Accident in the Czech Republic classification - Report

Tuan Kiet Nguyen, Dominik Fuka

Link to dataset: https://www.policie.cz/clanek/statistika-nehodovosti-900835.aspx

Description of the dataset

This analysis focuses primarily on the accident rate in the Czech Republic in 2023. The data is downloaded from the official website of the Police of the Czech Republic. The original dataset consists of 5 data files in .xsl format, readable in Microsoft Excel, and one file with explanations of abbreviations in the header and individual values. Most values are defined as a numeric key, so it may be suitable for machine reading, but it is harder for humans to quickly understand.

- *Inehody.xls* - (38 attributes, 94 945 records) - A master file that describes specific accident records. Each line contains a unique case ID, time, accident character, number of participants, consequences, causes, etc.
- *Ichodci.xls* – (14 attributes, 3 291 records) – The file describes the pedestrian who was the victim of an accident, i.e. only cases of collision with a pedestrian. It is assigned to accidents using the ID attribute, but this identifier is not unique because there can be multiple pedestrians in one accident.
- *IntGPS.xls* - (7 attributes, 94 945 records) – This file contains specific data about the location (GPS X, Y, state, …) of the accident site. It is useful for determining the specific district of the accident site or entering coordinates into maps. The number of lines and ID is identical to the main file.
- *IVozidla.xls* – (21 attributes, 153 843 records) – The file contains records of the vehicles involved in the accident. As well as pedestrian data, they are attached to the accident by ID in a 1:m relationship.
- *Inasledky.xls* – (9 attributes, 159 940 records) – This file expresses the consequences on vehicles. However, the attributes are not described in the explanation of the abbreviations.

Exploratory data analysis

A classification task is performed on this data, specifically on the consequences for a pedestrian when he/she is a victim of an accident. This question would be very useful for predicting whether a pedestrian would be injured or killed in a particular accident condition.

Data preprocessing

This process is very important both for getting the correct result and for performing machine learning. This work, of course, it takes most of the time. As described, the dataset is very large, it contains a lot of accident information, so it is necessary to remove problems such as unnecessary attributes (not related to the task), missing values, merging files, convert into pure readable file (.csv) etc. The cleaning is done exactly according to the following steps:

1. ***Convert .xls to .csv files*** - A Python script is created for automatic conversion and future reuse. The user puts the raw .xls files into the input folder and runs the script with the command "python parse.py -raw filename1 filename2 ...". The script converts and exports the data to a .csv file. The whole process is also described in the file "parse.py". Later, the script is updated with a dictionary so that the output is human readable.
2. ***Create dictionary*** for translating keys into text and define default values for some of columns.
3. ***Import files into Jupyter Notebook*** – Read-ready files are created from the previous step. First, the main accident file "Inehody" is imported. The 20 columns that may affect the result are selected. Unique on ID column is also checked.

   Imported 20 columns: ['id', 'Lokalita nehody', 'Druh nehody', 'Druh srazky jedouciho vozidel', 'Druh pevne prekazky', 'Zavineni nehody', 'Pritomnost alkoholu u vinika', 'Pritomnost drog u vinika', 'Hlavni priciny nehody', 'Typ povrchu silnice', 'Stav povrchu vozovky v dobe nehody', 'Stav komunikace', 'Povetrnostni podminky v dobe nehody', 'Viditelnosti', 'Rozhledove podminky', 'Deleni komunikace', 'Situovani nehody na komunikaci', 'Rizeni provozu v dobe nehody', 'Smerove pomery', 'Druh pozemni komunikace']

   In the next step, the pedestrian file "Ichodci" is imported. 10 columns are selected. Unique ID check is not needed for 1:m relationship.

   Imported 10 columns: ['id', 'Reflexni prvky u chodce', 'Chodec na osobnim prepravniku', 'Pritomnost alkoholu (Chodec)', 'Druhy drogy u chodce', 'Chovani chodce', 'Pohlavi chodce', 'Rok narozeni chodce (posledni dvojcisli)', 'Poskytnuti prvni pomoci', 'Nasledky pro chodce']

   The two files are merged using the function `pd.merge(data_accidents, data_pedestrians, on='p1', how='right')` with the method 'right join', where right is the pedestrian dataset to get only the accidents associated with them. The number of rows merged should be equal to the number of pedestrians.

4. ***Choose target column and split dataset*** – As stated above, the target is the Pedestrian Consequences, which are the key "p33g". The ID column is also omitted as it has no meaning in this case. The splitting result is 30% Testing / 70% Training with shape [988 rows x 27 columns] and [2303 rows x 27 columns].

5. ***Provide missing values treatment on training and testing dataset separately*** – First of all, treatment year of birth column is needed to be calculated into age. In case of missing age values, it is then replaced by the average age generated from the pandas `mean` strategy. The ages are later mapped into group with age intervals. The other columns are filled with the default value defined in the dictionary. Same treatment steps are provided on both sets, but each run independently.

Modeling

Two well-known algorithms are used to build the model - decision tree and logistic regression. Both successfully ran on same dataset.

Decision tree:

No specific parameters were set.

Logistic Regression:

There was Limit Iteration warning on first try for default limit, so number of max_iter is set to 10 000. It run later without warning.

Results and Evaluation

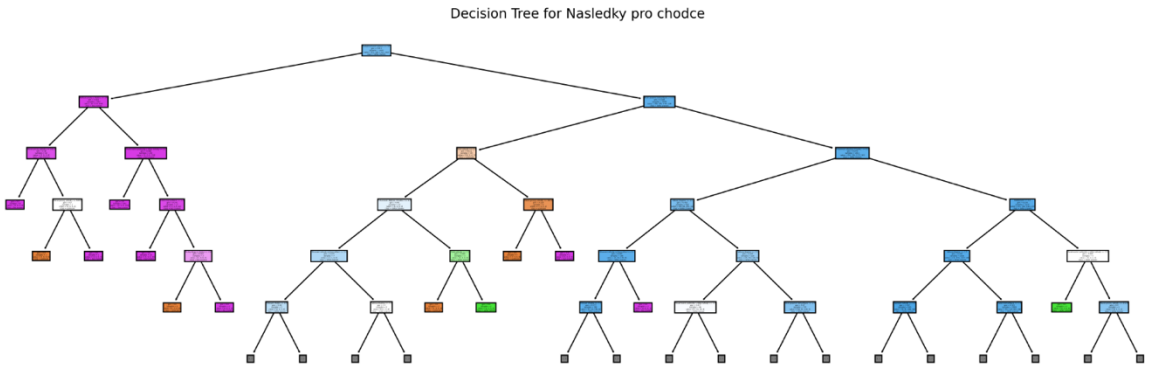
The metrics report for Decision Tree are:

```
Accuracy for Nasledky pro chodce: 0.7398785425101214

Confusion Matrix for Nasledky pro chodce:

[[  6   4  11   2]
 [  4  15  71   6]
 [ 15  81 598  43]
 [  2   2  16 112]]

Classification Report for Nasledky pro chodce:

              precision    recall  f1-score   support

           1       0.22      0.26      0.24        23
           2       0.15      0.16      0.15        96
           3       0.86      0.81      0.83       737
           4       0.69      0.85      0.76       132

    accuracy                           0.74       988
   macro avg       0.48      0.52      0.50       988
weighted avg       0.75      0.74      0.74       988
```

The final accuracy (0,74%) is quite decent, but there are still incorrect predictions (etc. 6 out of 96 for class 2 – Tezke zraneni (Seriously injured)).

However, by applying metaparameter tuning - GridSearchCV, a model with better accuracy was found. DecisionTreeClassifier(max_depth=4, min_samples_leaf=6) with accuracy = 0.8593117408906883.

The tree is quite large so for visualization max_depth is set to 5.



The metrics report for Logistic Regression are:

```
Accuracy for Nasledky pro chodce: 0.8370445344129555

Confusion Matrix for Nasledky pro chodce:

[[  4   1  16   2]
 [  2   0  91   3]
 [  2   0 711  24]
 [  0   0  20 112]]

Classification Report for Nasledky pro chodce:

              precision    recall  f1-score   support

           1       0.50      0.17      0.26        23
           2       0.00      0.00      0.00        96
           3       0.85      0.96      0.90       737
           4       0.79      0.85      0.82       132

    accuracy                           0.84       988
   macro avg       0.54      0.50      0.50       988
weighted avg       0.75      0.84      0.79       988
```

The logistic regression algorithm generated a more accurate model. The result is very similar to the best decision tree model.

Using GridSearchCV, the best model was: LogisticRegression(max_iter=5000, solver='saga') with an accuracy of 0.8380566801619433.

Conclusion

Both algorithms achieved very close results, but the average accuracy does not exceed 90%. Indeed, the consequences of pedestrian accidents can be influenced by various factors such as road conditions, vehicle speed, pedestrian behaviour, weather conditions and more. As a result, accurately predicting the consequences for pedestrians can be difficult.