1 Problem Definition

In the Set Cover Problem one is given a universe set $U = \{1, 2, \dots, n\}$, a set of Subsets $S = \{s_1, s_2, \dots, s_m\}$ and a set of costs $C = \{c_1, c_2, \dots, c_m\}$ that assign costs to each subset $s_k \in S$. We can represent the relationship between U and S in a matrix A which is defined in the following way

$$a_{ij} = \begin{cases} 1 & \text{if } i \in s_j \\ 0 & \text{otherwise} \end{cases}$$

To describe a solution X we will use a vector $\vec{x}$ with

$$x_j = \begin{cases} 1 & \text{if } s_j \in X \\ 0 & \text{otherwise} \end{cases}$$

The goal is now to find an X which

$$\text{minimizes } \sum_{j}^{m} x_j \cdot c_j \tag{1}$$

$$\text{with } \sum_{j}^{m} a_{ij} \cdot x_j \geq 1,\ i \in \{1, 2, \dots, n\} \tag{2}$$

2 Features

The table below shows an overview of the selected features

Group	Features	Description	Definition
Subset Size	mean standard deviation median absolute deviation mininum 0.25 - quantile median 0.75 - quantile maximum		$\left\{\frac{\sum_i^n a_{ij}}{\lvert U\rvert} \mid 1 \leq j \leq m\right\}$
Subset Size to Cost ratio	mean standard deviation median absolute deviation mininum 0.25 - quantile median 0.75 - quantile maximum		$\left\{\frac{\sum_i^n a_{ij} \cdot \sum_k^m c_k}{\lvert U\rvert \cdot c_j} \mid 1 \leq j \leq m\right\}$
Element Appearances	mean standard deviation median absolute deviation mininum 0.25 - quantile median 0.75 - quantile maximum		$\left\{\frac{\sum_j^m a_{ij}}{\lvert S\rvert} \mid 1 \leq i \leq n\right\}$
Costs	variation coefficient relative median absolute deviation quartile coefficient of dispersion		$\{c_j \mid 1 \leq j \leq m\}$
Singular elements	count	Elements that appear in only a single set	$\left\{i \mid \sum_j^m a_{ij} = 1\right\}$
Graph	number of connected components shortest cycle longest cycle		$G = (V, E)$ $V = U$ $E = \{(i, j) \mid \exists s \in S : i, j \in s\}$