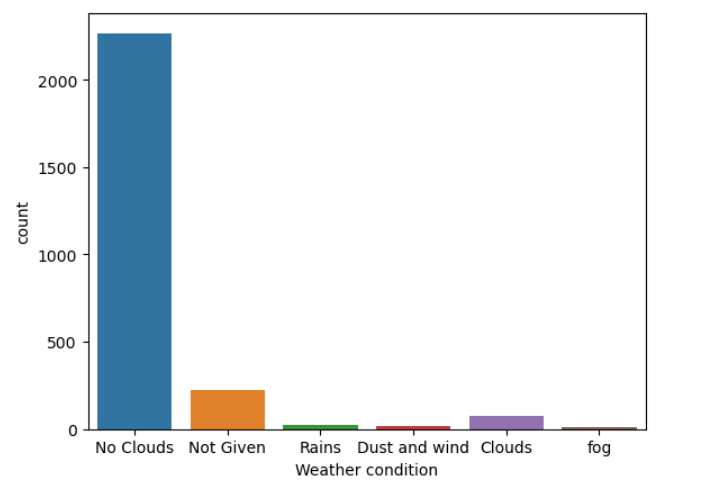
**Figure 8 (2013 - 2019)**

The above count plot displays the count of the accident size based on the category of accident size such as basic, medium, disaster, particle and limited. The basic type of accident size occurred the most based on the data set while no accident size of the disaster category occurred during the above time period.



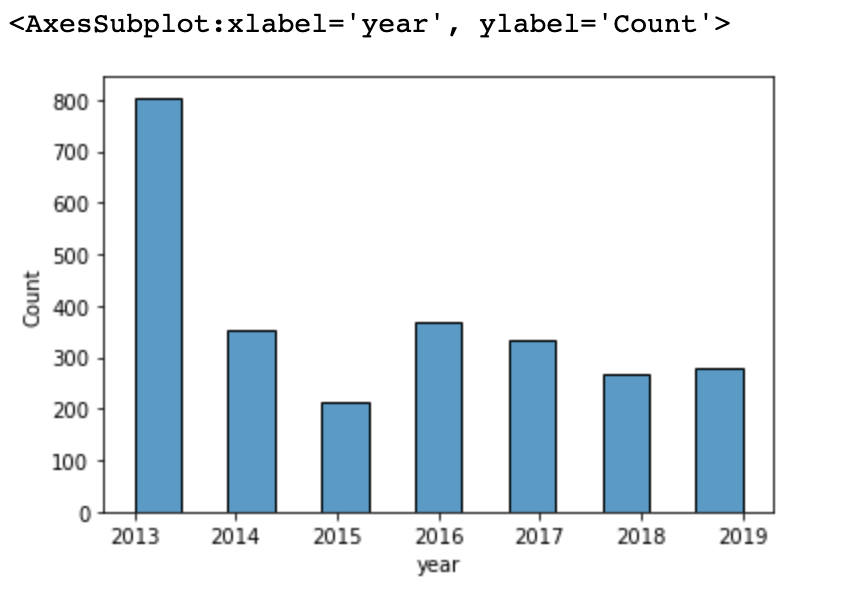
**Figure 9 (2013 - 2019)**

The above count plot displays the count of the road condition for which the accident occurred based on the various categories that are natural, unclassified, crowded, road works and unpaved methods. The natural type of road accident occurred the most based on the data set while road works and unpaved methods occurred the least.



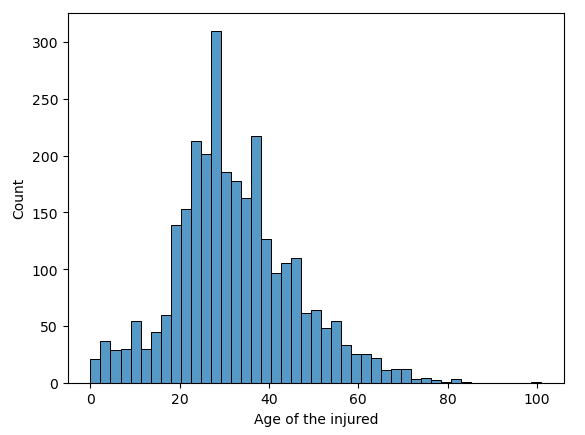
**Figure 10 (2013 - 2019)**

The count plot above displays the count of the weather condition for which the accident occurred based on the various categories of weather condition such as no clouds, did not specify, rains, dust and wind, clouds and fog. The no clouds type of weather occurred the most while the fog weather condition occurred the least.



**Figure 11 (2013 - 2019)**

The above histplot shows the occurrences of the accidents that took place in the years 2013-2019 in the UAE based on the count of the year from 2013 to 2019. The year 2013 showed that most accidents took place in that year and in 2015 the least accidents occurred. The histplot shows an overall decrease in the accidents from 2013 to 2019.



**Figure 12 (2021)**

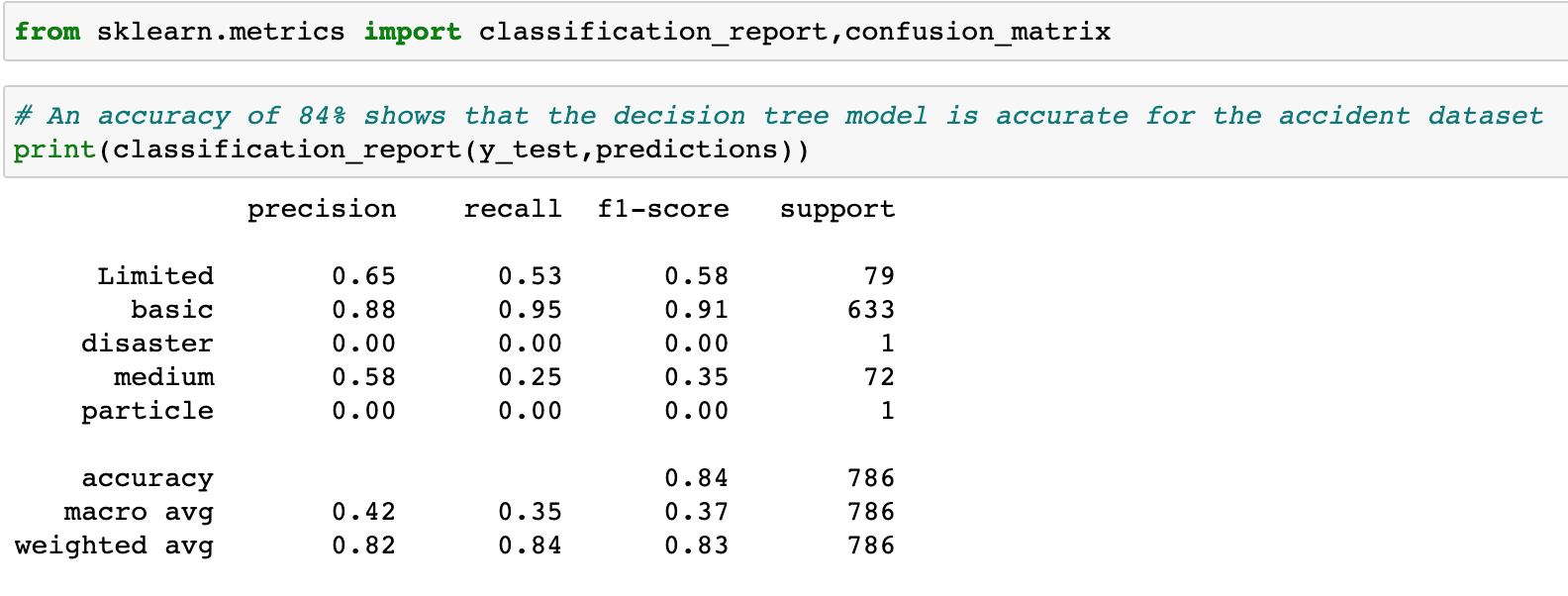
Figure 12 showed an interesting twist of 2021 where people in their 30s are the one who got into accidents. The mean age of people was 32.

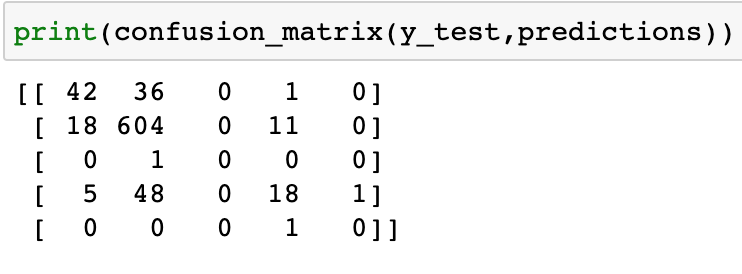
**6. Evaluation**

Evaluating the model of our machine, how did we improve it, what is good or bad and so on

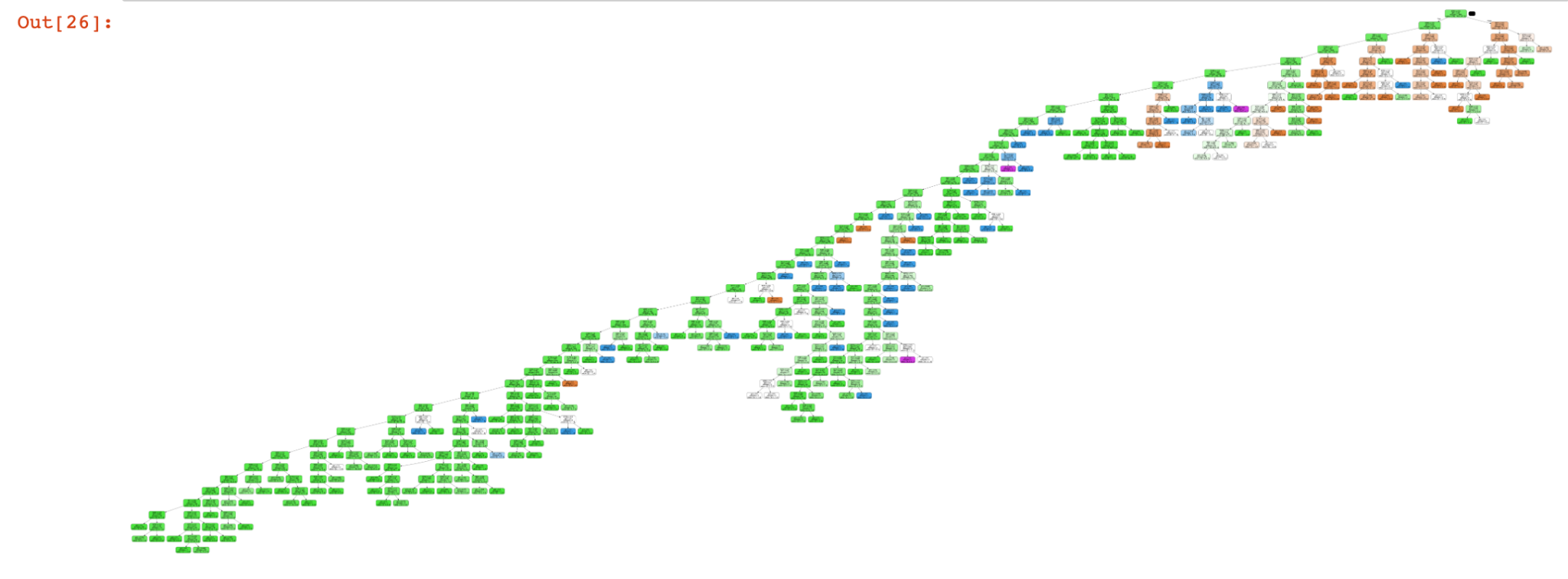
Since we had 2 datasets, our evaluation of the models were divided into 2 where we had used DecisionTreeClassifier() for 2 different datasets and predicting different things but having the same goal “Finding the causes of the accident”.

2013 - 2019 Dataset



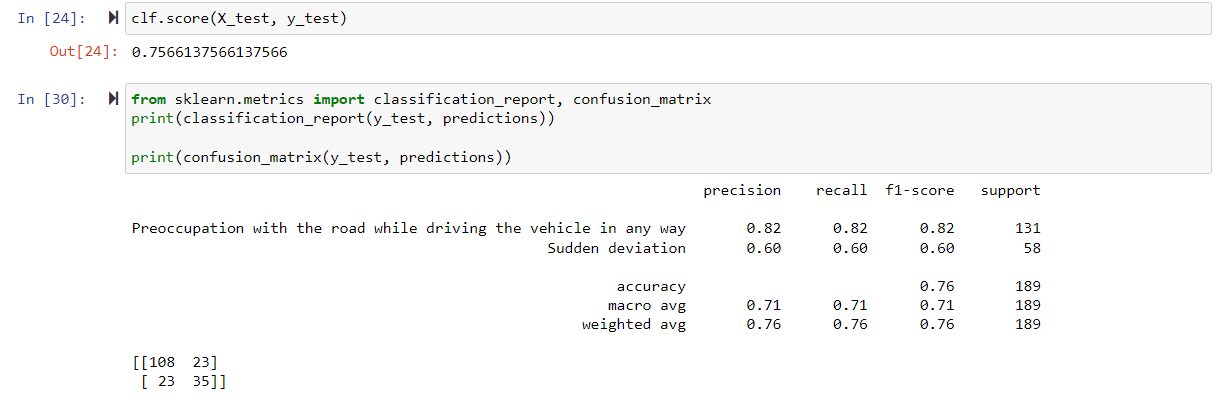


The decision tree machine learning model has displayed an accuracy of 84% to predict the accident size in the UAE from 2013 to 2019. This shows how accurately the decision tree has predicted the accident size. The weather condition, road condition, band’s work summary and the priority level were used to predict the accident size resulting in high accuracy and showing how highly they correlate to the accident size.

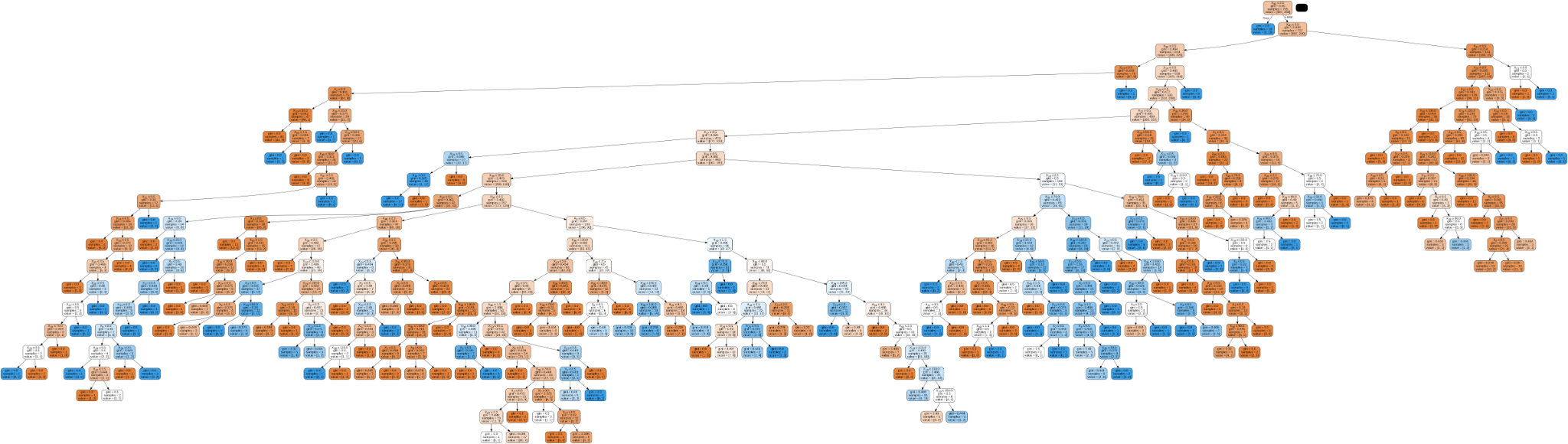


This shows a display of the decision tree model that was used to predict the accident size.

2021 Dataset



The Decision Tree Classifier which is the name of our Machine Learning Model gave us accuracy of about 76%. Though we had many multi class target variables, there were too many for us to make good predictions and a data imbalance which led to our initial accuracy of about 50% and even less than that. After focusing our efforts on the Top 2 reasons of why the accidents happen, our accuracy improved and our decision tree model also readable and made good connections as you can see below.



**7. Conclusion**

To conclude, we believe that the decision tree machine learning algorithm we have used has accurately predicted the cause of the accidents in UAE in 2021 as well as the accident size in UAE from 2013 to 2019. Though we are missing other reasons as well of the occurrence of accidents which we missed, we focused on the top number of reasons of why the accidents occur and how to better prepare for them. This will greatly help the UAE government to keep track of the accidents accurately using a machine learning model and determine what were the external factors that played their part in the cause of the accident which was our main goal along with the accident size.

**8. References**

1. Albuquerque, F., & Awadalla, D. (2019, May 26). *Characterization of road crashes in the emirate of Abu Dhabi*. Elsevier. Retrieved December 5, 2022, from <https://www.sciencedirect.com/science/article/pii/S2352146520305536#:~:text=Road%20crashes%20were%20more%20likely,minors%2C%20and%2For%20Emiratis>.
2. Amir, S.A. (2022) *Road accidents in UAE caused 381 deaths last year*, *The National*. The National. Available at: https://www.thenationalnews.com/uae/2022/05/26/road-accidents-in-uae-caused-381-deaths-last-year/ (Accessed: November 14, 2022).
3. Alexandersson, M. (2015, October 30). *How to change a figure’s size in Python Seaborn package*. Stack Overflow. Retrieved December 5, 2022, from <https://stackoverflow.com/questions/33446029/how-to-change-a-figures-size-in-python-seaborn-package>
4. *2.10. Decision Trees*. (n.d.). Scikit-learn. Retrieved December 5, 2022, from <https://scikit-learn/stable/modules/tree.html>
5. *scikit learn decision tree export graphviz - wrong class names in the decision tree*. (2016, December 18). Stack Overflow. Retrieved December 5, 2022, from https://stackoverflow.com/questions/41207923/scikit-learn-decision-tree-export-graphviz-wrong-class-names-in-the-decision-t
6. Police, D. (2014, August 5). *DubaiPulse*. DubaiPulse. Retrieved December 5, 2022, from <https://www.dubaipulse.gov.ae/main_search?searchword=Traffic+Accidents&as_sfid=AAAAAAV_-FCLlmZzUT-bHqaDA3xKqjN0zNEel20D3dPQA57YTiqczfofbhAQhBKNqjRZT2WUPWNaX0vv5eOUFZh_mq2p9GUt-TCfNGmBTW_O6XMW6RIgmhLiIxYeOp-fvU-MESA%3D&as_fid=3595de43568557390c7b1e147c91a6f2d371cea0>
7. Python pandas extract year from datetime: Df['year'] = df['date'].year is not working. (2015, June 5). Retrieved December 5, 2022, from https://stackoverflow.com/questions/30405413/python-pandas-extract-year-from-datetime-dfyear-dfdate-year-is-not-wo
8. Bureau Chief, H. T. (2017, November 23). *One traffic accident per minute in Saudi Arabia*. One Traffic Accident per Minute in Saudi Arabia | Saudi – Gulf News. Retrieved December 5, 2022, from https://gulfnews.com/world/gulf/saudi/one-traffic-accident-per-minute-in-saudi-arabia-1.2129414