Crime Prediction and Forecasting: Feature Selection

and Vulnerable Region Detection Models

by

Galym Bekmaganbet

Submitted to School of Engineering and Digital Sciences in partial fulfillment of the requirements for the degree of Masters of Data Science

at the

NAZARBAYEV UNIVERSITY

April 2021

Author ……

Galym Bekmaganbet

School of Engineering and Digital Science

April, 2021

Certified by ……

Adnan Yazici

Full Professor

Thesis Supervisor

Certified by ……

Enver Ever

Associate Professor

Thesis Co-supervisor

Accepted by ……

Crime Prediction and Forecasting: Feature Selection

and Vulnerable Region Detection Models

by

Galym Bekmaganbet
Submitted to School of Engineering and Digital Sciences
on April 2021, in partial fulfillment of the requirements
for the degree of Master of Science in Data Science

Abstract

Crime is one of the most negatively affecting destructive factor for society. The efforts of law enforcement bodies are mostly oriented to determine the criminals post factum. However, in order to reduce the crime growth tendency proactive measures are essential. Therefore, constructing an effective crime sensitive region prediction model along with identifying proper features (factors) would concentrate the efforts of governmental bodies on most vulnerable areas.

The objective of this research is to apply a suitable machine learning algorithm on crime, economic and social data to predict the likelihood of particular regions having low or high crimes levels with further defining main social and economical factors that correlate with crime growth in order to assist not only law enforcement bodies but whole governmental programs to solve related issues and improve crime prevention measures.

In current work the most accurate prediction models were compared and investigated. Tests on available open source data were made and acquired models were applied to available data from Kazakhstani officials.

During evaluation two main issues were faced: inconsistency and inadequacy of data. Consequently, data collection, exploration, preprocessing and normalization were significant steps.

Furthermore, the number of popular models with efficient methodology were compared, combined and the one, that proved to be appropriate for Kazakhstani situation was figured out.

Main prediction models based on Classification, Regression and Clustering techniques: Decision Tree, Random Forest, Naïve Bayesian, K-means, Support Vector Machine algorithms were selected.

They were tested applying both - data available from opensource materials and collected from Kazakhstani state bodies. As a result of tuning parameters and testing various types of feature selection techniques Random Forest model proved to be the most accurate (UCI Repository materials, Accuracy: 0.837, Precision: 0.884, Recall: 0.872, F1 score: 0.868) among listed models, whereas Decision Tree achieved the best result on Kazakhstani data (govstat.kz materials, Accuracy: 0.781, Precision: 0.801, Recall: 0.767, F1 score: 0.784).

Furthermore, statistical analysis were performed to define an appropriate threshold for classifying the high and low crime rate groups.

At final stage hypothesis of importance of a certain feature was tested and model proved that this feature correlates with target (crime rate) and its inclusion positively affected the accuracy of result. Therefore, it can be claimed that the more we acquire expertise in the field of important features, the better selected model will perform.

Thesis Supervisor: Adnan Yazici

Title: Full Professor

Thesis Co-supervisor: Enver Ever

Title: Associate Professor

Acknowledgements

I am delighted to express my gratitude to Professor Adnan Yazici for his assistance in conducting my Thesis work. He dedicated his time for my researches to be qualitative.

Besides, I am sincerely grateful to Professor Enver Ever for his help and being always available for contact and for his valuable feedback.

Also, I would like to thank Arkadiy Son for his support in forming, cleaning and preprocessing the Kazakhstani dataset from scratch for current research.

Contents

List of figures

List of Tables

Chapter 1

1. Introduction

Crime is a one of the most predominant and alarming aspects in any society and its prevention is a vital task for every governmental structure. Throughout the history civil societies attempted to determine appropriate solution to crime prediction issue and variety of approaches have been investigated and implemented.

The efforts of law enforcement bodies are frequently directed towards detecting criminals post factum, without determining specific patterns and anomalies preceding the occurred events. According to criminology principles, it is highly recommended to organize proactive activities and measures to keep the community secure. Therefore, the obvious demand for advanced systems and approaches that allow the prevention of crime constantly persists.

1.1 Motivations

Understanding the causes of crime is a longstanding issue in researcher's agenda. While it is a hard task to extract causality from data, several linear models have been proposed to predict crime through the existing correlations between crime and other statistical metrics (social, economic etc.). However, because of non-Gaussian distributions and multicollinearity in indicators, it is common to find controversial conclusions about the influence of some urban indicators on crime. Machine learning ensemble-based algorithms can handle well such problems.

A variety of researches have been made to define the most accurate and effective models of crime prediction. As the number of committed crimes permanently increases, the more data is available, so that dealing with them is very challenging. In particular, issues arise as to how to choose accurate techniques for analyzing data due to the inconsistency and inadequacy of these kinds of data. These issues motivate scientists to conduct research on these kinds of data to enhance crime data analysis [1].

So far this kind of researches have not been carried out using social, economic and crime data of Kazakhstan. Therefore it was decided to look into currently widespread techniques, tune and ensemble them with further application on Kazakhstani dataset in order to determine paramount model and foremost features.

1.2 Related work

According to [1] crime analysis is a systematic way of detecting and investigating patterns and trends in crime. The results of testing of three clustering approaches - K-Means clustering, Agglomerative clustering and Density Based Spatial Clustering with Noise (DBSCAN) to analyze the crime data of Tamilnadu (from National Crime Records Bureau (NCRB) of India) displayed that the DBSCAN clustering gives result with high accuracy and effectively forms clusters than the other two algorithms.

In their recent researches [2] attempted to construct a model based on repeat and near repeat victimization, which is based on analysis of repetitive behavior of criminals. Authors' approach contains interesting technique, such that dividing data into prior (capturing historical trends captured during past year) and post prior (capturing recent trends within last two month) sets and generate a spatial model by using Gaussian Mixture Models. Moreover, in this research it is suggested to combine results of different models (Dempster-Shafer theory method and Multikernel method) by setting the results to share the same domain and with further integration of components. For example, Dempster-Shafer theory method gives results of high level risky regions with no upper bound, whereas multikernel method results in probability. Authors integrated this two results by setting the results to classify every region (in this case every cell) in five levels using natural breaks and numbers from 1 to 5 and visualized it by putting on map. This model was accepted by Chilean police field use. However the project had some constraints due to lack of required data and restriction of predicting mostly burglary crimes.

This approach can be implemented to combine best models as [1], [4] and [7] and predict regions with high crime risk levels.

Crime data usually has non-Gaussian distribution and multicollinearity with urban data. This factors negatively affect the accuracy of linear models. Authors of [3] tested a

random forest regressor to predict crime and quantify the influence of urban indicators on homicides. Results of experiments illustrated that their approach can have up to 97% of accuracy on crime prediction, and the importance of urban indicators is ranked and clustered in groups of equal influence, which are robust under slightly changes in the data sample analyzed. Moreover, results determine the rank of importance of urban indicators to predict crime, unveiling that unemployment and illiteracy are the most important variables for describing homicides in Brazilian cities.

Review of related work displays that many authors claim Decision Tree to be effective in crime prediction.

For example [4], [5] and [6] made experiments and compared different types of models, among which decision tree models proved to have highest accuracy.

[4] study considered the development of crime prediction prototype model using decision tree (J48) algorithm because it has been considered as the most efficient machine learning algorithm for prediction of crime data as described in the related literature. From the experimental results, J48 algorithm predicted the unknown category of crime data to the accuracy of 94.25287% which is fair enough for the system to be relied on for prediction of future crimes.

[5] authors compared the two different classification algorithms namely, Naïve Bayesian and Decision Tree for predicting 'Crime Category' for different states in USA. The results from the experiment showed that, Decision Tree algorithm outperformed Naïve Bayesian algorithm and achieved 83.9519% Accuracy in predicting 'Crime Category' for different states of USA.

Taking into consideration the fact that previously Naïve Bayesian algorithm was accepted by some researches to be one of the most effective crime prediction models, outperformance of Decision Tree technique is a matter to be highlighted.

For instance, in [7] authors have applied their models on the same 'Crime Category' data and proved that in comparison to Back Propagation approach, Naïve Bayesian had better results with the accuracy level of 94.0822%.

In above mentioned researches authors' methods are based on either background historical knowledge or offenders' profiling, in [8] authors present a novel approach to predict crime in a geographic space from multiple data sources, in particular mobile phone and demographic data. The main contribution of the proposed approach lies in using aggregated and anonymized human behavioral data derived from mobile network

activity to tackle the crime prediction problem. Experimental results with real crime data from London displayed an accuracy of almost 70% when predicting whether a specific area in the city will be a crime hotspot or not. Although we cannot use mobile data in our research, this approach should be taken into consideration in future works of the project.

Another interesting factor that affect to crime rate is suggested by[9], in which besides regions with high crime density ('hotspots'), and forecasting based on historical data, author additionally used data derived from Web and Social Media.

For examples, the regions which contained the highest internet connections to pornographic sites, had higher number of sex crimes as well.

Additionally, post of social media can also be included into analytical and forecasting process (if data is available).

Finally, we test the method proposed by [10] and implement K-means algorithm by partitioning data into groups based on their means. K-means algorithm has an extension called expectation - maximization algorithm where we partition the data based on their parameters. This easy to implement data mining framework works with the geospatial plot of crime and helps to improve the productivity of the detectives and other law enforcement officers.

1.3 Significance

In current research, we compared the number of popular models with efficient methodology, ensembled them and figured out the best fit for Kazakhstani society.

Initial experiments were carried out using data from UCI Repository "community-crime-data", which already contained clean dataset.

Main prediction models based on Classification, Regression and Clustering techniques:

I. Decision Tree Classification

II. Random Forest Classification

III. Naïve Bayesian

IV. K-means

V. Support Vector Machine

Metrics of accuracy [11] that measure the performance of every model give percentage of features that are forecasted properly among total number of features:

- Precision [11] that is calculated as number of positive features classified by the model that are positive;
- Recall [11] that gives number of positive features classified correctly by the model;
- F1-score that is harmonic mean of mentioned Precision and Recall.

At every step of research, besides prediction and forecasting constituent importance of features were inspected.

Furthermore, experiments were followed by conducting CRISP-DM process model to understand the business processes of Kazakhstani law enforcement and governmental structures, collecting proper information to form a dataset, aggregate, clean and preprocess the data to apply the listed above selected models and as a result acquire best crime prediction models with the most important features that affect the increase in crime rate in Kazakhstan.

At the final stage hypothesis of importance of a certain feature was tested and the model proved that this feature correlates with target (crime rate) and its inclusion positively affected the accuracy of result. Therefore, it can be claimed that the more we acquire expertise information in the field of important features, the better selected model will perform.

1.4 Outline

Chapter 2: defines different approaches implemented for crime prediction and feature selection models, compares advantages and disadvantages of them, and explores various methods of parameter-tuning in order to optimize the efficiency of selected models.

Chapter 3: displays experimental results after implementation of parameter-tuning in to optimize the efficiency of selected models and compares the derived results of the experiment.

Chapter 4: explains implementation of the whole cycle of research stage by stage according to CRISP-DM model process to demonstrate how data set of Kazakhstani crime statistics was formed with further implementation of different prediction and feature selection models. Illustrates the results of hypothesis testing by adding a new feature to dataset, which is considered to be significant factor that causes immense crime rate.

Chapter 5: reviews the outcome of models and experiments and the effectiveness of a suggested alternative solution.

Chapter 2

2. Main models and metrics

2.1 Machine learning

Machine Learning is a brunch of AI which defines the patterns by analyzing (big) data. According to [12] ML techniques based on machine learning have been applied successfully in diverse fields ranging from pattern recognition, computer vision, spacecraft engineering, finance, entertainment, and computational biology to biomedical and medical applications.

Computational machines are capable to "learn" with further predictions based on derived data by using machine learning, and they do not require explicit programming. It is common to divide Machine learning into the following major categories:

- Supervised Learning
- Semi-supervised Learning
- Unsupervised Learning
- Reinforcement Learning.

In current research, we basically investigate supervised learning techniques to predict crime categories and select main features which correlate with target value.

2.2 Supervised Learning

Main difference between Supervised and Unsupervised learning is: former one requires specifically defined labeled dataset, whereas the latter can handle with unlabeled datasets to make predictions. Besides, supervised learning input object contains various number of features and usually is represented in a vector form.

Supervised learning models can be implemented on both classification and regression problems. In our research we are aimed to predict and classify whether particular region will have an increase in crime rate or not by using the crime dataset. As the crime categories are discontinuous, this is a supervised classification problem. There are different types of supervised classification models. In this research Decision Tree Classification, Random Forest Classification, Naïve Bayesian, K-means, Support Vector Machine models are used.

Decision Tree Classifier – according to [13] the most important feature of Decision Tree Classifier is their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret. It is considered to be non-parametric supervised learning technique and can be implemented in both - classification and regression solutions. As it is obvious from name of the model tree displays a piecewise constant approximation.

Gaussian Naive Bayes – calculates the probability that a given instance belongs to a certain class or group. For example, for an instance X, described by its feature vector (x1,…, xn), and a class target y, Bayes' theorem allows us to express the conditional probability P(y|X) as a product of simpler probabilities using the naïve independence.

One typical way to handle continuous attributes in the Naive Bayes classification is to use Gaussian distributions to represent the likelihoods of the features conditioned on the classes [14].

K-means – according to [15] k-means algorithm is an incremental approach to clustering that dynamically adds one cluster center at a time through a deterministic global search procedure consisting of *N* (with *N* being the size of the data set) executions of the *k*-means algorithm from suitable initial positions.

Briefly, k-mean is a technique of vector quantization, which is based on signal processing and its goal is to partition number (n) of observations in some (k) clusters with further determination of belonging of every observation to the cluster with the closest mean (centroid), which is denoted as a prototype of the cluster. Thus data is divided into clusters, that are partitioned to elements grouped to centroids regarding the closest one with respect to squared Euclidean distances.

Support Vector Machine - is a supervised ML method that is capable for both classification and regression case. Nevertheless, SVM is frequently implemented in classification issues. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Random Forest – is a composite model of numerous decision trees, which are built using samples drawn with reinstatement from the training dataset. Unlike decision tree, the splitting of every node is based on not on the best split of all features, but selecting the best split among a random set of features. The bias of the tree rises as a result of

random state, whereas taking mean assists to lessen the variance as well, so that this model generally achieves more advantageous outcome.

2.3 Performance Metrics

It is important to choose proper metrics to evaluate how well an algorithm is performing.

In our research four metrics are used to measure and compare performance of different models.

Accuracy - is a ratio of correctly predicted observation to the total observations.
Its formula is denoted as:

Accuracy = True Positive +True Negative / True Positive + False Positive + False Negative + True Negative.

Precision - is the ratio of correctly predicted positive observations to the total predicted positive observations.
Its formula is denoted as:

Precision = True Positive / True Positive + False Positive.

Recall (or Sensitivity) - is the ratio of correctly predicted positive observations to the all observations in actual class – Yes (or True).
Its formula is denoted as:

Recall = True Positive / True Positive + False Negative.

F1 score - is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

Its formula is denoted as:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

2.4 Overview of the Dataset

In current Thesis at this stage it was decided to use dataset from a UCI Repository – "Communities and Crime dataset" that consists of 1993 instances and 104 attributes of social and economic statistic information, as well as crime data.

Description is as follows (Fig.1):

t[2]:

	state	communityname	fold	population	householdsize	racepctblack	racePctWhite	racePctAsian	racePctHisp	agePct12t21	...	PctForeignBorn	PctBornSa
0	1	Alabastercity	7	0.01	0.61	0.21	0.83	0.02	0.01	0.41	...	0.03	
1	1	AlexanderCitycity	10	0.01	0.41	0.55	0.57	0.01	0.00	0.47	...	0.00	
2	1	Annistoncity	3	0.03	0.34	0.86	0.30	0.04	0.01	0.41	...	0.04	
3	1	Athenscity	8	0.01	0.38	0.35	0.71	0.04	0.01	0.39	...	0.03	
4	1	Auburncity	1	0.04	0.37	0.32	0.70	0.21	0.02	1.00	...	0.12	

5 rows × 104 columns

Fig.1 Description of Communities and Crime dataset

The attributes are considered to be actual and of multivariate characteristics. Dataset already contained uncleaned 'dirty' data and preprocessed (with reduction of null and empty values) 'clean' data (Fig. 2).

In [16] authors used both clean data (removal of missing values was needed to get an appropriate crime data set) and dirty data.

```
[(0, 'state'), (1, 'communityname'), (2, 'fold'), (3, 'population'), (4, 'householdsize'), (5, 'racepctblack'), (6,
'racePctWhite'), (7, 'racePctAsian'), (8, 'racePctHisp'), (9, 'agePct12t21'), (10, 'agePct12t29'), (11, 'agePct16t2
4'), (12, 'agePct65up'), (13, 'numbUrban'), (14, 'pctUrban'), (15, 'medIncome'), (16, 'pctWWage'), (17, 'pctWFarmSel
f'), (18, 'pctWInvInc'), (19, 'pctWSocSec'), (20, 'pctWPubAsst'), (21, 'pctWRetire'), (22, 'medFamInc'), (23, 'perCap
Inc'), (24, 'whitePerCap'), (25, 'blackPerCap'), (26, 'indianPerCap'), (27, 'AsianPerCap'), (28, 'OtherPerCap'), (29,
'HispPerCap'), (30, 'NumUnderPov'), (31, 'PctPopUnderPov'), (32, 'PctLess9thGrade'), (33, 'PctNotHSGrad'), (34, 'PctB
SorMore'), (35, 'PctUnemployed'), (36, 'PctEmploy'), (37, 'PctEmplManu'), (38, 'PctEmplProfServ'), (39, 'PctOccupMan
u'), (40, 'PctOccupMgmtProf'), (41, 'MalePctDivorce'), (42, 'MalePctNevMarr'), (43, 'FemalePctDiv'), (44, 'TotalPctDi
v'), (45, 'PersPerFam'), (46, 'PctFam2Par'), (47, 'PctKids2Par'), (48, 'PctYoungKids2Par'), (49, 'PctTeen2Par'), (50,
'PctWorkMomYoungKids'), (51, 'PctWorkMom'), (52, 'NumIlleg'), (53, 'PctIlleg'), (54, 'NumImmig'), (55, 'PctImmigRecen
t'), (56, 'PctImmigRec5'), (57, 'PctImmigRec8'), (58, 'PctImmigRec10'), (59, 'PctRecentImmig'), (60, 'PctRecImmig5'),
(61, 'PctRecImmig8'), (62, 'PctRecImmig10'), (63, 'PctSpeakEnglOnly'), (64, 'PctNotSpeakEnglWell'), (65, 'PctLargHous
eFam'), (66, 'PctLargHouseOccup'), (67, 'PersPerOccupHous'), (68, 'PersPerOwnOccHous'), (69, 'PersPerRentOccHous'),
(70, 'PctPersOwnOccup'), (71, 'PctPersDenseHous'), (72, 'PctHousLess3BR'), (73, 'MedNumBR'), (74, 'HousVacant'), (75,
'PctHousOccup'), (76, 'PctHousOwnOcc'), (77, 'PctVacantBoarded'), (78, 'PctVacMore6Mos'), (79, 'MedYrHousBuilt'), (8
0, 'PctHousNoPhone'), (81, 'PctWOFullPlumb'), (82, 'OwnOccLowQuart'), (83, 'OwnOccMedVal'), (84, 'OwnOccHiQuart'), (8
5, 'RentLowQ'), (86, 'RentMedian'), (87, 'RentHighQ'), (88, 'MedRent'), (89, 'MedRentPctHousInc'), (90, 'MedOwnCostPc
tInc'), (91, 'MedOwnCostPctIncNoMtg'), (92, 'NumInShelters'), (93, 'NumStreet'), (94, 'PctForeignBorn'), (95, 'PctBor
nSameState'), (96, 'PctSameHouse85'), (97, 'PctSameCity85'), (98, 'PctSameState85'), (99, 'LandArea'), (100, 'PopDen
s'), (101, 'PctUsePubTrans'), (102, 'LemasPctOfficDrugUn'), (103, 'ViolentCrimesPerPop')]
```

Fig.2 Attributes of Communities and Crime dataset

As a base model we selected the experiment of [16] with some alteration in principles of defining classification threshold and parameters of each model.

Based on previous experiments we added a new feature and named it as 'highCrime', which was calculated from feature 'ViolentCrimesPerPop' and gave it a value of True and False.

Authors in [16] gave a value '1' for 'ViolentCrimesPerPop' greater than 0.1 and '0' otherwise, which meant that they had a classification threshold of 0.1.

They decided the threshold of 0.1upon manual analysis of data by view-through process. In our research it was decided to investigate optimal ways to define a proper threshold for classification.

For this purpose we explored the distribution of 'ViolentCrimesPerPop' and in case it is normally distributed we could have used threshold as mean = 0.24.

Nevertheless visualization of graph displayed that the data is not normally distributed and more to the point the data is turned out to be more saturated towards 0. Hence taking mean as threshold is not a solution, so it was decided to declare median =0.15 as a threshold value.

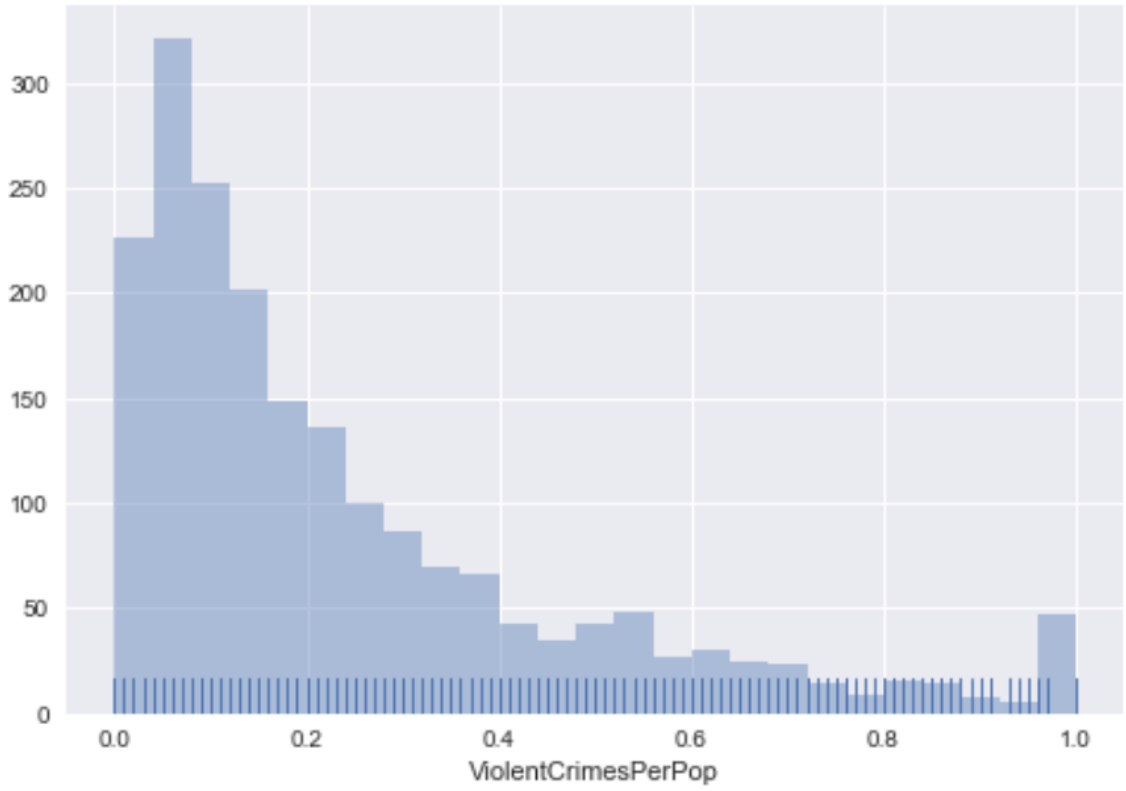

Fig. 3. Distribution of 'ViolentCrimesPerPop'

Further on, we defined a new added 'High crime' column as a target feature on which our proposed models would be based on.

Using Python pandas data frame the dataset was converted into appropriate data frame and the target feature ‘High crime’ was assigned to a variable ‘Target’ and the remaining features to ‘Features’.

Chapter 3

3. DESIGN AND IMPLEMENTATION

3.1 DECISION TREES

The dataset was divided into training and testing parts and in order to predict target column we designed a decision tree model. The splitting criterion remained default ('Gini'), however at the end of experiment we tested changing it to 'Entropy', which turned out to add no change into the acquired results.

The Gini impurity measures the frequency at which any element of the dataset will be mislabelled when it is randomly labeled.

According to [17], the entropy of a variable can be defined as, $\sum$ -P * log P, for the probability P of that variable taking values of 0,1,2.... n and, the sum of probabilities of all variables is 1. Dingsheng Wan [18] mentions in his paper that, smaller the value of entropy, better can the system be described.

To determine best depth for a decision tree we iterated through different depth and figured out optimal depth value for model's best performance and found the maximum depth of 3.

To avoid overfitting we also applied 10-fold cross-validation as initially it was obvious that the model was overfitting by demonstrating very high value of accuracy rather than in [16].

Moreover, taking into consideration the fact that one of the main aims of our experiment was determination of important features, we checked main features for classification and ranked them by their feature importance.

As a result we derived following outputs which are divided into base model results and results after making changes into model by altering the threshold and tuning the parameters of depth and cross-validation.

Evaluating Measure Decision Tree Classifier	10-fold Cross-Validation (%)
Accuracy	75.9%
Precision	80.62%
Recall	81.53%
F1 score	81,07%

Table 1: Base model Performance Measures- Decision Trees

Evaluating Measure Decision Tree Classifier	10-fold Cross-Validation (%)
Accuracy	79,8%
Precision	84,3%
Recall	83,9%
F1 score	83,6

Table 2: Performance Measures after model tuning- Decision Trees

The top 10 features extracted according to the feature importance scores were: PctKids2Par, 'racePctWhite', 'racePctHisp', 'HousVacant', 'LemasPctOfficDrugUn', 'PctEmplProfServ', 'NumUnderPov', 'PctPopUnderPov', 'PctLess9thGrade', 'PctNotHSGrad'.

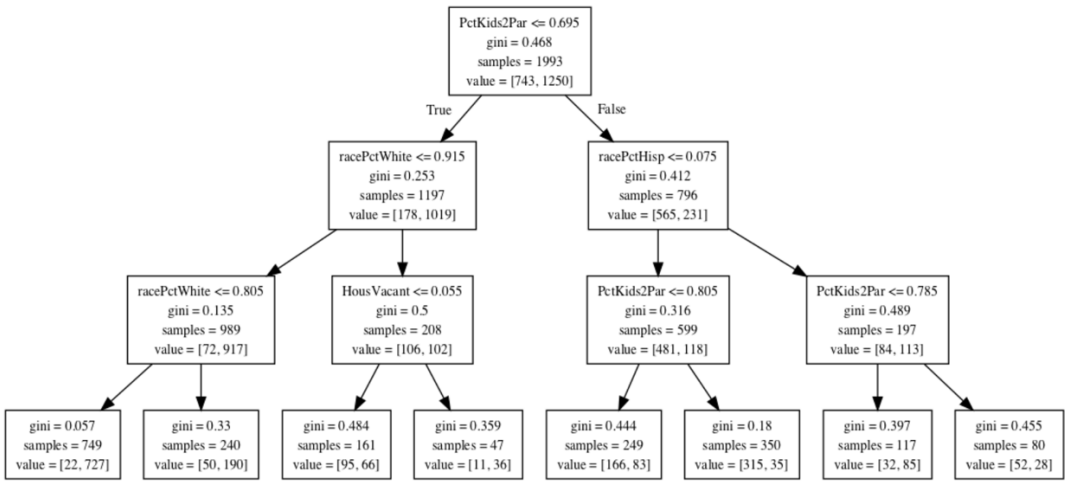

Fig.4 Tree of main features

Moreover, in order to reduce the multicollinearity we analyzed the correlation matrix and excluded the attributes that had high correlation with each other.

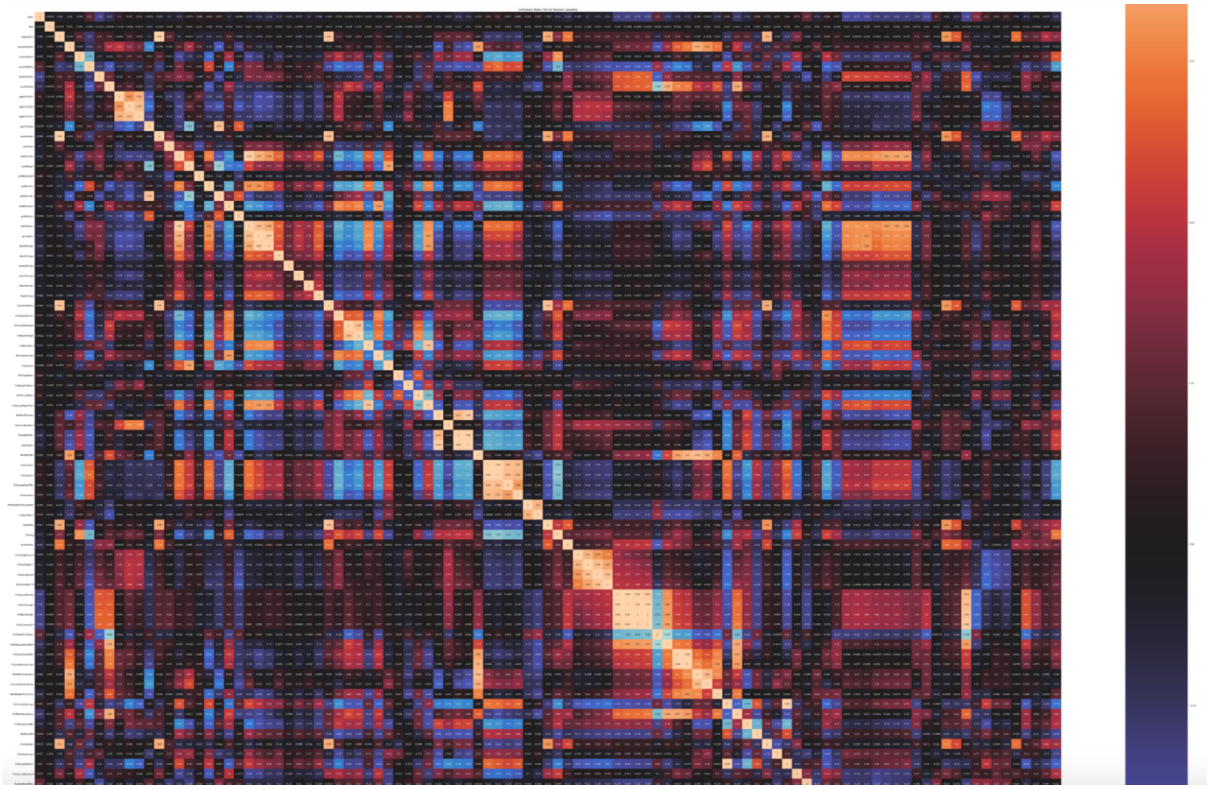

Fig. 5 Correlation matrix of Communities crime dataset

Our model of Decision Tree Classifier selected 'PctKids2Par' as its main implying the major predictive feature for 'High crime'.

Using Gini as a splitting criterion, the remaining features were extracted. Besides, performance of model using entropy as a splitting criterion had no effect on accuracy, precision, recall and f1 score, so it was decided to remain at default (Gini) criterion.

3.2 RANDOM FOREST CLASSIFICATION

Random Forests Classifiers adjust the Decision Trees' manner of overfitting while dealing with training data. Likewise in Decision Tree model Gini criterion was used as an impurity measure to form trees with in the Random Forest model.

Evaluating Measure Random Forest Classifier	10-fold Cross-Validation (%)
Accuracy	83.39%
Precision	88.30%
Recall	84.86%
F1 score	86,54

Table 3: Base model Performance Measures- Random Forest Classifier

Evaluating Measure Random Forest Classifier	10-fold Cross-Validation (%)
Accuracy	83,7%
Precision	88,4%
Recall	87,2%
F1 score	86,83%

Table 4: Performance Measures after model tuning - Random Forest Classifier

The top 10 features extracted according to the feature importance scores were:

'PctKids2Par',

'PctIlleg',

'racePctWhite',

'PctPersDenseHous',

'FemalePctDiv',

'TotalPctDiv',

'PctFam2Par',

'NumUnderPov',

'NumIlleg',

'PctTeen2Par'

Since the accuracy, precision and recall values for Random Forest Classifier is almost same on both base model and tuned model, this model perfectly fits for median

values for both with and without missing values. Also, the top features like 'PctKids2Par': Percentage of kids in family housing with two parents and 'racePctWhite': Percentage of population that is Caucasian are common to the decision trees all together.

3.3 NAÏVE BAYES CLASSIFICATION

Naïve Bayes Classification uses Bayes theorem [19] that describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It is given as:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

where P(A|B) = Probability of A given B is true P(B|A) = Probability of B given A is true P(A) and P(B) = Probability of A and B respectively A: Target and B: Remaining Features.

To construct this model in Python we implemented GaussianNB Classifier which is based on Naïve Bayes theorem to make prediction and find best features ranked according to their importance. To reduce overfitting, cross validation was used to measure accuracy, precision, recall and f1 score. The initial steps of taking target feature in Target and other features is a usual construction for each of the listed in current thesis model and method.

Evaluating Measure Naïve Bayes Classifier	10-fold Cross-Validation (%)
Accuracy	77.64 %
Precision	92.53 %
Recall	69.82 %
F1 score	79.58 %

Table 5: Baseline model Performance Measures- Naïve Bayes Classifier Clean Data

Evaluating Measure Naïve Bayes Classifier	10-fold Cross-Validation (%)
Accuracy	77,8 %
Precision	92,6 %
Recall	70,2 %
F1 score	79,85

Table 6: Performance Measures after model tuning - Naïve Bayes Classifier Clean Data

The top 10 features extracted according to the feature importance scores were:
'PctKids2Par', 0.8093364216318364
'PctFam2Par', 0.74516152011997
'racePctWhite', 0.7348840522379364
'PctIlleg', 0.7089291060645266
'FemalePctDiv', 0.6936040623575482
'TotalPctDiv', 0.6742823162675443
'PctYoungKids2Par', 0.6646705535981033
'pctWInvInc', 0.6607203219208594
'PctTeen2Par', 0.6426208020335665
'MalePctDivorce', 0.6165342437579264

3.4 K-MEANS

Besides mentioned above three algorithms that were selected as base models for our researches from [16] it was decided to experiment with few more models.

Taking into consideration that Decision Tree, Random Forest and Naïve Bayesian Classifiers are mostly supervised learning algorithms we tried to construct a model of K-means algorithm, which is considered to be based on unsupervised learning principle.

K-means is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k-means algorithm works is as follows:

- Specify number of clusters *K*.
- Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The objective of the function of k-means algorithm is as follows:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

where wik=1 for data point xi if it belongs to cluster *k*; otherwise, wik=0. Also, μk is the centroid of xi's cluster.

K-means Classifier - Clean Data	10-fold Cross-Validation (%)
Accuracy	53,67 %
Precision	72,07 %
Recall	52,24 %
F1 score	43,91 %

Table 7: Performance Measures- K-means Classifier

3.5 SUPPORT VECTOR MACHINE

Nonlinear classification: SVM can be extended to solve nonlinear classification tasks when the set of samples cannot be separated linearly. By applying kernel functions, the samples are mapped onto a high-dimensional feature space, in which the linear classification is possible.

By using the kernel function, a nonlinear version of SVM can be developed and the expression of the optimization problem for such SVM can be written (in the dual form).

The samples closest to the separating hyperplanes are those whose coefficients *ai* are nonzero. These samples are called support vectors. The support vectors include the necessary information to construct the optimal hyperplane while other samples lay no effects on it. This is the reason why SVM could be used for the classification tasks whose number of samples is limited.

There are different kernel functions used in SVM, like linear, polynomial and Gaussian RBF. In our research we preferred polynomial kernel function as it was appropriate for the performance of SVM, since the kernel defines the high-dimensional space where the samples will be classified. Although it is known that among the kernel functions, Gaussian RBF is the most commonly used in intelligent fault diagnosis.

Non-Linear SVM Classifier	10-fold Cross-Validation (%)
Accuracy	73,85 %
Precision	79,8 %
Recall	79,35 %
F1 score	78,84 %

Table 8: Performance Measures- non-linear SVM Classifier

3.6 COMPARISON OF RESULTS

The baseline models that were selected displayed that Random Forest Classifier has the most balanced outcome with respect to accuracy, precision, recall and F1 scoreout of all listed above models for prediction of ‘highCrime’ feature.

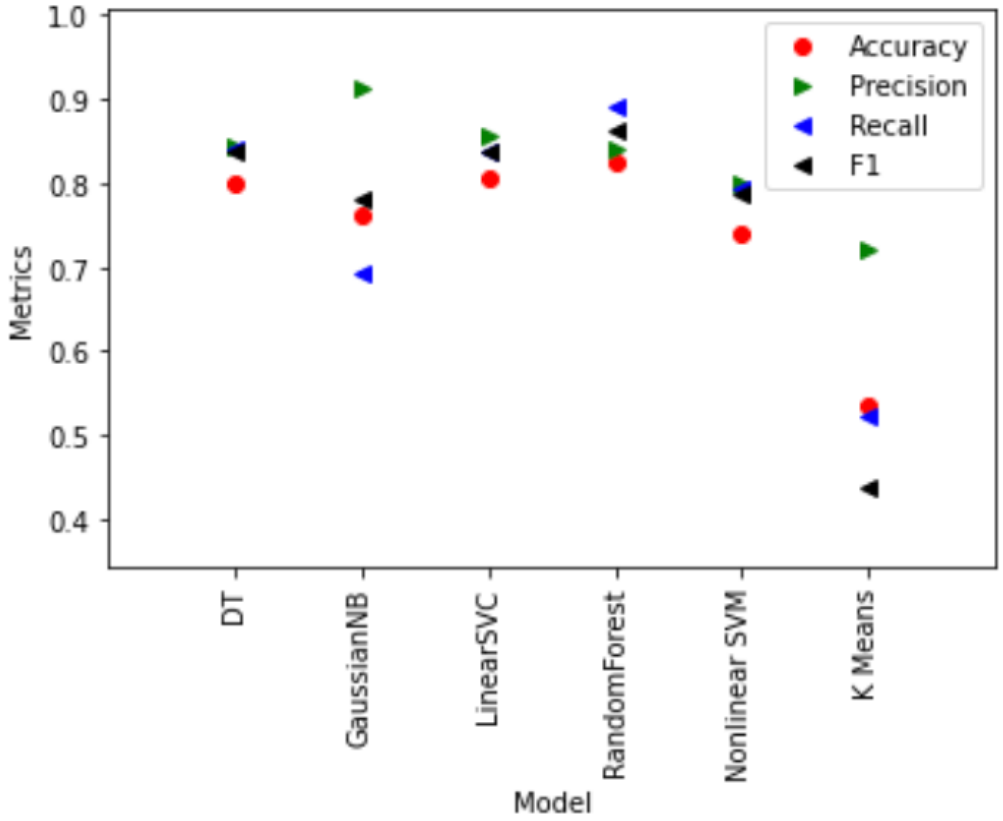

Fig. 6 Comparison of models

While k-means classifier gave the lowest values in these performance measures, the data could not be handled by their centroids and nearest elements as they were already too close to each other so that model had issues of dividing it to appropriate classes.

Random Forest Classifier takes multiple trees into account and gives an average of the result which proved to be perfect for this type of data. Naïve Bayes proved to be a balancing quotient for this crime data as it had values close to the Random Forest Classifier. Some common features having high importance scores that proved to be highly predictive of 'High crime' features are:

'PctKids2Par', 'racePctWhite' were important in all mentioned above models.

'NumUnderPoverty', 'MalePctDivorce', 'PctFam2Par', 'FemPctDiv', 'PctIlleg' are also common features for two or more models

Defining new threshold, optimal depth, reducing the possibility of overfitting by implementing cross validation and removing attributes based on analysis of correlation

matrix table improves performance by enough training and testing samples that seemed to help in this analysis by giving correct and consistent performance measures.

As we notice Random Forest has optimal results rather than other models used in experiment.

Chapter 4

4. CRISP-DM MODEL

In order to apply mentioned above models in prediction of crime rate in Kazakhstan and define the main social or economic feature that correlate with high crime rate it was decided to implement CRISP-DM model process.

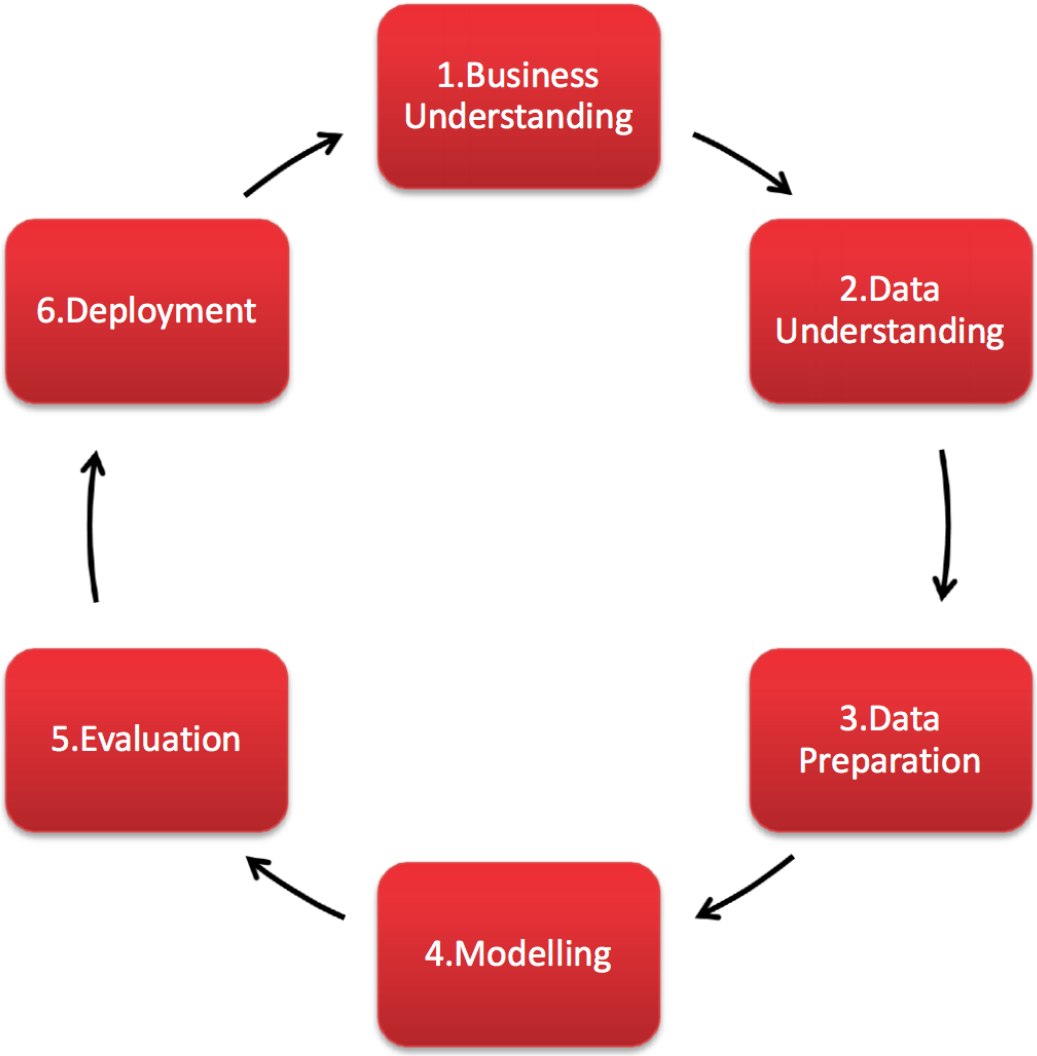

Fig. 7 CRISP-DM model process structure

Due to absence of statistical dataset regarding social and economic attributes along with crime information we collaborated with Kazakhstan governmental and law enforcement structures and followed CRISP-DM steps by understanding the business

process, gathering information about sources of data, preparation of required data (collecting, cleaning, preprocessing), implementing models and evaluating them.

However, decision about deployment of best model is long lasting for approvement and under consideration of management of several governmental bodies. Hence, this stage is a matter of future work.

4.1 BUSINESS UNDERSTANDING

It is well known fact that, along with crime investigation there are a variety of measures that are intended to public safety. Therefore, preventive measures are thought to be the most responsible and effective ones. It is postulated that prevention system that effectively addressed the goal of improving social safety would be uniform, consistent, and focused on identifying characteristics known to affect state well-being. We examined data on social, economic and crime data of Kazakhstani state bodies to determine whether the system there demonstrated such characteristics.

Regionwide listed above attributes were collected from responsible bodies along with exploration of stat.gov.kz website of Statistics Committee of Kazakhstan. We discovered that data from different governmental structures are formed in several different way: manually in MS Word or Excel sheets on regular or occasional basis, CSV files, downloaded from databases, stored in papers in archives, published in media.

Finding one central database or storage for this kind of data was a challenging task so it was decided to form our own dataset containing required historical information.

Moreover, regions are classified to high or low crime rated groups in comparison to each other and threshold did not exists.

On governmental (strategic) level almost all resources are concentrated on main attitude of improving economic indicators and scientific approaches in crime preventions are implemented in rare cases.

4.2 DATA UNDERSTANDING

Data was collected from following state bodies:

1) Ministry of Healthcare
2) Ministry of Education

3) Ministry of Labor and Social Care

4) Ministry of Economics and Industrial Development

5) Ministry of Transport and Communications

6) Attorney-General's office

7) Justice Ministry

8) Statistics Committee

Main issues:

- Same indicators were differently named
- Some regions and periods had missing data
- Databases were stored in different ways (manually in MS Word or Excel sheets on regular or occasional basis, CSV files, downloaded from databases, stored in papers in archives, published in media)
- Due to COVID-19 pandemic situation many responsible people were not available
- Contained wrong or incorrect information (noticed during visual exploration)

We merged all available data, aggregated them by their appropriate and common attribute names, by applying statistical methods and in collaboration with experts in specific fields replaced and filled missing values, as well as cleaning the faulty information. Finally we formed a clean and model-applicable Kazakhstani dataset, that contained:

Social indicators – number of schools, colleges, universities, marriages, divorces, birth, death, migration, population, hospitals, residential buildings, students, children.

Economic indicators – average income, minimal income, maximum income, prices in tenge and USD, employment, salary, gross regional product per 1000 people, industrial products, agricultural products, construction prices, transport and passengers.

Crime information – number of committed crimes.

All listed above information were described yearly and region wise.

4.3 DATA PREPARATION

Using Python Pandas tool dataset was converted into appropriate data frame. Description is as follows (Fig.7, 8):

	region #	city/region	year	population_1000_per	birth_1000_per	death_1000_per	increase_pop_1000_per	migration_1000_per	birth_coef_1000_per	death_coef_1
0	1	NurSultan	1991	298.671	4.801	2.283	2.518	0.769	16.2	
1	1	NurSultan	1992	292.172	4.367	2.330	2.037	9.600	14.8	
2	1	NurSultan	1993	294.556	4.003	2.638	1.365	2.610	13.6	
3	1	NurSultan	1994	293.155	3.608	2.800	0.808	4.979	12.3	
4	1	NurSultan	1995	289.715	3.248	3.112	0.136	4.776	11.1	

5 rows × 63 columns

Fig. 8. Sample of Kazakhstani social, economic and crime data set.

```
[(0, 'region #'), (1, 'city/region'), (2, 'year'), (3, 'population_1000_per'), (4, 'birth_1000_per'), (5, 'death_1000
_per'), (6, 'increase_pop_1000_per'), (7, 'migration_1000_per'), (8, 'birth_coef_1000_per'), (9, 'death_coef_1000_pe
r'), (10, 'marriage_coef_1000_per'), (11, 'divorce_coef_1000_per'), (12, 'hospitals'), (13, 'places_in_hospital'), (1
4, 'kindergarten'), (15, 'child_in_kindergarten_1000_per'), (16, 'schools'), (17, 'students_in_schools_1000_per'), (1
8, 'colleges'), (19, 'students_in_colleges_1000_per'), (20, 'university'), (21, 'students_in_university_1000_per'),
(22, 'crimes'), (23, 'income_ave'), (24, 'people_low_income_pct'), (25, 'min_income'), (26, 'min_income_usd'), (27,
'able_bodied_1000_per'), (28, 'working_1000_per'), (29, 'hired_1000_per'), (30, 'self_emp_1000_per'), (31, 'unemp_100
0_per'), (32, 'registered_unemp_1000_per'), (33, 'official_registered_unemp_1000_per'), (34, 'unemp_level'), (35, 'un
emp_level_youth_15_24'), (36, 'unemp_level_youth_15_28'), (37, 'ave_salary'), (38, 'consumer_price_idx'), (39, 'indus
try_product_price_idx'), (40, 'cunstruction_price_idx'), (41, 'agriculture_product_price_idx'), (42, 'logistics_price
_idx'), (43, 'gross_regional_product'), (44, 'gross_regional_product_usd'), (45, 'gross_regional_product_per_person_1
000'), (46, 'gross_regional_product_per_person_1000_usd'), (47, 'industry_product_volume_mln_tenge'), (48, 'mining_ml
n_tenge'), (49, 'manufactur_industry_mln_tenge'), (50, 'electrecity_gas_aircondition_mln_tenge'), (51, 'water_supply_
mln_tenge'), (52, 'agriculture_gross_output'), (53, 'main_capital_investment_mln_tenge'), (54, 'construction_volume_m
ln_tenge'), (55, 'residential_buildings_sqm'), (56, 'places_in_school'), (57, 'transport_communication'), (58, 'load_
transportation_mln_tons'), (59, 'passenger_transportation_mln_km'), (60, 'passenger_transportation_mln_person'), (61,
'retail_product_sell_mln_tenge'), (62, 'CrimePerPop')]
```

Fig. 9. Attributes of Kazakhstani social, economic and crime data set.

During data cleaning it was decided to implement three main approaches:

1. Replace missing values by mean value of a certain attribute (column).
2. Replace missing values by mean value of closest three regions (for example, for missing values in Pavlodar which is located in northern part of the country, it was decided to put mean values of North-Kazakhstan, Akmola and Kostanay regions as they had similar indicators).
3. Replace missing values by median of a certain attribute (column).
4. Replace missing values by median of closest three regions.
5. In some cases filled null values with the value of subsequent or preceding year (for example, for Nur-Sultan which in some cases contained information only since city foundation date, it was decided to put values of 1998 year).
6. Determining outliers and replacing them according to above described principle.

Implementation of each method of dealing with null or dirty values was decided based on consultation with experts in particular field.

Finally, we acquired an appropriate data set for the eventual testing of selected models.

Data consisted of 62 attributes (while initially contained 175 attributes) and 498 rows.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 498 entries, 0 to 497
Data columns (total 63 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   region #                                    498 non-null    int64
 1   city/region                                 498 non-null    object
 2   year                                        498 non-null    int64
 3   population_1000_per                         498 non-null    float64
 4   birth_1000_per                              498 non-null    float64
 5   death_1000_per                              498 non-null    float64
 6   increase_pop_1000_per                       498 non-null    float64
 7   migration_1000_per                          498 non-null    float64
 8   birth_coef_1000_per                         498 non-null    float64
 9   death_coef_1000_per                         498 non-null    float64
 10  marriage_coef_1000_per                      498 non-null    float64
 11  divorce_coef_1000_per                       498 non-null    float64
 12  hospitals                                   498 non-null    float64
 13  places_in_hospital                          498 non-null    float64
 14  kindergarten                                498 non-null    int64
 15  child_in_kindergarten_1000_per              498 non-null    float64
 16  schools                                     498 non-null    int64
 17  students_in_schools_1000_per                498 non-null    float64
 18  colleges                                    498 non-null    int64
 19  students_in_colleges_1000_per               498 non-null    float64
 20  university                                  498 non-null    int64
 21  students_in_university_1000_per             498 non-null    float64
 22  crimes                                      498 non-null    int64
 23  income_ave                                  498 non-null    float64
 24  people_low_income_pct                       498 non-null    float64
 25  min_income                                  498 non-null    int64
 26  min_income_usd                              498 non-null    float64
 27  able_bodied_1000_per                        498 non-null    float64
 28  working_1000_per                            498 non-null    float64
 29  hired_1000_per                              498 non-null    float64
 30  self_emp_1000_per                           498 non-null    float64
 31  unemp_1000_per                              498 non-null    float64
 32  registered_unemp_1000_per                   498 non-null    float64
 33  official_registered_unemp_1000_per          498 non-null    float64
 34  unemp_level                                 498 non-null    float64
 35  unemp_level_youth_15_24                     498 non-null    float64
 36  unemp_level_youth_15_28                     498 non-null    float64
 37  ave_salary                                  498 non-null    int64
 38  consumer_price_idx                          498 non-null    float64
 39  industry_product_price_idx                  498 non-null    float64
 40  cunstruction_price_idx                      498 non-null    float64
 41  agriculture_product_price_idx               498 non-null    float64
 42  logistics_price_idx                         498 non-null    float64
 43  gross_regional_product                      498 non-null    float64
 44  gross_regional_product_usd                  498 non-null    float64
 45  gross_regional_product_per_person_1000      498 non-null    float64
 46  gross_regional_product_per_person_1000_usd  498 non-null    float64
 47  industry_product_volume_mln_tenge           498 non-null    float64
 48  mining_mln_tenge                            498 non-null    float64
 49  manufactur_industry_mln_tenge               498 non-null    float64
 50  electrecity_gas_aircondition_mln_tenge      498 non-null    float64
 51  water_supply_mln_tenge                      498 non-null    float64
 52  agriculture_gross_output                    498 non-null    float64
 53  main_capital_investment_mln_tenge           498 non-null    float64
 54  construction_volume_mln_tenge               498 non-null    float64
 55  residential_buildings_sqm                   498 non-null    float64
 56  places_in_school                            498 non-null    int64
 57  transport_communication                     498 non-null    float64
 58  load_transportation_mln_tons                498 non-null    float64
 59  passenger_transportation_mln_km             498 non-null    float64
 60  passenger_transportation_mln_person         498 non-null    float64
 61  retail_product_sell_mln_tenge               498 non-null    float64
```

```
 62  CrimePerPop                                498 non-null    float64
dtypes: float64(52), int64(10), object(1)
memory usage: 245.2+ KB
```

Fig. 10. Attributes of Kazakhstani social, economic and crime data set after cleaning.

Further on as it was described in Chapter 2 we added two new features and named them 'CrimePerPop', which was calculated from feature ration of 'crimes' to 'population_1000_per' and the other column and 'highCrime' and gave it a value of True and False.

To define a proper threshold for classification we explored the distribution of 'CrimePerPop' and in case it is normally distributed we could have used threshold as mean = 0.12.

However, likewise it happened to our UCI Repository clean data set, visualization of graph displayed that the data is not normally distributed and turned out to be more saturated towards 0. Hence taking mean as threshold is not a solution, so it was decided to declare median =0.109 as a threshold value.

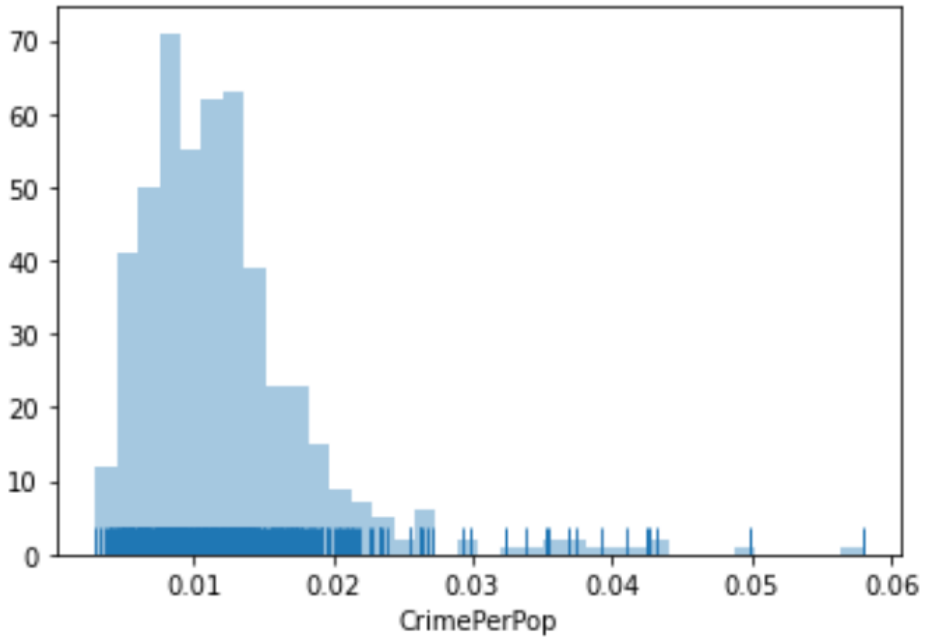

Fig. 11. Distribution of 'CrimePerPop'

Moreover, we defined a new added 'High crime' column as a target feature on which our proposed models would be based on.

Using Python pandas data frame the dataset was converted into appropriate data frame and the target feature ‘High crime’ was assigned to a variable ‘Target’ and the remaining features to ‘Features’.

Finally, we explored the crime rate of each region during 1991-2020 year.

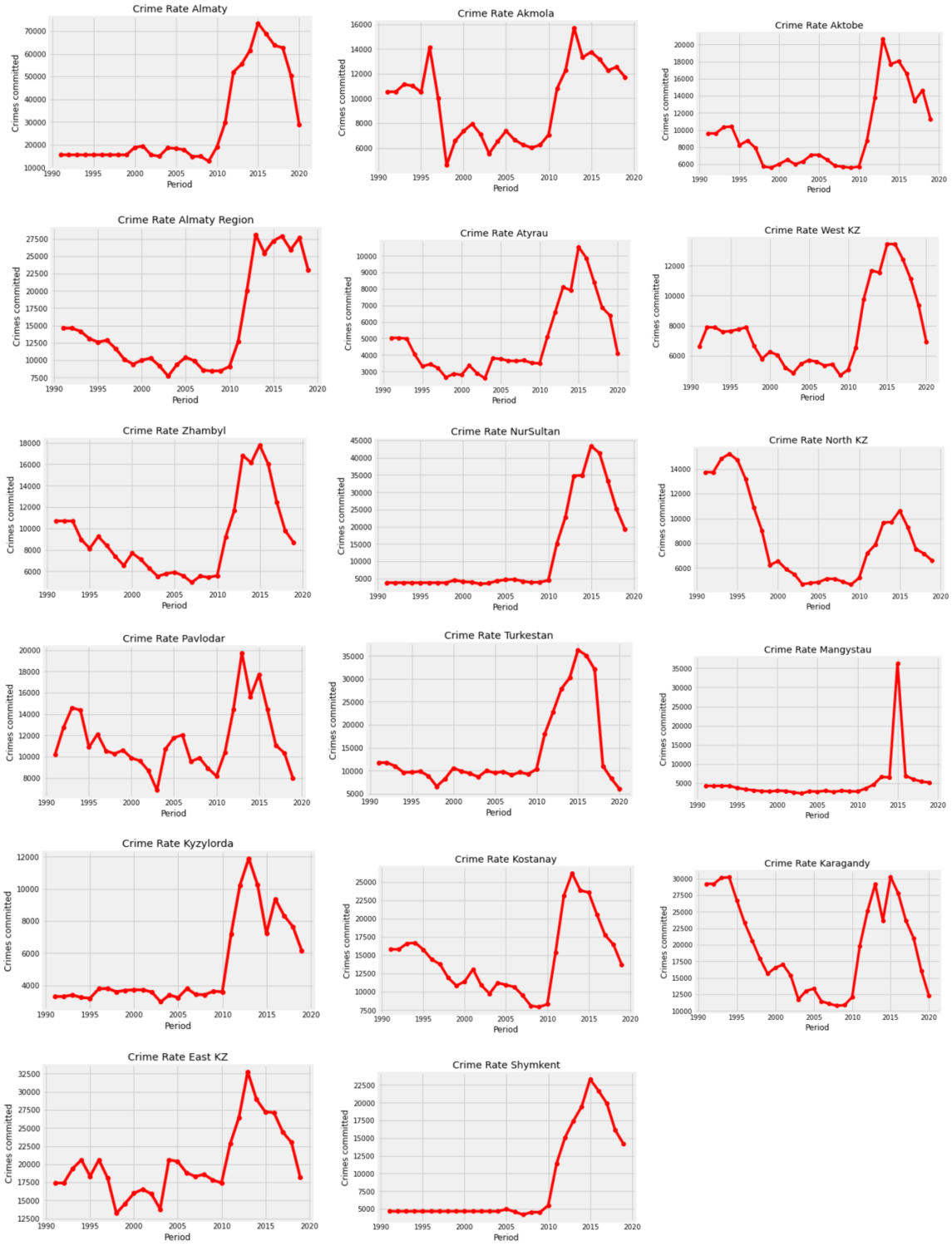

Fig. 12. Graph of Kazakhstan crime rate region wise

From graph above it is obvious that there was immense increase of committed crimes in 2010 across the whole country with further decline by the year 2020.

Experts claims that until current time the reason of such tendency has not been analyzed.

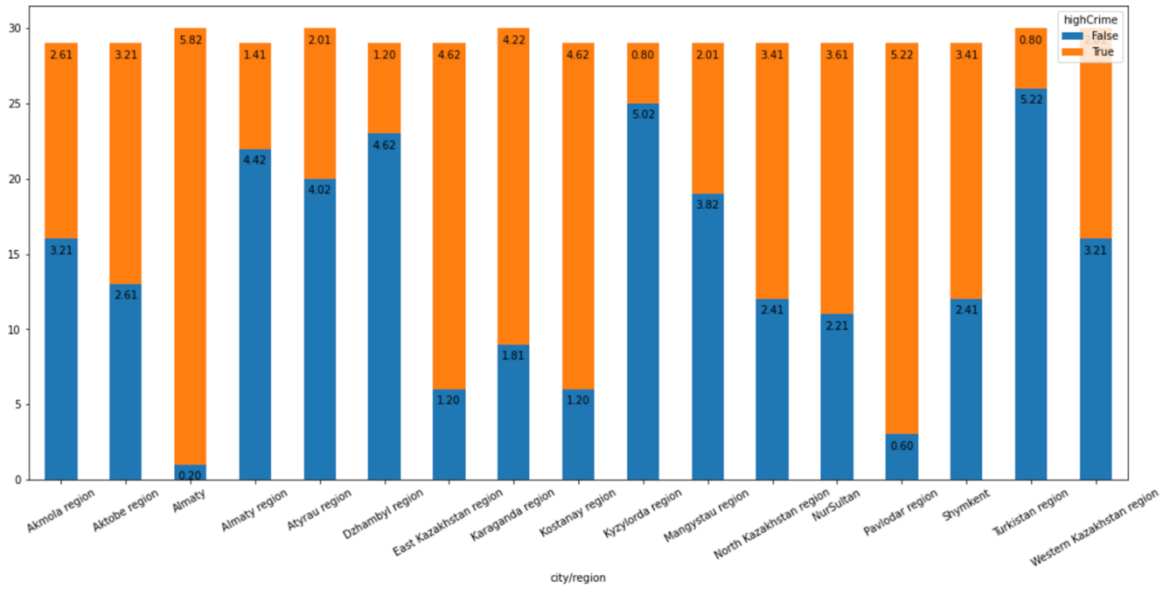

Fig. 13. Distribution of 'highCrime' True or False - Kazakhstan crime rate region wise

Besides, we explored the distribution of regions for having True and False values for 'highCrime' attribute and noticed that Almaty city is almost constantly keep the rate of highly criminal region, whereas Kyzylorda region is relatively remains in low criminal area.

4.4 MODELLING and EVALUATION

Similar to Chapter 3 we constructed the listed above 5 prediction and classification models upon our new Kazakhstani data set and came up to following results.

DECISION TREES

After dividing data set into random training and testing parts, defining proper depth and testing criterion of Gini and Entropy, applying 10-fold cross-validation as well as defining new threshold and excluding multicollinear attributes we derived following outputs.

Evaluating Measure Decision Tree Classifier	10-fold Cross-Validation (%)
Accuracy	78,11%
Precision	80,16%
Recall	76,74%
F1 score	78,41%

Table 9: Performance Measures Decision Trees Kazakhstani data

The top 10 features extracted according to the feature importance scores were: 'divorce_coef_1000_per', 'retail_product_sell_mln_tenge','students_in_schools_1000_per', 'self_emp_1000_per', 'hired_1000_per', 'working_1000_per', 'able_bodied_1000_per', 'min_income_usd', 'min_income', 'people_low_income_pct''.

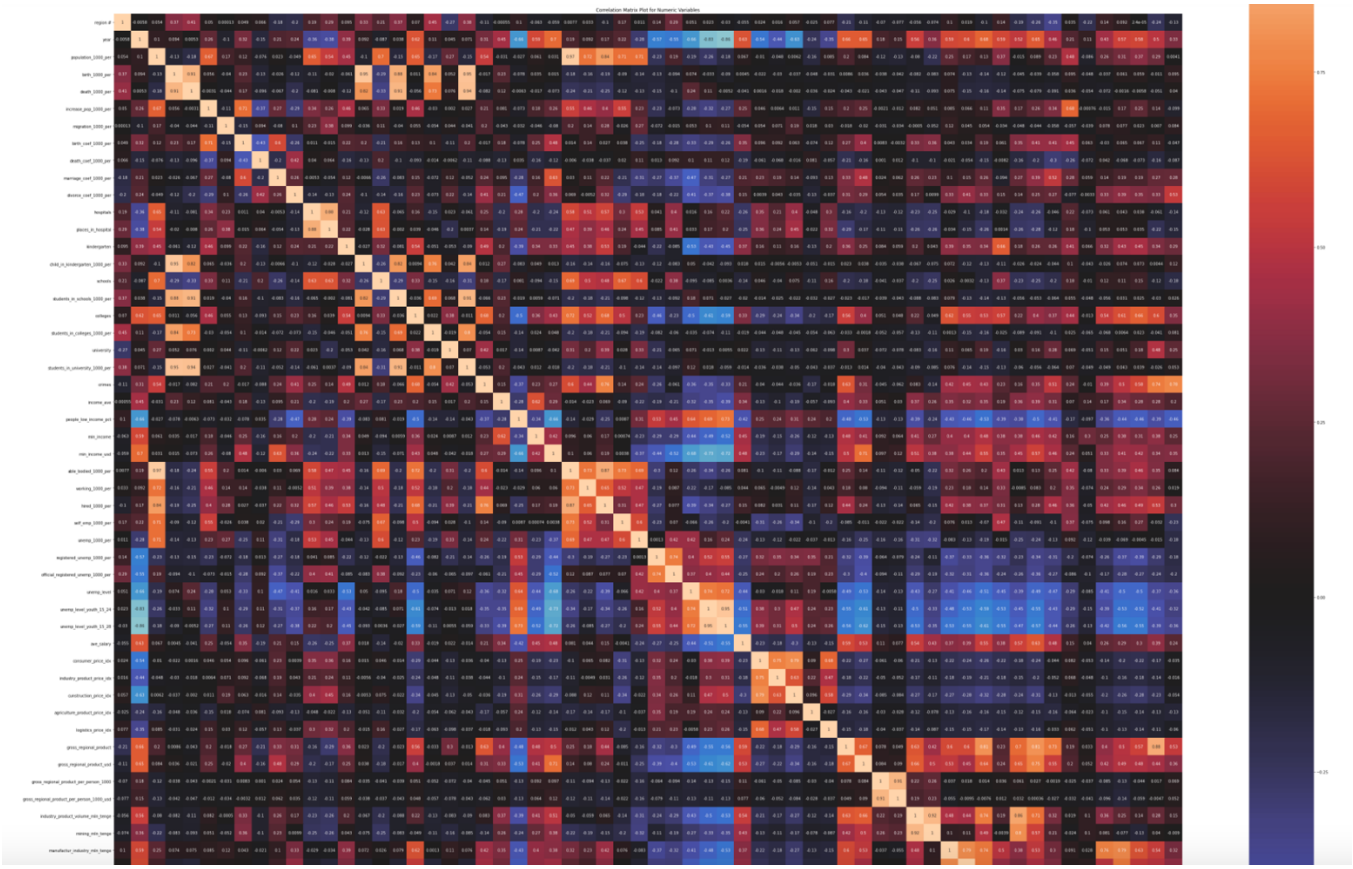

Fig. 14. Correlation matrix of Kazakhstani crime dataset

Our model of Decision Tree Classifier selected 'divorce_coef_1000_per' as its main implying the major predictive feature for 'High crime'.

RANDOM FOREST CLASSIFICATION

Random Forests Classifiers also demonstrated pretty balanced and accurate metrics, although a bit less than Decision Tree.

Evaluating Measure Random Forest Classifier	10-fold Cross-Validation (%)
Accuracy	71,87%
Precision	75,70%
Recall	67,78%
F1 score	70,99%

Table 10: Performance Measures Kazakhstani crime data - Random Forest Classifier

The top 10 features extracted according to the feature importance scores were: 'divorce_coef_1000_per', 'child_in_kindergarten_1000_per', 'year', 'people_low_income_pct', 'retail_product_sell_mln_tenge', 'increase_pop_1000_per', 'min_income', 'income_ave', 'kindergarten', 'birth_coef_1000_per'

NAÏVE BAYES CLASSIFICATION

We designed Gaussian NB model with the same parameters and derived following outputs:

Evaluating Measure Naïve Bayes Classifier	10-fold Cross-Validation (%)
Accuracy	56,77%
Precision	40,23%
Recall	66,55%
F1 score	47,80%

Table 11: Performance Measures- Naïve Bayes Classifier Kazakhstan crime Data

The top 10 features extracted according to the feature importance scores were:

('divorce_coef_1000_per', 0.6057519237372865)

('people_low_income_pct', 0.4781637426743349)

('kindergarten', 0.387368629522203)

('retail_product_sell_mln_tenge', 0.35184198607029726)

('electrecity_gas_aircondition_mln_tenge', 0.32596664463290403)

('passenger_transportation_mln_km', 0.32536553839393806)

('manufactur_industry_mln_tenge', 0.30308209883069925)

('passenger_transportation_mln_person', 0.29244806547784546)

('water_supply_mln_tenge', 0.29189384283055375)

('gross_regional_product', 0.28727327787321716)

K-MEANS and SVC

K-means Classifier	10-fold Cross-Validation (%)
Accuracy	54,73%
Precision	59,39%
Recall	22,26%
F1 score	24,11%

Table 12: Performance Measures- K-means Classifier

Non-Linear SVM Classifier	10-fold Cross-Validation (%)
Accuracy	59,02%
Precision	74,25%
Recall	30,6%
F1 score	42,68%

Table 13: Performance Measures- non-linear SVM Classifier

4.5 CONCLUSION

Experiments on Kazakhstan crime data with our tuned models displayed that Decision Tree Classifier has balanced and high performance metrics rather than others, whereas for UCI Repository materials Random Forest Classifier had better results.

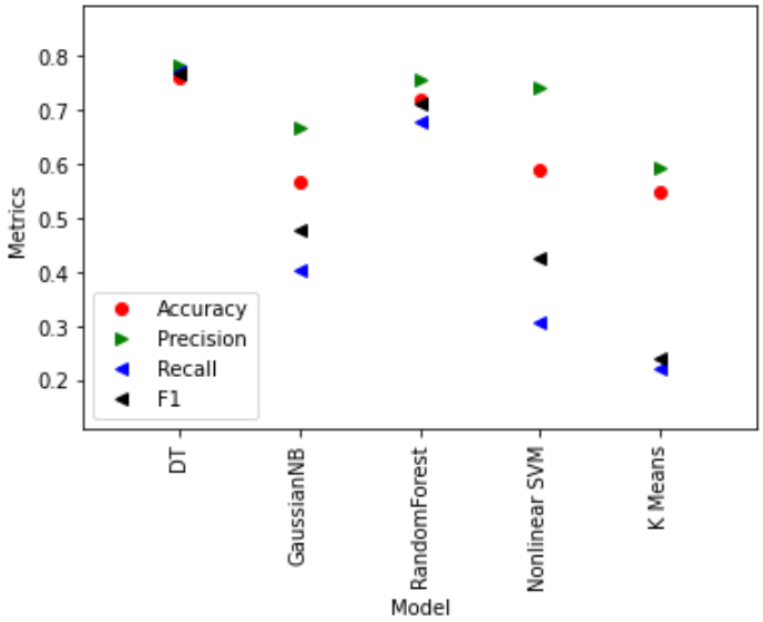

Fig. 15. Comparison of models (Kazakhstan Crime data)

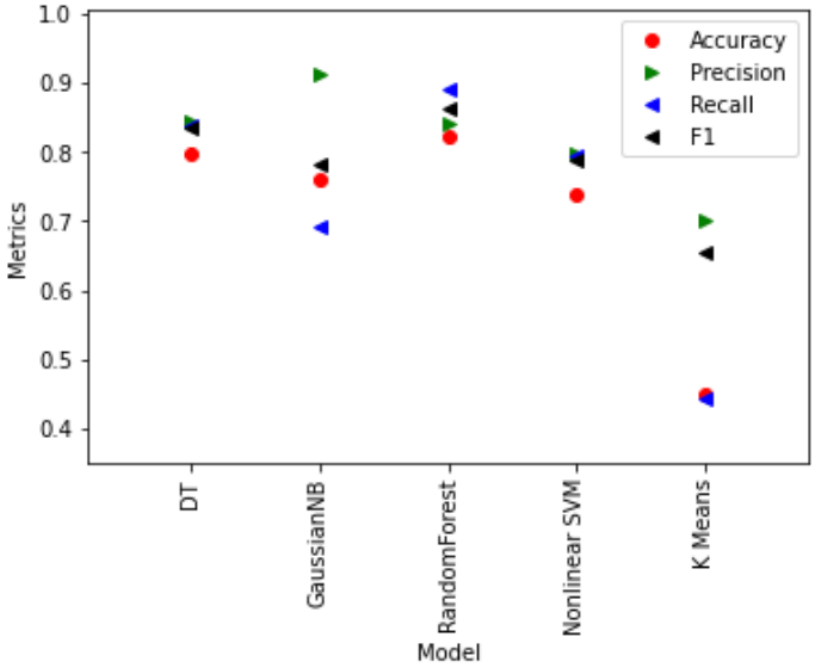

Fig. 16. Comparison of models (UCI Repository Crime data)

For all of listed models ‘divorce_coef_1000_per’ proved to be the top main feature and 'people_low_income_pct' is though not the most important, but is a common feature as well.

Chapter 5

5. REVIEW OF OUTCOME

Comparison of results of experiments on UCI Repository data and Kazakhstani crime data had close results with only difference of Decision Tree Classifier had slight better performance metrics in the latter one, while in former one Random Forest displayed best outcome.

This might be due to the fact that UCI Repository data was unbalanced and Kazakhstan crime data was relatively balanced.

In particular, 'community crimes data' demonstrated:

```
highCrime
False    37.280482
True     62.719518
dtype: float64
------------------
Percentage Positive Instance = 62.719518314099346
Percentage Negative Instance = 37.280481685900654
```

while 'Kazakhstan crime data' consisted of:

```
highCrime
False    48.192771
True     51.807229
dtype: float64
------------------
Percentage Positive Instance = 51.80722891566265
Percentage Negative Instance = 48.19277108433735
```

Consequently, if we apply a dumb model to UCI material that would give True value to whole data set its accuracy would demonstrate 62%. However, Kazakhstani data would have almost half (51%) accuracy.

Moreover, during consultation with experts it was suggested to test features importance ranking of our best model (Decision Tree Classifier) and hypothesis of influence of number of imprisoned criminals to the crime rate was tested.

It was decided to collect and include into data set new attribute 'number_of_imprisoned' and rerun our model.

As a result, our model's performance metrics had slight positive changes:

Evaluating Measure Decision Tree Classifier	10-fold Cross-Validation (%)
Accuracy	78,12% (formerly 78,11%)
Precision	80,17% (formerly 80,16%)
Recall	76,74% (formerly 76,74%)
F1 score	78,41% (formerly 78,41%)

Table 14: Performance Measures Decision Trees Kazakhstani data

The top 10 features extracted according to the feature importance scores were: 'divorce_coef_1000_per', 'retail_product_sell_mln_tenge','students_in_schools_1000_per', 'self_emp_1000_per', 'hired_1000_per', 'working_1000_per', **'number of imprisoned'**, 'able_bodied_1000_per', 'min_income_usd', 'min_income'.

Chapter 6

6. Conclusion and Future Work

Counteractions to crime is one of the major duties of government in general and law enforcement bodies in particular.

Therefore, not only investigation of committed crimes, but also prevention is a paramount task.

In current thesis we attempted to design a suitable machine learning algorithm on crime, economic and social data to predict the probability of regions having low or high crimes levels with further defining main social and economical factors that correlate with crime growth.

Prediction models based on Classification, Regression and Clustering techniques: Decision Tree, Random Forest, Naïve Bayesian, K-means, Support Vector Machine algorithms were selected.

They were tested applying both - data available from opensource materials and collected from Kazakhstani state bodies. After dividing data set into random training and testing parts, defining proper depth and testing criterion of Gini and Entropy, applying 10-fold cross-validation as well as defining new threshold and excluding multicollinear attributes we achieved new results outputs.

Random Forest model proved to be the most accurate (UCI Repository materials, Accuracy: 0.837, Precision: 0.884, Recall: 0.872, F1 score: 0.868) among listed models, whereas Decision Tree achieved the best result on Kazakhstani data (govstat.kz materials, Accuracy: 0.781, Precision: 0.801, Recall: 0.767, F1 score: 0.784).

At final stage hypothesis of importance of a certain feature was tested and model proved that this feature correlates with target (crime rate) and its inclusion positively affected the accuracy of result. Therefore, it can be claimed that the more we acquire expertise in the field of important features, the better selected model will perform.

Reduction of overfitting using cross validation improves performance by enough training and testing samples that seemed to help in this analysis by giving correct and consistent performance measures. These predicted features will be useful for the Police Department and other governmental structures to utilize their resources efficiently and take appropriate actions to reduce criminal activities in the society.

By maintaining dynamic databases with the criminal records across various fields, this technique can be implemented widely.

The present dataset consists of all types of crimes, this type of analysis can be narrowed down to a single category of crime.

Bibliography

1. S.Sivaranjani, S.Sivakumari, Aasha.M (2016), Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches, 2016 International Conference on Emerging Technological Trends [ICETT].
2. N.Baloian, E.Bassaletti, M.Fernández, O.Figueroa, P.Fuentes, R.Manasevich, M.Orchard, S.Peñafiel, José A. Pino, M.Vergara (2017). Cooperative Work in Design Crime Prediction using Patterns and Context, Proceedings of the 2017 IEEE 21st International Conference on Computer Supported
3. L.G.A.Alves, H.V.Ribeiro, F.A.Rodrigues, Crime prediction through urban metrics and statistical learning, Physica A 505 (2018) 435–443
4. E.Ahishakiye, D.Taremwa, E.O.Omulo, I.Niyonzima (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 06 – Issue 03, May 2017
5. R.Iqba, M.A.A.Murad, A.Mustapha, P.H.S.Panahy, N.Khanahmadliravi (2013, March). An Experimental Study of Classification Algorithms for Crime Prediction
6. S.Sathyadevan, M.S.Devan, S.Gangadharan (2014), Crime Analysis and Prediction Using Data Mining, 2014 First International Conference on Networks & Soft Computing (pp.406 – 412).
7. A.Babakura, N.Sulaiman, M.A. Yusuf (2014). Improved Method of Classification Algorithms for Crime Prediction, 2014 International Symposium on Biometrics and Security Technologies (ISBAST)
8. A.Bogomolov, B.Lepri, J.Staiano, N.Oliver, F.Pianesi, A.Pentland (2014, September). Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data.
   https://www.researchgate.net/publication/265554568_Once_Upon_a_Crime_Towards_Crime_Prediction_from_Demographics_and_Mobile_Data
9. J.Azeez, D.J.Aravindhar (2015). Hybrid Approach to Crime Prediction using Deep learning, 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), (pp.1701 – 1710). IEEE

10. V.Jain, A.Bhatia, Y.Sharma, V.Arora (2017). Crime Prediction using K-means Algorithm, GRD Journals- Global Research and Development Journal for Engineering, Volume 2, Issue 5, (pp. 206 - 209). ISSN: 2455-5703
11. Pang-Ning Tan, Michael Steinbach, Vipin Kumar (2006). Introduction to Data Mining, Pearson Addison Wesley, Page 187,227, 296, 297, 729
12. El Naqa I., Murphy M.J. (2015) What Is Machine Learning?. In: El Naqa I., Li R., Murphy M. (eds) Machine Learning in Radiation Oncology. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1
13. S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458.
14. Mitchell, T: Machine Learning. McGraw-Hill, 1997.
15. Aristidis Likas, Nikos Vlassis, Jakob J. Verbeek, The global k-means clustering algorithm, Pattern Recognition, Volume 36, Issue 2, 2003, Pages 451-461, ISSN 0031-3203, https://doi.org/10.1016/S0031-3203(02)00060-2.
16. Prajakta Yerpude, Vaishnavi Gudur (2017). Predictive modelling of crime dataset using data mining, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4
17. Xing Li-Ning (College of Information System and Management,National University of Defense Technology), Tang Hua, “Data mining algorithm based on genetic algorithm and entropy,” Journal of Computational Information Systems, Volume 3, May 2007
18. Wan Dingsheng et.al, “Data Mining Algorithmic Research and Application Based on Information Entropy”, 2008 International Conference on Computer Science and Software Engineering
19. Jeffreys, Harold (1973). ScientificInference (3rded.). Cambridge University Press. Page31.
20. Malathi, Dr. S. Santhosh Baboo (2011). Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters , Global Journal of Computer Science and Technology, Volume 11 Issue 11 Version 1.0
21. Ghada M. Tolan and Omar S. Soliman (2015). An Experimental Study of Classification Algorithms for Terrorism Prediction, International Journal of Knowledge Engineering, Vol. 1, No. 2

22. Sayali D. Jadhav, H. P. Channe (2013). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.
23. Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S (2014). Crime Analysis and Prediction Using Data Mining, 2014 First International Conference on Networks & Soft Computing.
24. Anisha Agarwal, Dhanashree Chougule, Arpita Agrawal, Divya Chimote (2016). Application for analysis and prediction of crime data using data mining, International Journal of Advanced Computational Engineering and Networking, issn: 2320-2106, volume-4, issue-5.
25. Shivani B. Mehta, Rushabh D. Doshi (2020). Automatic Clustering Crime Region Prediction Model using Statistical Method in Data Mining, International Journal of Engineering Research & Technology (IJERT), http://www.ijert.org ISSN: 2278-0181 Vol. 9 Issue 04
26. MohammadReza Keyvanpoura, Mostafa Javidehb, Mohammad Reza Ebrahimia (2011).Detecting and investigating crime by means of data mining: a general crime matching framework, Procedia Computer Science 3 (2011) 872–880, 1877-0509, 2010 Published by Elsevier Ltd.
27. Gondy Leroy, Juliette Gutierrez (2007). Predicting Crime Reporting with Decision Trees and the National Crime Victimization Survey, Proceedings of the Thirteenth Americas Conference on Information Systems, Denver, Colorado.
28. Chunxue Wu, Fang Yang, Yan Wu, Ren Han (2019). Prediction of crime tendency of high-risk personnel using C5.0 decision tree empowered by particle swarm optimization, Mathematical Biosciences and Engineering Volume 16, Issue 5, 4135–4150, pages 4135 – 4150.
29. Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O.,Akinjobi J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison, International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3.
30. A H Wibowo, T I Oesman (2020). The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency, Journal of Physics: Conference Series 1450 (2020) 012076 doi:10.1088/1742-6596/1450/1/012076.