Prediction of criminals' recidivism risk using Bayesian network

March 18, 2020

Sung-gon, Ha

Abstract. The Bayesian network is a graphical model of probability that can be useful for finding prior probabilities in situations where the posterior probability is known. Using this, it is possible to predict the probability of a result based on various variables. In this study, we will analyze the difference between the offender who recommits a crime and the first offenders by various factors and use the Bayesian network to create a model that predicts the probability of a criminal's recidivism. By using this, it is possible to find a factor that greatly affects the recidivism rate, which can be usefully used to lower the recidivism rate and may have an important effect on determining the inmate's parole.

1 Introduction

In today's society, various crimes are occurring, and investigative agencies also provide high-quality security services every year, recording an arrest rate of over 80%. However, on average, about 1.8 million crimes occur annually, and about 1.6 million cases occur in 2018. Due to this situation, the shortage of manpower of investigative agencies and the overcrowding of prisoners' prison facilities such as jails and prisons have become serious. The lack of manpower in the investigative agency leads to a drop in the arrest rate, and the overcrowding of prisoners' inmates may lead to new crimes in prisons. Therefore, measures are needed to reduce the manpower and funds required for criminal investigations and the acceptance of criminals. One way to do this is the parole system. The parole system is a system that conditionally releases prisoners before the expiration of the sentence if the prisoner's behavior is good and fully educated. At this time, the prisoner undergoes an eligibility screening to decide whether to receive parole. During this process, various factors such as the prisoner's age, criminal motive, crime name, prison sentence, corrective grades, health status, life ability after parole, living environment, risk of recidivism, etc. It is decided whether or not to parole. At this time, the factor of risk of recidivism is very ambiguous because it predicts the future potential. In the United States and Canada, an assessment tool called Level of Service Inventory-Revised (LSI-R) is used to predict the risk of criminals' recidivism. LSI-R consists of 10 subcategories (crime record, education and job, financial status, family/marriage relationship, residential environment, leisure and entertainment, peer/friendship, alcohol and drugs, emotions and personality, attitude and orientation) and consists of a total 54 questions. Each question consists of a sentence type that answers 'Yes' or 'No' and a categorical scale from 0 to 3. Using this, the inmate receives a response to each questions and sums the scores for each questions, so that the higher the overall score, the higher the recidivism risk.

In this study, the Bayesian network based on several factors such as LSI-R will be used to predict the risk of recidivism. Bayesian networks group known probabilities into conditional probability tables to help calculate the risk of recidivism using probabilities. This will reveal the risk of recidivism for parole decisions. In addition, it is effective to lower the overall recidivism rate by correcting treatment in the direction of lowering the risk of recidivism by

understanding the importance of each factor.

The data used for this was the statistical data from the Crime Statistics DB of Crime and Criminal Justice Statistical Information (CCJS). This data is a statistically aggregated data of various factors of all criminals and criminals in Korea in 2018. The data used are Characteristics of the recidivist and total offenders in 2018. Characteristics are sex and minority marital and parent relationships, living standards, age, type of crime, criminal history, and if the offender is the recidivist, the duration from the previous crime to the recidivism, education Degree, occupation, and motivation of crime. Here, the standard of living is the degree to which the investigating agency objectively judges the suspect's property status, family relations, status and social status, education and career. Regarding the age of the suspects, the youthful offender aged 10 to 14 was counted except for the fact that statistics were not accurately grasped because they were sent to the juvenile department of the court instead of being sent to the prosecution. In marital and unmarried relationships, criminals were largely classified as married, unmarried, and unknown, and then married again by marital relationship and unmarried by parental relationship. The meaning of the recidivism was aggregated, including cases where investigation or trial was in progress, or detention, fines, probation, prosecution, prosecution, and suspension of prosecution for the previous crime. In addition, the criminal record has been convicted of fines or more. Therefore, those who were detained, fined, protected, prosecuted, held pending, and suspended from prosecution were excluded. There is a case where there is unknown data for each variable, which means that the variable is not entered in the protocol of the case.

2 Prior Knowledge

• Conditional Probability

The conditional probability P(A = true | B = true) represents the probability that A is true if B is true. It represented as follows:

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

P (A = true, B = true) means the probability that A and B are true at the same time.

• Law of total probability

The law of total probability is $\sum_b P(a,b) = \sum_b P(a \mid b)P(b)$. If there are multiple variables, for example a,b,c,d, then P(b) = $\sum_a \sum_c \sum_d P(a,b,c,d)$ So using this and the conditional probability, then

$$\mathrm{P(c \mid b)} = \sum_a \sum_d P(a,c,d \mid b) = \frac{\sum_a \sum_d P(a,b,c,d)}{P(b)}$$

Using Conditional probability and Law of total probability, the total combined probability using chain rule can be expressed as follows:

$$\mathrm{P(a,b,c,...,z) = P(a|\ b,c,...,z)P(b|c,...,z)...P(z)}$$

• Conditional Independence

When P(a, b) = P(a)P(b), a and b are said to be independent. Conditional independence is when P (A|B, C) = P (A|C), that is, when C is given, A and B are independent, but without C, they are not independent.

• Bayesian network

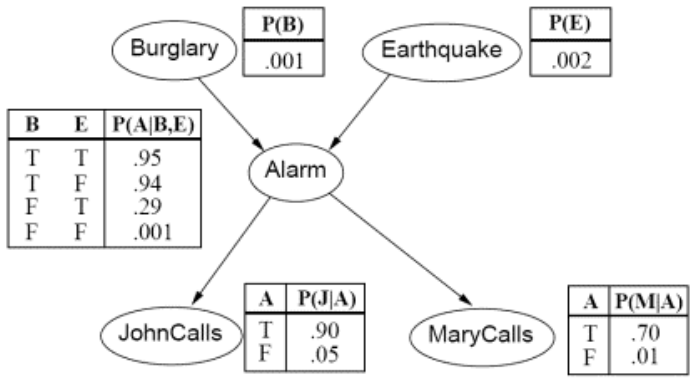

Figure 1: bayesian network

The Bayesian network is directed acyclic graphs that allow efficient and effective representation of the joint probability distribution over a set of random variables Each node in the graph represents a variable, and the arc connecting the nodes represents the dependency between variables. In particular, it is useful for reducing the number of required variables by using conditional independence. Each node is a probability value for each variable. Also, if there is another node connected to the node, it indicates the conditional probability of the node. At this time, the node on which the arrow exits is called the parent, and the node on the arrow receiving side is called the children.

Figure 1 is an example of a simple Bayesian network. The situation is that there is a machine in the house that sounds an alarm for a specific situation. This alarm machine activates when a burglar break in or an earthquake strikes. The owner of the house is currently out, but the neighbors John and Mary will contact the owner if they can contact the owner when the alarm goes off. The table next to each node shows conditional probabilities. The probability of a burglary in the Burglary node, the probability of an earthquake in the earthquake node, and the probability of an alarm in the alarm node. Since the alarm node is connected to the burglary node and the earthquake node, it represents the conditional probability for both. Johncalls and Marycalls represent the probability that John will contact and the probability that Mary will contact, respectively, since they are also associated with the alarm node, indicating the conditional probability. The total number of information required to obtain the total probability P(B, E, A, J, M) is the true or false probability of each of B, E, A, J, M, so a total of $2^5 - 1 = 31$ odds are required. However, for the above Bayesian network, the following equation can be used.

$$P(B,E,A,J,M) = P(B)P(E)P(A|B,E)P(J|A)P(M|A)$$

Therefore, in this case, 1+1+4+2+2= the total probability can be obtained with only 10 probability. Therefore, the number of required information can be reduced. In addition, in previous studies, the risk of recidivism was predicted only by the score. However, Bayesian networks can be used to predict the risk of recidivism as a probability. Also, when calculating the initial probability under a specific condition, the required condition can be reduced. For example, the probability of a burglar breaks, an alarm activates and mary contacting you can be found as follows.

$$\begin{gathered}\text{P(a=true, b=true, m=true)} = \sum_J \sum_E P(a, b, E, J, m) = \\ \sum_J \sum_E P(J \mid a) P(m \mid a) P(a \mid b, E) P(E) P(b) = \\ P(m \mid a) P(b) \sum_J P(J \mid a) \sum_E P(a \mid b, E) P(E)\end{gathered}$$

3 Method

The Bayesian model for predicting recidivism in this study is as follows.

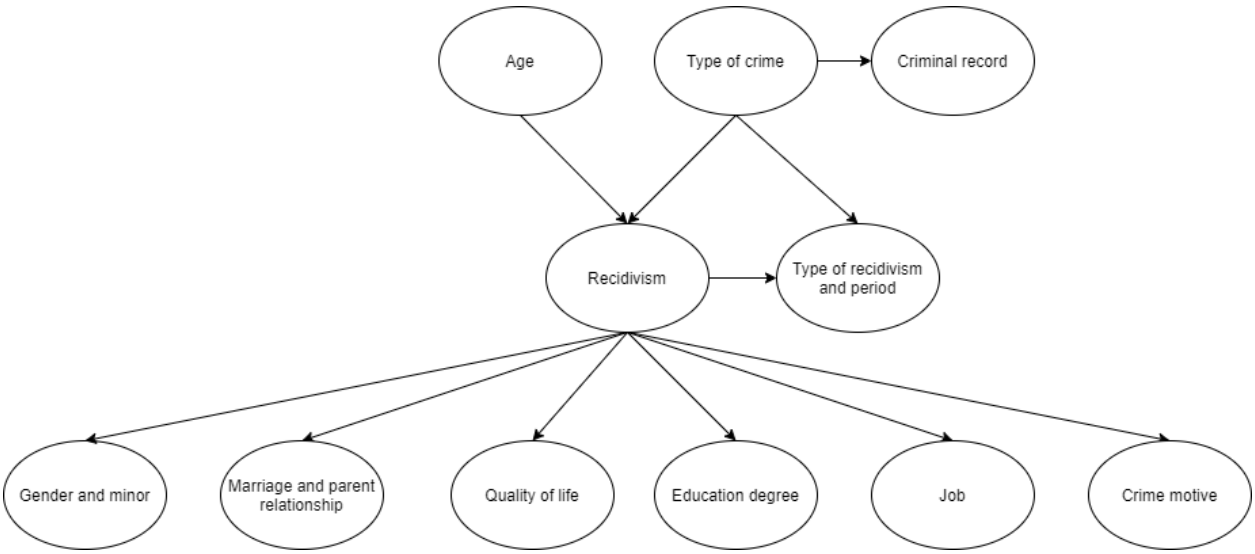

Figure 2: Bayesian model for predicting recidivism

There are 19 variables in the Age nodes: 14, 15, 16, 17, 18, 19, 20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-56, 56-60, 61-65, 66-70, 71 and older, unknown. The unknown data is about 0.1% There are 5 variables in type of crime node: rape and indecent assault, murder and attempted, robbery, assault, and theft. Although there are many types of crime, this study investigated only five crimes called the top five crimes in Korea. There are 10 variables in criminal record node: 1, 2, 3,. . . , 8, 9 or more, none. This represents a conditional probability since it is connected to the type of crime node. There are 2 variables in the recidivism node: recidivist, not recidivist. Since this is a child of the age node and the type of crime node, it represents the conditional probability for both. Each node connected to the recidivism node represents the conditional probabilities of each factor for the recidivism. There are 6 variables in the gender and minor node: adult male and female, minor male and female, and unknown male and female. The unknown is when the gender is entered in the

record but the age is not entered. Unknown data was included in this case, as it was less than about 0.1% of the total and had little effect on the results. There are 13 variables in the marriage and parent relationship nodes: married, cohabitation, divorced, bereavement, real parents, stepparents, real father and stepmother, real father and no mothert, real mother and stepfather, real mother and no father, stepfather and no mother, stepmother and no father, and no parents. Of these, parental relationships were investigated only for singles. The unknown data here were excluded because about 33% of the total was too large to interfere with the results. There are 3 variables in the quality of life node: low, middel, high. In this case, unknown data was also excluded because the unknown data were about 33%. There are 20 variables in education degree node: not attending school, elementary school drop out, graduation, in elementary school, middle school dropout, graduation, in middle school, high school dropout, graduation, in high school, junior college dropout, graduation, in junior college, college dropout, graduation, in college, graduate school dropout, graduation, in graduate school, others. Unknown data was excluded because it was about 36%. There are seven variables in the job node: professional, civil servant, other, employee, self-employed, unemployed, and unknown. Among them, professions include doctors, lawyers, professors, religionists, journalists, and artists. Others include students, housewives, public service personnel, etc. Employees include private teachers, technicians, clerks, factory workers, drivers, security guards, salespeople, employees of state-owned enterprises, general office workers, financial institution employees, entertainment workers, catering workers, and daily workers. Self-employed persons include agriculture, forestry, fisheries, mining, manufacturing, construction, wholesale, retail, trade, catering, lodging, entertainment, finance, real estate, medical health, vehicle maintenance, street vendors, and merchants. The unknown data is about 7%. There are 10 variables in the crime motive node for offenses: lust, meander, retaliation, family discord, curiosity, temptation, accidental, real dissatisfaction, carelessness, and others. Unknown data was about 34% and was excluded. The criminal record node and the type of recidivism and period node are connected to the type of crime node. There are 15 variables in type of recidivism and period node: homogeneous and heterogeneous, within 1 month, 3 months, 6 months, 1 year, 2 years, 3 years, over 3 years, none.

4 Result

Python was used to build the model. I used this because Python's pgmpy package contains functions to help build a Bayesian network model.Using pgmpy, we can create nodes and connect nodes to form a network. Then, by applying conditional probability to each node, we can create a Bayesian network. From the statistical data, we created a conditional probability table for the model above and coded it in Python. At the end of the code, the probability of recidivism comes out when several variables are included, so the probability of recidivism can be predicted according to the characteristics of the offender. We used it to predict the probability of recidivism of famous criminals. The entire code is attached to the appendix.

Lee Chun-jae, the culprit of the Hwaseong serial murder in the 1980s, was arrested on suspicion of attempted robbery in Suwon before being arrested for murder. At this time, he was released from custody as a probation because the charges of murder had not been discovered. Based on the information of Lee Chun-jae at this time : Adult (male), real parent,

26-30 years old, employee, low quality of life, robber, high school graduate, we predict the probability of recidivism. The probability of Lee Chun-jae's recidivism was 61.48%, far exceeding the average recidivism rate of 44.87%. In fact, Lee Chun-jae later committed the 9th and 10th crimes of the serial murder of Hwaseong, and was later arrested for his murder.

We also investigated Jeong Seong-hyun, the culprit of the Anyang elementary schoolchild kidnapping and murder in 2007. Jeong Seong-hyun was arrested for murder in 2004, four years ago in 2007. Jeong Seong-hyun assaulted and killed a woman, but released due to insufficient evidence. Based on the information of Jeong Seong-hyun at this time : Adult (male), real father and stepmother, low quality of life, 36 40 years old, college graduate, employee, murder, we predict the the probability of recidivism. The probability of Jeong Seong-hyun's recidivism is 78.45%, which also significantly exceeds the average recidivism rate. In fact, four years later, Jeong Seong-hyun committed the murder of Anyang elementary schoolchild.

Although the recidivism rate for actual recidivists was high, as shown above, the reliability of the model was tested using test data to more accurately verify the reliability of the model. Because the personal information of the criminals was difficult to obtain, we generate the test data of random criminals using the statistics of 2018, and each variable was randomly generated based on the statistics. The average of the probability of recidivism of bayesian model using 100,000 test data was 48.01%, the predicted recidivism rate was 48.01%, which was only 3% different from the average recidivism rate of 44.87%.

5 Conclusion

In modern society where various crimes are prevalent, it has become important to predict the likelihood of criminals' recidivism and reduce the manpower and funds needed to accommodate criminals.Therefore, the probability of recidivism of criminals can be identified by using the predictive tools for risk of recidivism to solve the overcrowding of prison facilities through parole systems, and furthermore, by analyzing factors contributing to the increase in recidivism probability, corrective treatment provided to criminals to lower the probability of recidivism is needed However, the wrong decision of parole can disturb the society and seriously increase the security problem. Therefore, the recidivism risk prediction tool must have high reliability based on objectivity.

The recidivism prediction model predicted the recidivism probability of the actual recidivists high, and predicted the recidivism probability for arbitrary criminals. However, this model has the following limitations. The first is that statistical data was used. What I felt while building the model is that it is better to use a larger number of individual data than statistical data. In the case of personal data, I can create a conditional probability table by first creating the desired network type and organizing the data accordingly, but in the case of statistical data, I need to create a network that fits the data. My model is bit complicated to increase the efficiency of the overall calculation, but since the criminal record node is conditional independent of the type of crime node and the information of the type of crime node is almost always given, it does not actually affect to the recidivism node. The crime record is highly relevant to the risk of recidivism so that it is used by all other investigations. However, due to the lack of information, the crime record nodes could not

affecting the recidivism node. The second is the unknown data present in each variable. In this study, for variables with more than 30% of unknown data, all unknowns were removed. However, depending on what variable the unknown data actually was, the overall result may vary. The error of the recidivism probability that occurred in the reliability survey is most likely due to an unknown. The third is the lack of diversity of variables. LSI-R, one of the tools for predicting the risk of recidivism, investigates 10 variables including criminal records, education and occupation, financial status, and family relations, etc, and 2 to 4 items for each variable to utilize a total of 54 information. In addition, when parole is judged in Korea, more than 10 information are used, including the prisoner's age, criminal motive, crime name, prison sentence, corrective record, health status, livelihood ability after parole, living environment, risk of recidivism, and other necessary circumstances. There are 10 variables used in this study without criminal record. It is less than another tool for predicting the risk of recidivism. Therefore, there is insufficient information to predict accurately.

On the other hand, the significance of this study was that the computational complexity was effectively reduced by using the Bayesian network, and the risk of recidivism was predicted by constructing the network according to the correlation between the characteristics of criminals that the previous studies did not. Previous studies asked prisoners in a questionnaire format and scored each response to predict the risk of recidivism based on the score. However, in this study, it is significant that the risk of recidivism was predicted by probability according to the characteristics of the prisoners. In addition, if the criminals' personal information can be obtained based on the method implemented in this study, the Bayesian network can be reconfigured to create a more efficient recidivism risk prediction tool.

What is being considered as a follow-up study is not to use statistical data, but rather to construct a network more efficiently by using personal information of criminals. However, since it is difficult to obtain personal information of criminals, it seems that efforts will be required, such as obtaining assistance from criminal prison facilities. In addition, in this study, I used crime statistics data for 2018, and I can use the data from other years to measure the reliability of the model again and multiply the probability of recidivism by the total number of criminals to predict the number of recidivists in the next year. It is also possible to think about which node has the most influence on the probability of recidivism in this model. If such a node can be found, it may be possible to contribute to lowering the recidivism rate of criminals by conducting corrective treatment in a direction that lowers the risk of recidivism.

6 References

손다래 (2019). 남성범죄자 가석방 결정을 위한 재범 위험성 척도, 석사학위, 동의대학교대학원.

정유희, 손외철 (2017) 성인 보호관찰 대상자들의 정적 재범예측요인 분석, 보호관찰, 17:1, 231-271

이민식, 김혜선 (2009). LSI-R을 이용한 성별·범죄유형별 재범유발요인, 형사정책연구 제20권 제1호

이윤 (2016). 범죄수사에서의 의사결정을 위한 베이지안 네트워크 활용 가능성 연구.

한국심리학회지: 법,7(3), 157-180
"범죄 통계 DB" (2020년 4월 27일), CCJS 범죄와 형사사법 통계정보,
https://www.crimestats.or.kr/portal/main/indexPage.do
경찰청 (2018). 경찰범죄통계
이민식, 김혜선 (2009). LSI-R 위험요인과 양형, 한국공안행정학회보, 18(2), 215-247
Friedman N, Geiger D and Goldszmidt M (1997). Bayesian network classifiers.
Machine Learning 29: 131-163

Appendix

A1:Code

```
from IPython.display import Image
```

```
from pgmpy.models import BayesianModel
from pgmpy.factors.discrete import TabularCPD

# Defining the model structure. We can define the network by just passing a list of edges.
model = BayesianModel()
model.add_node("S")
model.add_node("G")
model.add_node("M")
model.add_node("W")
model.add_node("A")
model.add_node("E")
model.add_node("J")
model.add_node("R")
model.add_node("C")
model.add_node("CR")
model.add_node("CV")

model.add_edge("A", "S")
model.add_edge("C", "S")
model.add_edge("S", "G")
model.add_edge("S", "M")
model.add_edge("S", "W")
model.add_edge("S", "E")
model.add_edge("S", "J")
model.add_edge("S", "R")
model.add_edge("C", "CV")
model.add_edge("S", "CV")
model.add_edge("C", "CR")
```

Figure 3: Whole code of Bayesian model for predicting recidivism

```
cpd_A = TabularCPD(variable='A', variable_card=19,
                   values=[[0.005255632],
                           [0.007309463],
                           [0.008390426],
                           [0.009702122],
                           [0.010927214],
                           [0.012175695],
                           [0.011260985],
                           [0.07652084],
                           [0.08184095],
                           [0.084097699],
                           [0.09935256],
                           [0.103889446],
                           [0.12507317],
                           [0.119737256],
                           [0.110696355],
                           [0.066438168],
                           [0.033586991],
                           [0.03215266],
                           [0.001592367]],
                   state_names={'A': ['14세', '15세', '16세', '17세', '18세', '19세', '20세', '21~25세', '26~30세', '31~35세',
                                      '36~40세', '41~45세', '46~50세', '51~55세', '56~60세', '61~65세', '66~70세', '71세이상', '미상']})

cpd_C = TabularCPD(variable='C', variable_card=5,
                   values=[[0.071940295],
                           [0.002642306],
                           [0.003554604],
                           [0.624881005],
                           [0.296981791]],
                   state_names={'C': ['강간,강제추행', '살인기수 및 미수', '강도', '폭행', '절도']})
```

Figure 4: Whole code of Bayesian model for predicting recidivism

```
cpd_S = TabularCPD(variable='S', variable_card=2,
                   values=[[0.094339623, 0.236245955, 0.232415902, 0.270833333, 0.25, 0.230443975, 0.245179063, 0.262516915, 0.335175681,
                            0.445544554, 0.513175525, 0.554041353, 0.587412587, 0.621247113, 0.625065274, 0.590222222, 0.577531646,
                            0.531291611, 0.136363636, 0, 0, 0, 0, 0, 0, 0.214285714, 0.310344828, 0.396551724, 0.46969697, 0.630952381,
                            0.604395604, 0.616666667, 0.688073394, 0.68, 0.528571429, 0.558823529, 0.382978723, 0, 0.407407407,
                            0.470588235, 0.612244898, 0.696428571, 0.583333333, 0.698113208, 0.865384615, 0.66, 0.639175258, 0.690721649,
                            0.706422018, 0.717171717, 0.786407767, 0.746835443, 0.813559322, 0.653846154, 0.75, 0.5, 0, 0.131771596,
                            0.206225681, 0.255673222, 0.273803397, 0.252869898, 0.221750663, 0.219694508, 0.226965803, 0.268185527,
                            0.33356623, 0.38061752, 0.401955434, 0.434151227, 0.444859503, 0.445805667, 0.447975182, 0.461564918,
                            0.413265306, 0.016877637, 0.280354351, 0.412816345, 0.468118196, 0.493861803, 0.483902922, 0.441311853,
                            0.504655493, 0.480116959, 0.508213403, 0.58264796, 0.622405372, 0.637058554, 0.648917197, 0.660874775,
                            0.646991272, 0.629334583, 0.580856541, 0.506462636, 0.125],
                           [0.905660377, 0.763754045, 0.767584098, 0.729166667, 0.75, 0.769556025, 0.754820937, 0.737483085, 0.664824319,
                            0.554455446, 0.486824475, 0.445958647, 0.412587413, 0.378752887, 0.374934726, 0.409777778, 0.422468354,
                            0.468708389, 0.863636364, 1, 1, 1, 1, 1, 0.785714286, 0.689655172, 0.603448276, 0.53030303, 0.369047619,
                            0.395604396, 0.383333333, 0.311926606, 0.32, 0.471428571, 0.441176471, 0.617021277, 1, 0.592592593,
                            0.529411765, 0.387755102, 0.303571429, 0.416666667, 0.301886792, 0.134615385, 0.34, 0.360824742, 0.309278351,
                            0.293577982, 0.282828283, 0.213592233, 0.253164557, 0.186440678, 0.346153846, 0.25, 0.5, 1, 0.868228404,
                            0.793774319, 0.744326778, 0.726196603, 0.747130102, 0.778249337, 0.780305492, 0.773034197, 0.731814473,
                            0.66643377, 0.61938248, 0.598044566, 0.565848773, 0.555140497, 0.554194333, 0.552024818, 0.538435082,
                            0.586734694, 0.983122363, 0.719645649, 0.587183655, 0.531881804, 0.506138197, 0.516097078, 0.558688147,
                            0.495344507, 0.519883041, 0.491786597, 0.41735204, 0.377594628, 0.362941446, 0.351082803, 0.339125225,
                            0.353008728, 0.370665417, 0.419143459, 0.493537364, 0.875]],
                   evidence=['A','C'],
                   evidence_card=[19,5],
                   state_names={'S': ['재범', '비재범'],
                                'A': ['14세', '15세', '16세', '17세', '18세', '19세', '20세', '21~25세', '26~30세', '31~35세', '36~40세',
                                      '41~45세', '46~50세', '51~55세', '56~60세', '61~65세', '66~70세', '71세이상', '미상'],
                                'C': ['강간,강제추행', '살인기수 및 미수', '강도', '폭행', '절도']})
```

```
cpd_G = TabularCPD(variable='G', variable_card=6, values=[[0.851224883, 0.686598655],
                                                          [0.117213728, 0.260769706],
                                                          [0.027532216, 0.039016213],
                                                          [0.003875614, 0.010851979],
                                                          [0.000132427, 0.001529643],
                                                          [0.000021132, 0.001233805]],
                   evidence=['S'],
                   evidence_card=[2],
                   state_names={'G': ['성년(남)', '성년(여)', '미성년(남)', '미성년(여)', '미상(남)', '미상(여)'],
                                'S': ['재범', '비재범']})
```

Figure 5: Whole code of Bayesian model for predicting recidivism

```
cpd_G = TabularCPD(variable='G', variable_card=6, values=[[0.851224883, 0.686598655],
                                                          [0.117213728, 0.260769706],
                                                          [0.027532216, 0.039016213],
                                                          [0.003875614, 0.010851979],
                                                          [0.000132427, 0.001529643],
                                                          [0.000021132, 0.001233805]],
                   evidence=['S'],
                   evidence_card=[2],
                   state_names={'G': ['성년(남)', '성년(여)', '미성년(남)', '미성년(여)', '미상(남)', '미상(여)'],
                                'S': ['재범', '비재범']})

cpd_M = TabularCPD(variable='M', variable_card=13,
                      values=[[0.479722005, 0.434165341],
                              [0.021339151, 0.012526745],
                              [0.136725073, 0.052574455],
                              [0.020129433, 0.021314121],
                              [0.227303983, 0.398914085],
                              [0.001281877, 0.001868271],
                              [0.002075622, 0.001642157],
                              [0.018455635, 0.019140595],
                              [0.001922816, 0.002035031],
                              [0.048762267, 0.040016619],
                              [0.000304198, 0.000282643],
                              [0.000611226, 0.000472014],
                              [0.041366713, 0.015047922]],
                      evidence=['S'],
                      evidence_card=[2],
                      state_names={'M': ['유배우자', '동거', '이혼', '사별', '실(양)부모', '계부모', '실부계모', '실부무모', '실모계부',
                                         '실모무부', '계부무모' ,'계모무부', '무부모'],
                                   'S': ['재범', '비재범']})

cpd_W = TabularCPD(variable='W', variable_card=3,
                      values=[[0.639555967, 0.534670367],
                              [0.348071745, 0.45053389],
                              [0.012372289, 0.014795742]],
                      evidence=['S'],
                      evidence_card=[2],
                      state_names={'W': ['하류', '중류', '상류'],
                                   'S': ['재범', '비재범']})
```

Figure 6: Whole code of Bayesian model for predicting recidivism

```
cpd_E = TabularCPD(variable='E', variable_card=20,
                   values=[[0.00839194, 0.009726884],
                           [0.000083, 0.000196744],
                           [0.013459415, 0.007932344],
                           [0.056246156, 0.036946062],
                           [0.005255708, 0.02738015],
                           [0.024242065, 0.008465937],
                           [0.096354649, 0.050917839],
                           [0.012158086, 0.048792412],
                           [0.049068097, 0.022094902],
                           [0.462998837, 0.365937126],
                           [0.002362549, 0.007479238],
                           [0.007026879, 0.005449201],
                           [0.056287656, 0.071885936],
                           [0.009761448, 0.048419791],
                           [0.033109776, 0.028837842],
                           [0.131195577, 0.212456254],
                           [0.001468812, 0.005514783],
                           [0.001496973, 0.001442786],
                           [0.014888209, 0.024679397],
                           [0.014144169, 0.015444372]],
                   evidence=['S'],
                   evidence_card=[2],
                   state_names={'E': ['불취학', '초재', '초중', '초졸', '중재', '중중', '중졸', '고재', '고중', '고졸', '전문대재',
                                      '전문대중', '전문대졸', '대재', '대중', '대졸', '대학원재', '대학원중', '대학원졸', '기타'],
                                'S': ['재범', '비재범']})

cpd_J = TabularCPD(variable='J', variable_card=7,
                   values=[[0.025858559, 0.047141437],
                           [0.00298666, 0.011520481],
                           [0.054978495, 0.1205964],
                           [0.382496763, 0.358728768],
                           [0.253325125, 0.201158355],
                           [0.228821833, 0.165436493],
                           [0.051532565, 0.095418066]],
                   evidence=['S'],
                   evidence_card=[2],
                   state_names={'J': ['전문직', '공무원', '기타', '피고용자', '자영자', '무직자', '미상'],
                                'S': ['재범', '비재범']})
```

Figure 7: Whole code of Bayesian model for predicting recidivism

```
cpd_R = TabularCPD(variable='R', variable_card=10,
                   values=[[0.137289533, 0.116454328],
                           [0.018407775, 0.012194047],
                           [0.000375815, 0.000099855],
                           [0.020208256, 0.018126604],
                           [0.005484322, 0.016288099],
                           [0.006516026, 0.006593361],
                           [0.258990616, 0.23650637],
                           [0.006391707, 0.00443767],
                           [0.177926103, 0.212497137],
                           [0.368409847, 0.376802528]],
                   evidence=['S'],
                   evidence_card=[2],
                   state_names={'R': ['이욕', '사행심', '보복', '가정불화', '호기심', '유혹', '우발적', '현실불만', '부주의', '기타'],
                                'S': ['재범', '비재범']})
```

```
cpd_CV = TabularCPD(variable='CV', variable_card=15,
                    values=[[0.005397357, 0.004264392, 0.039092055, 0.009797796, 0.041516884, 0, 0, 0, 0, 0],
                            [0.007723804, 0.008528785, 0.035308953, 0.018851119, 0.047570266, 0, 0, 0, 0, 0],
                            [0.009305788, 0.004264392, 0.023959647, 0.021802422, 0.045335171, 0, 0, 0, 0, 0],
                            [0.028010422, 0.010660981, 0.071878941, 0.062562316, 0.115833783, 0, 0, 0, 0, 0],
                            [0.022985297, 0.021321962, 0.029003783, 0.042355193, 0.056007748, 0, 0, 0, 0, 0],
                            [0.017960171, 0.025586354, 0.042875158, 0.038539769, 0.047216376, 0, 0, 0, 0, 0],
                            [0.039177368, 0.031982942, 0.0592686, 0.100769732, 0.070517238, 0, 0, 0, 0, 0],
                            [0.013958682, 0.008528785, 0.029003783, 0.007271905, 0.022257818, 0, 0, 0, 0, 0],
                            [0.027079844, 0.023454158, 0.036569987, 0.018159822, 0.029913018, 0, 0, 0, 0, 0],
                            [00.035548111, 0.036247335, 0.04665826, 0.027146674, 0.030117901, 0, 0, 0, 0, 0],
                            [0.127675414, 0.142857143, 0.197982346, 0.093803592, 0.090707594 ,0, 0, 0, 0, 0],
                            [0.103573423, 0.08315565, 0.098360656, 0.084364739, 0.059826035, 0, 0, 0, 0, 0],
                            [0.101805323, 0.1108742, 0.084489281, 0.082583321, 0.057255676, 0, 0, 0, 0, 0],
                            [0.459798995, 0.488272921, 0.20554855, 0.391991598, 0.285924491, 0, 0, 0, 0, 0],
                            [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]],
                    evidence=['C','S'],
                    evidence_card=[5,2],
                    state_names={'CV': ['동 1개월내', '동 3개월내', '동 6개월내', '동 1년내', '동 2년내', '동 3년초과',
                                        '이 1개월내', '이 3개월내', '이 6개월내', '이 1년내', '이 2년내', '이 3년초과',
                                        '없음'],
                                 'C': ['강간,강제추행', '살인기수 및 미수', '강도', '폭행', '절도'],
                                 'S': ['재범', '비재범']})
```

Figure 8: Whole code of Bayesian model for predicting recidivism

```
cpd_CR = TabularCPD(variable='CR', variable_card=10,
                    values=[[0.151868903, 0.125854993, 0.127516779, 0.137194867, 0.132314102],
                            [0.095527787, 0.095759234, 0.06903164, 0.100006514, 0.086050271],
                            [0.069220651, 0.060191518, 0.073825503, 0.079216029, 0.06075543],
                            [0.05075085, 0.05745554, 0.049856184, 0.062827469, 0.04709485],
                            [0.039460704, 0.039671683, 0.052732502, 0.050189385, 0.03790811],
                            [0.03003398, 0.042407661, 0.037392138, 0.040101254, 0.030986749],
                            [0.024169681, 0.027359781, 0.030680729, 0.032358333, 0.025807113],
                            [0.019511126, 0.021887825 ,0.035474593, 0.025657729, 0.019876144],
                            [0.116463882, 0.177838577, 0.29050815, 0.184210771, 0.176858977],
                            [0.402992437, 0.351573187, 0.232981783, 0.288237648, 0.382348254]],
                  evidence=['C'],
                  evidence_card=[5],
                  state_names={'CR': ['1범', '2범', '3범', '4범', '5범', '6범', '7범', '8범', '9범이상', '없음'],
                               'C': ['강간,강제추행', '살인기수 및 미수', '강도', '폭행', '절도']})
```

```
model.add_cpds(cpd_S, cpd_G, cpd_M, cpd_W, cpd_A, cpd_E, cpd_J, cpd_R, cpd_C, cpd_CR, cpd_CV)
model.check_model()
```

```
True
```

```
from pgmpy.inference import VariableElimination
infer = VariableElimination(model)
```

Figure 9: Whole code of Bayesian model for predicting recidivism

```
res = infer.query(['S'], evidence={'G': '성년(남)', 'M': '실부계모', 'W':'하류', 'A':'36~40세', 'E': '대졸', 'J': '피고용자',
                                   'C':'살인기수 및 미수'})
print(res)
```

```
Finding Elimination Order: : 100%|████████████████████████████████████| 3/3 [00:00<00:00, 2998.79it/s]
Eliminating: CV: 100%|████████████████████████████████████| 3/3 [00:00<00:00, 999.83it/s]
```

```
+--------+----------+
| S      |   phi(S) |
+========+==========+
| S(재범)  |   0.7845 |
+--------+----------+
| S(비재범) |   0.2155 |
+--------+----------+
```

```
from pylab import *
from numpy import *
from random import *
import pandas as pd
```

Figure 10: Whole code of Bayesian model for predicting recidivism

```
def test_model():

    gen_ran = rand()
    if gen_ran < 0.760467962:
        gender = '성년(남)'
    elif 0.760467962 <= gen_ran < 0.956822776:
        gender = '성년(여)'
    elif 0.956822776 <= gen_ran < 0.990686014:
        gender = '미성년(남)'
    elif 0.990686014 <= gen_ran < 0.998407633:
        gender = '미성년(여)'
    elif 0.998407633 <= gen_ran < 0.999310333:
        gender = '미상(남)'
    else:
        gender = '미상(여)'

    mar_ran = rand()
    if mar_ran < 0.464524566:
        marry = '유배우자'
    elif 0.464524566 <= mar_ran < 0.48292395:
        marry = '동거'
    elif 0.48292395 <= mar_ran < 0.591576865:
        marry = '이혼'
    elif 0.591576865 <= mar_ran < 0.612101503:
        marry = '사별'
    elif 0.612101503 <= mar_ran < 0.89665362:
        marry = '실(양)부모'
    elif 0.89665362 <= mar_ran < 0.898131115:
        marry = '계부모'
    elif 0.898131115 <= mar_ran < 0.900062136:
        marry = '실부계모'
    elif 0.900062136 <= mar_ran < 0.91874627:
        marry = '실부무모'
    elif 0.91874627<= mar_ran < 0.92070652:
        marry = '실모계부'
    elif 0.92070652 <= mar_ran < 0.966551289:
        marry = '실모무부'
    elif 0.966551289 <= mar_ran < 0.966848297:
        marry = '계부무모'
    elif 0.966848297<= mar_ran < 0.967413083:
        marry = '계모무부'
    else:
        marry = '무부모'
```

Figure 11: Whole code of Bayesian model for predicting recidivism

```
wel_ran = rand()
if wel_ran < 0.60457525:
    wealth = '하류'
elif 0.60457525 <= wel_ran < 0.986819458:
    wealth = '중류'
else:
    wealth = '상류'

age_ran = rand()
if age_ran < 0.005255632:
    age = '14세'
elif 0.005255632 <= age_ran < 0.012565095:
    age = '15세'
elif 0.012565095 <= age_ran < 0.020955521:
    age = '16세'
elif 0.020955521 <= age_ran < 0.030657643:
    age = '17세'
elif 0.030657643 <= age_ran < 0.041584857:
    age = '18세'
elif 0.041584857 <= age_ran < 0.053760552:
    age = '19세'
elif 0.053760552 <= age_ran < 0.065021537:
    age = '20세'
elif 0.053760552 <= age_ran < 0.141542377:
    age = '21~25세'
elif 0.141542377 <= age_ran < 0.223383327:
    age = '26~30세'
elif 0.223383327 <= age_ran < 0.307481026:
    age = '31~35세'
elif 0.307481026 <= age_ran < 0.406833586:
    age = '36~40세'
elif 0.406833586 <= age_ran < 0.510723032:
    age = '41~45세'
elif 0.510723032 <= age_ran < 0.635796202:
    age = '46~50세'
elif 0.635796202 <= age_ran < 0.755533459:
    age = '51~55세'
elif 0.755533459 <= age_ran < 0.866229814:
    age = '56~60세'
elif 0.866229814 <= age_ran < 0.932667982:
    age = '61~65세'
elif 0.932667982 <= age_ran < 0.966254973:
    age = '66~70세'
elif 0.966254973 <= age_ran < 0.998407633:
    age = '71세이상'
else:
    age = '미상'
```

Figure 12: Whole code of Bayesian model for predicting recidivism

```
cri_ran = rand()
if cri_ran < 0.071940295:
    crime = '강간,강제추행'
elif 0.071940295 <= cri_ran < 0.074582601:
    crime = '살인기수 및 미수'
elif 0.074582601 <= cri_ran < 0.078137205:
    crime = '강도'
elif 0.078137205 <= cri_ran < 0.703018209:
    crime = '폭행'
else:
    crime = '절도'
```

Figure 13: Whole code of Bayesian model for predicting recidivism

```
edu_ran = rand()
if edu_ran < 0.00883526:
    edu = '불취학'
elif 0.00883526 <= edu_ran < 0.008956034:
    edu = '초재'
elif 0.008956034 <= edu_ran < 0.020579969:
    edu = '초중'
elif 0.020579969 <= edu_ran < 0.070416777:
    edu = '초졸'
elif 0.070416777 <= edu_ran < 0.083019768:
    edu = '중재'
elif 0.083019768 <= edu_ran < 0.102022755:
    edu = '중중'
elif 0.102022755 <= edu_ran < 0.18328834:
    edu = '중졸'
elif 0.18328834 <= edu_ran < 0.207612282:
    edu = '고재'
elif 0.207612282 <= edu_ran < 0.247722879:
    edu = '고중'
elif 0.247722879 <= edu_ran < 0.678488591:
    edu = '고졸'
elif 0.678488591 <= edu_ran < 0.682550336:
    edu = '전문대재'
elif 0.682550336 <= edu_ran < 0.689053286:
    edu = '전문대중'
elif 0.689053286 <= edu_ran < 0.750520959:
    edu = '전문대졸'
elif 0.750520959 <= edu_ran < 0.773120416:
    edu = '대재'
elif 0.773120416 <= edu_ran < 0.804811529:
    edu = '대중'
elif 0.804811529 <= edu_ran < 0.962992881:
    edu = '대졸'
elif 0.962992881 <= edu_ran < 0.965805315:
    edu = '대학원재'
elif 0.965805315 <= edu_ran < 0.967284293:
    edu = '대학원중'
elif 0.967284293 <= edu_ran < 0.985424048:
    edu = '대학원졸'
else:
    edu = '기타'
```

Figure 14: Whole code of Bayesian model for predicting recidivism

```
    job_ran = rand()
    if job_ran < 0.037591613:
        job = '전문직'
    elif 0.037591613 <= job_ran < 0.04528289:
        job = '공무원'
    elif 0.04528289 <= job_ran < 0.13643593:
        job = '기타'
    elif 0.13643593 <= job_ran < 0.505829617:
        job = '피고용자'
    elif 0.505829617 <= job_ran < 0.730395683:
        job = '자영자'
    elif 0.730395683 <= job_ran < 0.924273763:
        job = '무직자'
    else:
        job = '미상'

    rea_ran = rand()
    if rea_ran < 0.13047014:
        reason = '이욕'
    elif 0.13047014 <= rea_ran < 0.146844153:
        reason = '사행심'
    elif 0.146844153 <= rea_ran < 0.147129645:
        reason = '보복'
    elif 0.147129645 <= rea_ran < 0.166656573:
        reason = '가정불화'
    elif 0.166656573 <= rea_ran < 0.175676988:
        reason = '호기심'
    elif 0.175676988 <= rea_ran < 0.182218326:
        reason = '유혹'
    elif 0.175676988 <= rea_ran < 0.433849815:
        reason = '우발적'
    elif 0.433849815 <= rea_ran <0.439601964:
        reason = '현실불만'
    elif 0.439601964 <= rea_ran < 0.628843216:
        reason = '부주의'
    elif 0.628843216 <= rea_ran < 1:
        reason = '기타'

    result =  infer.query(['S'], evidence={'G': gender, 'M':marry, 'W':wealth, 'A':age, 'C':crime, 'E': edu, 'J':job, 'R':reason})
    return result
```

Figure 15: Whole code of Bayesian model for predicting recidivism