

In the realm of cancer research, cell lines serve as common in-vitro tumor models. The utilization of genomic data and large-scale drug screening has expedited the selection of the right drugs for cancer patients. The accuracy of drug response prediction is paramount for the success of precision medicine. Given the diversity of data types and the vast volume of big data, few methods efficiently integrate and identify the principal low-dimensional manifold in high-dimensional cancer multi-omics data to predict drug response. To address this challenge, a novel k-means Ensemble Support Vector Regression (kESVR) method is developed. This method predicts drug response values for individual patients based on cell-line gene expression data. kESVR combines supervised and unsupervised learning, relying on embedded clustering (Principal Component Analysis and k-means clustering) and local regression (Support Vector Regression). It is entirely data-driven and is designed to handle missing data and noise from outliers. Comparative analyses are conducted, pitting kESVR against four standard machine learning regression models: simple linear regression, support vector regression, random forest (quantile regression forest), and backpropagation neural network. The results, based on drug response across 610 cancer cells from Cancer Cell Line Encyclopedia (CCLE) and Cancer Therapeutics Response Portal (CTRP v2), showcase kESVR's superior accuracy, as indicated by the smallest mean squared error (MSE) measure. Furthermore, kESVR is compared with 17 existing drug response prediction models, encompassing a range of methods such as regression, Bayesian inference, matrix factorization, and deep learning. Upon ranking these models based on prediction accuracy, kESVR emerges as the top performer in the majority (74%) of cases. Even in the remaining cases (26%), kESVR consistently ranks within the top five models. In conclusion, the introduced model, kESVR, proves to be a powerful tool for drug response prediction using high-dimensional cell-line gene expression data. It outperforms current existing prediction models in terms of accuracy and speed, overcoming overfitting challenges. The potential application of kESVR in developing a robust drug response prediction system for cancer patients, guided by cancer cell-lines and multi-omics data, is highlighted for future exploration.