Project Report

DATA MINING AMES SALMONELLA TEST RESULTS FOR CARCINOGENICITY AND MUTAGENICITY

TEAM MEMBERS:
Sucharu Gupta
Nagadhatri Chennavajula
Andrew Wong
Sprush Ujjwal

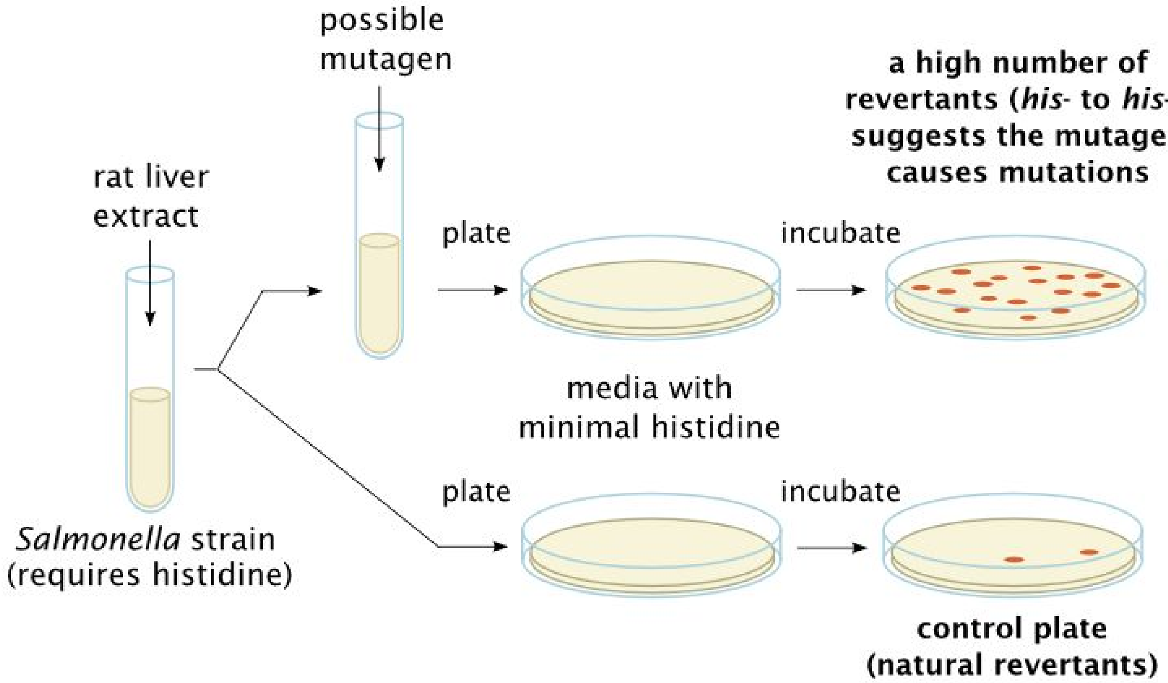
Abstract

Determining the toxicity of chemicals is necessary to identify their harmful effects on humans, animals, plants, or the environment. It is also one of the main steps in drug design. Animal models have been used for a long time for toxicity testing. However, in vivo animal tests are constrained by time, ethical considerations, and financial burden. Therefore, computational methods for estimating the toxicity of chemicals are considered useful. In silico toxicology is one type of toxicity assessment that uses computational methods to analyze, simulate, visualize, or predict the toxicity of chemicals. This report discusses the use of data mining techniques to study estimators of toxic potency in toxicological databases. Data Mining was used to uncover the patterns and present the knowledge discovery in the databases (KDD).

INTRODUCTION

In our everyday life we have to deal with an ever increasing number of new and different chemical compounds, such as food colourings and preservatives, drugs, dyes for clothes and ordinary objects, pesticides and many others: at present the number of registered chemicals is estimated at 28 million. It is well recognised that uncontrolled proliferation of new chemicals may pose risks to the environment and people; hence, their potential toxicity has to be considered. Biologically active chemicals interact with biomolecules, triggering specific mechanisms, such as the activation of an enzyme cascade or the opening of an ion channel, which lead to a biological response. These mechanisms, determined by the chemical properties, are unfortunately largely unknown; thus, toxicity tests are needed. Mutagenic toxicity, also called mutagenicity, can be assessed by various test systems. It is a property of high public concern because it has a close relationship with carcinogenicity and, in the case of germ cell mutations, with reproductive toxicity. For assessing the potential of a chemical to be toxic, a significant breakthrough was the creation of cheap and short-term alternatives to the rodent bioassay, the main tool of the research on chemical carcinogens. With this intent, Bruce Ames created a series of genetically engineered Salmonella Typhimurium bacterial strains, each strain being sensitive to a specific class of chemical carcinogens. As discussed in other papers, the estimated inter-laboratory reproducibility of this in vitro test is about 85%. This observation will be taken into account in the conclusive discussion. Alongside classical experiments for assessing toxicity, the use of computational tools is gaining more and more interest in the scientific community and in the industrial world as accompaniment to or replacement of existing techniques.

Whereas animal tests are very expensive and time consuming, high throughput computational approaches, otherwise known as in silico models, are broadening the horizons of experimental sciences: with increasing sophistication of such models, we are increasingly moving from experiments to simulations. A critical quantitative component of such studies is statistical characterization of the dose-response to the agent, and whether this response is indicative of a significant toxic effect in the system under study. Summary measures describing the potency of the agents toxicological activity are of interest, and the development of such measures is an area of active statistical research. In recent years, rapid technological and scientific developments have enhanced our ability to generate large amounts of toxicological data on a variety of endpoints, even when a laboratory's resources are limited. As a result, a wealth of archived toxicity database has emerged; Databases with such huge information represent excellent sources for studying the statistical characteristics of newly-developed (and some older) estimators of toxic potency. In effect, this represents a form of data mining, where information buried in large collections of data is quarried in a systematic fashion to uncover data characteristics. Data mining is often used in identifying/manipulating quantitative structure-activity relationships (QSARs) of hazardous chemicals, and there are strong overlaps between the use of data mining for potency estimation and that for QSAR identification. One can also use toxicological databases to study quantitative activity-activity relationships (QAARs), where, say, the mutagenic activity of a set of hazardous chemicals is compared for predictive purposes against their carcinogenic activity. Data mining is also growing in allied applications such as pharmacology and drug discovery, ecotoxicology, and in the burgeoning studies of gene expression, microarrays and toxicogenomics.



DESCRIPTION:

Determining the toxicity of chemicals based on Dose , count and other affecting factor. This is necessary to identify their harmful effects on humans, animals, plants, or the environment.

Project Purpose:

The purpose of this project is to give the users with a forum where they can check the toxicity of the chemicals which would help them to identify the factors they need to take care before distributing the chemical to the user base. This would help to prevent possibility restricting cancer mutagen chemical to the environment.

REQUIREMENTS:

Functional Requirements

Primary:

1. Provide a table to user to see from the test data if there are chemicals with matching inputs.
2. Users can check on the blog, toxicity of the chemical based on the given inputs.
3. Output should be mutagen or non-mutagen

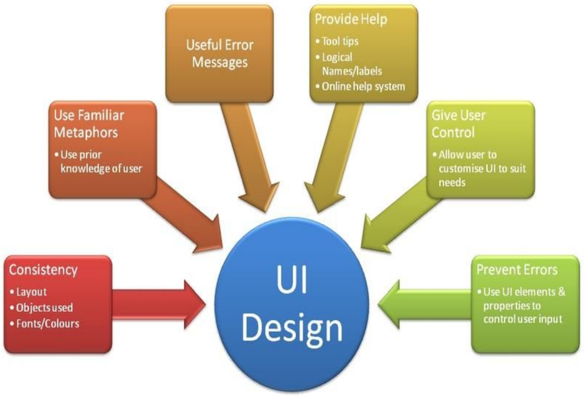
Secondary:

1. User can compare the outputs from different best ML algorithms.

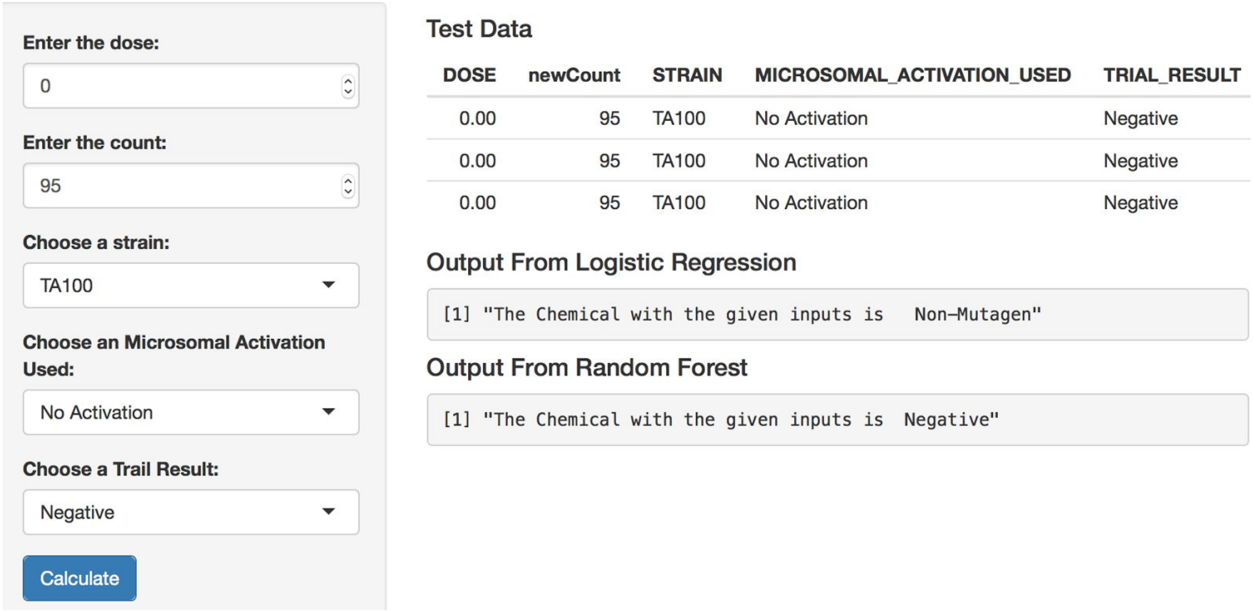
Technical Requirements:

1. Statistical Analysis and reporting
2. If possible system should be hosted on open source, cloud platform to allow users to access the application.

UI DESIGN PRINCIPLES:



Storyboard:



Enter the dose:

0

Enter the count:

95

Choose a strain:

TA100

Choose an Microsomal Activation Used:

No Activation

Choose a Trail Result:

Negative

Test Data

DOSE	newCount	STRAIN	MICROSOMAL_ACTIVATION_USED	TRIAL_RESULT
0.00	95	TA100	No Activation	Negative
0.00	95	TA100	No Activation	Negative
0.00	95	TA100	No Activation	Negative

Output From Logistic Regression

```
[1] "The Chemical with the given inputs is   Non-Mutagen"
```

Output From Random Forest

```
[1] "The Chemical with the given inputs is  Negative"
```

High Level Architecture Design

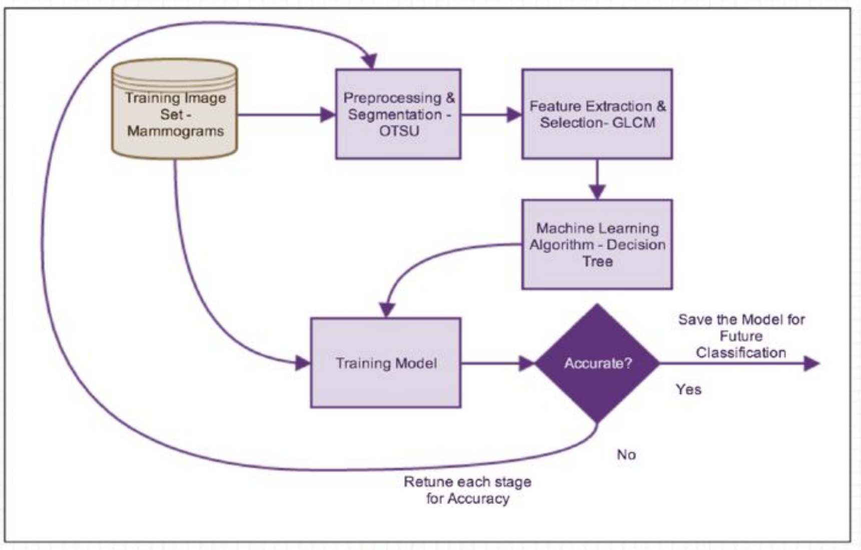

Training Image Set - Mammograms
Preprocessing & Segmentation - OTSU
Feature Extraction & Selection- GLCM
Machine Learning Algorithm - Decision Tree
Training Model
Accurate?
Save the Model for Future Classification
Yes
No
Retune each stage for Accuracy

Figure 2: Training Model Creation

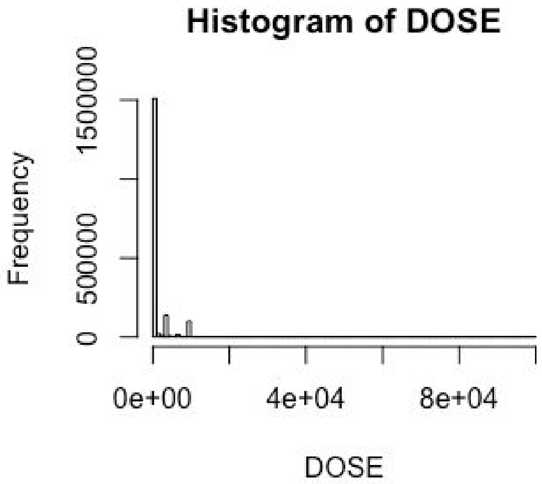
Datasets & Data patterns

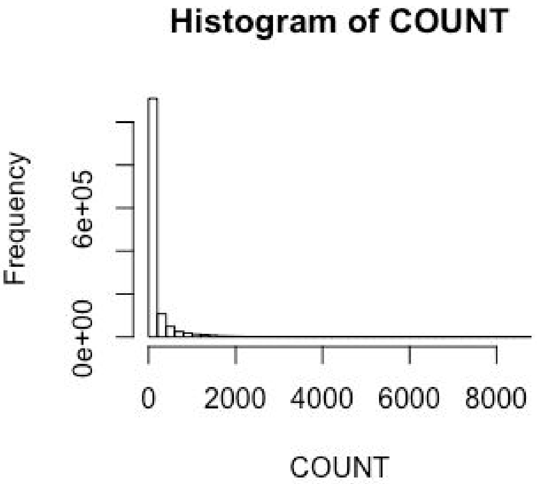
GENETOX_Bacterial mutagenicity_NTP data is in chemical-in-rows format, comma-separated values.
GENETOX_Bacterial mutagenicity_NTP.txt is 640.9MB .
final_project_data folder in Project directory has 3 files:

- GENETOX_Bacterial mutagenicity_NTP.txt
- gene.train.csv (training data, 1.5MB)
- Gene.valid.csv (validation data, 354 KB)
- Test.data.csv (test data, 459 KB)

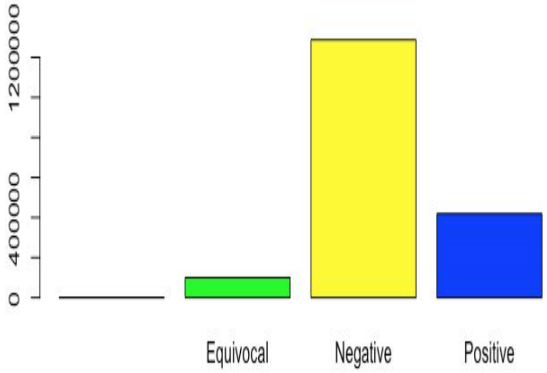
Data patterns

From the dataset the data is not normally distributed. When we look at the plots for Dose and Count which are our main predictors for the analysis , data looks more gamma distributed.

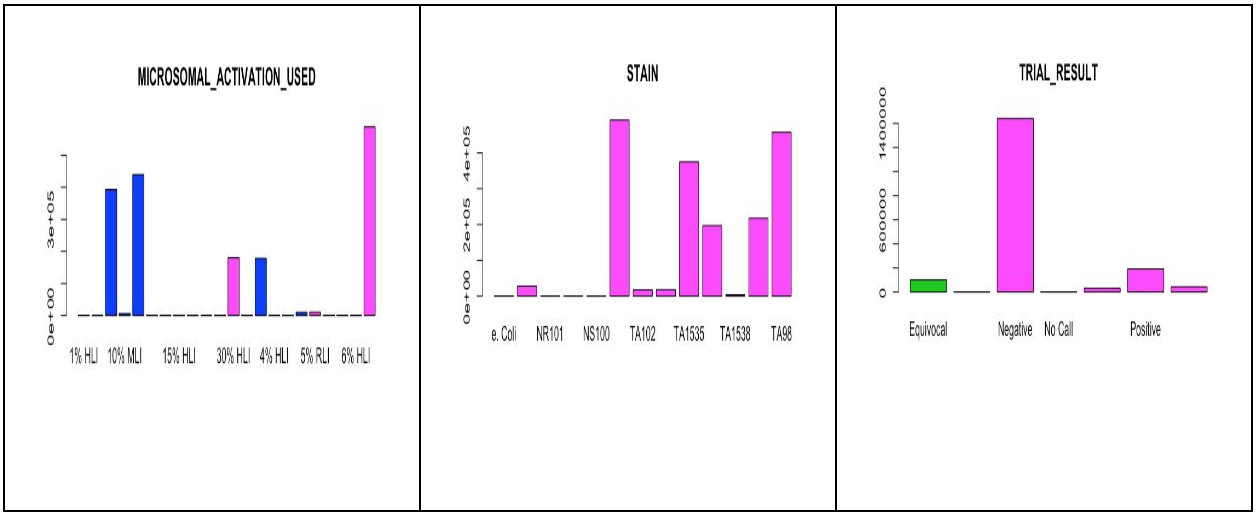




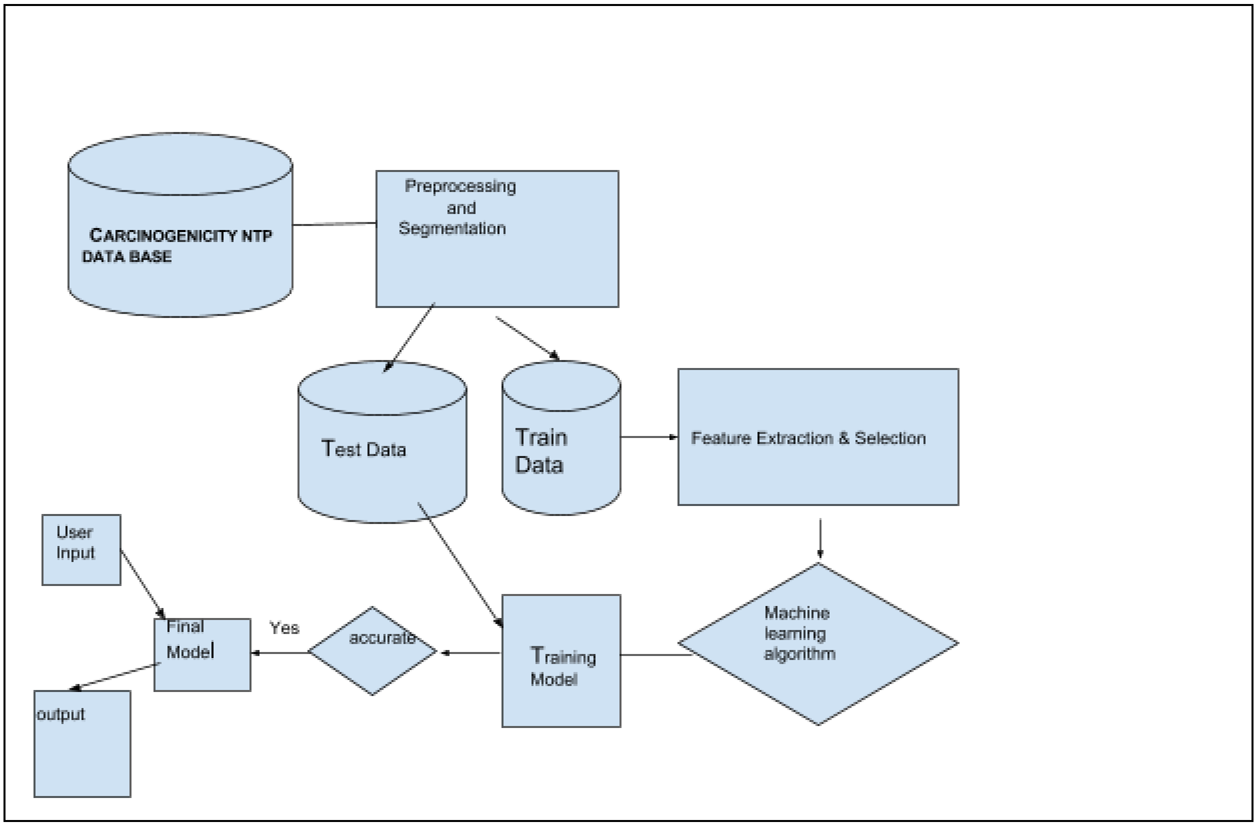
Plot of the class (STUDY_CONCLUSION)



Plot of MICROSOMAL ACTIVATION USE

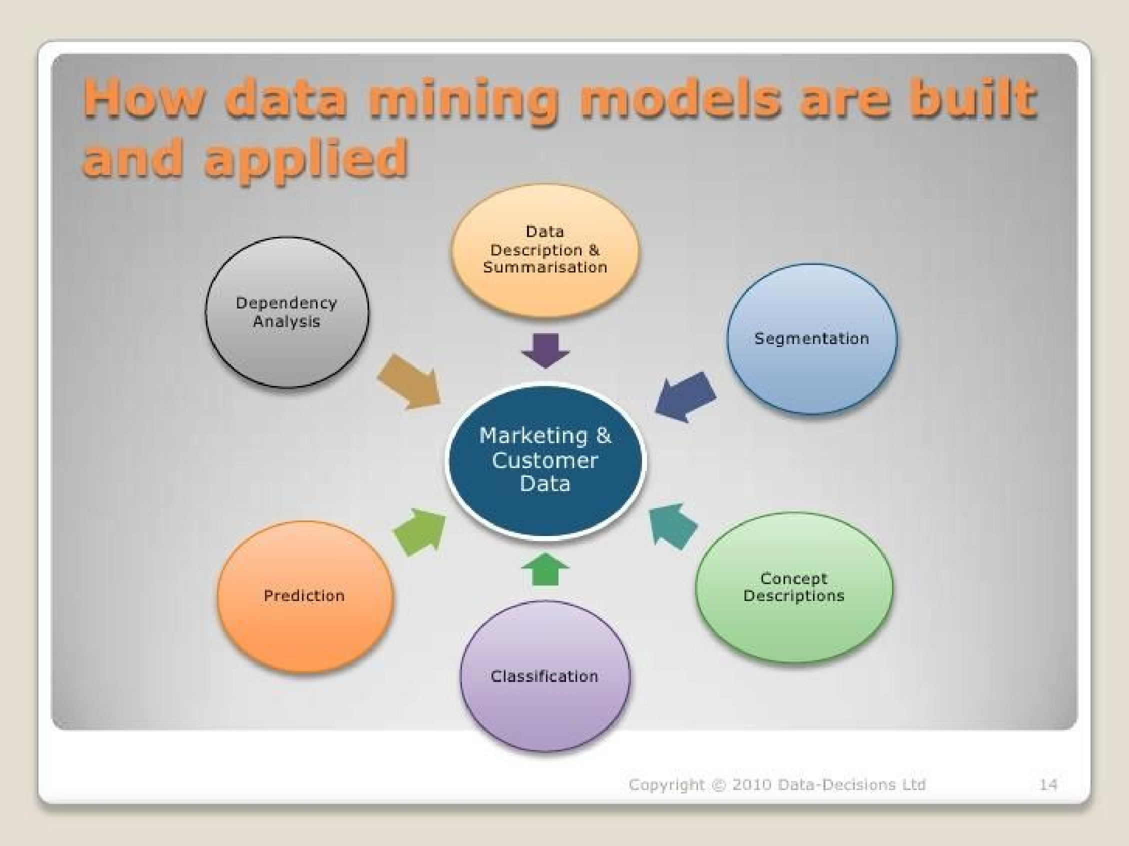


Data Flow Diagram & Architecture


CARCINOGENICITY NTP DATA BASE
Preprocessing and Segmentation
Test Data
Train Data
Feature Extraction & Selection
User Input
Machine learning algorithm
Final Model
Yes
accurate
Training Model
output

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Here we took NTP dataset and did cleaning and preprocessing.We divided the dataset into training,test and Validation dataset.Further extraction of features and selecting them for running various algorithms to see which one will work appropriately to this dataset. We tried all our models on training dataset, finalised our model and then we run our model on test dataset to see whether it's predicting correctly and run on validation dataset to see whether our model will work for future data or not?

DATA MINING PRINCIPLES



How data mining models are built and applied

Data Description & Summarisation

Dependency Analysis

Segmentation

Marketing & Customer Data

Prediction

Concept Descriptions

Classification

Copyright © 2010 Data-Decisions Ltd

14

Use the following in the Rstudio :-

- Logistic Regression
- Multinomial Regression
- Random Forest
- SVM

Logistic Regression:

This is a regression model where the dependent variable (DV) is categorical that is, it can take only two values, like whether the genes are mutagen or not. These are represented by “1” or “0”.Then regression analysis can be done by this. In our project we need to decide if the chemical is mutagenic or not.

Multinomial Regression

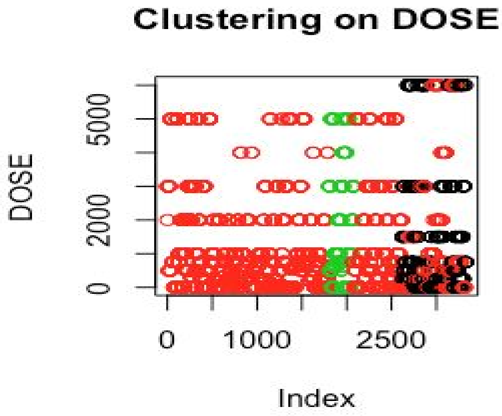
We have used Random Forest Algorithm to to identify whether chemical is mutagen or not

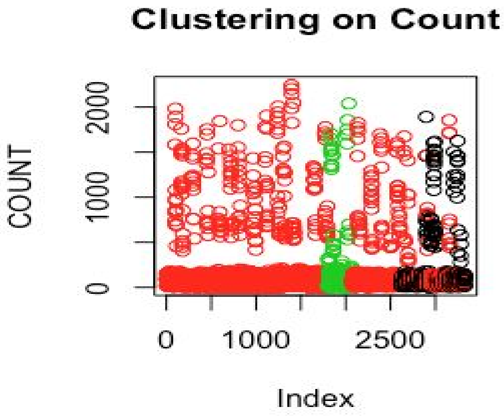
Random Forest

We have used Random Forest Algorithm to to identify whether chemical is mutagen or not.

Support Vector Machine

We have used support Vector Machine algorithm to identify whether chemical is mutagen or not. However, model selection focuses on finding one best classifier while genetic algorithms are based on the idea of recombining and mutating a large number of good candidate classifiers to realize further improvements. It is shown that the empirical estimator is the superior fitness criterion in this sense, leading to a greater number of promising models on average. From the Plots we see that we can perform SVM’s





KDD Principles

Data mining is also sometimes called Knowledge Discovery in Databases (KDD), because the purpose of KDD is to extract of interesting information from data in large database, and analyze them to find patterns for big data through techniques in statistics, artificial intelligence and database management. Knowledge Discovery Process is an iterator process with the aim of finding and interpreting patterns from data that leads to the following steps, and they are Data Cleansing, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern evaluation, Knowledge Presentation.

- **Development of Understanding**
    - To choose the domain with pre-existing knowledge in it and to create the goals of the project to end-user.
    - We choose the domain of Ames test which is a popular method that utilizes bacteria on a trial whether a given chemical that can cause mutations in the DNA of the organism. The goal of the project to uncover the patterns and to build an estimator of toxic potency from toxicological databases.
- **Target Data Set Creation**
    - To select a data set or subset of it.
    - We found the data source at National Institutes of Environmental Health Science, and they provide the dataset about Chemical Effects in Biological Systems.
- **Data Cleansing**
    - To remove noise and inconsistent data.
    - We have done data cleansing due to missing values to the selected dataset.
    - We have merged all the “Weakly Positive” study conclusions to “Equivocal” as it’s not well enough to conclude as purely Negative or Positive.
    - As there are missing values for COUNT variable we replaced the missing values using the mean of count.
    - There are 74 entries with no COUNT, COUNT_SEM, COUNT_MEAN, not sure if we can remove these rows count is not there , for now removed these rows assuming they provide no information without COUNT, this would need further investigation.
    - Removed the rows with TREATMENT_GROUP_TYPE is null.
    - Removed the rows whose TRIAL RESULTS are *Not a Valid Tes*t, *Failed Experiment and No call*

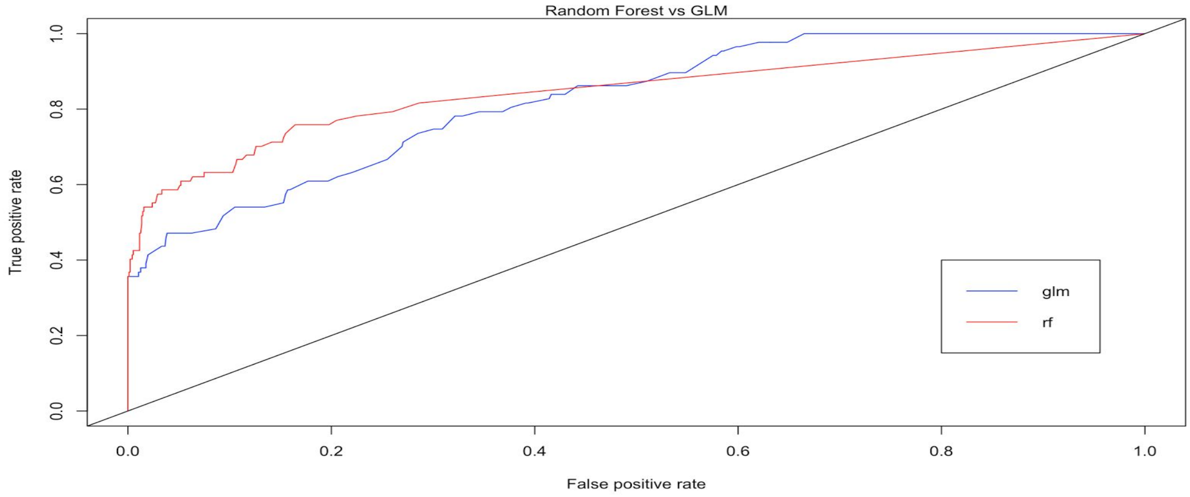
- **Data Integration**
    - To combine data where comes from multiple sources.
    - We only use the single data source, no need to merge or combine data.

- **Data Selection**
    - To retrieve relevant data for analysis from database. Data selection was made based on the p-values.
- **Data Transformation**
    - Transformed the DOSE and COUNT as 1/log(DUI) (DUI - Dose per unit count) to see if it's a prominent predictor in the analysis.
- **Data Mining**
    - To extract data patterns through intelligent statistical and machine learning approaches.

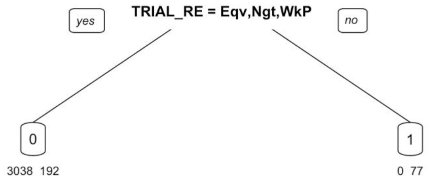
As the data is very complex, we designed to perform our analysis in two stages. We divided the data as "mutagen" and "non-mutagen" based on TRIAL RESULT. Data values which are either negative or equivocal are characterized as "Negative" and rest as "Positive". These results are used in the second level of model abstraction resulting in more precise estimates for the three levels of STUDY CONCLUSION (Positive, Negative, Equivocal) .

Step1 :

In the first step of analysis we performed logistic regression and random forest algorithm. Random forest was seemed to winner from the residual, AIC, BIC values and ROC Curve.



Random Forest Tree

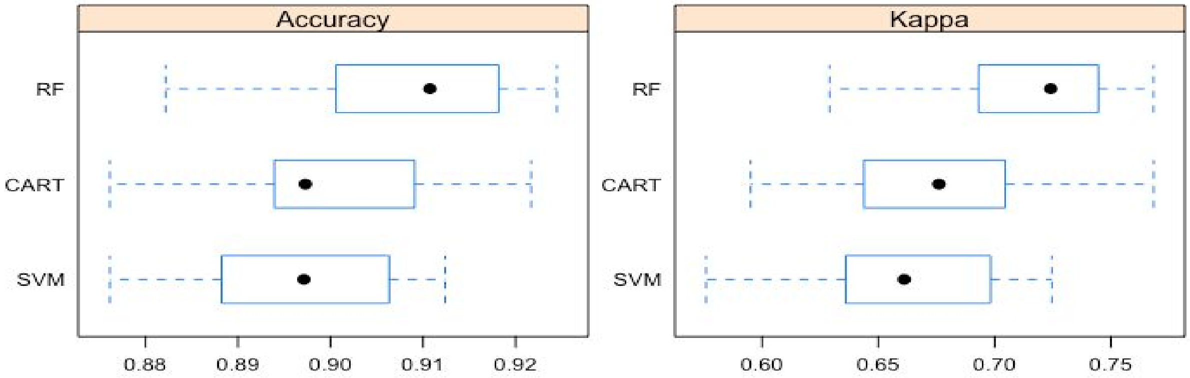


Step2:

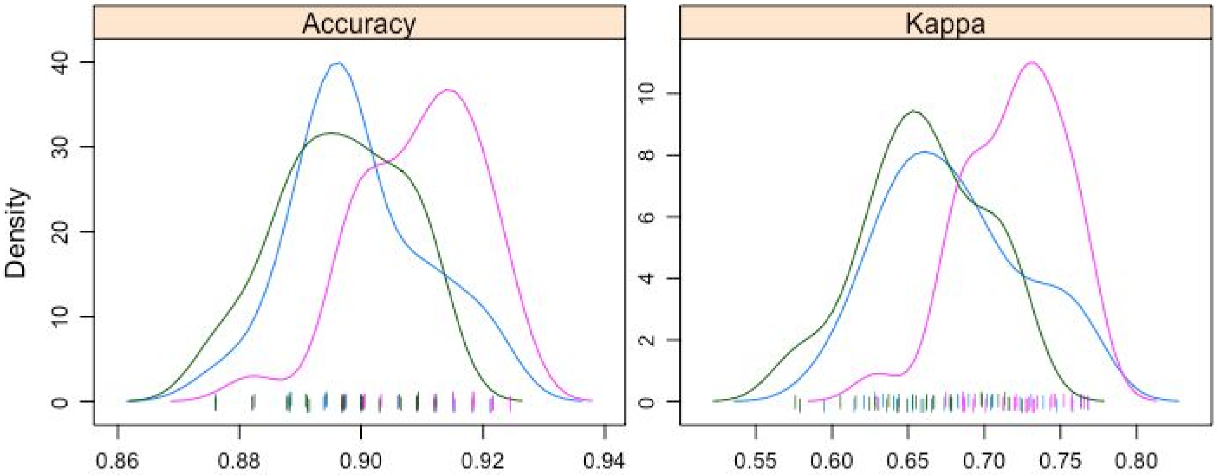
From the output that we got from the Step 1 we performed second level of modelling to get more precise estimates for the final output. We used CART, Support vectors and Random Forest algorithms. Based on the box plots, confusion matrix, dot plots etc, Random Forest showed prominent results.

Comparing the models (CART, SVM, RF)

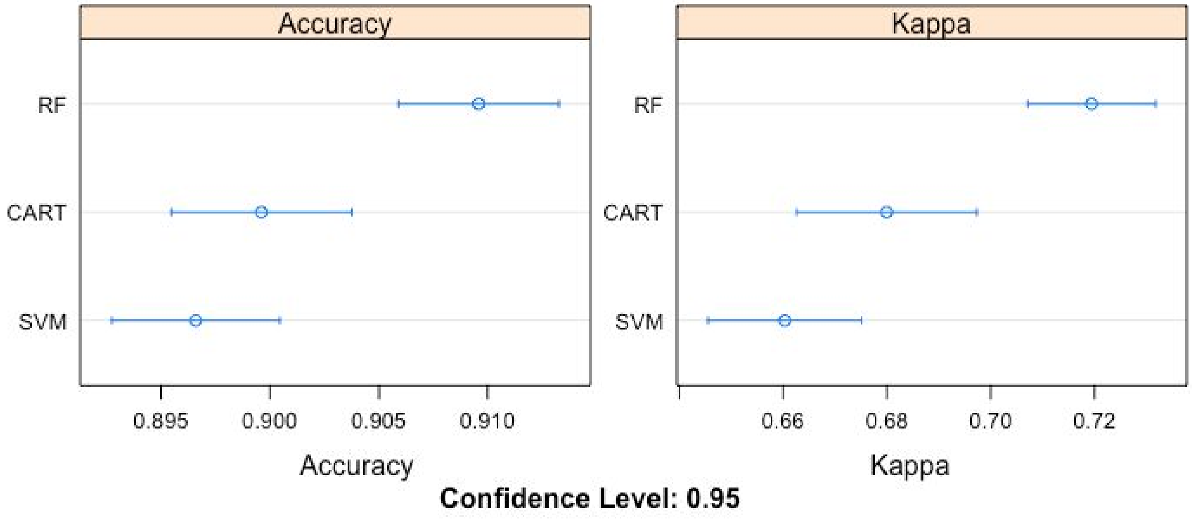
BOX PLOTS

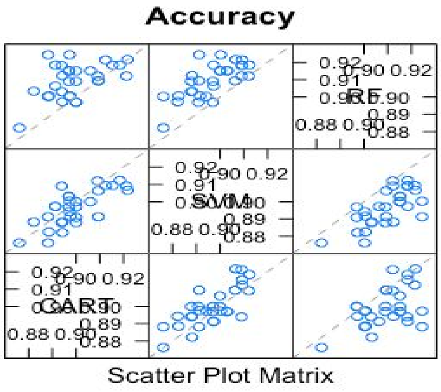


Density Plots

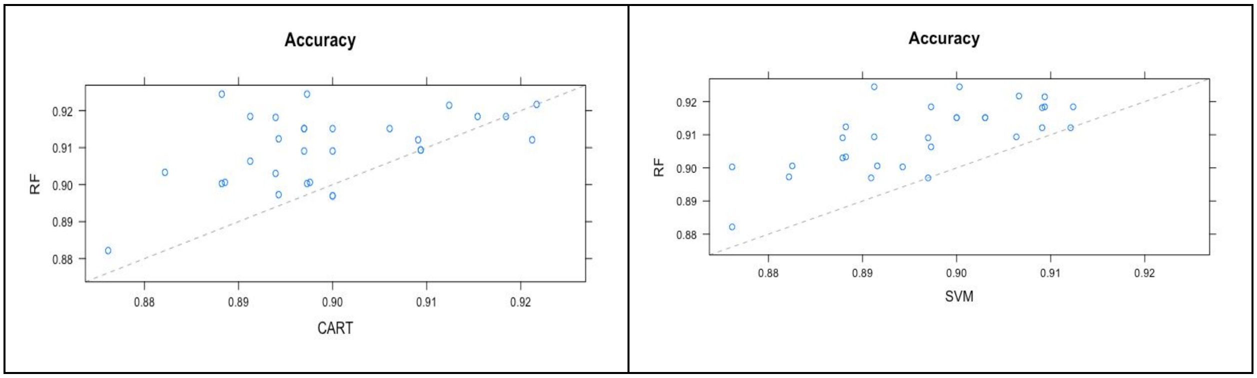


Dot plots of accuracy

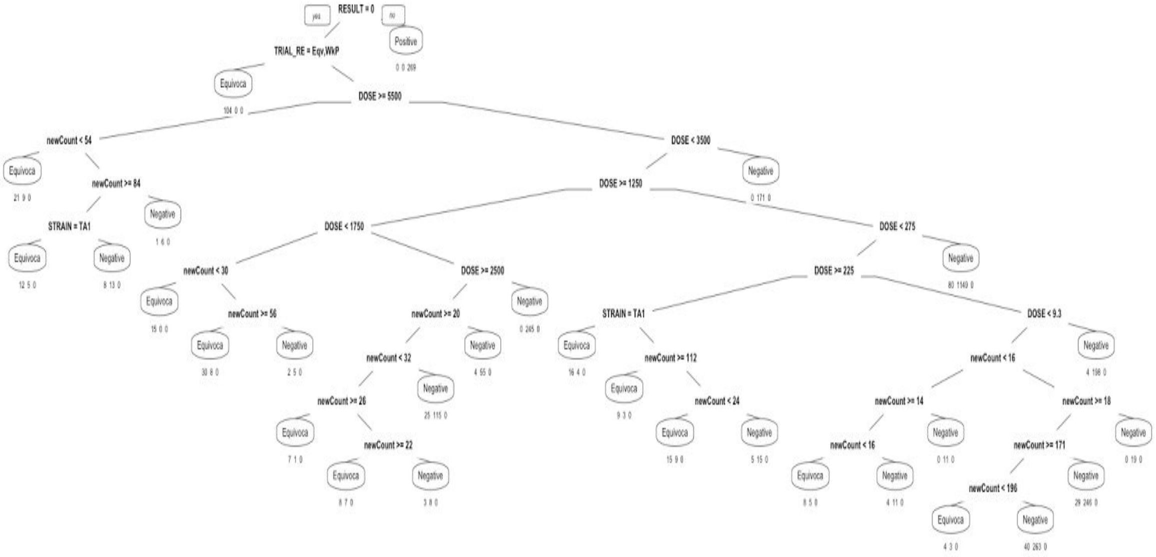




Summary of differences in the models



Random Forest Tree

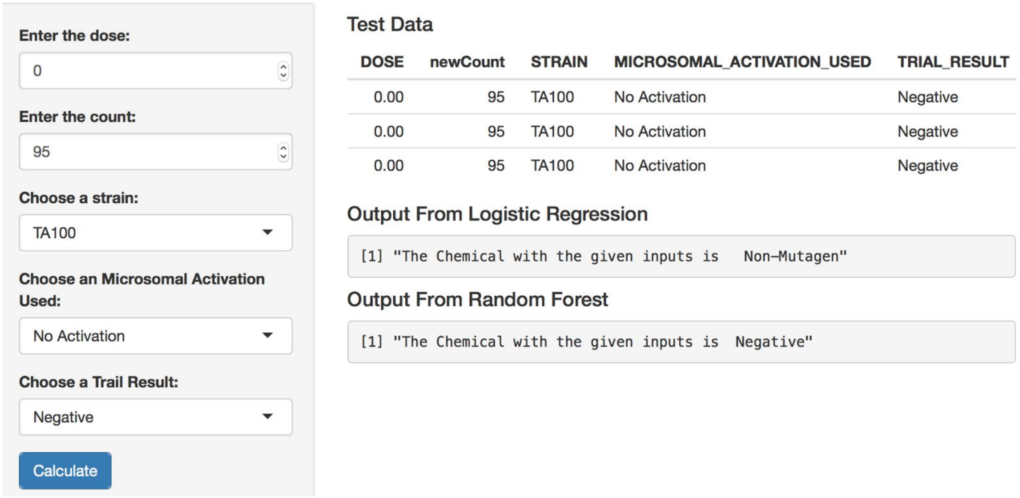


- **Pattern Evaluation**
  - To identify interesting patterns inside scope representing knowledge discovery.
- **Knowledge Presentation**
  - To visualize and present knowledge of what knowledge has been mined.

Data Tools

- R
- R Studio
- Shinny
- AWS

Client Side Design:



Enter the dose: 0

Enter the count: 95

Choose a strain: TA100

Choose an Microsomal Activation Used: No Activation

Choose a Trail Result: Negative

Test Data

DOSE	newCount	STRAIN	MICROSOMAL_ACTIVATION_USED	TRIAL_RESULT
0.00	95	TA100	No Activation	Negative
0.00	95	TA100	No Activation	Negative
0.00	95	TA100	No Activation	Negative

Output From Logistic Regression

```
[1] "The Chemical with the given inputs is   Non-Mutagen"
```

Output From Random Forest

```
[1] "The Chemical with the given inputs is  Negative"
```

Testing:

Testing of the model is done using 10 fold cross validation repeated 3 times.

As the problem is classification problems, we used stratified k-fold cross-validation, in which the folds are selected so that each fold contains roughly the same proportions of class labels.In repeated cross-validation, the cross-validation procedure is repeated n times, yielding n random partitions of the original sample. The n results are again averaged (or otherwise combined) to produce a single estimation.

Design Patterns Used

Lambda architecture

Except if you go for online learning (update the model after every example), need to retrain your model regularly and cross-validate its performance. On the other hand we need our model to be able to predict over real-time or simply incoming data. That means we need two layers. One where the data goes for predictions, one where it is historicized and used for regular trainings. There will be a lot of code in common, such as your processes for feature creation, data cleaning, configuration for the machine learning libraries and so on.The "fake" disclaimer is because we don't have a serving layer .

Project Team Contributions

Sucharu and Dhatri : Performed the data analysis and modelling.

Sprush and Andrew : UI design and Performed testing on the models