Collaborative Filtering Recommender System for Predicting Drugs for Prostate Cancer

By

Vishwaben Patel

A thesis submitted in partial fulfillment
of the requirements for the degree of
M.Sc. Computational Sciences

The Faculty of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis Titre de la thèse	Collaborative Filtering Recommender System for Predicting Drugs for Prostate Cancer
Name of Candidate Nom du candidat	Patel, Vishwaben
Degree Diplôme	Master of Science
Department/Program Département/Programme	Computational Sciences
Date of Defence Date de la soutenance	June 18, 2021

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Abdel Omri
(Committee member/Membre du comité)

Dr. Jinan Fiadhi
(External Examiner/Examinateur externe)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Tammy Eger, PhD
Vice-President Research
Vice-rectrice à la recherche
Laurentian University / Université Laurentienne

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

I, **Vishwaben Patel**, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.

Abstract

Prostate cancer is a common type of cancer found in men. Identifying drug targets and inhibitors in drug designing is a challenging task. The Recommender systems (RSs) are regarded as a useful tool and are further considered as optimistic method. The use of tool reflects unprecedented growth and development and a tremendous impact on e-commerce. In the research work, for making the prediction in context of cancer activity class (active/inactive) for compounds extracted from ChEMBL, the RS Methods was used. There are two RS approaches that: Collaborative filtering and Content-based Filtering. From these approaches Collaborative Filtering is applied and successfully conducted the investigation and evaluation for making effective prediction over classes for compounds. In the conducted research the interactions among some of the compounds are known.

Further this way prediction of interaction profiles could be conducted. The gathered result from classification is considered as relatively good prediction and maintains the quality. Then we applied various regression techniques on data set which are Lasso, EN (Elastic Net), CART (Classification and regression trees), KNN (k-nearest neighbors), SVR (Support vector regression), RFR (Random forest regression), GBR (Gradient boosting regression) and ETR (Extra tree regression). After analyzing the data set with regression techniques, we compare their results and then we get best results from SVR technique and this technique can be used to find compounds to fight against prostate cancer in lesser time with more efficiency.

Acknowledgements

It is a great pleasure for me to welcome my dissertation professor, Professor Dr. Kalpdrum Passi Sincerely thank you for supporting my master's degree and encouraging me to provide valuable advice, discussions and guidance during my studies. His guidance in researching and writing this book helped me. I couldn't imagine having more advisors and mentors for my master's degree. Without their guidance and encouragement, this article would not have been possible.

I want to thank my God for bringing me to this point and for giving me the right people to help me through the various stages of my knowledge. I am also grateful to all my friends here in Sudbury and to my friends in India for their encouragement and for changing the course of my career. I could not have achieved this without your help. I would like to express my heartfelt gratitude to my loving and caring parents and dear brothers and sisters, and thank you for your continued support and encouragement during my years of study, research and financial support. I pay this success would not have been possible without it. Thank you very much.

Abbreviations

BCL-2	B-cell lymphoma 2
CBF	Content-based filtering
BMTMKL	Bayesian multi-task multiple kernel learning
EGFR	Epidermal Growth factor receptor
KRAS	Kirsten's rat sarcoma
EGFR	Epidermal growth factor receptor
CwKBMF	Component-wise kernelized Bayesian matrix factorization
CGA	Cancer Genome Atlas
ICGC	International Cancer Genome Consortium
CCLE	Cancer Cell Line Encyclopedia
GDSC	Genomics of Drug Sensitivity in Cancer
DNA	Deoxyribonucleic acid
QSAR	Quantitative structure–activity relationship models
DHDPS	Dihydrodipicolinate synthase
SVM	Support Vector Machine
RNA	Ribonucleic acid
CF	Collaborative Filtering
CS	Cold state problem
SGIMC	Sparse-group inductive matrix completion

Table of Content

List of Tables

List of figures

Chapter 1

Introduction

Advanced algorithms and the rapid boom in available information are highly credited for the popularity of multitask learning. Multitask learning is advantageous in that it has provisions for an opportunity to utilize additional information from similar tasks, and make predictions for the same with limited information. The prediction method is effective in enhancing the prediction performance and is also known for effectively processing small ad erratic data sets prevalent in the drug discovery field. The mechanisms taken in multitask learning are varied and numerous in differing fields, with particular definitions and notations [1].

1.1 Recommender System

Chemoinformatics has begun utilizing the concept of multitask learning, and most notably, its realization in the Proteochemometric and "read across" approach. In e-commerce, Recommender System or RS is another name for multitask learning in factors is one of the approaches that is responsible for making apparent and reality multitask prediction. The enthusiasm recommended by the Recommender system is successfully initiated with announcement of a Netflix Prize competition [2]. It is further responsible for stating the accurate figures for user's preferences and based on these analyses' suggestions are recommended in context of content. This is also known as most accurate system [3].

Malignancy is a hereditary infection brought about by the gathering of cellular and molecular transmutations of the living body, going from direct changes towards duplicate number of varieties and primary modifications. Thus, it requires quality articulation of the physiology of the changes and ultimately adding to the signs of malignancy, including uncontrolled cell multiplication and metastasis. In contrast to ordinarily utilized malignancy medicines, for example, chemotherapy or radiotherapy, directed medications can be better at murdering tumor cells. They may also potentially have lesser poisonous to typical tissues [4]. Furthermore, it's evidenced that only one out of every odd number of patients reacts to medication treatment similarly, and atomic data, for example, change or quality articulation information can inform of patient reaction. For example, KRAS (Kirsten's rat sarcoma viral oncogene homolog) changes can be used as indicators of protection from treatment with Epidermal growth factor receptor (EGFR) Inhibitors, and focusing on over expressed B-cell lymphoma 2(Bcl-2), as seen in smaller cell molecular breakdown in the lungs, appear to give remedial advantages. These results emphasize the need for the use of sub-atomic data to predict drug reactions and tailor malignancy care accordingly. As the quantity of patients/tumors with sub-atomic information increases across malignant growth types, empowered especially by enormous scope studies, for example, The Cancer Genome Atlas (CGA) and The International Cancer Genome Consortium (ICGC), the ID of disease driver qualities as a result has profited significantly. Nonetheless, these information sources ordinarily require drug reaction data and are not reasonable for distinguishing drug reaction biomarkers [5][6][7][8]. However, drug screening in context of malignancy cell lines has been displayed, for instance, in the Cancer Cell Line Encyclopedia (CCLE) and the community-oriented genomics of Drug Sensitivity in Cancer (GDSC) ventures. These cell line data sets permit the use of genomics and apply numerical and measurable ways to

deal with interpreted useful connections. Moreover, they allow developing models that can anticipate tolerant explicit medication reactions [9]. Out of the few models proposed for predicting drug reactions that use genomic highlights, the most generally utilized feature for designing the recommender system is a drug specific model, which is prepared for each medication dependent on hereditary and drug reaction data from cell lines tried with each medication independently. For e.g., a straight relapse model using pattern quality articulation or relying on a mix of quality articulation and other genomics data, duplicate number changes and Deoxyribonucleic Acid (DNA) methylation, nonlinear models, such as neural organizations, arbitrary backend, support vector machines, and piece relapse relying on various kinds of relapse. The models used for the purpose of medication explicitly are restricted to some extent with the aid of quality. A Bayesian model performs different tasks, a multiple bit learning approach shown to have the best display for drug reaction expectation [10]. This will help in creating the amount of knowledge dependent upon vigorous and general models for drug reaction [11]. The work presented in [12], elaborates the significance of sharing data across drugs to improve the exactness of medication reaction forecast. Performing different learning tasks assigns equal significance to all drugs accordingly forecast for specified drug, but it is important to construct a model that organizes data from comparable medications, as is conceivable using collective methods of separation. In the area of recommender frameworks, communication sifting is a system that is responsible for breaking down connections between clients (cell-lines/patients) and conditions among things (drugs) that further aids in recognizing new client illness affiliations (quite explicit medication reaction). The two key groups of synergistic separation strategies are (i) neighborhood techniques that forecast customer association based on predefined customers and similarities, and (ii) idle factor models that use

matrix factorization to distinguish an idle space that captures customer affiliations with symptoms [13] [14]. Specially, matrix factorization strategies, proves to be more effective. Similarly, synergistic separating methods have been utilized for foreseeing quite explicit medication reactions in few studies. In light of a local methodology, Sheng et al [15] characterized drug-explicit cell line closeness and medication basic comparability, and afterward anticipated imperceptibly drug reactions by figuring a weighted normal of watched drug reactions as per both medication and cell line likeness. This model is completely founded on the presumption that the predefined likenesses can clarify drug reactions, but it didn't consider observed drug reaction data to characterize drug comparability [16]. Alternatively, utilizing the inactive factor approach, Khan et al. [17] built segment kernel zed Bayesian network factorization models to anticipate furtive drug reactions dependent on various cell line portions and monitored drug reaction information. Khan et al demonstrated that Component-wise kernelized Bayesian matrix factorization (cwKBMF) can recognize drug pathway affiliations and beat Bayesian multi-task multiple kernel learning (BMTMKL) in drug reaction predictions. Nonetheless, a typical constraint of the two models is a requirement for standardization of medication reaction information, with this preprocessing step prompting lost data on relative positioning of medications inside every cell line [18] [19].

1.2 Classification using RS methods

The information required for performing the classification by use of RS methods can be further classified into three main categories such as collaborative filtering (CF), content-based filtering (CBF), and hybrid approaches among others. In order to make a choice between the approaches to be used, there are two main things under consideration. They are prediction accuracy and

available computational resources. Collaborative Filtering (CF) is the approach that is most commonly used RS methods due to its simplicity. The method gained the popularity during the Netflix competition [20]. Users with similar profiles have preferences which are identical and vice versa. CF model forwards recommendations of new content based on evaluation. In drug discovery, similarly, CF methods predict interaction values and select compound-target. This is highly dependent upon similarity that one could locate in profiles of compound or target interaction.

Out of all RS methods CF has the easiest implementation but with an array of limitations. The main limitation is the cold-start problem (CS) the found interaction values are difficult to be predicted mainly for pairs that consist of new compounds or targets. This is due to inefficiency in finding the similarity during interaction profiles. The second one is the sparsity problem: calculation of similarity becomes more complicated as a result of fewer known interaction values [22] [23] [24] [25]. Further, the third limitation is the Scalability in which quadratic computational and memory complexities of CF algorithms exists. Compounds and targets are characterized by side-channel information. The Content - based Filtering (CBF) RS methods, is an advanced approach and uses interaction values for making prediction which is add on feature. Recommendations with CBF are based on content which have similarity to those chosen by a user previously [26]. In drug discovery, similarity may be employed based on features, or descriptors of compounds and targets. The disadvantages of CF methods are overcome with feature information, which involves prediction for new compounds and targets and very sparse data matrices. Moreover, using CF algorithms has one of a valuable advantage that interprets the things by analyzing different and important features [27]. The target characterization is counted as the main disadvantages in the method of CBF. Among the ever-growing number of multitask

prediction applications in drug discovery, only some of them are categorized under RS approaches. The study of licensed drugs and their potential side effects, drug re-purposing, drug-drug interactions, toxic genomic prediction, or guidelines for treatment were typically included in these studies. On comparison to various methods, an RS algorithm in drug discovery is considered as most successful option. Additionally, a publication describes new methods of matrix completion stating validation on drug databases [28] [29]. RS methods began being applied since 2008 in clinical medicine to improve recommendation schemes in treatment. RS methods, for example, have been used to automatically diagnose omissions in prescription lists in addition to improve care in the light of the issue of information overload, by recommending knowledge-based things of interest to clinicians for particular diseases. RS approach can be compellingly applied in the field of new antiviral searches [30].

1.3 Motivation

Antiviral activity is typically measured in far more complicated structures that comprises of viruses, cells, and compounds. The focus is mainly laid over individual targets that are very different from typical approaches, where targets are defined by individual proteins. One can successfully locate the issue during the current corona pandemic; the search against less studied viruses for broad-spectrum antiviral or therefore it decreases the use of common molecules or privileged groups, such as nucleosides. In existing review papers, analysis of various compilation of a large annotated data set of small-molecule antiviral activity, Viral ChEMBL v. 0.1 has been done [32]. After filtering, data on the activity and inactivity of about 250K compounds were represented as a sparse matrix of compound-virus interactions against 158 viral organisms containing only 400K data points of 40M possible. Typically, the sparsity of interaction matrix

M in the RS setting reaches 90−99%.This is generally calculated using predictive models. In the present study, a recommender system has not been applied for antiviral activity. This study tries to find the best RS method using matrix factorization methods.

1.4 Objectives

The objectives of this research work include: -

(1) To calculate RS approach as an effective approach for conducting the drug activity prediction for prostate cancer.

(2) To find the more accurate prediction approach for prediction of cancer drug.

(3) To have access to prediction results for new compounds or new viral species.

To tackle these challenges, different scenarios with the aim of prediction are developed using compounds and viruses, which were used for following

- Model building
- Prediction of interaction profiles for new compounds or viruses.

1.5 Proposed Methodology

The proposed methodology for recommender system for antiviral activity is shown in Fig1. This research work presents an attempt to apply RS approaches in context of antiviral drug discovery. CF algorithm is used for the completion of the antiviral activity matrix, implemented in Surprise package and sparse-group inductive matrix completion (SGIMC) implementation of CBF. Largely, the accessibility of restricted preparing information, with few cell lines tried with each

medication, speaks to a significant test for learning models that give important expectations in new datasets. Moreover, it helps in gaining natural experiences that is later investigated. To deliver these barriers and to grow more strong models dependent on data sharing over different medications, recommender system has been built up in the literature: the CaDRReS (for Cancer Drug Response forecast utilizing a Recommender System) structure. CaDRReS maps medications and cell lines into an inactive "pharmacogenomic" space to anticipate drug reactions for explicit inconspicuous cell lines and patients [33]. The standardizing examination uses openly accessible datasets (CCLE, GDSC) propositions that are permitted from the CaDRReS to have remarkably preferred prescient execution and superiority over other existing strategies. Correlations on inconspicuous patient-determined cell-line datasets similarly feature CaDRReS's capacity to sum up across datasets, significantly necessary for oncological applications [34].

The various steps of predicting cancer drug response using recommender as shown below in Figure 1 are: -

i) Data cleaning and preprocessing

ii) Data classification into active, inactive and intermediate state.

iii) Applying Lipinski rule on active data.

iv) Applying IC50 rule.

v) Collaborative filtering.

vi) Regression techniques

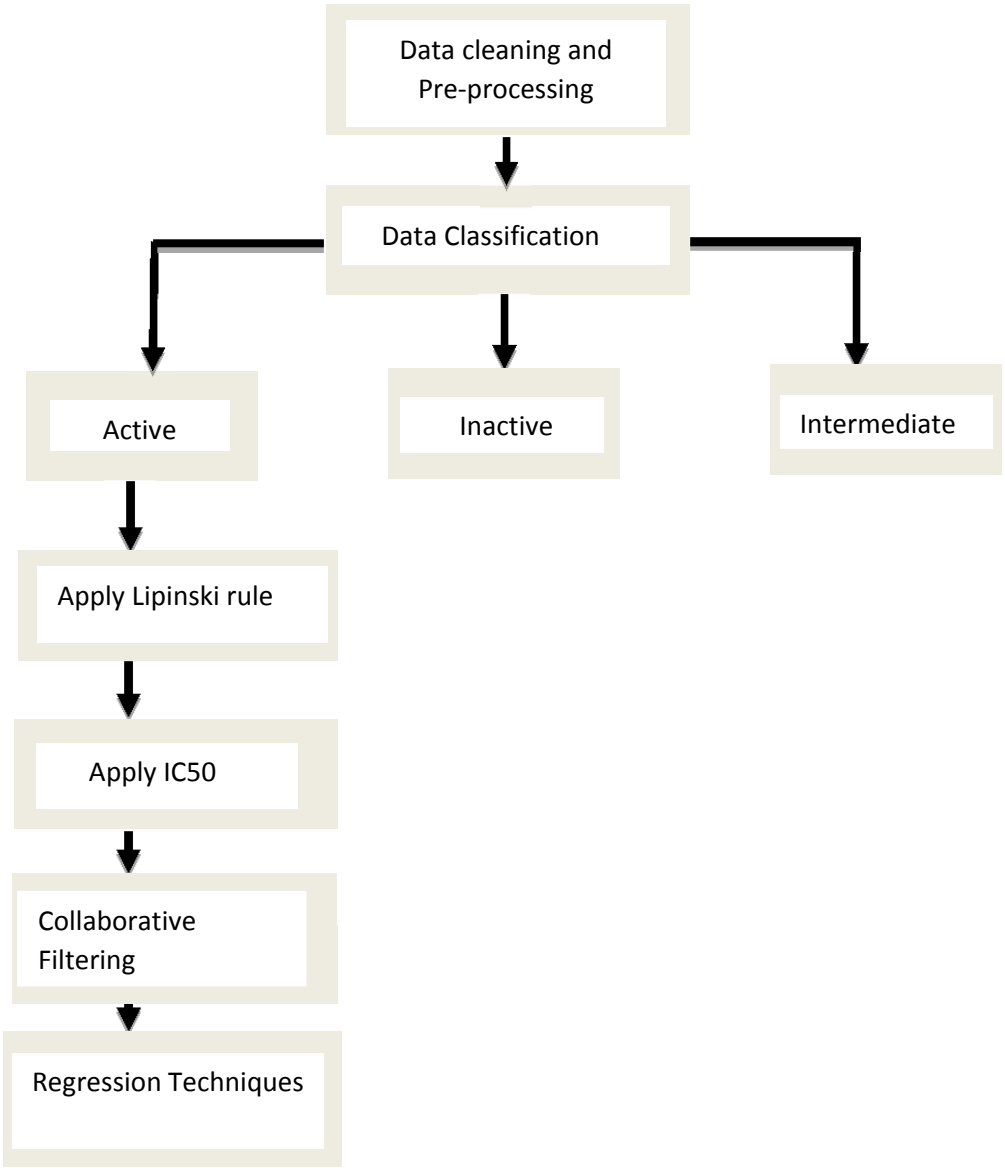

Figure 1.1 *Research Methodology of predicting cancer drug response using Recommender*

Chapter 2

Literature Review

2.1 Review of Existing Techniques

Begg et al. [35] discussed that Cancer is a genetic disorder caused by mutation accumulation, ranging from point mutations to changes in the number of copies and structural modifications. These in turn, influence the expression of genes and eventually lead to cancer hallmarks, including unregulated proliferation of cells and metastasis. Targeted medications are useful and effective in killing tumors cells and/or have less toxicity to normal tissues in comparison to most typical theories such as chemotherapy or radiotherapy. However, the reaction of every patient is not similar to drug therapy and molecular information. This will tell us about how the patient body system responds to a drug. For example, the epidermal growth factor receptor (EGFR) in inhibitor therapy, KRAS mutations can be used for this therapy as predictor of resistance. Also, it is observed in small cell lung cancer, targeting over expressed Bcl-2 has been seen to provide therapeutic benefits. Therefore, the result collected from these observations are than used to demonstrate the need of the use of molecular data to predict drug response and then provide cancer therapy specific to patient molecular structure.

Weinstein et al. [36] discussed that according to large scale studies, such as cancer genome atlas (TCGA) and international cancer genome consortium (ICGC), the number of molecular data,

patient/ tumor is increasing across all cancer types the identification of genes that drive cancer has greatly benefited. However, this data is not appropriate for identification of bio markers of drug responds.

Iorio et al. [37] proposed that on the other hand, drugs screening owned multiple panels of cancer cell lines was carried out under the cancer cell lines encyclopedia (CCLE) and the collaborative genomics of drugs sensitivity in cancer (GDSC). In order to decode functional relationships and create models that can predict patient-specific drug responses, these cell-line datasets are mainly depending upon genomic features. Apply mathematical and statistical approaches are used to draw the useful results.

Menden et al. [38] discussed that several types of models are recommended to use genomic features to predict drug responses. Based on genetic and drug response information from cell lines, a model of drug specific was trained independently for each drug. By using a baseline gene expression, some of the technique would work by including of model of linear regression. Moreover, to support the kernel regression and vector machines, some neural network, random forest and other combination of gene expression changes. However, it identified the variation in the neural network that highlights the information about drug property).

Costello et al. [39] proposed that in order to test the given drug, there are some determinants such as number of cell line that usually talk about the drug specific model. To obtain the best result in the DREAM challenge for drug respond prediction, a model that is known as (BMTMKL) Bayesian multitask multiple kernel learning was developed to generate accurate number of data points and correct result for drug response model. This conducted work

demonstrated the significance of exchanging drug data that improves the accuracy of the prediction.

Koren et al. [40] discussed that Multitask learning assigns equal importance to all drugs, but building a model that prioritizes information from similar drugs is probably more meaningful. The collaborative filtering techniques are used for the purpose. The definition of Collaborative filtering is to identify the connection between the users and items such as drugs in order to fulfill the recommendation system of new patients. (i) The neighborhood methods helped to judge the association between the two items and the two main classes of collaborative filtering method. (ii) The matrix factorization is used by latent factor model that captures the association of user item. In particular, matrix factorization methods successfully result in the Netflix Award, a competition for collaborative filtering methods to predict movie consumer ratings depend upon rating.

Sheng et al. [41] identified unnoticed drug reaction by calculating weighted average of perceived drug reaction based on the similarity of two that is drug and cell line. This approach is based on the predefined similarities which provide drug related response.

Ammad-ud-din et al. [42] developed component-wise kernelized Bayesian matrix factorization (cwKBMF) models to forecast the response of drug by using the various latent factor based on drug related data found with kernels of cell lines. Ammad-ud-din et al. showed the prediction of drug response; cwKBMF could provide information of association of drug pathway and drug-pathway associations and ingenerate BMTMKL. However, the need for normalization of drug response data is a common weakness of both models. The preprocessing stage results in the loss of information. A matrix factorization model based on cell line and drug similarities (SRMF)

was recently proposed by Wang et al., which may outperform cwKBMF. The model does not, however has a projection matrix, and is therefore not tailored to predict the drug reaction of unseen samples. Overall, with a small number of cell lines tested for each drug, the availability of minimal training seen a restrictive state. In addition, there has been no thorough exploration of the interpret ability of models and their use to gain biological insights in the field.

In response to some limitations, CaDRReS a new framework is developed that is regarded as robust model by using knowledge sharing. In order to estimate drug respond for specific and invisible lines of cells and patients, CaDRReS maps medicines and cell lines into concealed pharmacogenomic. Our publicly accessible datasets (CCLE and GDSC) bench marking research shows that this enables CaDRReS to provide substantially improved predictive efficiency and robustness than other current methods. Cell lines datasets derived by patients which cells are invisible often illustrate the robustness and generalization potential of CaDRReS into datasets, key prerequisite for oncology application accuracy. In addition, by logical interpretation fully adopted the unique pharmacogenomic model based upon CaDRReS is well adapted to biotic interpretation, that help in understanding different responses of drugs, (ii) differentiate cellular subtypes from drug response profiles, and (iii) characterize associations of drug pathways. A web server was developed by Singla et al to predict inhibitors against the bacterial target protein GlmU (N-Acetylglucosamine-Uridyltransferase). To build models for predicting inhibitory activity (IC50) of chemical compounds against GlmU protein, Quantitative structure-activity relationship (QSAR) and docking techniques have been used. These models were educated on 84 different compounds taken from the PubChem Bioassay (GlmU inhibitors). These inhibitors were docked at the active site of the GlmU protein (2OI6) C-terminal domain using Auto Dock. They developed a QSAR model using docking energies as descriptors and achieved a complete

correlation between the actual and projected pIC50. Later on, QSAR models were established using molecular descriptors and the full correlation was reached. Finally, the Department of Bioinformatics, KSWU 5 Development of Predictive Systems developed hybrid models using different types of descriptors to screen unknown bio active for anti-cancer against prostate cancer and achieved a high correlation between expected and actual pIC50.

Singla et al. [43] found that there was a high association between certain molecular descriptors used in the analysis and pIC50. Using the models built in the research, about 40 possible GlmU inhibitors were expected. The author's claim that drugs against Mycobacterium tuberculosis could be produced using these inhibitors. The research shows that better results will be obtained by docking energy-based descriptors along with widely used molecular descriptors for predicting inhibitory activity (IC50) of molecules against bacterial target GlmU. An open-source framework for the prediction of GlmU inhibitors available at http:/crdd.osdd.net/raghava/gdoqq was built based on the analysis.

Jamal et al. [44] developed the model with the help of machine learning techniques to predict the new antimalarial compound activities. Using high-throughput screens of anti-malarial agents which inhibited the development of the apicoplast in the Plasmodium malaria parasite, the classification models were developed.

Mishra et al. [45] found that as measured from ROC curve analysis, the Random Forest-based method provided better accuracy. The study indicates that powerful and precise predictive computational models could be designed to screen large data sets in silico and could be used to prioritize high-throughput screen molecules. A MetaPred web server has been created which predicts the metabolization of a drug molecule's isoform. The study shows that the SVM based

QSAR model can predict with high accuracy the substrate specificity of major Cytochrome P450 isoform. The metabolization of drug molecule can be predicted by the help of these molecules. These models also help to understand the molecules show the metabolized or not. Specificity of isoform can be predicted with the help of accurate result models. You can access the server at http:/:/crddosdd.net/raghava/metapred.

Garg et al. [46] developed A KiDoQ web server available at http:/crdd.osdd.net/raghava/kidoq has been developed by Garg enabling the prediction of the inhibitory value of a new ligand molecule against Dihydrodipicolinate synthase (DHDPS). By using docking created energy-based scores as descriptors for QSAR modeling, the study has incorporated both QSAR and docking approaches. QSAR models have been conditioned and tested on Dihydrodipicolinate syntheses inhibitors that have been experimentally confirmed. The inhibitors were docked using Auto Dock software at the DHDPS active site.

The QSAR models based upon Multiple Linear Regression and Support Vector Machine. The study suggests that for the Department of Bioinformatics, KSWU 6 Development of Predictive Systems to screen unknown bio active for anti-cancer against prostate cancer DHDPS using QSAR modeling, ligand-receptor binding interactions is most attractive approach that predicts antibacterial agents. Periwal et al developed binary classification models using four widely used state-of-the-art classifiers from high-performance full-cell screens of anti-tubercular agents, i.e., Naïve Bayes, Random Woodland, J48 and SMO, respectively.

Periwal et al. [47] concluded that predictive models can recognize new active scaffolds that can accelerate the process of Mycobacterium tuberculosis drug discovery when applied to virtual screening of large compound libraries. For plant viral RNA silencing suppressor proteins in plant

viruses, Zeenia Jagga and Dinesh Gupta developed classification models. There are four J48, Random Forest, LibSVM and Naive Bayes in the analysis. The Random forest-based model was best among the four classifiers and demonstrated maximum accuracy among other classifiers. The study further concludes that these models will help to discover novel suppressors of viral Ribonucleic acid (RNA) silencing that will contribute to the design of novel antiviral therapeutics. The authors believe this research will assist in the analysis.

Varnek et al. [48] discussed that Chemoinformatics is an enormous area that uses computer science and chemistry to solve chemistry problems such as molecular graph mining, searching for compound databases, retrieval and extraction of chemical knowledge. Chemo informatics also involves chemical space exploration, scaffold analysis, library architecture, pharmacophores, and computer-aided drug synthesis in drug discovery. Chemo informatics is now recognized as an area that lies at the crossroads of other areas such as computer science and chemistry. Many of the chemo informatics activities in the past were dependent on proprietary or commercial software and proprietary data.

Ramesh [49] Chemo informatics techniques are used commonly in academic field due to the availability of open-source software such as PubChem and ChEMBL. Machine learning is an exciting method for the mining of knowledge from broad compound databases for design drugs with important biological properties, as a large amount of data is accessible in many databases.

Ali et al. [50] discussed that Machine learning helps to understand the links between chemical structures and also their biological activities in drug discovery. The compound structure needs to be translated into chemical knowledge for machine learning, which can be achieved through multiple computational processing using descriptor generation, chemical graph retrieval, etc. It is

essential to evaluate the fundamentals of chemical graph theory to understand how the structure of a chemical compound influences its biological activity. A chemical graph is also referred to as a 'structural graph' or 'molecular graph' and is essentially a construct of mathematical expression consisting of an ordered pair G = (V, E), where V denotes atom vertices connected by a set of edges or bonds E. The chemical structure is fully specified in a graph representation; it includes the necessary data to model and provide insights into a wide range of biological phenomena.

García-Domenech et al. [51] projected several chemical graph variations, like weighted chemical graphs and chemical pseudo graphs. Self-loops and multiple edges are used by chemical pseudographs (reduced graphs) to imprison detailed bond valence information while weighted chemical graphs assign edges and vertices values to indicate atomic properties and bond lengths. Chemical descriptors are a set of numerical values obtained for diversity analysis of compounds, molecular data mining, and activity prediction of compounds from chemical structures.

Hansch et al. [52] classified chemical descriptors as one-dimensional (0D or 1D), 2D, 3D, or 4D. The CaDRReS' predictive performance and robustness is compared with other existing methods, that follows the model depending upon elastic net regression, cwKBMF, SRMF and a control method based on random drug sensor permutations (Control).

2.2 Comparative analysis of existing techniques

The analysis of existing techniques has been demonstrated in the below Table 2.1:

Table 2.1. *Comparison of various Existing Techniques*

Author's Name	Methodology	Technique	Finding	Limitation
Begg et al. [35]	Targeted medications are useful and effective in killing tumor cells.	inhibitor therapy, KRAs mutations can be used for the therapy as predictor of resistance	Less toxic to normal tissues, small cell lung cancer.	The reaction of every patient is not similar to drug therapy and molecular information, provide cancer therapy specific to patient molecular structure.
Iorio et al. [37]	drugs screening owned multiple panels of cancer cell lines was carried out under the cancer cell lines encyclopedia (CCLE)	Collaborative genomics of drugs sensitivity in cancer (GDSC) and mathematical and statistical approaches are used to draw the useful results.	In order to decode functional relationships specific drug responses, cell-line is used.	Datasets are mainly depending upon genomic features.
Menden et al. [38]	Based on genetic and drug response information.	linear regression, vector machines, neural network, random forest	variation in the neural network highlight the information about drug property	A model of drug specific was trained independently for each drug.

Koren et al.[40]	Multitask learning assigns equal importance to all drugs.	collaborative filtering	identify the connection between the users and items such as drugs in order to fulfill the recommendation system of new patients	Does not build a model that prioritizes information from similar drugs.
Sheng et al. [41]	This approach is based on the predefined similarities which provide drug related response.	collective filtering	Identify unnoticed drug reaction by calculating weighted average of perceived drug reaction based on the similarity of two that is drug and cell line.	Does not work on dissimilarities.
Ammad-ud-din et al. [42]	forecast the response of drug by using the various latent factor	component-wise kernelized Bayesian matrix factorization	provide information of association of drug pathway and drug-pathway associations	Problem of normalization.

Chapter 3

Data and Preprocessing

3. 1 Data Preparation

The data contained in the database of CHEMBL is obtained from over 42,500 publications. In the database in total, over 1 million distinct compound structures are presented, with 5.4 million activity values. The data are comprised of 8200 targets that include 5200 proteins (of which 2388 are human). Targets are the different selections in the database with respect to organisms, cancers, protein types, species, etc.

The CHEMBL database is available on https://www.ebi.ac.uk/chembldb. The interface is simple and easy to use and offers all the information with much ease. The user can fetch useful information from the interface such as compound, targets in the desired manner.

The above fetching operation could be understood through the following example: in order to locate the protein of the prostate cancer from the database different pathways can be used on the database. A keyword search of the database can be performed using a protein name, synonym, UniProt accession or ChEMBL target identifier of interest. Another way to fetch the protein is by browsing according to protein family (e.g., to retrieve all chemokine receptors), or organism (e.g., to retrieve all Plasmodium falciparum targets). As another method, BLAST search could be used since the database includes protein targets for which bio activity data are available. Users

perform search on ChEMBL target dictionary with a protein sequence of interest. This is one of the convenient ways to locate closely related proteins with activity data, even if the sequence of interest is not represented in the database.

Users can search for compounds using a keyword search with names/synonyms or ChEMBL identifiers that are developed for this purpose. The properties of compounds can be obtained such as potency, selectivity, ADMET (absorption, distribution, metabolism, and excretion - toxicity) information or closely related compounds. The use of ChEMBL identifier is seen as an effective strategy for searching the required compound. The interface provides a choice of several different drawing tools, allowing users to sketch a suggested structure or substructure of interest. The concerned data for the desired target is achieved through drug-screening data by conducting large-scale studies; from ChEMBL web resource and during this study all the content was kept protected with baseline gene expression data.

The target search is the “Prostate cancer” in the ChEMBL data source. As an output it will display the data of prostate cancer. For the further filtering, the target protein taken is a simple protein or complex protein and the target organism taken is Homo sapiens. As a result, it provides the compounds of all human prostate cancer. An approach known as the Bayesian sigmoid curve fitting approach is implemented to gather the data by calculating raw intensity. Various types of data are obtained and then each compound is tested with each drug at 8 different concentrations Bayesian sigmoid curve fitting is used to draw a curve of drug response for each pair. This was done with the clear motive of re-computing at IC50 (half maximal inhibitory concentration) by considering variation in drug dosages. IC50 is the informative measure of a drug efficacy. It is used to indicate that what amount of drug is needed to inhibit

biological process by half and provide measure of potency of a drug. IC50 can be calculated by plotting x-y graph and fit data in the straight line and thus its value can be calculated with the help of fitted line. Drug dosage of IC50 is a minimal concentration that ensures 50% cell death and the drug dosage obtained in this way is easily comparable across data set. A low IC50 is used that is in minimal amount to obtain 50% inhibition of the target protein. If the Drugs with median IC50 <1 μM then such drug is considered as cytotoxic drugs that is responsible for creating high toxicity across cell-lines and as a result conducting drug response prediction problem becomes easy. Our final data set contained 3373 types of the target protein and these final data sets are used in the models for the purpose of training and validation. Compounds having values of less than 1000 nM are considered to be active while those greater than 10,000 nM are considered to be inactive. As for those values in between 1,000 and 10,000 nM will be referred to as intermediate.

3.2 Data Preprocessing

The source of information is gathered from ChEMBL and is responsible for compound–virus interactions. The data sets create a model for training and cross-validation. The ChEMBL data set as per ICTV (International Committee on Taxonomy of Viruses) taxonomy comprises large activity data points that are extracted from the ChEMBL web source, annotated and standardized by virus species. As data processing output, large data set for cross-validation and training comprised approximately 40,000 interaction values and more than 3000 compounds.

3.2.1. Compound Features

PSA (Prostate-specific antigen) image is used for designing. Two-dimensional descriptors, chemical structures and various features are displayed. Descriptors were selected and calculated using default options, excluding the following set: (1) round descriptor values, (2) precluding descriptors with all missing values (3) precluding descriptors with constant and near-constant values, and (4) precluding descriptors with a standard deviation (SD) of 0.0001. In accordance with the SGIMC (Sparse Group Inductive Matrix Completion) requirements, the features with "NaN" (Not a Number) values were omitted. The features were reduced to investigate their effect on the production for further cross-validation tests. Lipinski's rule is used to define that whether a compound has pharmacological or biological properties that would make it a drug. This rule was given by Christopher A. Lipinski in 1997. The Lipinski's Rule states that to be an orally effective drug, it has to meet the following criteria:

Hydrogen bond acceptors < 10 (all nitrogen or oxygen atoms)

Hydrogen bond donors < 5(the total number of nitrogen–hydrogen and oxygen–hydrogen bonds)

Octanol-water partition coefficient (LogP) < 5

The Molecular weight < 500 Dalton

To ensure the soundness of selection, three random replicates used for given data set are stated below.

- Prediction of interaction profiles for compounds
- Prediction of new interactions for compounds

- Prediction of interactions for viral species

Prediction of various compounds which are targeting various molecules associated with prostate cancer. As a graph standard Value all data is one side with wide range but after applying Lipinski rule all data distributed from 0-10 in x-axis. So now, it is clearly visible that active compounds are between 400-500. The range was 3000 before applying Lipinski but after applying that it's in between 400-500. Since they possess only the interaction matrix, Collaborative filtering (CF) algorithms cannot execute because of its limitation. The main limitation is cold-start problem (CS): interaction values cannot be reliably predicted for pairs consisting of new compound or targets due to an inability to calculate similarity of their empty pairs. CF algorithms were used to handle the prediction. Hyper parameters for the best models were selected by k fold cross-validation and grid search in the model selection module. External validation was carried out on the Extra data set and Models.

For solving all the challenges, the sparse-group inductive matrix completion (SGIMC) content-based filtering (CBF) algorithm was implemented. With side feature, matrices models were trained on the interaction matrix. Using the interaction matrix, external validation was carried out. Model selection was performed on the basis of grid search of hyper parameters and cross-validation. Stratified fold is variation of K Fold that gives stratified folds using cross-validation object they are made by sample percentage of each class.10 cross validation is used to evaluate models by generating training data using partitioning of original data for training of model.

Using hyper parameters and data sets of the finest model, models for resolving species-wise CS problems were formed without cross-validation for the prediction of the latest interactions between viral species and between known compounds. Model evaluation and model building

were conducted where each species and their activity profiles are not included. These are implemented as external test sets.

3.3 Number of Features and its influence

The selected algorithm has the predictive ability that is based upon additional test that is responsible for calculating the effect of features. The model is created that represents different number of features using feature matrices. The prediction evaluation on reduced matrices was focused on a k-fold stratified cross-validation of hyper parameters from the best interaction prediction model.

On right side various compounds are shown and their values are shown in graph of Compounds targeting molecules associated to prostate cancer. In the table given below graph is shown which is representing the solubility vs. molecular weight. In the above graph, the different color dots are the different target molecule. The graph has different molecule in one compound of prostate cancer which is listed on side.

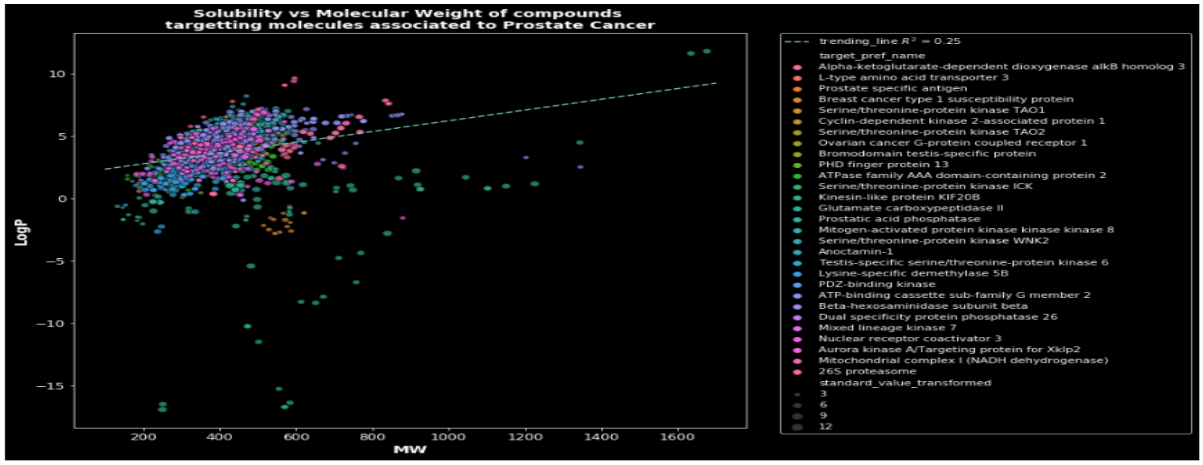

Figure 3.2. Solubility vs. Molecular Weight

Chapter 4

Collaborative Filtering Methods

4.1 Overview of methodology

Figure 4.1 shows the flow of methodology which describes how various methods are applied to get desired output of the recommendation system. The rectangles are showing processes and arrows are showing flow of the data from one step to another step.

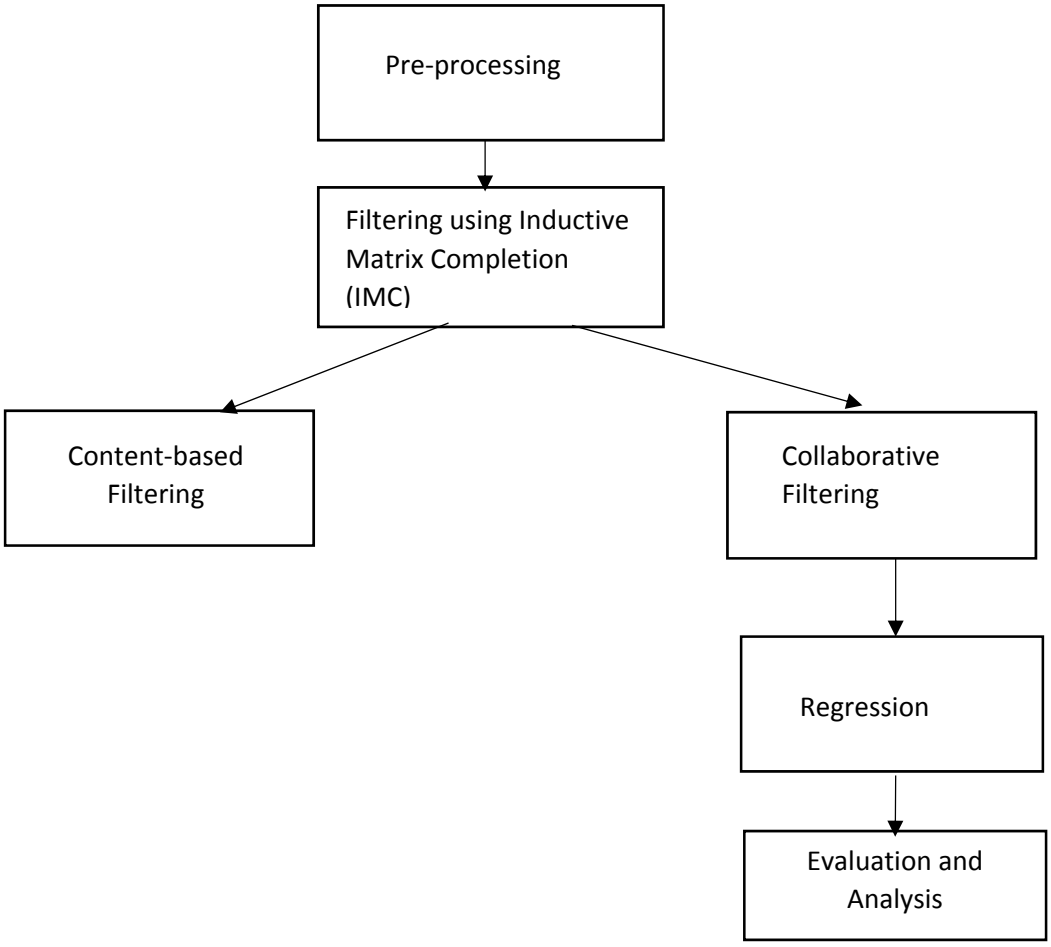

Figure 4.1. *Flow diagram of methodology*

Before applying the methods first data was pre-processed so that unwanted data and noisy data can be removed. For this the data is collected from CHEMBL and then separated active data from non-active data. Next, data filtering is applied using Inductive Matrix Completion (IMC) and then we consider two different types of recommendation systems namely, Collaborative Filtering (CF) and Content-based Filtering (CBF). From both these systems we choose CF over CBF. In CF various techniques are there and we used regression techniques for our prediction. Multiple regression techniques were applied and compared and finally, Singular Value Decomposition (SVD) was the best choice as it gave the least error rate in the results.

4.2 Filtering Methods

A number of filtering methods were applied and they are discussed here.

4.2.1 Inductive Matrix Completion (IMC)

In this we define a cancer associated matrix P $\in R^{N_g \times N_d}$,where each row consists of total number of genes N_g and each column consists of total number of diseases N_d for training, such that $P_{ij} = 1$ if gene $\boldsymbol{i}$ is linked to disease $\boldsymbol{j}$ and 0 for the unobserved relationship. Given a sample Ω for matrix $M \in R^{mxn}$, the main motive is to identify entries which are missing under the structure of the matrix. A common assumption is made that matrix should be of low-rank. By applying low rank standard model for association matrix $P \approx WH^T$, we solve the minimization problem in the equation given below:

$$\min_{W \in R^{N_g X\, k}, H \in R^{N_d x k}} \sum_{(i,j)\in\Omega} (P_{ij} - W_i^T H_j)^2 + \frac{1}{2}\lambda(\left|\left|W\right|\right|_F^2 + \left|\left|H\right|\right|_F^2)$$

In this λ is a regularization parameter, W_i and H_j are latent factors of i and j respectively, it should be noted that W $\in R^{N_g \times k}$ and H $\in R^{N_d \times k}$ factors have value closer to the observed value and the rank of WH^T is small. The associated matrix P is typically much dispersed.

Matrix completion is used to get the loss information by using machine learning techniques. It deals with complex matrices. Using this method, missing data can be found with the help of low rank matrix. The matrix completion problem is resolved by the matrix factorization method by finding latent features that establish the internal relationship in the data (between chemical compounds and cancer).

4.3 Recommender System Approaches

There are two types of approaches that are used in recommendation system as mentioned below.

i. Content-based Filtering (CBF)
ii. Collaborative Filtering (CF)

4.3.1 Content-based Filtering

This approach is based on a description of symptoms and user preference profile. This approach can be used when symptoms data is known but not of patient. In this, keywords are used to describe items and user profiles which can be further used to show the liking of patient. Content-based Filtering follows the idea by recommending the item to user K that is like previous value and is highly rated. Content-based Filtering follows TF-IDF (Term frequency — inverse

document frequency) that has potential to state the importance of document/word/movie etc. There exists transparency in the concept, but the concept becomes ineffective for large data.

Limitations of CBF include:

1. Content analyzed in this method is very limited, hence, not enough to discriminate various drugs for prostate cancer.
2. More advanced techniques like Collaborative Filtering method are better suited in searching a particular drug for prostate cancer.
3. CBF does not have any built-in packages to support development hence handwritten code will be written.

4.3.2 Collaborative Filtering

CF assumes that a person having some kind of symptoms in the past are likely to have particular symptoms in the future also and for them similar kind of treatment will be given. Recommendations are generated by analyzing user information related to symptoms. In this patient or symptoms are grouped together that matches with current user so that recommendation can be given to them based on past experience.

CF has various methods, but the most applied methods are matrix factorization and regression. In our thesis we used regression because it uses data to provide future patterns and the type of statistical calculation it provides makes it easy to analyze which one is best. General view of matrix factorization is given below but however this method is not used in this process.

Matrix Factorization

Matrix Factorization is an embedded model in which feedback matrix $A \epsilon R^{mxn}$ where m denotes number of users and n denotes number of items. The method learns:

A user matrix $U \epsilon R^{mxi}$ in which row m is devoted to user i

An item matrix $V \epsilon R^{nxj}$ in which row n is devoted to user j

These things are learned such that product of UV^T gives good approximation of feedback matrix A. UV^T is just dot product of (U_i, V_j) for user i and item j, that was expected to be close to $A_{i,j}$

Reasons to use CF

1. This method does not depend upon machine analysable content hence complex data can also be identified like appropriate drug for prostate cancer
2. Domain knowledge is not needed in this case because automatic embedding is there.
3. In this only feedback matrix is needed to train the system. There is no need to provide each minute details as these can be learnt automatically.

4.4 Regression

It is a statistical method which helps in analyzing the relationship between two or more interest variables. Regression helps to understand the important factors, the factors that can be ignored and what is the impact of one on another. In this we have one dependent variable and one or more independent variables. The independent variable regresses the value of dependent variable which means how dependent variable is impacted by change in independent variable.

As discussed earlier, regression is used to evaluate the relationship between two or more variables, and it helps to understand what data points represent and which one is the most appropriate approach to find less error rate to find the most appropriate drug for the prostate cancer.

There are various regression techniques and some of them are described below.

Lasso Regression

In this penalty is added for non-zero coefficients to the sum of absolute values. Thus, many coefficients are exactly zeroed for high values. Lasso function is given by:

$$L_{lasso} = \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} |\beta_j|$$

Where, y is observations of response variable, x is a linear combination of predictor variables, β is the estimation of sum of square of residuals.

A tuning parameter, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

When λ = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.

As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when λ = ∞, all coefficients are eliminated).

Elastic Net

Due to critique on lasso, elastic net emerged. Lasso is unstable because its value is also dependent on data. Penalties of both lasso and ridge are combined to form this solution to get best results. Elastic Net aims at minimizing the following loss function:

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

Where α denotes mixing parameter of ridge and lasso.

Decision tree regression

Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. It keenly observes features of the object and tree structure is made to train the model so that it can give meaningful outcomes in future. Its training process takes long time but after that it makes prediction very fast.

$$\bar{y}_i = \frac{1}{m}\sum_{j} y_j$$

In this m is the total number of predictions generated and y_j is the vector of predicted value and $\bar{y}_i$ denotes the predicted values

K-nearest Neighbors Regression

It stores all available cases and then with similarity measure it predicts numerical target. KNN regression uses same function as KNN classification. The KNN regression is used for estimating continuous variable.

$$\sum_{i=1}^{k} |x_i - y_i|$$

In this x_i and y_i are the two real vector points or training points.

Support Vector Regression (SVR)

Like classification it is known for the use of kernels, sparse solution and control of the margin and support vectors. It is a supervised machine learning model that predicts real values rather than class. This model can help in prediction where data is non-linear.

$$y = \sum_{i=1}^{N} (a_i - a_i^*).\langle x_i, x\rangle + b$$

In this $a_i - a_i^*$ denotes the difference between the weights of line draw from input values and $\langle x_i, x\rangle$denotes two input points and b denotes bias.

Random forest Regression

This technique is capable of performing both regression and classification by using multiple decision trees and this technique is called Bootstrap and aggregation. In these multiple trees are combined for determining output rather than depending on one tree. Row sampling and feature sampling in performed randomly from the data sets to form data sets of every model.

$$MSE = \frac{1}{n}\sum_{i=1}^{N} (fi - yi)^2$$

In this N denotes the total data points number, fi denotes returned value of model, yi actual value of data point.

Gradient boosting regression

It is a machine learning technique for regression and classification in which it produces prediction model in the form of group of weak prediction model like decision tree. It depends on the intuition that best model can be obtained by combining the next model with previous model and it also lowers error rate.

$$F_{m+1}(x) = F_m(x) + h_m(x) = y$$

Here m denotes the stages of the algorithm, $F_m(x)$ denotes the imperfect model generated on each stage and $h_m(x)$ denotes new estimator to improve the result of the generated model.

Extra tree regression

This technique uses Meta's estimator that can fit in randomized decision trees for various sub samples of the dataset. It also uses average error rate which should be low and fitting can be controlled. It creates large number of decision trees with the help of training data sets. Prediction is made by calculating average of predictions made by various decision trees.

Gain (S, A) = Entropy(S) - $\sum_{veValues(A)} \frac{|S_v|}{S} Entropy(S_v)$

Here Entropy(S) denotes entropy of the provided data, $Entropy(S_v)$ denotes entropy of data after transformation.

Singular Value Decomposition (SVD)

It is a method of decomposition of one matrix A ϵ C^{mxn} into three matrices as shown below.

$A= U\Sigma V^T$

U denotes m x m unitary matrix, Σ denotes m x n diagonal matrix, V denotes n x n unitary matrix, V^T denotes transpose of V

SVD calculation finds eigen values and eigenvectors of AA^T and A^TA. Columns of V are made by A^TA and columns of U are made by AA^T. In Σ singular values are square root of Eigen values. Singular values are diagonal entries and can be put in descending order. If matrix A is real then U and V are also real.

4.5 Regularization

This technique is used to decrease the error rate by applying a function on training set and to avoid over fitting. In this an additional penalty term is added to an existing function to decrease its error rate. Fluctuation in the result produced by functions is also controlled by it. In simple words we can say that this technique avoids more complex model to be implemented so that over fitting can be avoided.

L1 regularization

In this objective function is minimized which can is computed by adding penalty term to the sum of absolute values of coefficients. It uses minimum absolute values of coefficients. L1 regularization includes Lasso, Elastic Net, Classification and regression trees, K- nearest neighbor.

Cost function for this is:

$$\text{Min } (||Y - X(\Theta)||^2 + \lambda||\Theta||)$$

λ denotes hypermeter and its value equals to αin lasso regularization. It is used to select features from a large number of features.

L2 regularization

In this objective function is minimized by adding penalty term to the sum of coefficient squares. L2 regularization includes Support vector regression, Random forest regression and Extra tree regression.

Cost function for this

$$\text{Min}(||Y - X(\Theta)||^2 + \lambda||\Theta||^2)$$

Λ denotes αparameter. If we change value of αpenalty term will also get affected. αand penalty are directly proportional i.e., if value of αincreases, then penalty value will also increase and magnitude of coefficients will decrease.

4.6 Model evaluation

In this 10-fold cross validation model is evaluated in which data set is partitioned into k equal size data sets from which one set is taken to be validation set and another is taken to be training set. Training of data and validation continues till all sets are covered. Average is taken from each data set error to calculate final error. This algorithm is used to retain a constant proportion of inactive and active class assignments to prevent potential errors. Further grid scan is performed to optimize the algorithm. After that scan is completed then the test data which have labeled data in it and is used to check whether model is generating correct output or not and training data is compared based on similarities and differences so that eligibility of the model can be ensured. Then distance between test and training data is measured so that they can group the elements

which have less distance between them. Mean Squared Error is used to find average of set of errors. It tells the closeness of line of regression to the set of points. This thing is calculated by taking distance from points to regression line and these distances basically represent error points and then after obtaining them squaring of them is done. In this squaring is done because it will remove negative signs as negative distance does not make any sense. It is a risk function which gives expected value of error loss.

Chapter 5

Results and Discussion

5.1 Data Set Description

Many online sites provide variety of datasets which are suitable for various researches and experiments. The dataset is taken for this CHEMBL containing drugs which can have an impact on cancer. The dataset for recommender system for prostate cancer is installed from CHEMBL using 'pip install chembl_webresource_client' command. Figure 5.1 shows part of the dataset which has been installed from CHEMBL.

	cross_references	organism	pref_name	score	species_group_flag	target_chembl_id	target_components	target_type	tax_id
0	[]	Homo sapiens	Alpha-ketoglutarate-dependent dioxygenase alkB...	18.0	False	CHEMBL3112376	[{'accession': 'Q96Q83', 'component_descriptio...	SINGLE PROTEIN	9606.0
1	[{'xref_id': 'O75387', 'xref_name': None, 'xre...	Homo sapiens	L-type amino acid transporter 3	17.0	False	CHEMBL4148	[{'accession': 'O75387', 'component_descriptio...	SINGLE PROTEIN	9606.0
2	[]	Rattus norvegicus	Prostate	17.0	False	CHEMBL613656	[]	TISSUE	10116.0
3	[{'xref_id': 'P07288', 'xref_name': None, 'xre...	Homo sapiens	Prostate specific antigen	16.0	False	CHEMBL2099	[{'accession': 'P07288', 'component_descriptio...	SINGLE PROTEIN	9606.0
4	[]	Homo sapiens	Prostate cells	15.0	False	CHEMBL614850	[]	CELL-LINE	9606.0
...	...	...	...	...	...	...	...	...	...
70	[]	Homo sapiens	Mucin-1	4.0	False	CHEMBL3580494	[{'accession': 'P15941', 'component_descriptio...	SINGLE PROTEIN	9606.0
71	[]	Homo sapiens	Interleukin 13 receptor	4.0	False	CHEMBL3831285	[{'accession': 'P24394', 'component_descriptio...	PROTEIN COMPLEX	9606.0
72	[]	Homo sapiens	Aurora kinase A/Targeting protein for Xklp2	3.0	False	CHEMBL3883304	[{'accession': 'O14965', 'component_descriptio...	PROTEIN COMPLEX	9606.0
73	[]	Homo sapiens	Mitochondrial complex I (NADH dehydrogenase)	0.0	False	CHEMBL2363065	[{'accession': 'P03923', 'component_descriptio...	PROTEIN COMPLEX	9606.0
74	[]	Homo sapiens	26S proteasome	0.0	False	CHEMBL2364701	[{'accession': 'Q99460', 'component_descriptio...	PROTEIN COMPLEX	9606.0

Figure 5.1. *Dataset related to prostate cancer*

After installing the data sets, the target search query is implemented to find those drugs which have a relation with the treatment of prostate cancer. After applying this target search, a target ID is obtained for the compound for treating prostate cancer. Figure 5.2 shows the CHEMBL target IDs for the compounds which are related to prostate cancer.

```
0     CHEMBL3112376
1        CHEMBL4148
3        CHEMBL2099
5     CHEMBL3712961
6     CHEMBL4295936
7        CHEMBL5990
8     CHEMBL3712956
10       CHEMBL5261
11       CHEMBL5475
12       CHEMBL5578
13    CHEMBL1075195
14    CHEMBL3713916
24    CHEMBL1795185
25       CHEMBL2106
27    CHEMBL1764945
28    CHEMBL2150837
29    CHEMBL3580482
30    CHEMBL4105704
31    CHEMBL4296022
36       CHEMBL5660
37    CHEMBL1163126
38    CHEMBL1795198
39    CHEMBL2021752
40    CHEMBL3120039
41    CHEMBL3632454
42    CHEMBL3879825
43    CHEMBL4295952
44       CHEMBL1892
45       CHEMBL2633
47       CHEMBL4899
48       CHEMBL5639
49    CHEMBL2046267
```

Figure 5.2. *CHEMBL target IDs for compounds related to prostate cancer*

After applying target search to obtain target IDs of compounds related to prostate cancer, further target search is applied according to Half-maximal inhibitory concentration (IC50). The most commonly used and insightful indicator of a drug's efficacy is the half-maximal inhibitory concentration (IC50). It indicates the amount of drug needed to inhibit a biological process by half and thus serves as a measure of antagonist drug potency in pharmacological research. A low IC50 means that the drug concentration needs to be low to obtain 50% inhibition of the target

protein. Thus, according to it search query is implemented to get compounds which has low value for IC50. Figure 5.3 shows the list of compounds that have low IC50 value which is required to obtain target protein.

```
['Alpha-ketoglutarate-dependent dioxygenase alkB homolog 3',
 'L-type amino acid transporter 3',
 'Prostate specific antigen',
 'Breast cancer type 1 susceptibility protein',
 'Serine/threonine-protein kinase TAO1',
 'Cyclin-dependent kinase 2-associated protein 1',
 'Serine/threonine-protein kinase TAO2',
 'Ovarian cancer G-protein coupled receptor 1',
 'Bromodomain testis-specific protein',
 'PHD finger protein 13',
 'ATPase family AAA domain-containing protein 2',
 'Serine/threonine-protein kinase ICK',
 'Kinesin-like protein KIF20B',
 'tRNA-dihydrouridine(20) synthase [NAD(P)+]-like',
 'Glutamate carboxypeptidase II',
 'Prostatic acid phosphatase',
 'Mitogen-activated protein kinase kinase kinase 8',
 'Serine/threonine-protein kinase WNK2',
 'Anoctamin-1',
 'Testis-specific serine/threonine-protein kinase 6',
 'Lysine-specific demethylase 5B',
 'PDZ-binding kinase',
 'ATP-binding cassette sub-family G member 2',
 'Beta-hexosaminidase subunit beta',
 'Dual specificity protein phosphatase 26',
 'Mixed lineage kinase 7',
 'Nuclear receptor coactivator 3',
 'Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN',
 'Aurora kinase A/Targeting protein for Xklp2',
 'Mitochondrial complex I (NADH dehydrogenase)',
 '26S proteasome']
```

Figure 5.3. *List of compounds with low IC50 values required for target protein*

5.2 Results of Data cleaning and Pre-processing

The ChEMBL compound interactions are mainly divided into two categories, active compounds and inactive compounds and they are encoded as 1 and 0, respectively. Compounds having less than 1000 nM(nanometer) are considered to be active while those greater than 10,000 nM are considered to be inactive. Values between 1,000 and 10,000 nM are referred to as intermediate. Figure 5.4 shows the frequency of the active and inactive compounds of prostate cancer.

Figure 5.4. *Bio activity class v/s Frequency*

To check which compound is hydrogen bond acceptor and which compound is hydrogen bond donor Lipinski's rule is applied. The Lipinski's Rule states that to be an effective drug, the following criteria should be met:

Molecular weight < 500 Dalton

Octanol-water partition coefficient (LogP) < 5

Hydrogen bond donors < 5

Hydrogen bond acceptors < 10

Dalton is the unit used to express molecular weight of compound. Octanol-water partition coefficient is defined as ratio concentrated solute in water to its concentration in aqueous phase. Hydrogen bond donors and hydrogen acceptors are the donators and acceptors of hydrogen bond respectively. Data obtained after applying Lipinski's rule is checked for normal distribution. To visually explore the distribution of data, inspect different properties of data and the attribute required to transform. Figure 5.5 shows the compound values after applying Lipinski's rule, e.g.,

MW (molecule weight), LogP, number of hydrogen bond acceptors and number of hydrogen bond donors.

Table 5.1shows the standard values for the compounds Table 5.2 shows the transformed values after applying Lipinski's rule. Data obtained after applying Lipinski's rule is checked for normal distribution. To visually explore the distribution of data, inspect different properties of data and the attributes required to transform.

	target_pref_name	molecule_chembl_id	canonical_smiles	standard_value	bioactivity_class	MW	LogP	NumHDon
0	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145420	Cc1ccc2[nH]c(-n3[nH]c(C)c(Cc4ccccc4)c3=O)nc2c1C	10000.0	inactive	332.407	3.55796	
1	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145438	Cc1[nH]n(-c2nc3ccccc3[nH]2)c(=O)c1Cc1ccc(Cl)cc1Cl	10000.0	inactive	373.243	4.24792	
2	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145437	Cc1[nH]n(-c2nc3ccccc3[nH]2)c(=O)c1Cc1ccc(Cl)cc1	10000.0	inactive	338.798	3.59452	
3	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145418	Cc1[nH]n(-c2nc3ccccc3[nH]2)c(=O)c1Cc1ccccc1Cl	10000.0	inactive	338.798	3.59452	
4	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145434	O=c1c(Cc2ccccc2)c(-c2ccccc2)[nH]n1-c1nc2ccccc2...	10000.0	inactive	366.424	4.29970	
...	...	...	...	...	...	...	...	
3370	26S proteasome	CHEMBL491636	CC[C@H](C)[C@@H]1C(=O)O[C@H]1C(=O)N[C@H]1C[C@H...	1440.0	intermediate	369.418	-0.61450	
3371	26S proteasome	CHEMBL3237860	CC[C@H](C)[C@@H]1C(=O)O[C@H]1C(=O)N[C@H]1C[C@@...	5.7	active	563.695	3.90140	
3372	26S proteasome	CHEMBL3237861	CC[C@H](C)[C@@H]1C(=O)O[C@H]1C(=O)NCCC(COCc1cc...	29.0	active	453.579	4.13020	
3373	26S proteasome	CHEMBL3291290	CC(=O)N[C@@H](CC(=O)NCCC(COCc1ccccc1)COCc1cccc...	2300.0	intermediate	609.764	3.32620	

Figure 5.5. *Showing dataset compound values after applying Lipinski's rule*

Table 5.1. *Standard values*

Hydrogen bond donors	Hydrogen bond acceptors
3000	**0**
100	**10000**
90	**15000**
50	**17000**

Table 5.2. *Standard values transformed after applying Lipinski's rule*

Hydrogen bond donors	Hydrogen bond acceptors
100	**4**
300	**5**
280	**6**
500	**7**
50	**8**

After the transformation step is taken, the data gets uniformly distributed in which pIC50 which is known as negative logarithm of IC50 is computed. pIC50 is used when compound is converted into molar. The compounds' standard values get transformed. Figure 5.6 shows standard values that get transformed after applying pIC50. Figure 5.7 shows data set get transformed after applying negative logarithm of IC50.

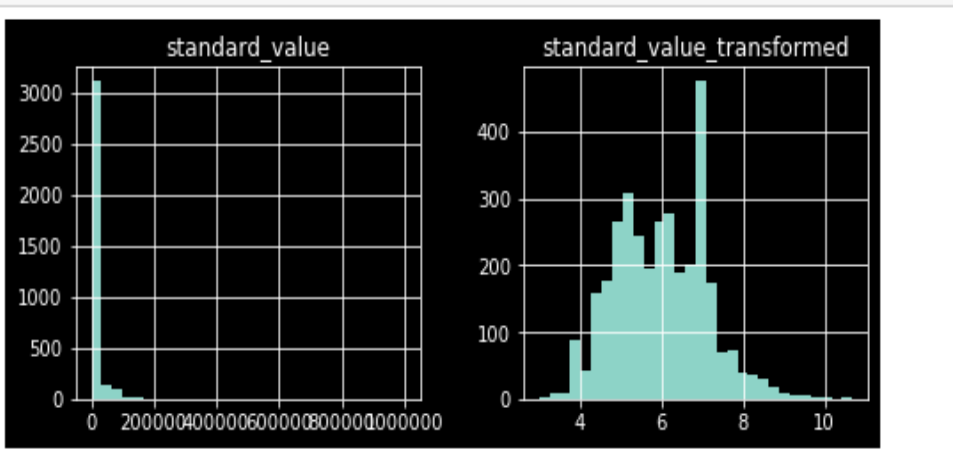

Figure 5.6. *Standard value vs standard value transformed*

	target_pref_name	molecule_chembl_id	canonical_smiles	bioactivity_class	MW	LogP	NumHDonors	NumHAcceptors	stand
0	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145420	Cc1ccc2[nH]c(-n3[nH]c(C)c(Cc4ccccc4)c3=O)nc2c1C	inactive	332.407	3.55796	2.0	3.0	
1	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145438	Cc1[nH]n(-c2nc3ccccc3[nH]2)c(=O)c1Cc1ccc(Cl)cc1Cl	inactive	373.243	4.24792	2.0	3.0	
2	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145437	Cc1[nH]n(-c2nc3ccccc3[nH]2)c(=O)c1Cc1ccc(Cl)cc1	inactive	338.798	3.59452	2.0	3.0	
3	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145418	Cc1[nH]n(-c2nc3ccccc3[nH]2)c(=O)c1Cc1ccccc1Cl	inactive	338.798	3.59452	2.0	3.0	
4	Alpha-ketoglutarate-dependent dioxygenase alkB...	CHEMBL3145434	O=c1c(Cc2ccccc2)c(-c2ccccc2)[nH]n1-c1nc2ccccc2...	inactive	366.424	4.29970	2.0	3.0	

Figure 5.7. *Dataset gets transformed*

Next, the compounds in dataset are sorted on the basis of bioactivity class. Firstly, the active compounds are sorted and then inactive compounds with their respective name, mean_pIC50, number of potential drugs, and standard deviation. Figure 5.8 shows sorted dataset with respect to bioactivity class.

	target_name	bioactivity_class	mean_pCI50	number of potential drugs	std
0	ATP-binding cassette sub-family G member 2	active	6.542876	514	0.395154
1	Lysine-specific demethylase 5B	active	7.022058	434	0.450795
2	Mitogen-activated protein kinase kinase kinase 8	active	7.253563	191	0.692827
3	Glutamate carboxypeptidase II	active	7.648947	141	1.054492
4	Serine/threonine-protein kinase TAO1	active	6.915112	119	0.488263
5	PDZ-binding kinase	active	6.706240	68	0.591532
6	ATPase family AAA domain-containing protein 2	active	6.789362	47	0.386870
7	26S proteasome	active	7.354350	45	0.821879
8	Bromodomain testis-specific protein	active	6.655502	38	0.485977
9	Anoctamin-1	active	6.550490	30	0.370112
10	Mixed lineage kinase 7	active	7.589279	19	0.757974
11	Alpha-ketoglutarate-dependent dioxygenase alkB...	active	6.185997	18	0.076142
12	Nuclear receptor coactivator 3	active	6.542665	15	0.452254
13	Aurora kinase A/Targeting protein for Xklp2	active	6.979054	14	0.791235
14	Prostatic acid phosphatase	active	7.712754	10	0.611243
15	Mitochondrial complex I (NADH dehydrogenase)	active	6.048346	2	0.017160
16	Prostate specific antigen	active	6.645892	2	0.000000
17	Serine/threonine-protein kinase TAO2	active	6.676133	2	0.565103
18	Breast cancer type 1 susceptibility protein	active	6.000000	1	NaN
19	Serine/threonine-protein kinase ICK	active	7.386581	1	NaN
20	ATP-binding cassette sub-family G member 2	inactive	4.550439	197	0.379131
21	Mitogen-activated protein kinase kinase kinase 8	inactive	4.572589	99	0.337852
22	Nuclear receptor coactivator 3	inactive	4.777133	62	0.204941
23	ATPase family AAA domain-containing protein 2	inactive	4.348843	56	0.417396
24	Alpha-ketoglutarate-dependent dioxygenase alkB...	inactive	4.575892	38	0.334190

Figure 5.8. *Sorted dataset w.r.t. bio activity class*

After that, data set obtained after transformation of standard values and application of pIC50 is split. The data gets separated on the basis of variance threshold value = (.8 * (1 - .8))). Figure 5.9 shows compounds data that is obtained after data splitting

	0	1	2	3	4	5	6	7	8	9	...	PDZ-binding kinase	PHD finger protein 13	Prostate specific antigen	Prostatic acid phosphatase	Serine/threonine-protein kinase ICK	Serine/threonine-protein kinase TAO1	Serine/threonine-protein kinase TAO2	Serine/threonin protein kina WNI
0	0	0	0	0	1	1	1	1	1	0	...	0	0	0	0	0	0	0	
1	0	1	0	0	1	1	1	0	0	1	...	0	0	0	0	0	0	0	
2	0	1	0	0	1	1	1	0	0	1	...	0	0	0	0	0	0	0	
3	1	1	0	0	1	1	1	0	0	1	...	0	0	0	0	0	0	0	
4	0	1	0	0	1	1	1	1	1	0	...	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3370	1	1	1	1	0	0	0	1	1	1	...	0	0	0	0	0	0	0	
3371	1	0	1	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
3372	1	0	1	1	0	0	0	0	0	1	...	0	0	0	0	0	0	0	
3373	1	0	1	1	0	0	0	0	0	1	...	0	0	0	0	0	0	0	
3374	1	0	1	1	0	0	0	0	0	1	...	0	0	0	0	0	0	0	

3375 rows × 207 columns

Figure 5.9. *Data splitting from transformed values*

5.3 Comparison of Regression techniques

Collaborative Filtering (CF) technology was used to evaluate recommendation methods for effectiveness. Various filtering techniques were investigated: Lasso(least absolute shrinkage and selection operator), EN (Elastic Net), CART (Classification and regression trees), KNN (*k*-nearest neighbors), SVR (Support Vector Regression), RFR (Random Forest Regression), GBR

(Gradient Boosting Regression) and ETR (Extra Tree Regression). The performance of the models is given in Figure 5.10 and Table 5.3.

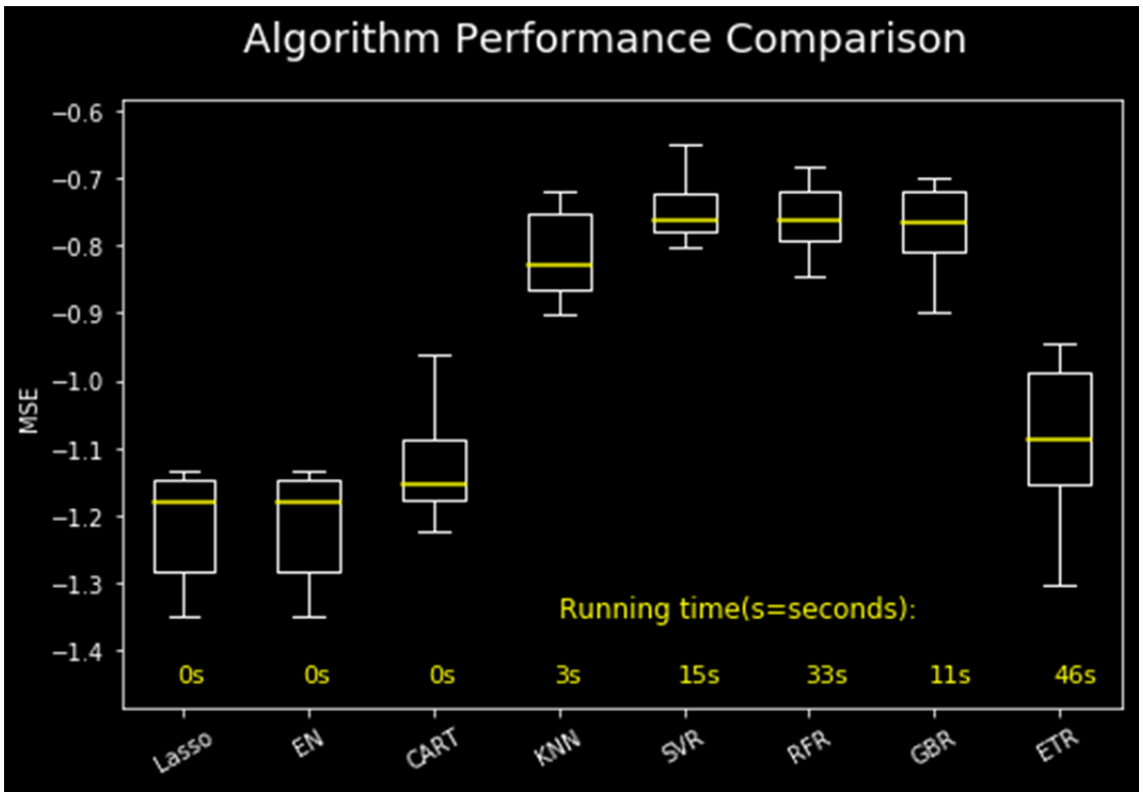

Figure 5.10. *Algorithm performance comparisons*

Table 5.3 shows the Mean Squared Error (MSE), Standard Deviation (STD) and running time for all the algorithms. Lasso regression is a regularization technique which uses shrinkage which means data values are shrunk towards the central point. After running Lasso, output is -1.2 MSE value, 0.08 STD value and it took 2 ms(millisecond)to run. Elastic Net Regression combines penalties of ridge regression and lasso regression to get best from both. After running Elastic Net, output is -1.2 MSE value, 0.08 STD and it took 1 ms to run. Classification and Regression Tree refers to the algorithm in which firstly data is classified and then on the target variable regression algorithm is performed to predict the value. After running Classification and

Regression Tree, output is -1.1 MSE value, 0.09 STD value and it took 2 ms to run. K-nearest Regression is a method that follows non-parametric approach and works in intuitive manner which approximates the value associated with independent variable and continuous outcome by taking average of observations. After running K-nearest Neighbor, output is -0.8 MSE value, 0.06 STD value and it took 3000 ms to run. Support Vector Regression uses same principle as Support Vector Machine but with some minor changes like margin of tolerance is set in this so that only finite possibilities are there. After running Support Vector Regression, output is -0.8 MSE value, 0.06 STD value and it took 15000ms to run. Random Forest Regression uses ensemble learning method for regression. It operates by making multiple decision trees during training time and gives output as the mean of all predictions. After running Random Forest Regression, output is -0.8 MSE value, 0.08 STD value and it took 33000ms to run. Gradient Boosting Regression performs both classification and regression to produce prediction model which is in the form of ensemble of weak prediction model mainly containing decision tree. After running Gradient Boosting Regression, output is -0.8 MSE value, 0.06 STD value and it took 11000ms to run. Extra Tree Regression combines predicted output from multiple decision trees. After running Extra Tree Regression, output is -1.1 MSE value, 0.01 STD value and it took 46000ms to run.

Table 5.3. *Predictivity of various algorithms*

Algorithm	**MSE**	**STD**	**Running Time**
Lasso	-1.2	0.08	2 ms
EN	-1.2	0.08	1 ms

CART	-1.1	0.09	2 ms
KNN	-0.8	0.06	3000ms
SVR	-0.8	0.06	15000ms
RFR	-0.8	0.08	33000ms
GBR	-0.8	0.06	11000ms
ETR	-1.1	0.01	46000ms

After comparing all the algorithms on the basis of MSE, STD and running time we can conclude that Support Vector Machine (SVM) or SVR and RFR as the most promising models.

SVR is a powerful method for building a classifier. It aims to create a decision boundary between two classes that enables the prediction of labels from one or more feature vectors. This decision boundary, known as the hyper plane, is orientated in such a way that it is as far as possible from the closest data points from each of the classes. These closest points are called support vectors.

Similar compounds possess comparable properties. The correspondence of compound is computed for the interaction profiles. The functioning of models varies on the direction of similarity calculation: compound-based similarity. Compound-based models demonstrate enhanced predictive power. Even so, the similarity calculation is important for SVM algorithm. Basically, due to increase in information capacity of similarity matrix, there is increase in the number of interaction profiles N, the predictive power of the model also increases, but at the same time, space and time complexity is $O(N^2)$. It rates the applicability of similarity-based

Collaborative Filtering methods limited for large data sets. The considered data set required at least 1700 GB RAM and 2 h or 2500 GB and 6 h for an *msd (Most Significant Digit)* or cosine calculation of 250K compounds, respectively.

Co-clustering and matrix factorization-based methods are based upon profile similarity; accordingly, the calculation does not need large memory spaces. Interaction matrix rows and columns are grouped for comparing of profiles and to complete the missing values. The best co-clustering model shows a cross-validation median ROC AUC of 0.81, which is compound (0.86) based SVM and RFR. Compound-based similarity methods perform better than co-clustering based methods in cross-validation (median AUCs of 0.83 and 0.86).

Cancer genomic data are high-dimensional, heterogeneous and noisy. The application of SVR learning in cancer genomics is popular and successful. The appeal of SVR approach is due in part to the power of the SVM algorithm, and in part to the flexibility of the kernel approach to representing data. SVR can be robust, even when the training sample has some bias.

Although SVR with non-linear kernels are extremely powerful classifiers, they do have some downsides as following: 1). Finding the best model requires testing of various combinations of kernels and model parameters; 2). It can be slow to train, particularly if the input dataset has a large number of features or examples; 3). Their inner workings can be difficult to understand because the underlying models are based on complex mathematical systems and the results are difficult to interpret. The success or failure of machine learning approaches on a given problem may vary strongly with the expertise of the user. Of special concern with supervised applications is that all steps involved in the classifier design (selection of input variables, model training, *etc*.) should be cross-validated to obtain an unbiased estimate for classifier accuracy. For instance,

selecting the features using all available data and subsequently cross-validating the classifier training will produce an optimistically biased error estimate.

We carried out 10-fold cross-validation and by grid search enhanced the hyper parameters of the models. y-scrambling test was performed to prove that lack of data set impact imbalances the prediction results, this test is performed on the 10 best models for each scenario in both external validation and cross-validation. Models' quality gets decreased by using y-randomization hence giving compelling proof of importance of prediction model.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

This chapter provides a summary of results obtained and focuses more on the importance of work. It helps to know the findings of the study and provide the conclusion of the study. Prostate cancer is the second leading cancer. This cancer has become a burden on today's society.

Finding drugs for prostate cancer is a complex and time-consuming task that takes years to find, right from the initial stage to the final stage of marketing. In this process, testing is done in which potential inhibitors of the particular target are tested for their bio activity. Millions of drug molecules are needed at the initial stage for the discovery of drug and design. Hence, using computational methods for the early prediction of a drug's biological activity saves time and money.

Considering the present scenario of prostate cancer cases and death rates and also the complexity that is involved in designing the drug, an attempt is made to develop a prediction system to predict the drug's biological activities by using regression models.Experiments performed shows that recommended system algorithm based on collaborative filtering of activity data has achieved a great result in predicting activity class. It is shown that collaborative filtering is having high prediction capability in cross-validation

The collaborative algorithm is preferred because of its capability in which feature information for compounds can be used. The main difficulty in this approach is the requirement of generating

and processing information of additional features, which is a challenging task in the case of prostate cancer and thus require a lot of computational resources. Using IMC, we differentiate data into a new set that contains more uniformly distributed values. By developing correct features can help in solving the problem, but this can be tricky.In collaborative filtering, various regression techniques have been used from which best results was given by Support Vector Machine (SVM) and Random Forest Regressor (RFR). Thus, two prediction models are available Support Vector Machine and Random Forest Regressor from which any method can be used by the user according to his choice. The cut-off points for defining potent compounds arepIC50>=6. Micro molar is the measuring unit for pIC50. If pIC50 is having a higher value then it shows greater potency.

This research shows that computational tools are a more effective alternative in drug screening so that lead compounds can be identified. It is believed that this prediction tool can help scientific society with drug discovery. This research will encourage bio-informatics scientists to develop free software's and servers to facilitate drug prediction system. The work done in this research can help to prioritize active molecules which could help in prostate cancer drug discovery.

6.2 Future Work

1. Bio active agents those are having therapeutic potential for prostate cancer can be identified using developed prediction system.
2. Further user interface can be added so that adding further details regarding compounds properties and functions in future.

3. Other targets of prostate cancer like vitamin D3 receptor, androgen and estragon receptor can be considered for research.
4. Bio active chosen through prediction tool can be used in in-vitro settings.
5. The best method discovered is SVD through which we have found 36 drugs which can be used to fight with prostate cancer but in future we will found 1 or 2 drugs which can fight with prostate cancer.
6. Regression is a supervised learning method while Matrix factorization is an unsupervised method. In future, comparison of these two methods can be done to find the better out of the two.

References

[1] Caruana, R. (1997). Multitask Learning,DOI: 10.1023/A:1007379606734

[2] Lipinski, C. F., Maltarollo, V. G., Oliveira, P. R., da Silva, A. B., & Honorio, K. M. (2019). Advances and perspectives in applying deep learning for drug design and discovery. *Frontiers in Robotics and AI*, *6*, 108.

[3] Norinder, U., & Svensson, F. (2019). Multitask Modeling with Confidence Using Matrix Factorization and Conformal Prediction. *Journal of chemical information and modeling*, *59*(4), 1598-1604.

[4] Zubatyuk, R., Smith, J. S., Leszczynski, J., & Isayev, O. (2019). Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science advances, 5(8),* eaav6490.

[5] Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., & Pande, V. (2017). Is multitask deep learning practical for pharmacy. *Journal of chemical information and modeling, 57(8),* 2068-2076.

[6] Sosnin, S., Vashurina, M., Withnall, M., Karpov, P., Fedorov, M., & Tetko, I. V. (2019). A survey of multi□task learning methods in chemoinformatics. *Molecular informatics, 38(4),* 1800108.

[7] Sosnin, S., Karlov, D., Tetko, I. V., & Fedorov, M. V. (2018). Comparative study of multitask toxicity modeling on a broad chemical space. *Journal of chemical information and modeling*, *59*(3), 1062-1072.

[8] Van Westen, G. J., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W., & Bender, A. (2011). Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm*, *2*(1), 16-30.

[9] Subramanian, V. (2016). Field-based Proteochemometric Models Derived from 3D Protein Structures: A Novel Approach to Visualize Affinity and Selectivity Features.

[10] Schaduangrat, N., Anuwongcharoen, N., Phanus-umporn, C., Sriwanichpoom, N., Wikberg, J. E., & Nantasenamat, C. (2019). Proteochemometric Modeling for Drug Repositioning. In *In Silico drug design* (pp. 281-302). Academic Press.

[11] Alves, V. M., Golbraikh, A., Capuzzi, S. J., Liu, K., Lam, W. I., Korn, D. R., & Tropsha, A. (2018). Multi-descriptor read across (MuDRA): A simple and transparent approach for developing accurate quantitative structure–activity relationship models. *Journal of chemical information and modeling*, *58*(6), 1214-1223.

[12] Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender systems handbook* (pp. 1-34). Springer, Boston, MA.

[13] Marín Velásquez, T. D. (2017). Modelo matemático para la predicción de la viscosidad de crudos pesados muertos producidos en el Estado Monagas, Venezuela. *Enfoque UTE*, *8*(3), 16-27.

[14] Bennett, J., Elkan, C., Liu, B., Smyth, P., & Tikk, D. (2007). Kdd cup and workshop 2007. *ACM SIGKDD Explorations Newsletter*, *9*(2), 51-52.

[15] Amatriain, X., & Basilico, J. (2015). Recommender systems in industry: A netflix case study. In *Recommender systems handbook* (pp. 385-419). Springer, Boston, MA.

[16] Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer*

Applications, *110*(4), 31-36.

[17] Aggarwal, C. C. (2016). *Recommender systems* (Vol. 1). Cham: Springer International Publishing.

[18] Sanghavi, B., Rathod, R., & Mistry, D. (2014). Recommender systems-comparison of content-based filtering and collaborative filtering. *International Journal of Current Engineering and Technology*, *4*(5).

[19] Aggarwal, P., Tomar, V., & Kathuria, A. (2017). Comparing content based and collaborative filtering in recommender systems. *International Journal of New Technology and Research*, *3*(4), 263309.

[20] Ariff, N. M., Bakar, M. A. A., & Rahim, N. F. (2018, October). Comparison between content-based and collaborative filtering recommendation system for movie suggestions. In *AIP Conference Proceedings* (Vol. 2013, No. 1, p. 020057). AIP Publishing LLC.

[21] SuXiao-yuan, E., & TaghiM, K. A. (2009). Survey of Collaborative Filtering Techniques. *Proc. Of Conference on Advances in Artificial In*.

[22] Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291-324). Springer, Berlin, Heidelberg.

[23] Sharma, M., & Mann, S. (2013). A survey of recommender systems: approaches and limitations. *International Journal of Innovations in Engineering and Technology*, *2*(2), 8-14.

[24] Cacheda, F., Carneiro, V., Fernández, D., & Formoso, V. (2011). Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web (TWEB)*, *5*(1), 1-33.

[25] Bokde, D., Girase, S., & Mukhopadhyay, D. (2015). Matrix factorization model in collaborative filtering algorithms: A survey. *Procedia Computer Science*, *49*, 136-146.

[26] Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325-341). Springer, Berlin, Heidelberg.

[27] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Springer, Boston, MA.

[28] Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., & Xiao, W. (2016). Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing*, *173*, 979-987.

[29] Fan, J., Yang, J., & Jiang, Z. (2018). Prediction of Central Nervous System Side Effects Through Drug Permeability to Blood–Brain Barrier and Recommendation Algorithm. *Journal of Computational Biology*, *25*(4), 435-443.

[30] Wang, H., Gu, Q., Wei, J., Cao, Z., & Liu, Q. (2015). Mining drug–disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies. *Clinical Pharmacology & Therapeutics*, *97*(5), 451-454.

[31] Sosnina, E. A., Sosnin, S., Nikitina, A. A., Nazarov, I., Osolodkin, D. I., & Fedorov, M. V. (2020). Recommender systems in antiviral drug discovery. *ACS omega*, *5*(25), 15039-15051.

[32] Yang, J., Li, Z., Fan, X., & Cheng, Y. (2014). Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization. *Journal of chemical information and modeling*, *54*(9), 2562-2569.

[33] Galeano, D., & Paccanaro, A. (2018, July). A recommender system approach for predicting drug side effects. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[34] Yu, H., Mao, K. T., Shi, J. Y., Huang, H., Chen, Z., Dong, K., & Yiu, S. M. (2018). Predicting and understanding comprehensive drug-drug interactions via semi-nonnegative matrix factorization. *BMC systems biology*, *12*(1), 14.

[35] Begg, A. C. (2012, April). Predicting recurrence after radiotherapy in head and neck cancer. In Seminars in radiation oncology (Vol. 22, No. 2, pp. 108-118). WB Saunders.

[36] Weinstein John, N., & Collisson Eric, A. (2013). Mills, Ozenberger Brad A, Ellrott Kyle, Shmulevich Ilya, Sander Chris, Stuart Joshua M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113-1120.

[37] Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., & Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. Cell, 166(3), 740-754.

[38] Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, *8*(4), e61318.

[39] Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., & Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, *32*(12), 1202-1212.

[40] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30-37.

[41] Huang, L., Li, F., Sheng, J., Xia, X., Ma, J., Zhan, M., & Wong, S. T. (2014). DrugComboRanker: drug combination discovery based on target network analysis. *Bioinformatics*, *30*(12), i228-i236.

[42] Ammad-Ud-Din, M., Khan, S. A., Malani, D., Murumägi, A., Kallioniemi, O., Aittokallio, T., & Kaski, S. (2016). Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*, *32*(17), i455-i463.

[43] Chauhan, J. S., Dhanda, S. K., Singla, D., Agarwal, S. M., Raghava, G. P., & Open Source Drug Discovery Consortium. (2014). QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. *PloS one*, *9*(7), e101079.

[44] Jamal, S., Periwal, V., & Scaria, V. (2013). Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. *BMC bioinformatics*, *14*(1), 1-8.

[45] Singh, N., Shah, P., Dwivedi, H., Mishra, S., Tripathi, R., Sahasrabuddhe, A. A., & Siddiqi, M. I. (2016). Integrated machine learning, molecular docking and 3D-QSAR based approach for identification of potential inhibitors of trypanosomal N-myristoyltransferase. *Molecular biosystems*, *12*(12), 3711-3723.

[46] Hansch, C., Kurup, A., Garg, R., & Gao, H. (2001). Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chemical Reviews*, *101*(3), 619-672.

[47] Periwal, V., Rajappan, J. K., Jaleel, A. U., & Scaria, V. (2011). Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. *BMC research notes*, *4*(1), 1-10.

[48] Varnek, A., & Baskin, I. I. (2011). Chemoinformatics as a theoretical chemistry discipline. *Molecular Informatics*, *30*(1), 20-32.

[49] Gillet, V. J. (2019). Applications of Chemoinformatics in Drug Discovery. *Biomolecular and Bioanalytical Techniques: Theory, Methodology and Applications*, 17-36.

[50] Ali, N., Cobb, S. L., & Mowbray, C. (2020). Introduction to the themed collection on 'Neglected tropical diseases'. *RSC Medicinal Chemistry*, *11*(10), 1098-1099.

[51] García-Domenech, R., Gálvez, J., de Julián-Ortiz, J. V., & Pogliani, L. (2008). Some new trends in chemical graph theory. *Chemical Reviews*, *108*(3), 1127-1169.

[52] Hansch, C., & Fujita, T. (1964). p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616-1626.