CASES OF DEATHS IN INDONESIA

BACKGROUND

The challenge is to classify the types of disasters in Indonesia. Classifying the types of disasters in Indonesia is crucial it helps in optimizing disaster management efforts, reducing the impact of disasters, and ultimately protecting lives and infrastructure. By utilizing the features of total death and year in the open dataset "Penyebab Kematian di Indonesia yang Dilaporkan", we build a predictive model that aims to provide insights and aid decision-making in disaster management.

DATA DESCRIPTION

Contains data compiled from the Indonesian Health Profile from 2000 to 2022 and COVID-19 data. The URL of the data source is listed in the "Source URL" column. The "Type" column (Type of Cause of Death) is filled in by the author, not from the sources mentioned, but inspired by the 2019 Indonesian Health Profile which divides Health Crises by Disaster Type, namely Social Disasters, Natural Disasters, and Non-Natural Disasters. In the context of this dataset, author used those 3 types but modified them slightly, which became: "Social Disasters", "Natural Disasters", and "Non-Natural Disasters and Diseases".

DATASET CHARACTERISTICS

	variable	type	levels	topLevel	topCount	topFrac	missFreq	missFrac
1	Cause	character	181	Tetanus Neonatorium	22	0.032	0	0.000
2	Type	character	3	Bencana Non Alam dan Penyakit	512	0.753	0	0.000
3	Year	integer	23	2005	83	0.122	0	0.000
4	Data.Redundancy	integer	1	1	680	1.000	0	0.000
5	Total.Deaths	integer	303	0	106	0.156	0	0.000
6	Source	character	20	Profil Kesehatan Indonesia Tahun 2005	82	0.121	0	0.000
7	Page.at.Source	character	109	Lampiran 3.4	38	0.056	2	0.003
8	Source.URL	character	19	https://pusdatin.kemkes.go.id/download.php?file=downloa...	82	0.121	3	0.004

APPLIED METHOD

Our approach involved utilizing the input features of total death, year, and population. We employed the decision tree algorithm to build a model that can accurately predict the type of disaster based on these features. To evaluate the model's performance, we used the accuracy metric, which measures the proportion of correctly classified instances. And we explored the impact of different input features alongside total death and year to assess their influence on the model's accuracy. By applying this methodology, we aimed to develop an effective and interpretable model.

RESULT & DISCUSSION

Because there is the same significance/importance value due to their perfect correlation coefficient, and it doesn't benefit the model, so we try to create a new decision tree model that uses Year and Total.Deaths variables.

The model shows a relatively good accuracy of 73% on the test set with both fits, but it is important to consider the possibility of overfitting or underfitting.

FIT ①

Total.Deaths	Population	Year
27.79444	27.48164	27.48164

FIT ②

Total.Deaths	Year
27.79444	27.48164

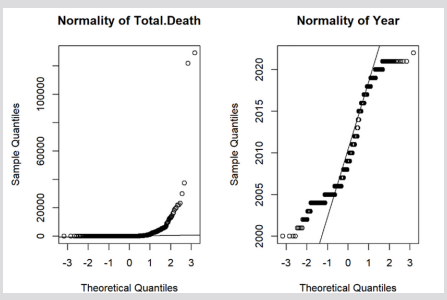
Confusion Matrix

Actual \ Predicted	Bencana Alam	Bencana Non Alam dan Penyakit
Bencana Alam	TP 7	FN 16
Bencana Non Alam dan Penyakit	FP 20	TN 89

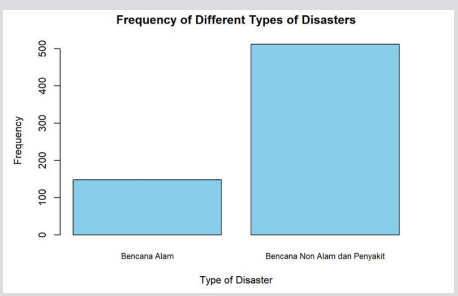
MODEL EXPLANATION

The model's performance can be considered quite good, as it achieves an **accuracy of 73%** on evaluation using 80% train and 20% validation.

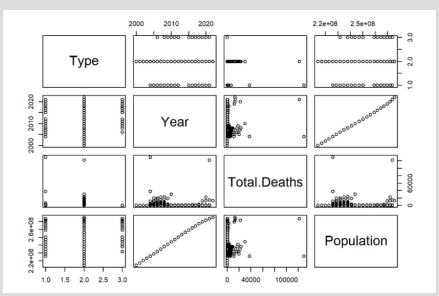
VISUALIZATION



The variable "total.death" is not normally distributed and The variable "year" shows a normal distribution pattern but there are many outliers.



Provides a visual representation of the frequencies, allowing us to identify which types of disasters occur more frequently. This information can be valuable for understanding the dataset and identifying patterns or trends related to different disaster types.



The average variable has a non-linear relationship, only the Year and population variables have a correlation with each other

Conclusions

The final model utilized the "Total Death" and "Year" variables as input features, which proved to be efficient for predicting the outcomes. Evaluation using the confusion matrix revealed that the model achieved 7 True Positives, correctly predicting instances of "Natural Disasters." However, it also had 16 False Negatives, misclassifying some instances of "Natural Disasters" as "Non-Natural Disasters and Diseases." Additionally, there were 20 False Positives, incorrectly categorizing some instances as "Natural Disasters" when they should have been labeled as "Non-Natural Disasters and Diseases." On the positive side, the model achieved 89 True Negatives, accurately identifying instances of "Non-Natural Disasters and Diseases."

Overall, the model's performance demonstrates its potential for predicting the given outcomes, but further analysis is needed to assess any potential issues like overfitting or underfitting.

GROUP MEMBERS:

- **Marchel Christhoper Wuisang-2501991611 - Data Science - marchel.wuisang@binus.ac.id**
- **Muhammad Athariq Naufal - 26021625524 - Data Science - muhammad.naufal025@binus.ac.id**
- **Muhammad Kaisar Novrenza - 2602113546 - Data Science - muhammad.novrenza@binus.ac.id**
- **Meisa Kamilia - 2602135446 - Data Science - meisa.kamilia@binus.ac.id**