Improving the Methodology for Determining Injury Rates in Road Accidents using Machine Learning - The Random Tree Method

Pugachev Igor Nikolaevich[1][a], Grigorov Denis Yevgenyevich[2][b], Skripko Pavel Borisovich[3][c], Sheshera Nikolay Gennadyevich[3][d]

[1]*Khabarovsk Federal Research Center of the Far Eastern Branch of the Russian Academy of Sciences (HFRC FEB RAS), Dzerzhinsky street, Khabarovsk, Russia*

[2] *Khabarovsk Institute of Infocommunications SibGUTI, Lenin street, Khabarovsk, Russia*

[3]*Far Eastern Law Institute of the Ministry of Internal Affairs of Russia named after I.F. Shilova, lane Barracks, Khabarovsk, Russia*

Keywords: Injury rates, machine learning method – random forest, data analysis, characteristic curve, forecasting, correlation analysis.

Abstract: Every year in Russia there are a large number of road traffic accidents (RTA), in which more than 100 thousand people are injured and over 10 thousand die. (https://rg.ru/2023/11/18/ezhegodno-v-avtomobilnyh-avariiah-gibnet-bolee-milliona-chelovek-chto-mozhet-izmenit-pechalnuiu-statistiku.html, 2023). This has an impact on the demographic and economic situation of the country. The aggravated problem forces the Government to develop various programs and strategies to improve road safety. They are primarily aimed at preventing accidents with the help of sanctions policy. But the practice of holding drivers accountable is not always objective and leads away from the truth of the origin of accidents, especially with victims (Pugachev, I., Furman, B., Umanets, I., 2023).

The article proposes a comprehensive approach to identifying accident blackspots with a high risk of injury. The study includes an analysis of the dependencies of all elements of the Driver-Vehicle-Road-Environment (DVRE) system. Regression analysis and machine learning libraries of the random forest method in the Python programming language are used.

1 INTRODUCTION

The event – a traffic accident – is not random and is accompanied by a certain set of characteristics of the street and road network (SRN). The driver is the final element of the DVRE system and has a certain psychophysical potential that allows them to interact to a certain extent with the environment, the road and the car (Pugachev, I., Kulikov, Y., Markelov, G., Ostapenko, A., 2021). Research by Professor V.F. Babkov explains the influence of accompanying transport, road and natural characteristics on a person through their capabilities or peculiarities of perception of the surrounding environment. Nevertheless, there are negative combinations of factors of the DVRE system, with the simultaneous influence of which the average driver will not be able to ensure safety when driving their vehicle due, for example, to a large number of irritating factors and high speed. Information overload will distract their attention from dangerous elements, and high speed will not leave them the opportunity to stop or make a safe maneuver (V.F.Babkov, 1991).

To determine the causes of road accidents with injuries, it is necessary to investigate each case, to describe the surrounding situation, and then compare the accident with and without victims (Pugachev, I.N., Lopashuk, V.V., Kulikov, Y.I., Vasilev, A.Y., Barsukova, N.V., 2021). In the 60s, Professor V.F. Babkov developed the “Methodology of Accident Rates”, thanks to which, by generalizing the road

[a] https://orcid.org/0000-0003-0345-4350
[b] https://orcid.org/0009-0005-4049-9488
[c] https://orcid.org/0009-0006-0923-7081
[d] https://orcid.org/0009-0006-3302-5572

transport characteristics of each individual case of an accident, it became possible to predict accident rates using coefficients (Babkov V.F., 1991). Based on the works of the author, his explanation of the causes and consequences of the interaction of the subsystems of the DVRE system, it was decided to continue research in this direction to determine injury rates in road accidents using regression analysis and machine learning in order to generalize characteristics, determine their dependencies and build a model.

The research took place on the SRN of the city of Khabarovsk, Russia.

2 DEVELOPMENT OF A METHOD FOR DETERMINING INJURY RATES IN ROAD ACCIDENTS USING LOGISTIC REGRESSION

In order to collect statistical data, more than 400 km of roads were surveyed. The studies of the characteristics of the SRN took place at the scene of accidents, for which a tabulated template was prepared in advance. The following indicators were analyzed: intensity, number of lanes, width of the road, roadside, slopes (longitudinal, transverse), radius of the curve, visibility in plan and profile, as well as crossings at the same level, type of crossing, regulation of crossroads and crosswalks, type of a crosswalk, marking, dividing strips, additional and main lanes, distance from the road to buildings, pavement, fences, heaviness, sidewalks, bus stops, curbs, poles, time of day, advertising banners, trees along the road, etc.

The data were grouped into specified intervals and binarized according to the principle: 1 - if the factor was present and 0 - if it was absent. This is necessary for regression analysis in the Statistica program (Pugachev I.N., Kamenchukov A.V., Kapsky D.V., Kot E.N., Burtyl Yu.V., Shcheglov, 2022).

Using correlation analysis, as well as the Pearson criterion, features not exceeding the correlation coefficient of 0.7 and the probability of error according to the Pearson criterion of 0.05 I were selected. (I. Pugachev, Y. Kulikov, G. Markelov, N. Sheshera, 2017).

In the Statistica 6.0 program, as a result of experimental studies, a prognostic scale was developed; the best predictive model was selected using the variable selection method, which consists of 9 SRN factors: 2 traffic lanes, roadway width is 18-18.9 m, presence of a 3 m wide roadside, visibility limitation in plan ≥ 200 m, rough (new) coating, tire-to-surface friction coefficient is 0.75, presence of a sidewalk ≥ 4 m wide, average speed is 60 km/h, crosswalk, curve radius in plan is 200-349.9 m.

Each factor is characterized by a rate (severity rate Sr1,Sr2…Sr9).

The logistic regression method is based on the equation:

$$Y= b_0+b_1X_1+ b_2X_2+\ldots+ b_iX_i \tag{1}$$

where Y is a dependent feature, X1, X2,...,X9 are independent features, b1,b2...,bi are coefficients, b_0 is a constant.

Taking into account the logical transformation, the final injury rate will be calculated using the formula:

$$SrTOTAL= Sr_1+Sr_2+Sr_3+\ldots+Sr_9 \tag{2}$$

where – Sr1 – 2 lanes (-0.34556), Sr2 – the width of the road of 18-18 m. (1.44674), Sr3 – the presence of a roadside with a width of 3 m (2.9149), Sr4 – limitation of visibility in plan of 200≤ m (1.312754), Sr5 – rough coating (new), tire-to-surface friction coefficient of 0.75 (0.937902), Sr6 – the presence of a sidewalk with a width of ≤ 4 m (1.262661), Sr7 – average speed of 60 km/h (-0.04846), Sr8 – crosswalk (1.59205), Sr9 – curve radius in plan of 200-349.9 m (2.09848).

The quality of the model was confirmed by the characteristic ROC curve (Figure 1). The area under the curve was 0.82, which confirms the high quality of the prognostic scale.

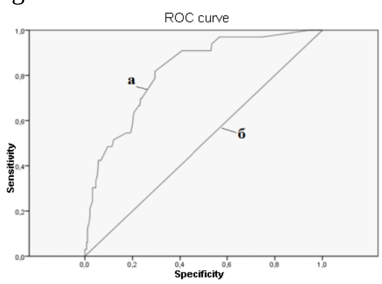

Figure 1: Characteristic curve (a – ROC curve (characteristic curve); b – null hypothesis).

Sensitivity and specificity studies have allowed us to develop a threshold value exceeding which increases the risk of injury in an accident by up to 70%. It was 2.09.

The obtained results were tested on the accident blackspot in Khabarovsk, Suvorova Street, from

Pavla Morozova Street to Malinovskogo Street. Similar to the Methodology of Accident Rates of Professor V.F. Babkov, the analysis involves calculating the injury rate in road accidents using a linear graph and the derived formula 2 (Figure 2).

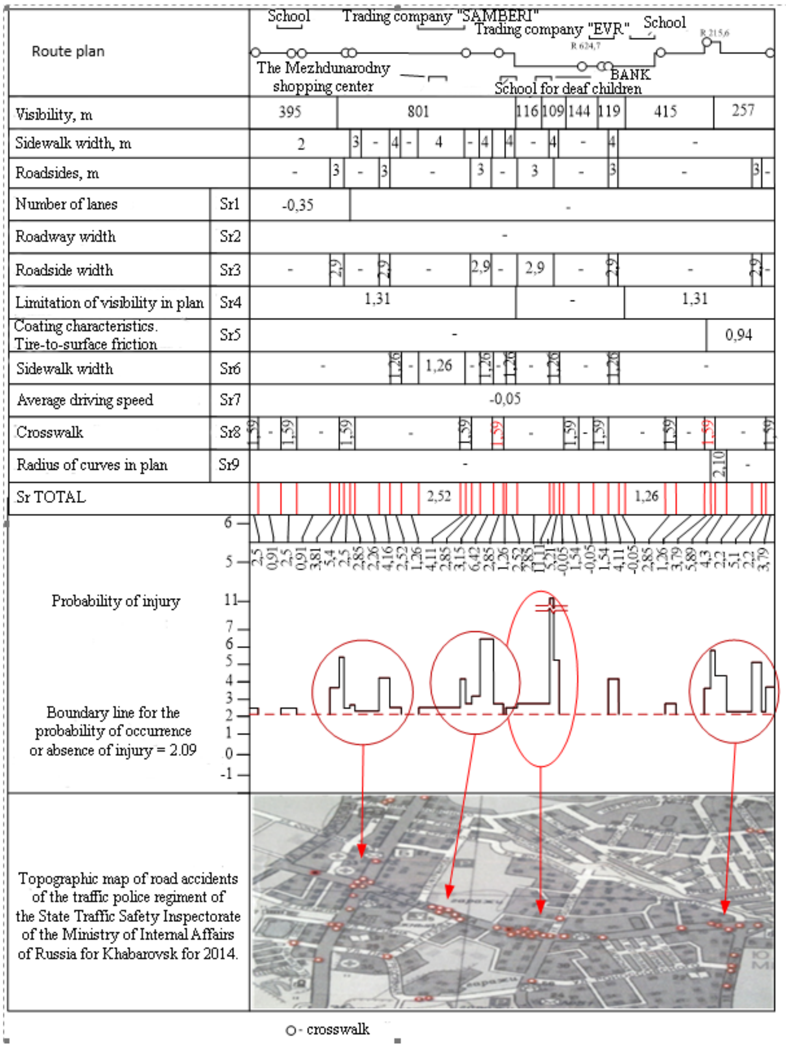

Figure 2: Linear graph of injury rates on Suvorova Street, from Pavla Morozova Street to Malinovskogo Street.

3 ANALYSIS OF INJURIES IN ROAD ACCIDENTS USING MACHINE LEARNING

Despite the efficiency of the prognostic scale created using logistic regression, a large number of features were not significant. However, Professor V.F. Babkov, when studying the psychophysical properties of a driver, considered it natural that all characteristics manifest themselves to varying degrees. Moreover, he spoke about the need for additional research of new ones on a par with the old ones, taking into account the changing transport situation (I.N. Pugachev, A.V. Kamenchukov, N.S. Nesterova, 2022).

The difficult task of studying the degree of influence of various combinations of DVRE subsystems has been solved using machine learning – the Random Forest Method.

3.1 Data collection and preparation

In order to increase the representativeness of the data, most of it was collected by software and hardware systems. For example, the intensity of traffic flow was formed using the INTEGRA KDD traffic flow registration, accounting and analysis system, data on accidents were taken from the open portals of the State Road Safety Inspectorate of the Russian Federation (http://stat.gibdd.ru/), geometric elements of roads were measured in-situ or using modern geoinformation systems, etc. (Pugachev I.N., Skripko P.B., Sheshera N.G., 2023). Additionally, information on weather and climatic conditions provided by the Khabarovsk Center for Hydrometeorology and Environmental Monitoring with the functions of the regional specialized meteorological center of the World Weather Service (Khabarovsk station No. 4853511) was added to the features under consideration.

When collecting data, special attention was paid to an underestimated indicator – speed. In the first part of the study, only the maximum permitted speed in the territory of Khabarovsk, established by state and local authorities, was taken into account. It was assumed that drivers observe it when driving, and accordingly, the value was taken as a constant (Rudenko N.V., 2022). Further observations forced us to doubt this. In the territory of the Russian Federation, in order to be held administratively liable for exceeding the maximum permissible speed, for example 60 km/h, the driver must drive at a speed of 81 km/h. This is necessary to level out various equipment errors and, accordingly, to increase confidence in the commission of an administrative violation. Knowing this, drivers increase their speed to the maximum permissible level, and in places where there are no means of recording violations, the speed is limited only by road conditions and the driver's personal confidence in safety. At the same time, on uncovered main streets in places where people are attracted, there is a high concentration of cars parked along the road. In conditions of insufficient parking spaces, people park on the roadway, violating the rules. This leads to difficulties in traffic flow (Batishcheva O.M., Ganichev A.I., Starikova A.G., 2023). In Khabarovsk there are areas where this happens systematically. The speed rate there is below the permissible limit.

Taking into account the above, it was decided to collect data on the maximum speed allowed in the areas where accidents occurred during the relevant time period. For this purpose, a GPS tracker was used, which recorded data as the vehicle moved along the studied road sections over a period of 2 years. They were processed and each accident was assigned the maximum speed provided in accordance with the place and time. It is worth noting that the research vehicle with a GPS tracker was moving at the maximum speed that the road conditions allowed.

The collected data amounted to over 75 thousand. Road accidents that occurred in the city of Khabarovsk (from 2016 to 2021). They included features that hypothetically influence the injury rates or are secondary, but necessary to prepare the data: Date, Time, Type of accident, Place (administrative districts of the city), Street, House, People died, Children died, People wounded, Children wounded, Grouped by injuries (0 - no victims, 1 - injured, 2 - died), Lane where the accident occurred, Width of the sidewalk, Width of the dividing strip, Type of pavement, Vehicle type (1 - bicycles; 2 - motorcycles; 3 - cars; 4 - buses; 5 - special equipment; 6 - trucks), Vehicle color, Purpose of use, Transportation, Transportation type, Damage, Technical problems, Category, Gender, Severity of consequences, Social characteristics of a driver (place of work), Direct traffic violations, Associated traffic violations, Driving license type, Experience (including grouped by 5-year intervals), Seat belts (whether the person was strapped during the accident and whether the vehicle design included seat belts in general), Chemical test, Traffic flow intensity (including at crossings and at intervals of 100, 250, 500 and 1000 vehicles/hour), Month, Day of week, Number of lanes in the forward direction, Number of additional left and right lanes (their number, type and length), Longitudinal slope (including intervals of 30‰ and 50‰), Road width (including intervals of 5 and 10 meters), Type of traffic (one-way - 1, two-way - 2, mixed - 1.2), Parking (distance to it and type), Type of dividing strip (marking only – 0, dividing strip – 1, barrier fences – 2), Bus stop (type and distance to it), the Type of intersection, Traffic light control, Presence of a crosswalk within a 50-meter radius, Maximum permissible speed, Temperature (air, soil and dew point (°C)), Partial pressure of water vapor (PA), Relative humidity (%), Saturation deficiency ($g/m^{3),}$ Atmospheric pressure at station and sea level (hPa), Visibility (Code – VV), Weather (Code – ww), Wind direction (degrees), Wind speed (m/s), Precipitation (including intervals of 5 and 10 mm), Part of the day (day, night, morning and evening twilight).

3.2 Random Forest Method

It was decided to carry out further data processing to determine dependencies and predict injury rates in road accidents (absence of victims, victims and fatalities) using an ensemble machine learning method - the Random Forest Method.

Random forest ensemble uses multiple trees, which helps reduce the likelihood of model overfitting. Each tree is trained independently and makes predictions, after which the results are combined to obtain the final probability. This helps to improve the generalization ability of the model and reduce the impact of outliers and noise in the data. By combining predictions from multiple trees, an ensemble of random forest can achieve high prediction accuracy.

The machine learning project uses the Scikit-Learn library (also known as sklearn), which is a machine learning library for the Python programming language. It provides a wide range of tools for the development and application of machine learning models, including classification, regression, clustering, data preprocessing and model selection.

Scikit-Learn (https://scikit-learn.org/stable/) has a simple and uniform API interface that facilitates the use and combination of various algorithms and methods. The library also comes with many functions for data preprocessing, model selection, evaluation, and validation.

3.3 Preparing data and installing dependencies

After studying the data processing features of the Scikit-Learn method, the general population was cleaned and converted to a numeric format. For example, all missing values were analyzed and, in accordance with the semantic load, replaced with “0” if the quantitative indicator was constantly present in the DVRE system (air temperature, partial pressure of water vapor, relative air humidity, etc.) or “-1” if the indicator was qualitative or quantitative, but the absence of a value meant the absence of this indicator in a given period of time (wind speed, weather code, visibility, etc.). Letters were also replaced with numbers (Baklanova K.V., 2023).

To implement the idea in the Python programming language, a virtual environment was created in the PyCharm development environment in order to modularly separate operations for reading data from files, learning operations, and forecasting operations. It included 4 files with the py extension, hereinafter referred to as modules.

- trainer.py – model training;
- config.py – list of variables;
- predict.py – prediction;
- utils.py – encapsulation of the implementation of some functions used in several places of the program, as well as for applying the DRY (Don't Repeat Yourself) principle.

3.4 Training

Taking into account the peculiarities of working in Python, libraries, commands and functions for machine learning and data analysis were imported into the trainer.py module:

`import joblib` –saving and loading machine learning models;

`import numpy as np` – working with multidimensional arrays and matrices;

`import pandas as pd` – working with tabular data;

`import seaborn as sns` – data visualization;

`from matplotlib import pyplot as plt` – creating graphs and visualizations;

`from sklearn.multiclass import OneVsOneClassifier` – multiclass classification;

`from config import cols` – column configuration parameters;

`from config import primary_column` – main column configuration parameters;

`from sklearn.ensemble import RandomForestClassifier` – random forest classifier;

`from sklearn.metrics import r2_score, roc_curve, auc` – metrics for evaluating the quality of models;

`from sklearn.model_selection import train_test_split` – splitting data into training and test sets;

`from sklearn.preprocessing import OneHotEncoder` – transformation of categorical features;

`from sklearn.metrics import classification_report` – output of the classification report;

`from sklearn.metrics import confusion_matrix` – creating an error matrix;

`from sklearn.metrics import RocCurveDisplay` – ROC curve display;

`from sklearn.preprocessing import LabelBinarizer` – binarization of labels;

`from sklearn.inspection import permutation_importance` – feature importance assessments;

`from utils import flatten_cols, encoder` – user utilities for working with data.

The data analysis included the use of graphs, so their style had been determined in advance:

```
plt.style.use('ggplot')
```

The pre-prepared data was downloaded from a CSV file named dataset.csv as a pandas DataFrame using the pd.read_csv() function:

```
severity_df = pd.read_csv('./datasets/dataset.csv', encoding="utf-8", delimiter=";")
```

The encoding="utf-8" parameter indicates the encoding of the file, which allows special and non-English characters to be interpreted correctly, delimiter=";" specifies the column delimiter in the CSV file.

One of the ways to work with data in pandas DataFrame is to change the column structure using the flatten_cols() function. In this case, the code takes the cols variable, which contains the column names, and applies the flatten_cols() function to it with the 'name' argument:

```
cols = flatten_cols(cols, 'name')
```

The cols list is located in the config.py module of this virtual environment, and is converted using the flatten_cols function with the 'name' argument into a one-dimensional list by the given key. It contains a list of keys to be analyzed (independent features). Also in the config.py module, the primary_column variable is assigned the dependent feature general_injuries:

```
cols = [
    {"name": "time"},
    {"name": "driving_experience_5"},
    {"name": "week_day"},
    {"name": "month"},
    {"name": "type_of_road"},
    {"name": "traffic_direction_separator"},
    {"name": "intersection_type"},
    {"name": "traffic_light"},
    {"name": "availability_of_a_pedestrian_crossing_within_a_radius_of_50_m"},
    {"name": "visibility,_code_(VV)"},
    {"name": "part_of_the_day"},
    {"name": "int_250"},
    {"name": "road_width_5"},
    {"name": "number_of_forward_lanes_(total_number_in_all_directions)"},
    {"name": "number_of_left_lanes_(in_all_directions:_oncoming_and_parallel)"},
    {"name": "number_of right_lanes_(in_all_directions)"},
    {"name": "left_stripe_length"},
    {"name": "longitudinal_slope"},
    {"name": "right_stripe_length"},
    {"name": "distance_to_parking"},
    {"name": "distance_to_bus_stop_from_the_center_of_the_house"},
    {"name": "max_speed"},
    {"name": "air_temperature"},
    {"name": "soil_temperature"},
    {"name": "dew_point_temperature"},
    {"name": "partial_pressure_of_water_vapor"},
    {"name": "relative_humidity"},
    {"name": "saturation_deficit"},
    {"name": "atmospheric_pressure_at_station_level"},
    {"name": "atmospheric_pressure_at_sea_level"},
    {"name": "wind_speed"},
    {"name": "precipitation_mm"},
    {"name": "vehicle_type_category"},
    {"name": "general_injuries"},
    # {"name": "color"},        #
        # {"name": "Color"},
    # {"name": "driving_experience_years"},
    # {"name": "location", "need_encode": True}
]
primary_column = 'general_injuries'
```

Looking ahead, it is worth noting that the number of signs included in the cols list at the learning stage is significantly less than what was collected at the first stage of the study. Through experimental selection of data, focusing on the values of the coefficient of determination (r2) and accuracy, the best predictive model was selected (Accuracy =97.99, r2 = 0.8212487585413919). The remaining features were excluded as insignificant.

To ensure the purity of the experiment, accidents involving drivers in a state of intoxication were excluded from the trainer.py module. The indexes of rows in the severity_df DataFrame that have a value in the 'results_of_medical_examination' column equal to "positive" have been deleted. The psychophysical apparatus of these drivers does not react correctly to external stimuli and is abnormal, therefore they should be excluded (Mosin K.K., Kovalevsky V.E., Zhukova N.A., 2023).

```
index_names = severity_df[severity_df['results_of_medical_examination'] == 'положительный'].index

severity_df.drop(index_names, inplace=True)
severity_ml = severity_df[cols]
severity_ml = severity_ml.dropna()
```

The rows with the found indexes from the severity_df DataFrame have also been deleted. The inplace=True argument specifies that the changes should be applied to the original DataFrame rather than returned separately.

In the newly created severity_ml DataFrame, which contains only the columns from the cols list, rows with missing values (NaN) have been removed from the severity_ml DataFrame.

If it is necessary to binarize the values of variables, a cyclic function is used:

```
for col in cols:
    if "need_encode" in col:
        severity_ml[col["name"]] = severity_ml[col["name"]].astype("string")
        severity_ml = encoder(severity_ml, col["name"], True)
```

It checks for the presence of the need_encode string in the variables. If found, binarizes the values according to the conditions. In this work, there was no need for it, but to ensure the accuracy of the prediction, "need_encode": True was added to each variable, after which the accuracy of the model was re-checked.

After binarization, rows with missing values are excluded and the dependent variable is separated from the independent ones.

```
severity_ml = severity_ml.dropna()
severity_train = severity_ml.drop(primary_column, axis=1)
```

The cleared data is divided into training and test samples for training. The size of the test set is defined as 20% of the total amount of data (test_size=0.2).

```
X_train, X_test, y_train, y_test = train_test_split(severity_train.values,
    severity_ml[primary_column].values,
    test_size=0.2,random_state=42)
```

Random_state sets the initial state for the random number generator so that the result is reproducible (Kononchuk E.V., Ermolenko T.V., Shishunov T.O., 2022).

A RandomForestClassifier class object is created, which trains the model on the training dataset (X_train, y_train) using the fit method, then the probabilities are predicted for the test dataset (X_test) using the predict_proba method, which creates a Random Forest model with 200 trees.

```
random_forest = RandomForestClassifier(n_estimators=200)
y_score = random_forest.fit(X_train, y_train).predict_proba(X_test)
result = random_forest.predict(X_test)
```

The model is saved in a virtual environment.

```
joblib.dump(random_forest, 'model.sav')
```

3.5 Model verification

The main indicators of the model's accuracy were the determination coefficient (r2) and accuracy.

```
acc_random_forest = round(random_forest.score(X_test, y_test) * 100, 2)
print("Accuracy", acc_random_forest)
print(r2_score(y_test, result))
```

As mentioned above, by experimentally selecting features, grouping into intervals, increasing or decreasing them, and binarization, the best predictive model was selected where Accuracy =97.99, r2 = 0.8212487585413919.

The resulting indicators influence the injury rate in road accidents to varying degrees, as evidenced by the correlation matrix (Figure 3), which was obtained using the code:

```
corr = severity_ml.corr()
plt.subplots(figsize=(20, 15))
sns.heatmap(corr)
plt.show()
```

Additionally, the significance of the features was analyzed using the permutation importance method for a random forest model trained on training data (X_train and y_train).

```
result = permutation_importance(random_forest, X_train, y_train, n_repeats=10, random_state=42)
importances = result.importances_mean
indices = np.argsort(importances)[::-1]
ar_f = []
for f, idx in enumerate(indices):
ar_f.append([round(importances[idx], 4), cols[idx]])
ar_f.sort(reverse=True)
print(ar_f)
```

As a result of the above command, the following significance coefficients were obtained: [[0.1658, 'max_speed'], [0.0044, 'availability_of_ a_pedestrian_crossing_within_a_radius_ of_50_m'], [0.0041, 'part_of_the_day'], [0.0036, 'time'], [0.0036, 'driving_experience_5'], [0.0032, 'road_width_5'], [0.0032, 'longitudinal_slope'], [0.003, 'int_250'], [0.0025, 'vehicle_type_category '], [0.0019, 'atmospheric_pressure_at_station_level'], [0.0018, 'atmospheric_pressure_at_sea_l evel'], [0.0017, 'air_temperature'], [0.0016, 'relative_humidity'], [0.0014, 'soil_temperature'], [0.0014, 'distance_to_ bus_stop_from_the_center_of_the_house'], [0.0013, 'number_of_forward_lanes _(total_number_ in_ all_directions)'], [0.0012, 'week_day'], [0.0012, 'distance_to_parking'], [0.0011, 'saturation_deficit'], [0.0011, 'dew_point_temperature'], [0.001, 'partial_ pressure_of_water_vapor'], [0.001, 'left_stripe_l ength'], [0.0007, 'traffic_light'], [0.0007, 'intersection_type'], [0.0006, 'right_stripe_length'], [0.0005, 'wind_speed'], [0.0005, 'month'], [0.0004, 'visibility,_code_(VV)'], [0.0004, 'type_of_road'], [0.0004, 'number_of_left_lanes_(in_all_directions:_ oncoming_and_parallel)'], [0.0003, 'precipitation_ mm'], [0.0, 'traffic_direction_separator'], [0.0, 'number_of right_lanes_(in_all_directions)']].

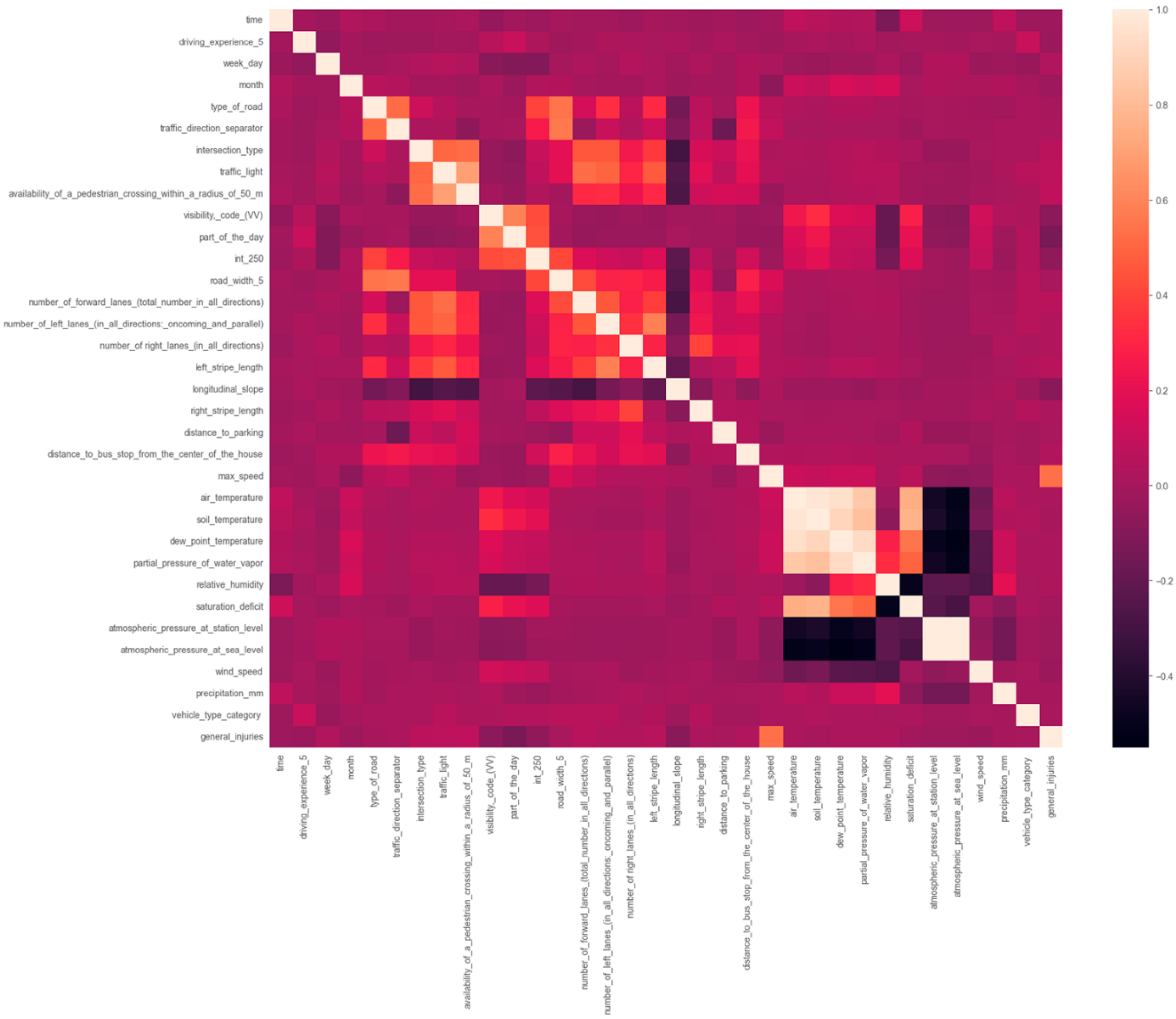

Figure 3: Correlation matrix of the influence of features.

Based on the result, it can be concluded that the indicator of the maximum speed has a strong influence. But this does not mean that the other features are not effective; on the contrary, together they create a complex effect on the injury rates in road accidents.

In order to further verify the accuracy, a matrix of the number of correctly and incorrectly guessed classes was constructed using the following line of code (Figure 4):

```
matrix = confusion_matrix(y_test, random_forest.predict(X_test))
matrix = matrix.astype('float') / matrix.sum(axis=1)[:, np.newaxis]
# Build the plot
plt.figure(figsize=(16, 7))
sns.set(font_scale=1.4)
sns.heatmap(matrix, annot=True, annot_kws={'size': 10},
            cmap=plt.cm.Greens, linewidths=0.2)
# Add labels to the plot
class_names = ["0", "1", "2"]
tick_marks = np.arange(len(class_names))
tick_marks2 = tick_marks + 0.5
plt.xticks(tick_marks, class_names, rotation=25)
plt.yticks(tick_marks2, class_names, rotation=0)
plt.xlabel('Predicted classes')
plt.ylabel('Actual classes')
plt.title('Confusion Matrix for Random Forest Model')
plt.show()
```

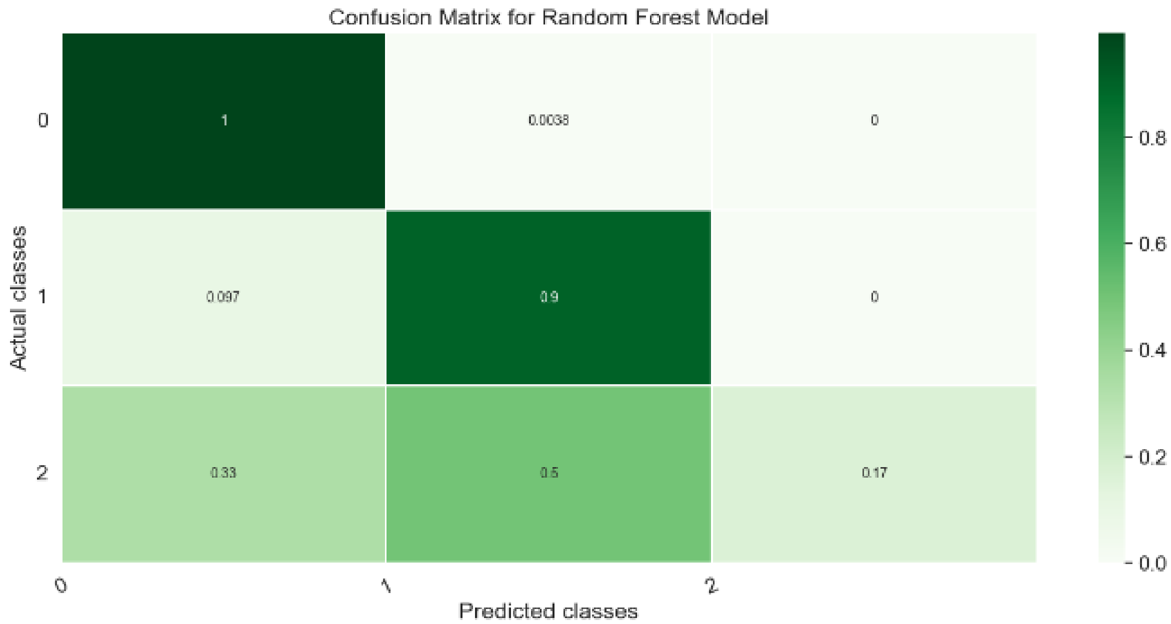

Figure 4: Error matrix

The confusion matrix is a square matrix where each element (i, j) is the number of objects that actually belong to class i but were predicted as class j. Thus, the confusion matrix provides information about correctly classified and misclassified objects for each class (Harutyunyan M.A., 2023).

The confusion_matrix method takes as arguments the true class labels (actual classes) and predicted class labels (predicted classes), and returns a confusion matrix.

As can be seen from Figure 4, the model makes errors, but they are insignificant for determining two injury rate values: 0 - not injured, 1 - injured. Predictions of the third value "2 - died" are not accurate and require further research.

4 PREDICTION

For prediction, data on 27 road accidents from the general population that did not participate in training were selected (dataset_predict.csv). The dependencies were prepared in a similar manner, and the data was initially processed to eliminate errors. The code is written in the predict.py module:

```
import joblib
import numpy as np
import pandas as pd
import warnings
from matplotlib import pyplot as plt
from config import cols, primary_column
from utils import flatten_cols, encoder
plt.style.use('ggplot')
df = pd.read_csv('./datasets/dataset_predict.csv', encoding="utf-8", delimiter=";")
cols = flatten_cols(cols, 'name')
df_test = df[cols]
df_test = df_test.drop(primary_column, axis=1)
df_test = df_test.dropna()
df = df[cols].dropna()
for col in cols:
    if "need_encode" in col:
        df_test[col["name"]] = df_test[col["name"]].astype("string")
        df_test = encoder(df_test, col["name"], True)
```

The machine learning model is loaded from the "model.sav" file using the joblib.load() function.

```
model = joblib.load('model.sav')
result = model.predict(df_test.values)
r = pd.DataFrame(np.array(result), columns=['severity'])
```

```
g = pd.concat([df.reset_index(), r], axis=1)
pd.set_option('display.max_rows', None)
g
```

The results are stored in the result variable, and then packed into a DataFrame r with the column name "severity".

A new DataFrame g is created by combining the original DataFrame df with the index prediction results (row index values). The combined DataFrame g is output, which contains the original data and the prediction ("severity").

After pre-processing, 20 of the 27 accidents remained, the rest did not include the necessary information. The prediction results are summarized in Table 1.

Table 1: Results of the prediction of road traffic accident injury rates.

No.	general_ injuries	severity	No.	general_ injuries	severity
1	1	1	11	0	0
2	0	0	12	1	1
3	0	0	13	0	0
4	0	0	14	0	0
5	0	0	15	1	1
6	1	1	16	2	1
7	1	1	17	2	2
8	0	0	18	0	0
9	0	0	19	0	0
10	1	1	20	1	1

5 CONCLUSIONS

These studies are a new step in improving the road safety system, which aims to achieve zero fatalities on the roads. The complex solution, in the version proposed by the authors of the article, has a direct impact on the management of traffic flows in an accident-free mode, which has not been previously developed or applied, this is the scientific novelty of the work.

Based on the Methodology of Accident Rates by Professor V.F. Babkov, a fundamentally new approach was developed, which was tested on a blackspot of the city of Khabarovsk, as a result of which a high accuracy of convergence of the real situation of injuries in road accidents with the prediction was demonstrated (Figure 2). The reliability of the prognostic scale was 70% (Pugachev I.N., Sheshera N.G., 2020).

To improve the accuracy of the prediction, additional studies were carried out and new parameters were included taking into account the logical justification and experimental observation. For example, the maximum permitted speed was replaced by the maximum provided speed. The intensity of traffic flows was not increased by means of control measurements with subsequent reduction, but by processing actual data from software and hardware systems. The number of classes of the dependent feature – injuries – was increased. In the first study it included: 0 – no victims and 1 – with victims; in the second study there are already three: 0 – no victims, 1 – with victims and 2 – fatalities.

Recent studies on the influence of SRN parameters have been carried out using machine learning – the Random Forest Method. After preparing the feature values, a predictive model was trained. It was tested on pre-prepared data that was not used in the training. The result is shown in Table 1. Out of 20 features, the model correctly classified 19. This indicates the high quality of the prognostic scale.

The second study gives a much better result than the first one, but this does not mean that it should be completely excluded. The two approaches have fundamentally different methods of application. In the first case, it is enough to have a pen and a piece of paper on hand, which simplifies the task of assessing road safety in the field (Pugachev I.N., Sheshera N.G., 2016). In the second case, computing technology is required, but this limitation is offset by high precision and greater detail.

The completed work solves a large number of tasks related to the design, construction and operation of transport facilities, increasing road safety, for the solution of which an algorithmic toolkit has been created, based on the theoretical results obtained and software prototypes of their main components, methods and analytical algorithms.

The full code of machine learning – the Random Forest Method and the results of the study can be found at: https://github.com/glowfisch8lan/3691308f2a4c2f6983f2880d32e29c84.git.

REFERENCES

Rg.ru. (2023, 18 noyabrya). Ezhegodno v avtomobil'nyh avariyah gibnet bolee milliona chelovek: chto mozhet izmenit' pechal'nuyu statistiku? [Online]. Dostupno: https://rg.ru/2023/11/18/ezhegodno-v-avtomobilnyh-avariiah-gibnet-bolee-milliona-chelovek-chto-mozhet-

izmenit-pechalnuiu-statistiku.html. [Data dostupa: 28 noyabrya 2021 g.].
The future of the transport and logistics complex in the light of the initiated transport strategy of the Russian Federation until 2030 with a forecast for the period up to 2035 / Pugachev, I., Furman, B., Umanets, I. // AIP Conference Proceedings , 2023, 2624(1), 040033.
Pugachev, I., Kulikov, Y., Markelov, G., Ostapenko, A., 2021. Peculiarities of strategic transport development in the Russian Far East and the Arctic. Transportation Research Procedia, 57, stranicy 511–517.
Babkov, V. F., 1993. Dorozhnye usloviya i bezopasnost' dvizheniya. M.: Transport, 271 s.
Pugachev, I.N., Lopashuk, V.V., Kulikov, Y.I., Vasilev, A.Y., Barsukova, N.V., 2021. Qualimetric Road Assessment in Strategic Development. Lecture Notes in Civil EngineeringEta ssylka otklyuchena., 150 LNCE, stranicy 108–114
Pugachev I.N., Kamenchukov A.V., Kapskij D.V., Kot E.N., Burtyl' YU.V., SHCHeglov V.I., 2022. Upravlenie bezopasnost'yu dorozhnoj infrastruktury. Habarovsk.
Pugachev, I., Kulikov, Y., Markelov, G., Sheshera, N., 2017. Factor Analysis of Traffic Organization and Safety Systems.12th International Conference «Organization and Traffic Safety Management in large cities», SPbOTSIC-2016, 28-30 September 2016, St. Petersburg. P. 529-535.
Ekspluataciya avtomobil'nyh dorog : uchebnoe posobie / sost.: I.N. Pugachev, A.V. Kamenchukov, N.S. Nesterova. – Habarovsk : Izd-vo DVGUPS, 2022. – 168 s.
Pugachev, I.N., Skripko, P.B., SHeshera, N.G., 2023. Programmnyj podhod k kompleksnomu sboru i podgotovki dannyh ob intensivnosti dvizheniya transportnyh sredstv, pogodnyh uslovij i estestvennoj osveshchennosti v chasovyh intervalah. T-Comm: Telekommunikacii i transport. 2023. Tom 17. №10. S. 43-51.
Rudenko, N.V., 2022. Harakteristiki dorozhnogo dvizheniya po avtomobil'nym dorogam i metody ocenki bezopasnosti dorozhnogo dvizheniya. *Mezhdunarodnyj zhurnal informacionnyh tekhnologij i energoeffektivnosti.* T. 7. № 3-3 (25). S. 100-106.
Batishcheva, O.M., Ganichev, A.I., Starikova, A.G., 2023. Faktory, vliyayushchie na obespechenie bezopasnosti dorozhnogo dvizheniya v krupnyh gorodah. Vestnik transporta Povolzh'ya. № 2 (98). S. 110-115.
Scikit-learn. Oficial'nyj sajt. [Online]. Dostupno: https://scikit-learn.org/stable/. [Data dostupa: 27.03.2024 g.].
Baklanova, K.V., 2023. Vliyanie parametrov transportnogo potoka i harakteristik dorog na bezopasnost' dorozhnogo dvizheniya. Intellekt. Innovacii. Investicii. № 2. S. 99-110.
Mosin, K.K., Kovalevskij, V.E., ZHukova, N.A., 2023. Ispol'zovanie metodov avtomatizirovannogo mashinnogo obucheniya dlya klassifikacii dorozhno-transportnyh proisshestvij. Intellektual'nye tekhnologii na transporte. № 2 (34). S. 15-26.
Kononchuk, E.V., Ermolenko, T.V., SHishunov, T.O., 2022. Modeli mashinnogo obucheniya dlya ocenki veroyatnosti poyavleniya DTP i ego ser'eznosti. Problemy iskusstvennogo intellekta. № 1 (24). S. 4-12.
Arutyunyan, M.A., 2023. Razrabotka algoritmicheskogo apparata po obespecheniyu bezopasnosti dorozhnogo dvizheniya. *Modelirovanie, optimizaciya i informacionnye tekhnologii.* T. 11. № 3 (42). S. 9-10.
Pugachev, I. N., SHeshera, N. G., SHCHeglov, V. I., 2021. Analiz geometricheskih elementov dorog s pomoshch'yu sovremennyh geoinformacionnyh sistem pri ocenke ih avarijnosti. *Vestnik grazhdanskih inzhenerov.* № 3 (86). S. 127–133.
Pugachev, I. N., SHeshera, N. G., 2020. Primenenie metodov statisticheskogo analiza dlya ocenki parametrov transportnyh potokov i harakteristik ulichno-dorozhnoj seti. Dal'nevost. yurid. in-t MVD Rossii. Habarovsk : RIO DVYUI MVD Rossii. 108 s.
Pugachev, I. N., SHeshera, N. G., 2016. Primenenie metodiki koefficientov travmatizma v celyah kontrolya kachestva budushchih i ekspluatiruemyh dorog. *Kachestvo i zhizn'*. № 1 (9). S. 58–61.
GitHub. Repozitorij glowfisch8lan/3691308f2a4c2f6983f2880d32e29c84. [Online]. Dostupno: https://github.com/glowfisch8lan/3691308f2a4c2f6983f2880d32e29c84.git. [Data dostupa: 28 noyabrya 2021 g.].