Correlation analysis between element and aromatic bond count attributes of illicit drugs

D Pon Mary Pushpalatha¹, D Joseph Pushparaj² and R Gayathri³

¹Department of Digital Sciences, Karunya Institute of Technology and Sciences, Coimbatore, India,
²Department of Information Technology, Francis Xavier Engineering college, Tirunelveli, India,
³Department of Aerospace Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamilnadu, India,

E-mail : ponmarymca@gmail.com, er.joseph.raj@gmail.com, gayathri@karunya.edu

Abstract : World has large amount of data. Data has to be segregated according to the discipline. Biological data has many number of division. This research work the drug data has been taken as input. In this illicit drug data has been taken from data.world website. There are many attributes are available, But the element count and aromatic bond count has been taken from the data set and the data has been statistically analyzed.

1. INTRODUCTION

Data can be represented as individual units of information. Data can be measured qualitative and quantitative way. A datum describes a single quality or quantity of some object or phenomenon. In analytical processes, data are represented by variables. Although the terms "data" and "information" are often used interchangeably, these terms have distinct meanings. Information has to be answered for what an entity is. Information have different meaning in different contexts [1-5].

A drug is taken into the body and it is the substance which is used to cure the body disease. A drug may be the legal or illegal drugs. Illicit drugs are substances that bind the central nervous system and cause hallucinogenic effects and it is banned globally. Data are stored in relational database management system. RDBMS contains rows and column. Every row is called as record and every column is called as attributes. Information can vary for different contexts. Biological data are represented in various formats. Text, Sequence data, protein struct and links are the various formats. There are various databases are used in Biology i.e. GenBank, Protein Data Bank, Scop etc., Biological are displayed in various formats. There are various formats which are included in the biological data are as follows: text, sequence data, protein structure and links. Each of these data can be available from the various sources, for example: Text formats are provided by PubMed and OMIM. Sequence data is provided by GenBank, in terms of DNA, and UniProt, in terms of protein. Illicit Drug Database has been developed by Juan Jose and that data set has been used here for the research [6-9].

2. METHODOLOGY

Illicit drug database descriptors was given by Juan Jose in the data.world website [10] and the dataset
has been taken as input in Figure 1. Except element count and aromatic bond count, other attributes are eliminated. The atomic properties like element identity, isotope, formal charge and hydrogen count are represented using smiles notation. The SMILES atom specification sublanguage represents the atomic properties element identity, isotope, formal charge, and implicit hydrogen count. Atoms are represented by the standard abbreviation of the chemical elements.

Aromatic compounds are called as arenes or aromatics. The examples of aromatic compounds are benzene and toluence (Figure 2). Drug is the substance, which changes the body functionalities either physically or psychologically. Drugs can be classified as legal and illegal drugs. Illicit drugs inhibit into the central nervous system and it is debarred globally. Illicit drugs are substances that can be either stimulate (such as cocaine or amphetamines) or inhibit (such as heroin or sedative-hypnotics) the central nervous system or cause hallucinogenic effects (such as marijuana or LSD), therefore the usage of this has been prohibited globally.

There are various statistical analysis have been made. Mean is the one which calculates the average of columns. Median is calculated from the list of numbers i.e. the middle number. The numbers are arranged in a sorted order. The probability mass function has the maximum value in the mode. It appears most often. Ranking is the relationship between various items. It forms the order.

3. RESULT AND DISCUSSION

Summary statistics are calculated using Excel. The sum of Element count is greater than the sum of Aromatic count that was shown in Table 1. Range values of Element Count is greater than the Aromatic Bonds Count. Percentile and Ranking was calculated and displayed on Table 2. Element count 75 has the highest ranking as hundred percentage. The Aromatic bonds count 18 has highest ranking of 98%. Regression and Statistics have been calculated and the values are acceptable since the R square value is within the range zero to one and it was shown in Table 3. ANOVA Table has been created and it was depicted on Table 4. P-value has the range from 0 to one and it was shown on
Table 5. t-Test: Two-Sample Assuming for mean and Equal Variances has been calculated, the correlation value is within one, the aromatic bonds count and element count values are acceptable, this was shown on Table 6 and Table 7.

**Table 1. Summary Statistics.**

| Aromatic Bonds Count | Element Count |
|----------------------|---------------|
| **Mean**             | **Mean**      |
| 7.256578947          | 41.50657895   |
| **Standard Error**   | **Standard Error** |
| 0.413637419          | 1.017235433   |
| **Median**           | **Median**    |
| 6                    | 40            |
| **Mode**             | **Mode**      |
| 6                    | 53            |
| **Standard Deviation** | **Standard Deviation** |
| 5.099664592          | 12.54132069   |
| **Sample Variance**  | **Sample Variance** |
| 26.00657895          | 157.2847246   |
| **Kurtosis**         | **Kurtosis**  |
| -0.950401138         | -0.736564093  |
| **Skewness**         | **Skewness**  |
| -0.065796145         | 0.108776666   |
| **Range**            | **Range**     |
| 18                   | 65            |
| **Minimum**          | **Minimum**   |
| 0                    | 10            |
| **Maximum**          | **Maximum**   |
| 18                   | 75            |
| **Sum**              | **Sum**       |
| 1103                 | 6309          |
| **Count**            | **Count**     |
| 152                  | 152           |

**Table 2. Percentile and Ranking.**

| Point | Aromatic Bonds Count | Rank | Percent  | Point | Element Count | Rank | Percent |
|-------|-----------------------|------|----------|-------|---------------|------|---------|
| 3     | 18                    | 1    | 98.00%   | 66    | 75            | 1    | 100.00% |
| 67    | 18                    | 1    | 98.00%   | 122   | 71            | 2    | 99.30%  |
| 95    | 18                    | 1    | 98.00%   | 95    | 66            | 3    | 98.60%  |
| 122   | 18                    | 1    | 98.00%   | 93    | 64            | 4    | 98.00%  |

**Table 3. Regression Statistics**

|                          |               |
|--------------------------|---------------|
| **Multiple R**           | 0.304971574   |
| **R Square**             | 0.093007661   |
| **Adjusted R Square**    | 0.086961045   |
| **Standard Error**       | 4.872886173   |
| **Observations**         | 152           |

**Table 4. ANOVA**

|                      | Df | SS     | MS    | F      | Significance F |
|----------------------|----|--------|-------|--------|----------------|
| Regression           | 1  | 365.2405 | 365.2405 | 15.38177 | 0.000133       |
| Residual             | 150 | 3561.753 | 23.74502 |        |                |
| Total                | 151 | 3926.993 |       |        |                |
Table 5. Intercept

|        | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
|--------|--------------|----------------|--------|---------|-----------|-----------|-------------|-------------|
| Intercept | 2.109336318  | 1.370639       | 1.538943 | 0.125925 | -0.59892 | 4.81759 | -0.59892 | 4.81759 |
| Element Count | 0.124010284  | 0.031619       | 3.92196  | 0.000133 | 0.061533 | 0.186487 | 0.061533 | 0.186487 |

Table 6. t-Test: Paired Two Sample for Means.

|                  | Aromatic Bonds Count | Element Count |
|------------------|----------------------|---------------|
| Mean             | 7.256579             | 41.50658      |
| Variance         | 26.00658             | 157.2847      |
| Observations     | 152                  | 152           |
| Pearson Correlation | 0.304972           |               |
| Hypothesized Mean Difference | 0          |               |
| Df               | 151                  |               |
| t Stat           | -35.1542             |               |
| P(T<=t) one-tail | 6.69E-75             |               |
| t Critical one-tail | 1.655007         |               |
| P(T<=t) two-tail | 1.34E-74             |               |
| t Critical two-tail | 1.975799         |               |

Table 7. t-Test: Two-Sample Assuming Equal Variances.

|                  | Aromatic Bonds Count | Element Count |
|------------------|----------------------|---------------|
| Mean             | 7.256579             | 41.50658      |
| Variance         | 26.00658             | 157.2847      |
| Observations     | 152                  | 152           |
| Pooled Variance  | 91.64565             |               |
| Hypothesized Mean Difference | 0          |               |
| df               | 302                  |               |
| t Stat           | -31.1897             |               |
| P(T<=t) one-tail | 9.51E-97             |               |
| t Critical one-tail | 1.649915         |               |
| P(T<=t) two-tail | 1.9E-96              |               |
| t Critical two-tail | 1.96785         |               |

4. CONCLUSION

The aromatic Bonds count and element count of illicit drugs are taken from data.world website. These attributes are analyzed and the correlation values are calculated. The anova table has been created. The mean, standard error and T-test values are calculated. The correlation values are within the range of 0 to 1, so the values are acceptable. The other attributes will be analyzed in the future research work.
REFERENCES

[1]. Princy Suganthi Bai, S., Ponmary Pushpa Latha, D.(2019) Impacts of PCA on face recognition system, in Proc. Int. Conf. on Trends in Electronics and Informatics, ICOEI, pp. 94-97.

[2]. Ahari, V., Venkatesan, R., Latha, D.P.P., (2019) A survey on task scheduling using intelligent water drops algorithm in cloud computing, in: Proc. Int. Conf. on Trends in Electronics and Informatics, ICOEI, pp. 39-45.

[3]. Princy Suganthi Bai, S., Ponmary Pushpa Latha, D. A vital SVKPCA feature set for robust FRS with ensemble neural network classifier, International Journal of Recent Technology and Engineering, 7 (6), 471-479 (2019).

[4]. Patricia Annie Jebamalar, J., Paul, S., Ponmary Pushpa Latha, D., Classifying Road Traffic Data Using Data Mining Classification Algorithms: A Comparative Study, Advances in Intelligent Systems and Computing, 750, 197-205 (2019).

[5]. Princy Suganthi Bai, S., Ponmary Pushpa Latha, D.: Improved facial feature sets for enhancing face recognition system, (2018) International Journal of Engineering and Technology(UAE), 7 (4), 3058-3064.

[6]. Jihad, A.J., Mathew, S.S., Paul, S., Pushpalatha, D.P. (2017) Continuous health monitoring using smartphones-A case-study for monitoring diabetic patients in UAE In Proc. 12th Int. Conf. on Innovations in Information Technology, IIT 2016, art. no. 7880027.

[7]. Ponmary Pushpa Latha, D., Joseph Pushpa Raj, D.: Quantitative structure activity relationship study for the prediction of inhibitory concentration 50 for 5-N-acetyl beta-d-neuraminic acid structurally similar compounds using neural net Asian Journal of Pharmaceutical and Clinical Research, 7 (4), pp. 173-176 (2014).

[8]. Latba, D.P.P., Sharmila, D.J.S.:QSAR study for the prediction of half maximal inhibitory concentration of compounds structurally similar to glycerol [Article@Gliserole yap{dotless} sal olarak benzeyen bileşiklerin yari {dotless} maksimal inhibisyon konsantrasyonları{dotless}ni{dotless}n QSAR çali{dotless}sması{dotless} ile tahmini ], Turkish Journal of Biochemistry, 35 (4), pp. 287-292 (2010).

[9]. Latba, D.P.P., Raj, D.J.P., Sharmila, D.J.S. (2008):Generation of unified data structure and data warehouse for Protein Data Banks In Proc. Int. Conf. on Computational Intelligence and Multimedia Applications, ICCIMA pp. 3-7.

[10]. http://data.world/dataset.