Experiment No. 2
Predict the survival on the Titanic using Bayesian Network
Date of Performance:
Date of Submission:

Aim: Predict the survival on the Titanic using Bayesian Network

Objective: Ablility to implement Bayesian Network for prediction

Theory:

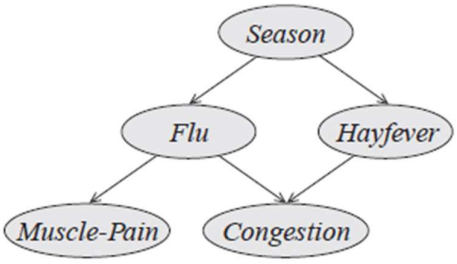
Bayesian Network is directed probability Graphical Model, used to depict cause and effect type of relation between random variables as here arrow is drawn from independent variable to dependent variable.

Using the relationships specified by our Bayesian network, we can obtain a compact, factorized representation of the joint probability distribution by taking advantage of conditional independence.

Bayesian Network is directed acrylic graph where each edge correspond to conditional dependencies and each node correspond to unique random variable.

The Conditional Probability Table associated with each node capture the likelihood of occurrence of that random variable.

Graph Representation

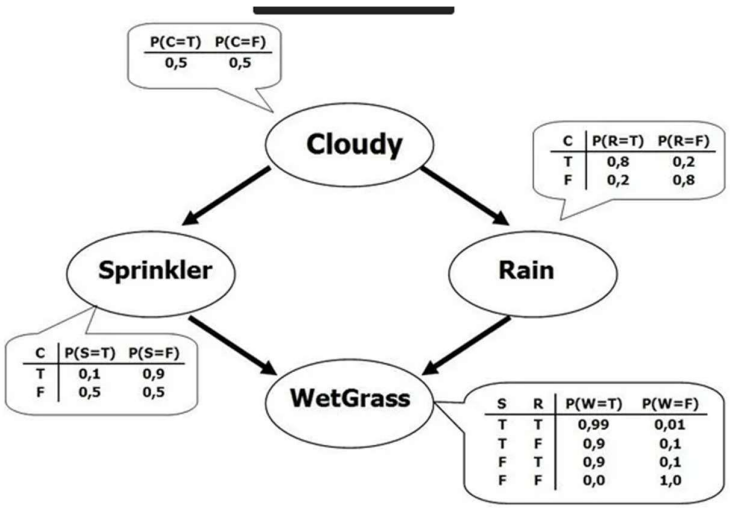


Independencies

$$(F \perp H \mid S)$$
$$(C \perp S \mid F, H)$$
$$(M \perp H, C \mid F)$$
$$(M \perp C \mid F)$$

Factorization

$$P(S, F, H, C, M) = P(S)P(F \mid S)$$
$$P(H \mid S)P(C \mid F, H)P(M \mid F)$$



Applications

1. Medical Diagnosis: BNs are used for diagnosing medical conditions by modeling the relationships between symptoms, test results, and diseases. They can incorporate expert knowledge and update probabilities based on new diagnostic information.

2.Risk Assessment: In fields like finance and insurance, BNs are employed to assess and manage risk. They can model the dependencies between different risk factors, helping decision-makers understand the likelihood and impact of various events.

3. Environmental Modeling: BNs can model complex environmental systems, incorporating variables such as pollution levels, weather conditions, and ecological factors. This is valuable for predicting the impact of changes and making informed decisions in areas like environmental policy.

4. Genetics and Bioinformatics: In genetics, BNs can model the interactions between genes, proteins, and other biological entities. This aids in understanding genetic pathways, predicting the effects of mutations, and analyzing complex biological systems.

5. Manufacturing and Quality Control: BNs are employed in manufacturing to model the relationships between various factors affecting product quality. They can help optimize processes, reduce defects, and improve overall quality control.

6. Fraud Detection: In finance and cybersecurity, BNs can be used for fraud detection. By modeling the relationships between various transactional and behavioral variables, BNs can identify suspicious patterns and activities.

Implementation:
Code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import bnlearn as bn

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

import warnings
warnings.filterwarnings("ignore")

train = pd.read_csv('../input/titanic/train.csv').set_index('PassengerId')
test = pd.read_csv('../input/titanic/test.csv').set_index('PassengerId')

train.head(3)

# Dropping columns with many unique values
drop_list = ['Name', 'Age', 'Cabin', 'Ticket', 'Fare']
train = train.drop(columns=drop_list)
test = test.drop(columns=drop_list)

# Data preparing
dfhot_train, dfnum_train = bn.df2onehot(train)
dfhot_test, dfnum_test = bn.df2onehot(test)

dfnum_target = dfnum_train.pop('Survived')

Xtrain, Xval, Ztrain, Zval = train_test_split(dfnum_train, dfnum_target, test_size=0.2,
random_state=0)
valid = pd.concat([Xval, Zval], axis='columns')
dfnum = pd.concat([Xtrain, Ztrain], axis='columns')

# Get score
def get_acc(model, df, col):
    # Get accuracy score by the model for the validation dataset df with target col
    pred = bn.predict(model, df, variables=[col])
    print(pred)
```

```
    acc = accuracy_score(df[col], pred[col])
    print('Accuracy -', acc)
    return acc

# Structure learning
DAG = bn.structure_learning.fit(dfnum, methodtype='hc', root_node='Survived',
bw_list_method='nodes', verbose=3)

# Plot
G = bn.plot(DAG)

# Parameter learning
model = bn.parameter_learning.fit(DAG, dfnum, verbose=3);

# Get score of the model1
acc1 = get_acc(model, valid, 'Survived')
```

Output:

```
     Survived         p
0           0  0.725084
1           0  0.725084
2           0  0.725084
3           1  0.662098
4           0  0.507407
..        ...       ...
174         0  0.507407
175         0  0.725084
176         1  0.662098
177         0  0.725084
178         0  0.725084

[179 rows x 2 columns]
Accuracy - 0.8156424581005587
```

Conclusion:

Using a Bayesian Network to predict Titanic survival involves considering various factors like age, gender, class, and family size. The network structure should reflect dependencies among these variables. Accuracy depends on data quality, feature selection, and model complexity. Regular validation against ground truth ensures reliability and effectiveness.