Crime Type Classification / Project Final Report

SERHAT HAKKI AKDAĞ, Middle East Technical University, Turkey
MERT ANIL YILMAZ, Middle East Technical University, Turkey

The aim of this paper is to classify type of crime that is likely to occur given circumstances. As data, Los Angeles Crime data set and Los Angeles Census data set is used. First, different models are created and comparatively tested on using only information gained from Los Angeles Crime data set. Then Los Angeles Census data set is correlated to the Los Angeles Crime data set, where improvements on classification models after correlation is tested and compared to single data set version of the models.

CCS Concepts: • **Information systems** → Data mining.

Additional Key Words and Phrases: classification, data mining, naive bayes, gradient boosting machine

ACM Reference Format:
Serhat Hakkı Akdağ and Mert Anıl Yılmaz. 2021. Crime Type Classification / Project Final Report. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 1, Article 1 (May 2021), 15 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In 2000s, with increasing use of technological advancements in surveillance systems and communication infrastructures, rate of criminal activities are significantly decreased. In addition to this, data science started to play important role in understanding crime patterns to take necessary precautions even before criminal activities occur. In this paper, using criminal activity record and census information of Los Angeles city, we are aiming to classify type of crime that is likely to occur given circumstances. Using proposed model, more detailed precautions can be put in place by police forces and security agencies. For example, if chance of robbery is highest given time of day, and type of premise, police can take higher measurements against any incident of the type robbery if the circumstances hold.

Crime records and census information of Los Angeles City are taken from data sets provided by Los Angeles Police Department. Crime data includes records from 2010 to 2017, while census information are given for 2010 statistics. We experimented with three different classification algorithms, while also applying different preprocessing steps. Light Gradient Boosting Machine gives the best results.

First, related work in the area of crime type classification are provided. Then, in Methods section, details of the data sets and applied preprocessing steps are pointed out. After these, experiments are given. In the end, conclusion and future work is presented.

Authors' addresses: Serhat Hakkı Akdağ, serhat.akdag@metu.edu.tr, Middle East Technical University, Ankara, Turkey; Mert Anıl Yılmaz, mert.yilmaz_01@metu.edu.tr, Middle East Technical University, Ankara, Turkey.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
2476-1249/2021/5-ART1 $15.00
https://doi.org/10.1145/1122445.1122456

2 RELATED WORK

Examples seen in this area are mostly focused on predicting if crime will occur or not. Not many articles and examples are found in crime type prediction. Some of the recent articles focusing on crime type classification mainly focus on extracting and visualizing repeating patterns instead of discussing how model for crime type classification is generated in more detail. Example for such an article is the works of Gupta et al. in "A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA" [3]. They provide statistical information extracted from Denver City crime data set, which helped us to modify the ways in which we apply preprocessing on our data. In their discussions, most descriptive parts of the data set are highlighted and visualized, giving us broader view of where to focus more. In addition, they worked on 6 different classification models and compared the results without delving into details of these models and how data is processed to utilize them. Work of Albahli et al. approached to crime type classification in similar manner [1]. Their study investigates influencing factors that impact crime rates in Saudi Arabia. They point out the importance of time information to classify type of crime that is likely to occur.

Another crime type classification related study can be seen in work of Ratul in University of Ottawa [5]. His work describes the data set and the way problem is tried to be tackled in more detail. Denver City crime data is used to conduct experiments. This data set does not include characteristics of victim such as age, sex, descent. These victim related features were also used classify type of crime along with the other attributes similar to the attributes used in this work. In addition, paper includes discussion on positive impact of dimensionality reduction and sampling. These approaches are utilized and improved to increase performance of our models.

We also benefited from work of Grandini et al. in their paper "Metrics for Multi-Class Classification: an Overview" [2]. This paper provides insightful information on how to evaluate multi-class classification models. It gives detailed discussion of how each metric can be utilized and what are each of the metrics' advantages and disadvantages. Using these information, we detailed our evaluation process, which really helped us in our comparative studies between models.

The aim is to provide deeper understanding of elements that describe distinct crime types. The statistical information provided in related works were utilized to modify the preprocessing steps and classification algorithms in this work. In addition to results concluded by the works of others, this work proposes wider range of preprocessing steps. Furthermore, census information was utilized to see its effects on classification capabilities of our learners.

3 METHODS

3.1 Data sets & Preprocessing

Mainly, Los Angeles Crime data set is used to demonstrate our model. In addition to this, we utilized Los Angeles Census data set to see whether our model would improve. Both of these data sets are taken from official website of Los Angeles city and it is pointed out by the publishing authority that these data sets are based on paper work from police stations, therefore, it can contain problems. Detailed information can be seen the subsections below.

3.1.1 Crime Code. This column is the label column that the model predicts. Following problems are dealt with regarding crime code attribute:

- Some of the crime types had little examples in data set. We removed data instances with crime codes that has less than 500 instances. In the Figure 1, demonstrates that the data set is unbalanced since CrimeCode samples are not equally obtained.

Fig. 1. The distribution of the crime codes in the crime data set.

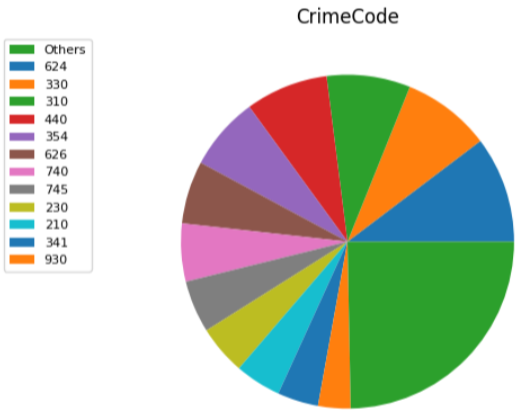


Table 1. 4-Group Crime Types and the Number of Instances

Crime Type	THEFT	ASSAULT	MINOR OFFENSE	SEXUAL & CRIMES REGARDING CHILDREN
Number of instance	664815	317756	247259	57078

In addition, we also experimented with this threshold and compared the results. This discussion can be seen under Categorical Naive Bayes section.

- Crime code for 'Vehicle Stolen' category had no victim information. As we want to find crime types related to victim age, sex, descent information, this category of crime code is removed.
- Crime codes for 'Other Miscellaneous Crime' and 'Other Assault' was assigned to crime codes with unknown crime codes, so these categories are also removed.

After the first experiment, it was predicted that having so many crime codes would cause different labeling of related instances with feature values that are very close to each other, thus reducing the performance of the model. Therefore, the next experiment was to group crime codes of the same type and have a more compact labeling. At first, crime codes were divided into 5 different groups. Since there are many crime codes related to theft, they are separated as THEFT_1 and THEFT_2. Thus, the number of instances in the groups was balanced. Apart from that, THEFT_1 and THEFT_2 groups were combined and 4 groups were formed and this was one the experiments. The distribution of instances in the groups in the 4-group experiment is shown in Table 1 Finally, random sampling is applied to find a solution to the unbalanced number of instances in the 4-group experiment, and it is also one of the experiments. The instance number of each group was determined based on the group with the least number of instances. After all this grouping was done, the experiments with 4-group crime codes were focused on in the section Experiments, since the 4-group crime codes gave the best results.

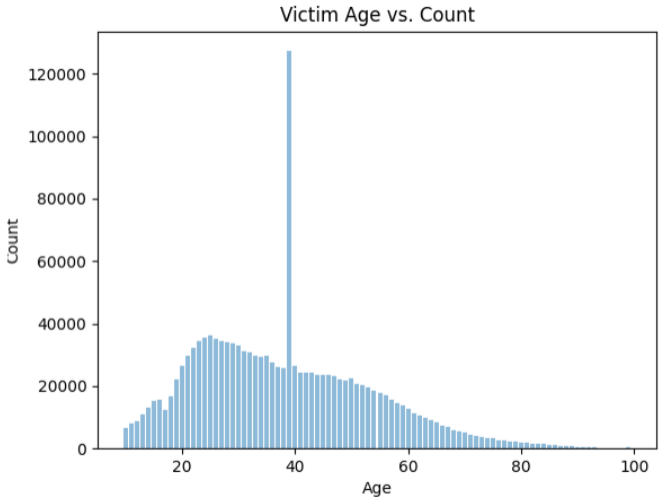
3.1.2 Date Occurred. This attribute holds the date in which crime occurred, in mm/dd/yyyy format. As we are interested in repeating patterns inside the data set, we extracted following attributes from this attribute:

- **Month Occurred**: We extracted this column to see if month of crime has effect on its type.
- **Day of Week**: This attribute is also created to see if day of week has any relation with the type of crime.

3.1.3 Time Occurred. This attribute was exact time of crime in military hours. We are not interested in exact time, therefore, we wanted to categorize this attribute. Therefore, we divided 24 hours to 8 intervals each with 3 hours and created categorical attribute for Time Occurred attribute.

3.1.4 Victim Age. Victim Age had many invalid and empty cells. To solve the issues related to it, we experimented with three approaches. The distribution of this attribute is shown in the Figure 2.

Fig. 2. The distribution of the victim ages in the crime data set.



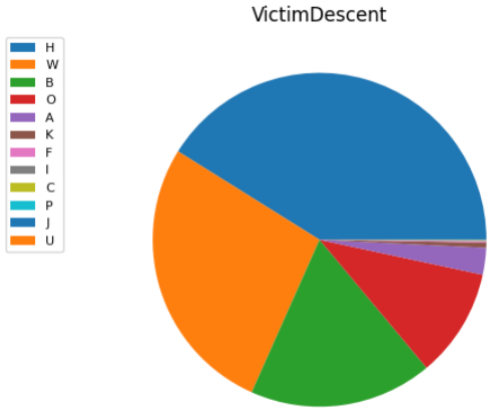
- We removed data samples that had these problematic Victim Age cells.
- We imputed these empty sets with average of the other Victim Age values.
- We tried to create knn model to impute Victim Age information using sklearn. KNNImputer package. However, as size of our data set is too high. Because imputations with other methods which are mode and median did not affect the results positively and preprocessing with this library took too long, this library has been abandoned.

3.1.5 Victim Sex, Victim Descent and Premise Code. These three attributes are each categorical attributes and similar problems are faced. For instance, the column VictimDescent shown in the Figure 3, is dominated by H - Hispanic/Latin/Mexican. Therefore, learning for the other descent may remain low. Some of the categories under these attributes had 1-100 data samples. These categories are removed. In addition, all of these attributes contained some undefined and empty cells. Following preprocessing steps are tried to solve these problems.

- We removed data samples that had these problematic cells.
- We imputed these empty sets with median of the other values of same attribute.
- We imputed these empty sets with mod of the other values of same attribute.

3.1.6 Location. This attribute is of type string and holds comma separated Longitude and Latitude information. First, these two attributes are extracted and added as separate attributes. We want to correlate crime data set with census data set, which contains only zip code information. To correlate these two data sets, we need to merge these data sets by converting location information to zip code information. However, while longitude and latitude describe exact location information, zip code is dependent on postal services of the city and is not exactly

Fig. 3. The distribution of the victim descents in the crime data set.



correlated with area information. From SimpleMaps [4], we obtained zip code to location map, which gives close middle point location of each zip code. Then, using Haversine distance calculation formula, each data sample in crime data set is assigned zip code information, which is saved to new attribute 'Zip Code'. As census data set also contain 'Zip Code' attribute, these two data sets are merged via this attribute.

3.1.7 Other Attributes. Following attributes are removed from the crime data set:

- **Date Reported**: We are utilizing 'Date Occurred' column for time related similarities.
- **Area ID, Area Name, Reporting District, Address, Cross Street**: We are planning to use zip code information for location related similarities. In addition zip code is also given in census data set, therefore, it is also suitable to create correlation between data sets.
- **MO Codes**: This attribute describes MO codes of the guilty. Even though some of the data samples have really descriptive information, knowing MO codes of the guilty before crime happened does not makes sense. This attribute is not added to model as it is not police forces' disposal before crimes occur.
- **Weapon Used Code, Weapon Description**: If crime is of type homicide, assault, etc. this column holds the weapon related information. This attribute is not suitable to every crime code, and gives too much information if it is not nan, therefore, it is removed.
- **Status Code, Status Description** These columns describe internal status of the case, which is not related to our problem.
- **Crime Code 1, 2, 3, 4** These columns hold additional crime code information. As we are interested in primary crime code, these are not necessary for us.

3.1.8 Los Angeles Census Data Set Attributes. This data set is important to research whether census specific information improves the model that classifies the crime types in specific circumstances. However, this data set had no common columns with the crime data set. 'Zip Code' column is created for the crime data set to merge these two data set as described in section Location. After merging these two data sets, the columns Zip Code, Total Households, Average Household Size, Latitude, and Longitude are removed. The new features which come from the census data are Total Population, Median Age, Total Males, and Total Females. Median Age and Total Population which are numerical columns were categorized. The ratio of Total Males and Total Females columns was taken and these ratios were categorized and added to the feature list.

3.2 Optional Preprocessing Steps

Since Categorical Naive Bayes needs categorical feature and Victim Age is not categorical, it is mapped to following category of values. With this mapping, 'Victim Age' is also included in created model.

- **Childhood**: ages between 0 and 14
- **Adolescence**: ages between 14 and 21
- **Youth**: ages between 22 and 35
- **Maturity**: ages between 36 and 49
- **Aging**: ages between 50 and 63
- **Old Age**: ages between 64 and max

There are two main experiments done with LightGBM. First of all, the data set used in Categorical Naive Bayes experiments was used without any changes. The other setup which was combined with one hot encoding is mapping Time Occurred and Month Occurred columns to the following category values. After this process, the preprocessing of the second experiment was done.

Time Occurred:

- **Morning**: between 6 AM and 12 AM
- **Afternoon**: between 12 AM and 6 PM
- **Evening**: between 6 PM and 12 PM
- **Night**: between 12 PM and 6 AM

Month Occurred to Season Occurred:

- **Winter**: month 12, 1, and 2
- **Spring**: month 3, 4, and 5
- **Summer**: month 6, 7, and 8
- **Fall**: month 9, 10, and 11

3.3 Classification Algorithms

3.3.1 Categorical Naive Bayes. As described in sections above, most of the features of the crime data set is categorical apart from Victim Age. We wanted to use simple approach as a start, therefore, picked Naive Bayes. Due to dominance of categorical attributes, we opted for Categorical Naive Bayes implementation under Scikit-Learn Naive Bayes library [6]. This implementation assumes each feature has its own categorical distribution. For the attributes taken from census data set, described categorizations are utilized to include them in Categorical Naive Bayes experiments.

3.3.2 Light Gradient Boosting Machine. After obtaining the results of Categorical Naive Bayes, we wanted to construct a model with a different approach, gradient boosting. Gradient boosting uses weak learners to improve the model incrementally. Gradient boosting uses decision trees as weak learners. At each step, a new decision tree is constructed with a new base assumption value for each sample. Light Gradient Boosting Machine is a gradient boosting framework that uses decision trees as weak learners. Light Gradient Boosting Machine supports numeric attributes, therefore, categorical attributes in crime data set are encoded using label encoding and one-hot encoding and results from both approaches are discussed.

3.3.3 K-Nearest Neighbors. Different from other approaches, we also experimented with algorithm that utilizes lazy learning approach: K-nearest neighbors classification. In this algorithm, data instances in training set are not used to calculate necessary information to test the learning algorithm. Instead, each test instance are compared to training instances to find their k nearest neighbors to classify the test instance. K-Nearest Neighbors algorithm in Scikit-learn package expects numeric attributes similar to Light Gradient Boosting Machine. Therefore, similar

setup is utilized to experiment with K-nearest neighbors classifier. In addition, different k numbers are tested to select appropriate one to optimize the learner. Details on these decisions are given in the next chapter.

4 EXPERIMENTS

In this section, experiments with described classification algorithms are given. We obtained results for most of the combinations of the preprocessing steps described in section Data Sets & Preprocessing. Using these results, we will provide effects of each preprocessing decision on each classification algorithm. Then best combination of preprocessing steps for the described algorithm will be selected to give more detailed results including classification reports and confusion matrices.

4.1 Categorical Naive Bayes

4.1.1 Experiment Setup. Categorical features that already reside in crime data set are: 'Time Occurred', 'Victim Sex', 'Victim Descent', 'Premise Code', 'Day of Week'. 'Victim Age' is the only numeric attribute left in the crime data set. To utilize this attribute, it is categorized as described in the section Optional Preprocessing Steps. There are also attributes coming from the census data set, which are 'Total Population', 'Total Males', 'Total Females' and 'Median Age'. These attributes are all numeric, again to utilize them categorization is done as described in the section Data Sets & Preprocessing. In the end, 3 more categorical attributes: 'Total Population Category', 'Median Age Category' and 'Female To Male Ratio Categorized' are added to the data set.

In terms of 'Crime Code' attribute, we experimented with 3 different orientations. In the first orientation, 'Crime Code' groups that have more than 500 instances are utilized, where 76 different crime codes are there in the data set. In the second one, 'Crime Code' groups that have more than 50000 instances are used. 11 different crime codes remain the data set in this orientation. As the last orientation, merging of crime codes described in section Preprocessing is utilized, where 4 crime code groups are created. These groups are: Theft, Assault, Minor Offense and Sexual & Crimes Regarding Children. Results for all these orientations are compared.

For the attributes that have invalid or missing values, removal of these problematic instances are selected instead of imputation options described in section Data Sets & Preprocessing. As we have large number of instances, imputation is removed as it decreased performance of the created models. Each of the categorical features are required to be described with number from 0 to n-1, representing n categories. Categorical features are encoded with Label Encoder to comply with this rule. In addition, 3-fold cross validation is used in this experiment. Different orientations of preprocessings are tested and their results are compared.

4.1.2 Results. First, different crime code orientations are experimented with where attributes from census data set are not utilized. Results showed that version where crime codes are merged to obtain 4 crime code groups gave the best results. Accuracy, precision, recall and f1-score metrics for this experiment are given in Table 2. In addition, confusion matrix for this experiment is given in Figure 4. To show how our model improved with the merge operation, comparison for 3 different crime code orientations are given in Table 3. It can be clearly seen that smaller number of classes produce better results.

Table 2. Categorical Naive Bayes model results with 4 different crime code groups

Crime Code Groups	Precision	Recall	F1-Score	Support
Theft	0.624	0.824	0.710	664815
Assault	0.447	0.459	0.453	317756
Minor Offense	0.874	0.224	0.357	247259
Sexual & Crimes Regarding Children	0.497	0.165	0.247	57078
Accuracy			0.589	1286908
Macro Average	0.610	0.418	0.442	1286908
Weighted Average	0.623	0.589	0.558	1286908

Fig. 4. The Confusion Matrix of Categorical Naive Bayes with 4 different crime code groups

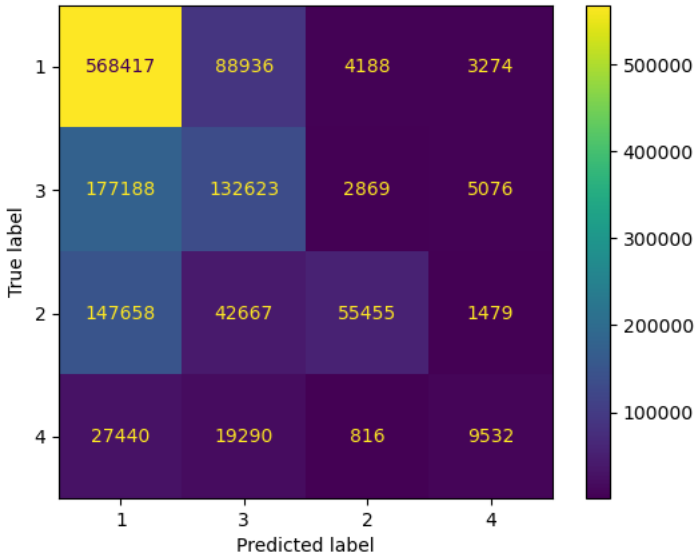


Table 3. Categorical Naive Bayes comparison of different crime code orientations

	Merged Crime Codes	Crime Codes with >50000 Instances	Crime Codes with >500 Instances
Accuracy	0.595	0.411	0.295
Precision Macro Average	0.613	0.407	0.111
Precision Weighed Average	0.625	0.380	0.079
Recall Macro Average	0.416	0.404	0.268
Recall Weighed Average	0.595	0.411	0.295

Table 6. Categorical Naive Bayes Changes after Utilizing Census Data Set

	Before Utilizing Census Dataset	After Utilizing Census Dataset
Accuracy	0.595	0.589
Precision Macro Average	0.613	0.610
Precision Weighed Average	0.625	0.623
Recall Macro Average	0.416	0.418
Recall Weighed Average	0.595	0.589

Note that information on what crime codes reside in each group can be seen in section Preprocessing. Looking at the given results, it can be seen that groups that model tries to predict does not contain similar number of instances. To solve this unbalanced data problem, we applied undersampling, where groups for Theft, Assault and Minor Offense are sampled to have same number of instances as Sexual & Crimes Regarding Children, which is 57078. Classification report obtained after utilizing undersampling can be seen in Table 4. In addition, changes to the results after utilizing undersampling are compared in Table 5. Results show that undersampling does not improve our model, however, results for undersampling version of our data set give useful insights on success of our merging operation.

Table 4. Categorical Naive Bayes model results with 4 different crime code groups after utilizing undersampling

Crime Code Groups	Precision	Recall	F1-Score	Support
Theft	0.466	0.539	0.500	57078
Assault	0.420	0.454	0.436	57078
Minor Offense	0.573	0.330	0.419	57078
Sexual & Crimes Regarding Children	0.511	0.606	0.554	57078
Accuracy			0.482	228312
Macro Average	0.492	0.482	0.477	228312
Weighted Average	0.492	0.482	0.477	228312

Table 5. Categorical Naive Bayes comparison for utilization of undersampling

	Before Utilizing Undersampling	After Utilizing Undersampling
Accuracy	0.595	0.482
Precision Macro Average	0.613	0.492
Precision Weighed Average	0.625	0.492
Recall Macro Average	0.416	0.482
Recall Weighed Average	0.595	0.482

As the next experiment, census data set attributes are added. Changes to precision, recall and accuracy scores are given in Table 6. Categorized census data set attributes does not necessarily improve our model, therefore, they are not utilized in the best model.

4.1.3 Discussion. Even though Naive Bayes assumes that each attribute is independent of each other, it still gave reasonable results. We also worked with some preprocessing approaches to see how much they affect our results. Despite having played around different changes, only some of them are given here.

Utilization of census data set was expected to improve performance of our models, however, results showed opposite. There may be two reasons why census data set was not able to improve our model:

- Latitude and longitude information of each crime instance is used to get a zip code and correlate according to the obtained zip code. However, some zip codes take huge amount of crime instances while others take little to no instances. Same 'Total Population', 'Total Males', 'Total Females' and 'Median Age' values are moved to instances that have the same zip codes. Therefore, instances that are not in the same group receives 4 more attributes that are same, which diminishes model's ability to divide these instances properly.
- Attributes taken from census data set are not categorical and manually divided to categories, which may influence effectiveness of these attributes.

4.2 Light Gradient Boosting Machine

4.2.1 Experiment Setup. There are two main experiments done with LightGBM. First of all, the data set used in Categorical Naive Bayes experiments are used without any changes except extra categorizations such as Victim Age categorization which is explained in the section Optional Preprocessing Steps. The other setup is preprocessing Time Occurred and obtaining Month Occurred columns are explained in the section Optional Preprocessing Steps. This preprocessing reduced the new feature columns which are generated by one hot encoding technique. After this process, the second experiment setup is done. In the last experiment, label encoder is used for Victim Sex, and Victim Descent.

In terms of 'Crime Code' attribute, 3 different orientations which is the same with Categorical Naive Bayes was utilized. Similarly, for the attributes that have invalid or missing values, removal of these problematic instances are selected and the details can be found in the section Experiment Setup for Categorical Naive Bayes.

4.2.2 Results. First, different crime code orientations are experimented with where attributes from census data set are not utilized. Results showed that version where crime codes are merged to obtain 4 crime code groups gave the best results. In addition, it is seen that utilization of census data set improves model performance for LGBM. Accuracy, precision, recall and f1-score metrics for the experiment where 4 different crime groups are used and census data is utilized are given in Table 7. In addition, confusion matrix for this experiment is given in Figure 5. Similar to categorical Naive Bayes, merged crime codes version gives significantly better results than other two crime code orientations.

Table 7. LightGBM model results with 4 different crime code groups by utilizing Census Data Set

Crime Code Groups	Precision	Recall	F1-Score	Support
Theft	0.631	0.888	0.738	664815
Assault	0.516	0.418	0.462	317756
Minor Offense	0.860	0.230	0.363	247259
Sexual & Crimes Regarding Children	0.593	0.290	0.390	57078
Accuracy			0.619	1286908
Macro Average	0.650	0.456	0.488	1286908
Weighted Average	0.645	0.619	0.582	1286908

Fig. 5. The Confusion Matrix of LightGBM with 4 different crime code groups by utilizing Census Data Set

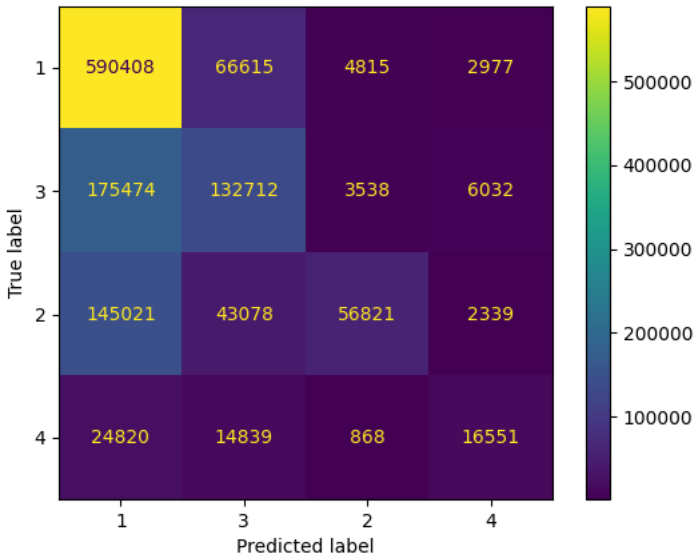


Table 8. LightGBM Changes after Utilizing Census Data Set

	Before Utilizing Census Dataset	After Utilizing Census Dataset
Accuracy	0.616	0.619
Precision Macro Average	0.647	0.650
Precision Weighed Average	0.625	0.645
Recall Macro Average	0.453	0.456
Recall Weighed Average	0.616	0.619

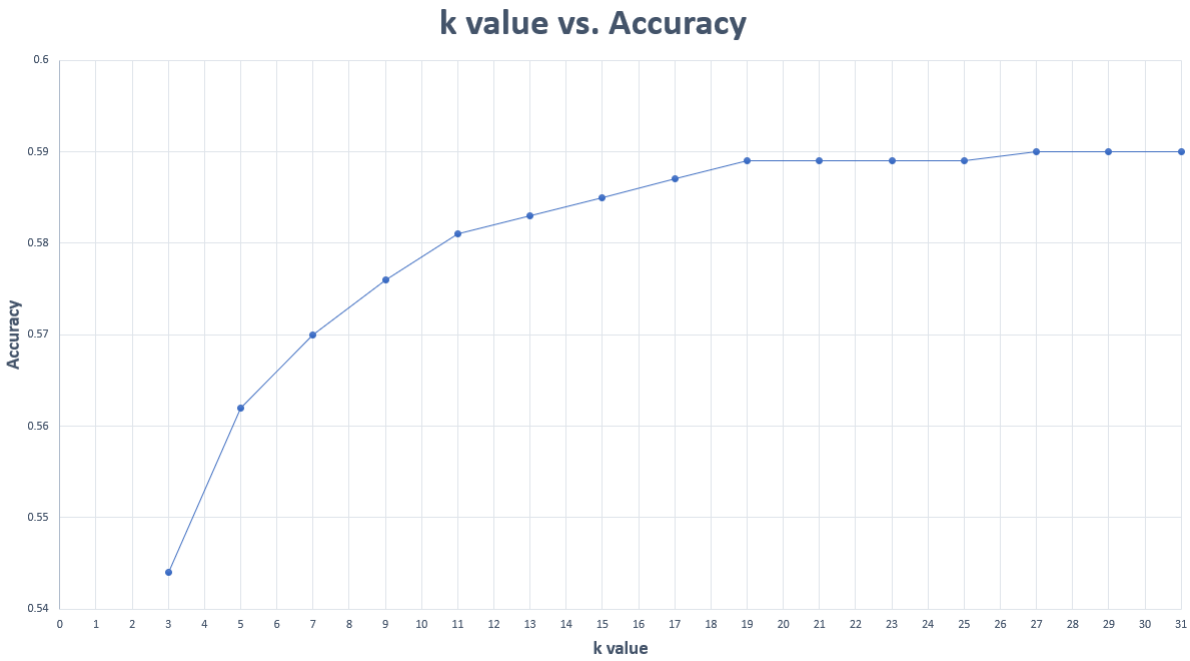
4.2.3 Discussion. When we compare the results of Categorical Naive Bayes and Light Gradient Boosting Machine, we observe that LGBM gives the better accuracy result than Naive Bayes. Furthermore, with small differences, LGBM has better precision and recall values than Naive Bayes. In addition, census data set improves results of LGBM, while decreasing performance of the categorical Naive Bayes models. This may happen due to categorization step done in categorical Naive Bayes. In lGBM raw versions of the census data set features are utilized instead of categorized counterparts.

When undersampling is utilized, LGBM gives worse results. This behavior is similar to categorical Naive Bayes, therefore comparison charts are not given. Similarly, comparisons for different crime code orientations are not given, as changes are similar to categorical Naive Bayes.

4.3 K-Nearest Neighbors

4.3.1 Experiment Setup. With the exact same setup used in LGBM, experiments are also done with K-Nearest Neighbors classifier. Different from other experiments, KNN is lazy learning algorithm, where for each testing instance, k closest training instances are used to decide on the label. Deciding on value of k is important while utilizing KNN. Experiments with different k choices are conducted to decide on the best value for k to use in the rest of our experiments. Accuracy of the models with changing k values can be seen in Figure 6. Scores seem to increase little after k is 9, therefore k is selected as 9. k is not chosen too big such as 31, as with increasing number of k, model becomes slower and neighborhoods become not well defined as they contain instances from more than one classes. Apart from selection of k, same setup is utilized with LGBM.

Fig. 6. Changes in the accuracy with the different k values

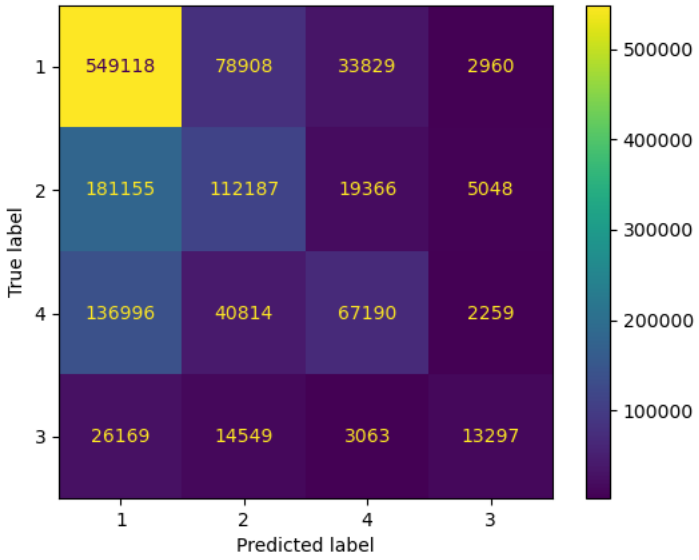


4.3.2 Results. The version where crime codes are merged to obtain 4 crime code groups gave the best results. Accuracy, precision, recall and f1-score metrics for this experiment are given in Table 9. In addition, confusion matrix for this experiment is given in Figure 7.

Table 9. KNN model results with 4 different crime code groups

Crime Code Groups	Precision	Recall	F1-Score	Support
Theft	0.614	0.826	0.705	664815
Assault	0.455	0.352	0.397	317756
Minor Offense	0.544	0.272	0.363	247259
Sexual & Crimes Regarding Children	0.566	0.233	0.330	57078
Accuracy			0.576	1286908
Macro Average	0.545	0.421	0.448	1286908
Weighted Average	0.559	0.576	0.546	1286908

Fig. 7. The Confusion Matrix of KNN with 4 different crime code groups



4.3.3 Discussion. When we compare the results of KNN and Light Gradient Boosting Machine, we observe that LGBM gives the better accuracy result than KNN. Although the KNN data preprocessing part is exactly the same with LGBM data preprocessing, there is a serious difference between them. As k value increases in KNN, accuracy increases, but the values that it approaches to converge are also much lower than LGBM.

LGBM and KNN use very different approaches. LGBM increases accuracy by adding new decision trees at each level. However, KNN looks at the k nearest neighbors and decides accordingly, and the effect of outliers will be higher. Therefore, the LGBM results are much more satisfactory.

5 CONCLUSION AND FUTURE WORK

In this paper, models for classification of crime types using crime and census data sets are given. Different orientations of preprocessing steps and different classification algorithms are comparatively tested to reach the best combinations. In addition, by manually merging different crime types, it is shown that model performances can be increased significantly. In the end, results showed that Light Gradient Boosting Machine where census data set is also utilized gives the best results.

All these experiments showed that our main problem is to specify similar crimes with different crime codes. For this reason, crimes with very close features can be specified with different labels, which reduces the accuracy of the model. Manually grouping crime codes has increased the accuracy significantly, but there are still crime groups with low precision. For this reason, the best crime code grouping can be done by establishing a neural network model, and then the logically incorrectly grouped crimes can be put into different groups manually. Thus, better grouped crimes can be obtained, which can increase the accuracy of the model.

6 THE DISTRIBUTION OF THE WORK

We were always available at similar times because we were colleagues and took exactly the same graduate courses. That's why we've always worked together. We worked synchronously by exchanging ideas, including the literature search and preprocessing parts. While writing code, we have adopted the principle of pair programming. After the preprocessing parts were finished, we shared the Categorical Naive Bayes and LGBM methods during the modeling phase. Serhat Hakkı Akdağ made experiments with Categorical Naive Bayes, while Mert Anıl Yılmaz made experiments with LGBM. We also shared our results with each other and talked about how we can improve the models. In addition, while Serhat Hakkı Akdağ wrote the code that reports the model results, Mert Anıl Yılmaz worked on the visualization parts. Furthermore, we also used a new categorization model, KNN. We worked together while making this decision and evaluating the results. We also decided together on the manual grouping of crime codes. As a result we cannot make a clear distinction about the distribution of work, but this is because we work together completely.

REFERENCES

[1] Albahli, S., Alsaqabi, A., Aldhubayi, F., Rauf, H. T., Arif, M., Mohammed, M., and Tayyab, H. Predicting the type of crime: Intelligence gathering and crime analysis. *Cmc -Tech Science Press- 66* (12 2020), 2318–2341.
[2] Grandini, M., Bagli, E., and Visani, G. Metrics for multi-class classification: an overview.
[3] Gupta, A., Mohammad, A., Syed, A., and Halgamuge, M. A comparative study of classification algorithms using data mining: Crime and accidents in denver city the usa. *International Journal of Advanced Computer Science and Applications 7* (08 2016).
[4] Pareto Software, L. Us zip codes database. https://simplemaps.com/data/us-zips, 2021. [Online; accessed 19-May-2021].
[5] Ratul, M. A comparative study on crime in denver city based on machine learning and data mining.
[6] Scikit-Learn. Naive bayes. https://scikit-learn.org/stable/modules/naive_bayes.html, 2007. [Online; accessed 15-May-2021].