CS F407 – Artificial Intelligence

PROJECT – Different Disease Prediction Systems

Group – 13

Yash Pandey, 2021A7PS0661P

Ohiduz Zaman, 2021A7PS2005P

GITHUB LINK FOR CODE: yashpandey474/AI-Disease-Prediction-System: AI-Disease Prediction System (github.com)

Overview

This report focuses on the model of predictive diagnosis involving data munging, feature selection, model training and accuracy calculation. We try to introduce and validate methods that improve certain aspects of this process.

We start with the overall problems identified with the predictive diagnosis models and possible areas of improvement identified in the data munging and feature selection process primarily, such as alternate methods of replacing values to avoid bias and a non-constant p-value for feature selection.

The trained model used throughout this paper is linear regression. For data munging, we propose improvements through using the method of k-nearest-neighbors and extensively compare the results with data munging done by replacing with mean.

For the feature selection, we use the simulated annealing algorithm to set up an environment of varying p-value within a given upper bound and lower bound, defined initial temperature and cooling rate. Evaluation of the model has been done by testing on the training set rather than dividing the sets, although results from cross-validation have been provided at numerous places throughout the report. This has been supported by the fact that the 90/10 Train-test method gave stale accuracies for most p-values although they had different attribute sets and considerable differences when predicting on the training set.

For data augmentation, to improve the diversity of the dataset - we have proposed a few data augmentation techniques such as adding noise, scaling etc.

All the results of the report have been presented in the form of graphs, comparing the different accuracies, errors etc. Moreover, at the end of the report a table listing all our defined functions along with their description has been provided to avoid confusion.

Background

Brief details about the concepts involved

With a rapidly growing world population and pandemics becoming a modern day havoc, disease prediction has become utmost important with a lack of medical professionals.

Generally, disease prediction models for accurate predictions have been being built for decades now, with research into the field using various machine learning algorithms. Heart Diseases represent a broad class of cardiovascular diseases becoming rampant in the modern world with the spread of diabetes and smoking [4]. Methods such as KNN & decision trees, genetic algorithms have been used for classifying the severity of the disease as well [4]

Breast cancer on the other hand, has become one of the most deadly cancers with the death risk rising exponentially [5] and an early prediction is crucial in a possible safe recovery. The wisconsin breast cancer dataset obtained from the UCI machine learning repository and the processed cleveland heart disease dataset from the repository are very well known and popular datasets in the field of disease prediction using machine learning, the same datasets have been used for our experiments here.

K Nearest Neighbors:

K Nearest Neighbors (KNN) is a popular machine learning algorithm used for both classification and regression tasks. It is a non-parametric algorithm that relies on instance-based learning, which means that it stores all the training data and makes predictions based on the similarity between the new data point and the existing data points.

Data munging using mean and KNN:

Data munging, also known as data wrangling, is the process of cleaning and transforming raw data into a more useful format. Mean imputation is a common data munging technique used to fill in missing values in a dataset. KNN imputation is another method used to fill in missing values by using the KNN algorithm to find the k most similar data points and use their values to fill in the missing data.

Feature selection using p-value significance limit:

Feature selection is the process of selecting a subset of relevant features from a larger set of features in a dataset. One way to do this is by using the p-value significance limit. In this approach, a statistical test is performed on each feature to determine its significance in relation to the target variable. Features with p-values below a certain significance level are considered relevant and selected for the final model. This approach helps to *reduce the dimensionality of the dataset and improve the performance of the model.*

Data augmentation for disease prediction:

Data augmentation is a technique used to increase the size of a dataset by creating new data points from existing ones. This can be useful in disease prediction, where the dataset may be limited in size. Data augmentation can be done by adding noise, rotating, flipping or changing the color of images or by generating synthetic data using generative adversarial networks (GANs).

Literature Review

This research paper primarily focuses upon problems identifies from the research paper: Application of Machine Learning in Disease Prediction.

The paper consists of applying different machine learning algorithms for prediction of diagnosis on three datasets -

1. Wisconsin Breast Cancer dataset
2. Cleveland Heart Disease dataset
3. Pima Indians Diabetes dataset

The datasets were retrieved from the UCI machine learning repository, however Pima Indians Diabetes dataset is no longer available.

The authors have proposed a general model for the prediction involving the following steps:

1. Data Munging - replacing missing values of attributes if any. This step is achieved by replacing missing values by the mean if continuous and mode if categorical.
2. Feature Selection - eliminating insignificant features to improve model performance. This is implemented using backward selection and elimination of attributes with p-value lower than the significance level of 0.05.
3. Classification Algorithm to create model: Further, a model is trained based upon one of the classification algorithms: Logistic Regression, Decision Trees, Random Forest, Support Vector Machine(SVM) and Adaptive Boosting and tested for accuracy using Train/Test split with fixed test size of 10% and training size of 90% of dataset.

The following comparisons between prediction accuracy of various methods were received on the heart disease dataset & breast cancer dataset:

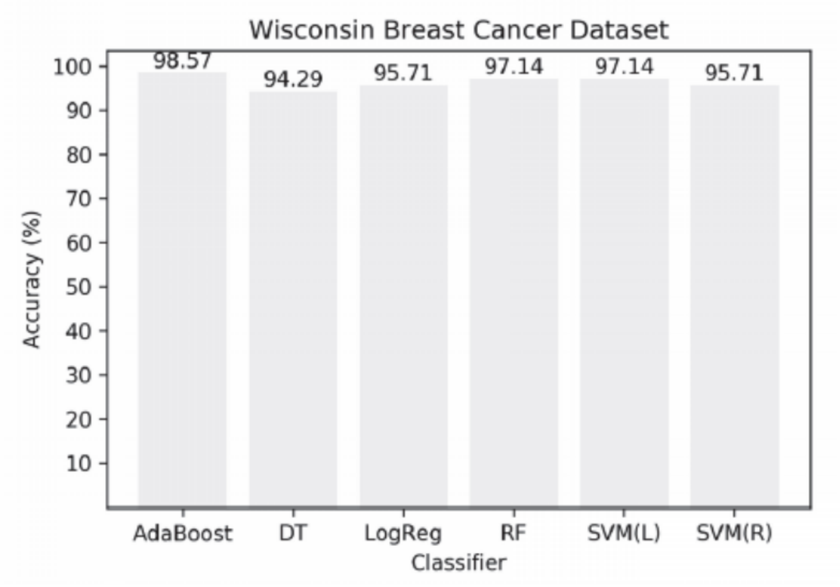


Fig. 2. *Comparison of different algorithm for Breast Cancer Dataset*

This figure illustrates the machine learning methods

used for breast cancer prediction and their accuracies

using the Wisconsin dataset. [TAKEN FROM RESEARCH PAPER]

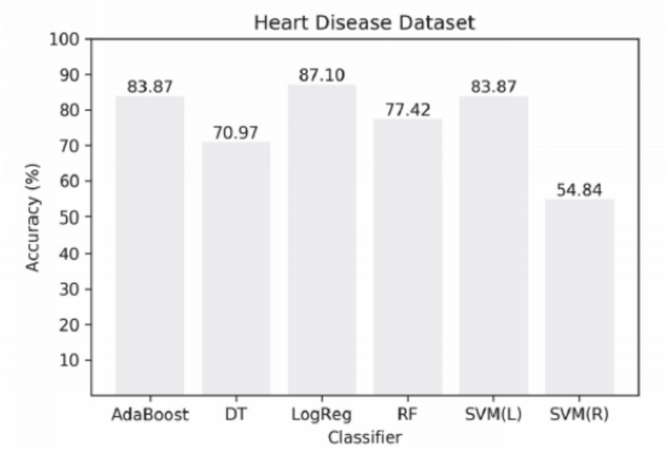


Fig. 4. *Comparison of different algorithm for Heart Disease Dataset*

Figure - Illustrates the different machine learning methods used for prediction of heart disease in the research paper along with their accuracies.[TAKEN FROM RESEARCH PAPER]

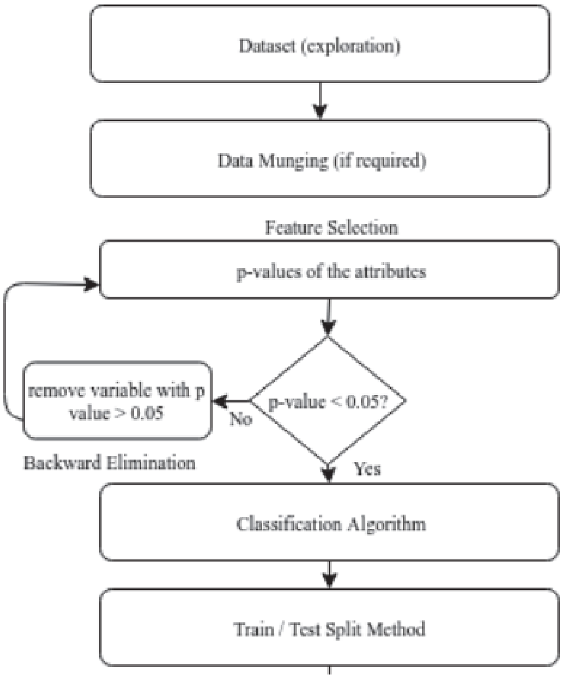

Dataset (exploration)
Data Munging (if required)
Feature Selection
p-values of the attributes
remove variable with p value > 0.05
No
p-value < 0.05?
Yes
Backward Elimination
Classification Algorithm
Train / Test Split Method

Fig. 1. *Proposed Method*

Figure: Proposed model of steps for disease prediction using machine learning algorithms in the research paper.

Overall Problems Identified In the Model:

Data Munging - A major problem in the research papers was to handle the records with missing values for certain attributes. Papers used mean/mode techniques to replace missing values for attributes, replacing missing values with the mean of the whole dataset in case of continuous variables and with mode in case of categorical variables which may lead to a huge loss of information and biased estimates. [Different conditions may consist of a large amount of missing data]

Feature Selection -

A.) Different statistical tests have not been considered in evaluating the significance of attributes in the feature selection process, tests like chi-squared values and f-tests may provide better results.

B.) The feature selection involved in ‘Application of Machine learning’ paper uses a constant p-value to eliminate attributes with p=values above that significant level – 0.05 p. However, this might lead to elimination of significant variables and different p-values may suit different datasets.

C.) Interaction between symptoms not considered in feature selection – significance may increase when features considered in combination for a disease. Example - headache alone may not be as significant as when considered along with high fever for viral fever.

Generalisation - As disease prediction needs to be fair and generalised, since the dataset can suffer from minority classes and less generalizability upon certain groups of people

Past History: Limited use of factors like medical history and lifestyle factors – (EMR used by Chen et. al.) might be important for certain diseases. For example, the heart disease dataset used is ‘processed’ and doesn’t include factors like smoking, etc. which might prove to be important.

with two most optimal k-values to create two separate records.

Assignment Problem Statement & Details

Problem Statement [3-PART]:

1. The data munging process introduces significant bias depending on majority of dataset by replacing with mean.

2. The feature selection process suffers from lack of adaptability by keeping a constant p-value significance limit of 0.05 without variance depending on the disease.

3. The datasets for diagnosis prediction must be general & not bias, with enough data for each class of features.

Details: What is implemented & what results are measured

1. Data Munging

K-NEAREST NEIGHBOURS:

We have experimented with using K-nearest-neighbours process for replacing missing values for the breast cancer dataset and the heart disease dataset with the following steps:

A.) Predict the value of missing attribute for complete records (where value is known) using value of k from 1 to 26 and measure overall sum of squared differences/mean squared error for all complete records between the known and predicted values. [Where value of bare nuclei is known for calculating accuracy] [Squared difference is used rather than difference as negative predicted values may give an incorrect suggestion of accuracy.]

B.) Choose the value of K with least sum of squared differences/least mean squared error for completing values in records with missing values for barenuclei.

C.) Complete values of missing records by finding the average of bareNuclei values of its k-nearest neighbors using the optimal k value

D.) Comparison: - Compute the sum of squared differences and mean squared differences between known and predicted values when replacing bareNuclei values with mean on the complete set of records. Compare this result with those computed for KNN with optimal K.

.2. **Analysis of Linear Regression for predicting diagnosis in breast cancer dataset**: Since the research paper experiments with other machine learning techniques, we have used the method of linear regression to train a model for prediction for all experiments involved here.

NOTE: In the research paper, the training set is 90% of the records and 10% is testing set. However, we have used the training set and testing set as 100% of the record set to provide different results. The results from this method have also been compared with the original method.

A) The following methods of evaluating the model’s accuracy have been used in our project:

I) Testing on training set with Accuracy = total number of correct predictions/(number of records)

II) Testing on testing set using 5 fold cross validation with Accuracy = total number of correct predictions/(number of records) and taking the average over 5 folds.

B) We created two arrays - the first, a two dimensional array with each column for a separate attribute and each row for a separate records, the second, a one-dimensional array with each row for a record’s predicted column value [nums column]

C) Trained the linear regression model on the attribute set resulting from different feature selection processes [Elaborated upon in next section].

D) Predicted the values for either the training set for direct accuracy calculation or using cross-validation - for the testing set into an array

E) Calculate accuracy through the direct method and the cross-validation method [which shuffles the arrays for random division into testing & training set]

Learning Curve for linear regression model for breast cancer dataset:

To extensively evaluate the accuracy of the linear regression model for the heart disease dataset and the breast cancer dataset, learning curves for the model were made by following these steps independently for the two datasets:

I) Create the two arrays of other attribute values and the predicted attribute values for all records and randomly shuffle the values (but aligned so that records don’t become skewed)

II) Create an array of training sizes to be used, ranging from no of attributes + 1 to total no of records - 1.

III) Iterate over the training sizes array and divide the overall record set into a training set of the training size and the remaining records as the testing set.

IV) Tran the linear regression model using the training X & Y arrays and predict values for records in the testing X array.

V) Finally, check the accuracy of the predictions made on the testing Y array. Output the training size and the accuracy, to collect data for plotting the learning curve.

NOTE: Accuracy Calculation - the predicted values from linear regression were classified using thresholds according to the following scheme:

Classification of predicted values

Breast Cancer Dataset: Predicted Value > 3 = 4 [Malignant]
Predicted Value <= 3 = 2 [Benign]

Heart Disease Dataset: Predicted Value = 0 : 0 [Absence]
Predicted Value >= 1: 1,2,3,4 [Presence]

2. Feature Selection & Prediction

For the feature selection process, we have tried using simulated annealing algorithm to shift from a high initial p-value limit to lower limits and calculate accuracy for each to arrive at the most optimal one

A. Forming an array of attribute values for all the records except of the predicted one, which is kept in a separate array. The first array is a two-dimensional array with each row for a different record and each column for a different attribute.

B. Analysing p-values of attributes for setting a range for experiments: We extracted the initial p-values of various attributes in the breast-cancer dataset & heart disease by training a linear regression model on the dataset & then retrieving the p-values.

C. Normal Feature Selection: As the control experiment, keeping a constant p-value significance limit - we eliminated attributes having p-values above the limit - then re-calculated the p-values and reiterated until all the attributes had p-values below the limit - imitating the process used in the research paper. Finally, calculated the accuracy of the linear regression model on training with the arrays containing only these attributes values - with cross-validation & direct accuracy calculation.

D. Feature Selection with Simulated Annealing: Implementing the idea of a variable p-value for different diseases with the following steps:

I) Setting an upper & lower limit on the values to be generated randomly. Initial p-value is the upper limit of range

II) Setting an initial temperature and constant cooling rate to generate the schedule -> Temperature cools down by the cooling rate on each iteration.

III) Generating random values in the range of upper-limit to lower-limit and shifted to the new value if (I) it was less than current or (II) with an exponential acceptance probability [E^(E2-E1)/T] that would decrease as temperature decreased [Algorithm’s p-value moves towards lower limit]

IV) On each iteration, used the p-value limit to iteratively eliminate attributes until all had p-values below the p-value significance limit

V) Training the linear regression model using the resulting attribute set and diagnosis values for each iteration

VI) Calculate the accuracy of prediction on the training set using the resulting trained linear regression model and store in a hashMap

VII) When temperature reaches 1, simulated annealing terminates and we iterate through the hashMap to find the p-value with maximum accuracy.

NOTE: An experiment using simulated annealing over a wide range of p-values using the 90-10 train-test method for evaluating accuracy was also performed with as explained, a almost constant accuracy limiting our choice of the ideal p-value shown in the result section with a graph

Prediction using KNN

Considering the drastic improvements to data munging resulting with significant reductions in sum of squared differences and mean squared error from using KNN algorithm with optimal K-value, we experimented with using KNN for prediction for the breast cancer dataset and the heart disease dataset with the following steps:

CHOOSING MOST OPTIMAL K:

I) Iterating over values of k over a range of 2 to 30.

II) Running k-nearest neighbors over each record of the complete set of records to find the k number of records closest to it in terms of a defined distance metric and storing their classification in an arrayList

III) With a plurality test, predicted value being the classification (absence or presence) having most occurrence in the arrayList of classifications of k nearest neighbors.

IV) If the predicted value is the same classification as known, increment a counter. Final accuracy = counter/number of records

V) Storing the accuracy and value of k in a hashMap and finally choosing the k-value with maximum accuracy.

FINDING ACCURACY WITH CROSS-VALIDATION:
Then, we found out the accuracy of KNN with optimal value of k by dividing the set of records into 10% for testing size and 90% for training

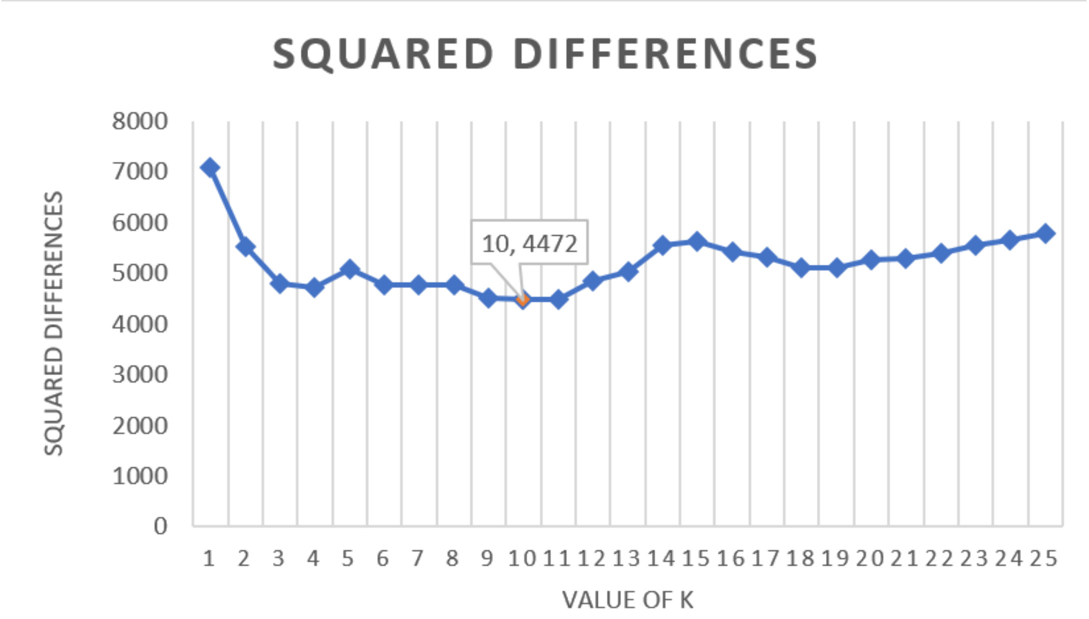
Data Augmentation

In the papers, a major weak-point is the lack of data as diagnosis datasets are hard to extract and another problem that stems from this is that some minority classes may not be represented aptly in the dataset and the model trained on it may not generalise well for these minority classes.

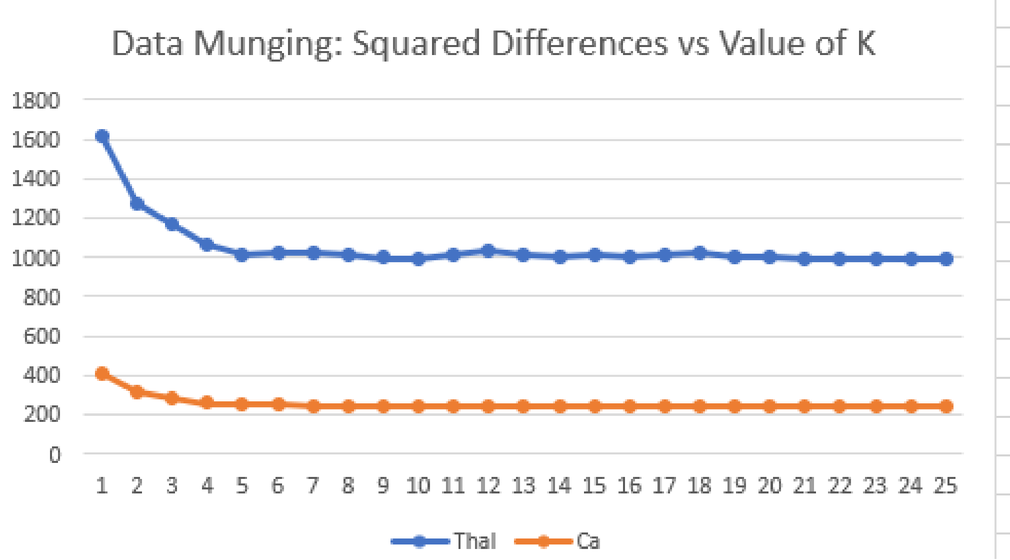
In our project, we use augmentation techniques of adding noise – adding a random value multiplied by a common deviation to all values and creating a new records, scaling – multiplying all values with a common factor to create a new records, combining – creating a new record by using maximum of values of two records.

RESULTS AND ANALYSIS

1. Sum of squared differences for data munging with KNN



Graph – 1: Values of K versus the sum of squared differences between predicted and known values for bare-nuclei value of breast cancer dataset.

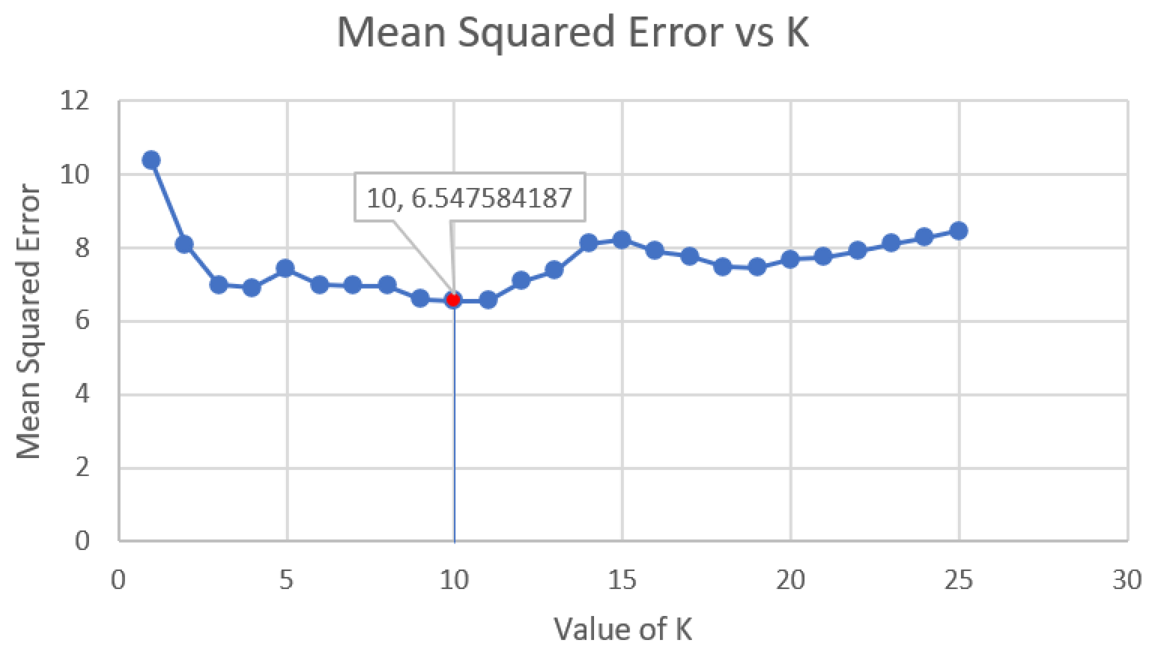


Graph – 2: Trend of total squared differences for different values of k in k nearest neighbours for replacing Thal & Ca values in heart disease dataset.

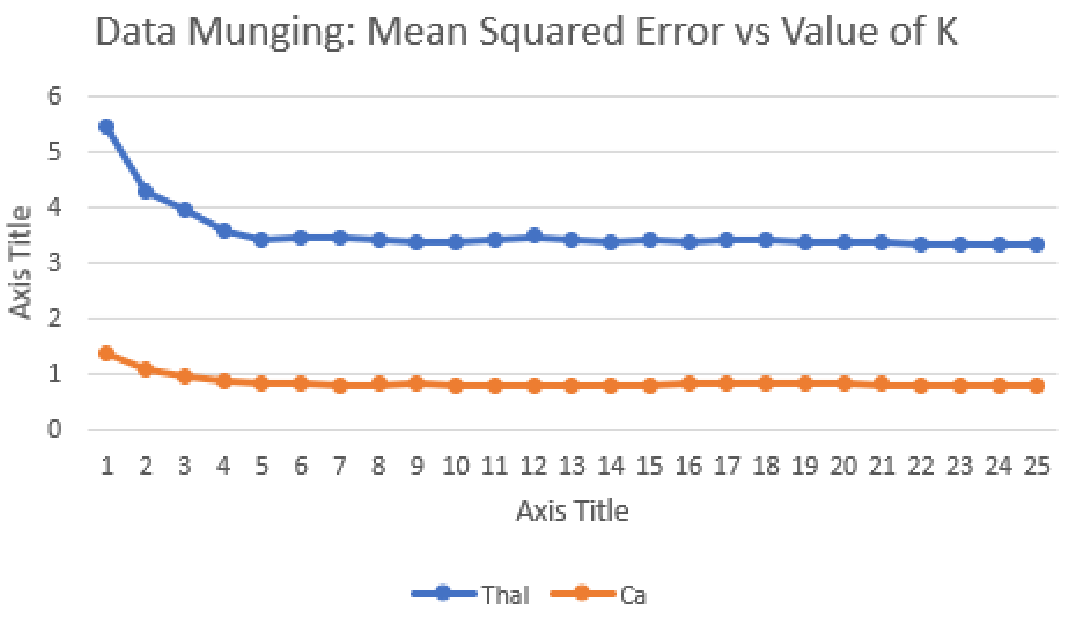
Analysis: For the breast cancer dataset, the sum of squared differences between known and predicted values for the attribute with missing values - bare nuclei - is minimised at K = 10.

For the heart disease dataset, the sum of squared differences between known and predicted values for Thal attribute is minimum at K = 13 and for Ca attribute it is minimum at K = 24.

2. Mean squared error for data munging with KNN



Graph – 2: Values of K versus the mean squared error for bare-nuclei value of breast cancer dataset[10 still appears to be most accurate value of k with least MSE]



Graph – 4: Trend of mean squared error for different values of in data munging by k nearest neighbours for replacing Thal & Ca values in heart disease dataset.

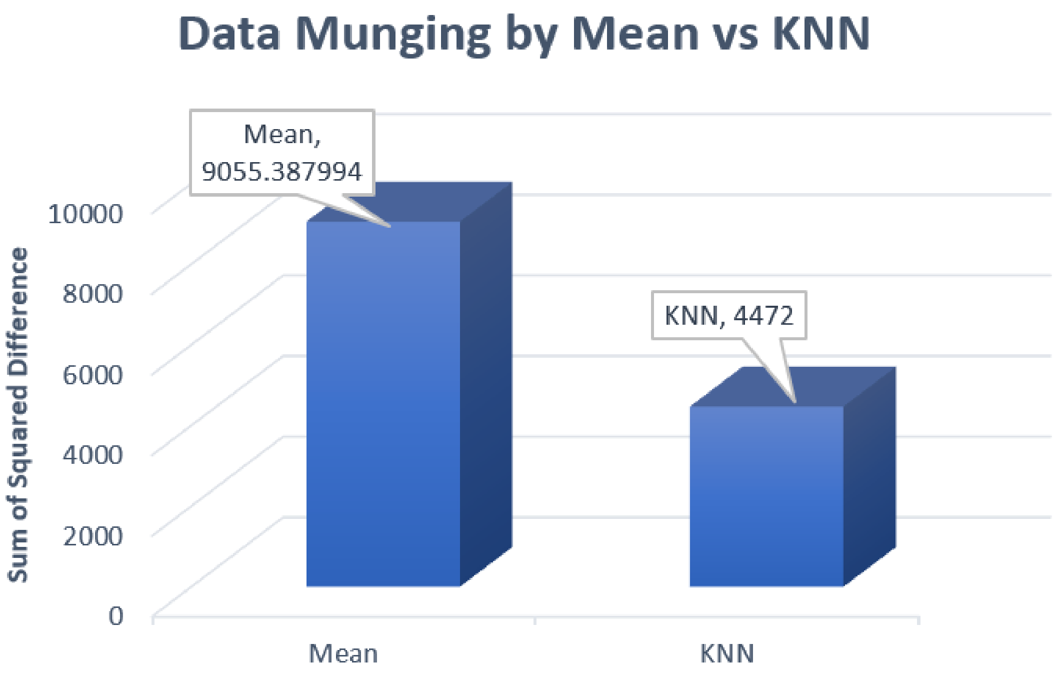
Analysis: For the breast cancer dataset, the mean squared errror = sum of squared difference/total records is also minimum at K = 10 for the bare nuclei attribute.

For the heart disease dataset, the MSE is minimum at K = 13 for Thal attribute and at K = 24 for Ca attribute.

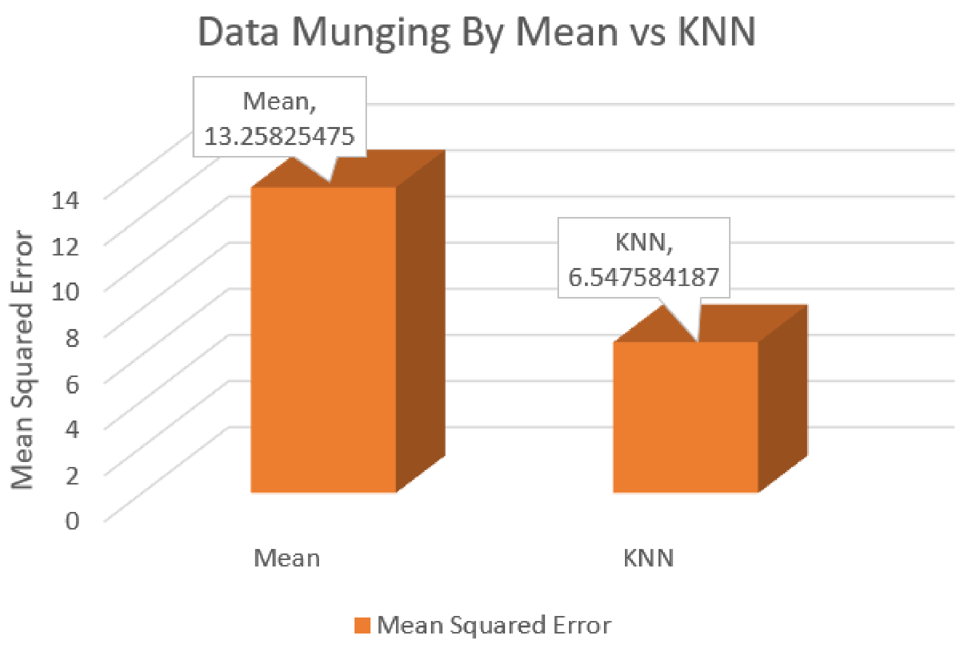
From the above graphs, most optimal values of K:

I) K = 10 for bare nuclei field of breast cancer dataset
II) K = 13 for Thal attribute of heart disease dataset.
III) K = 24 for Ca attribute of heart disease dataset.

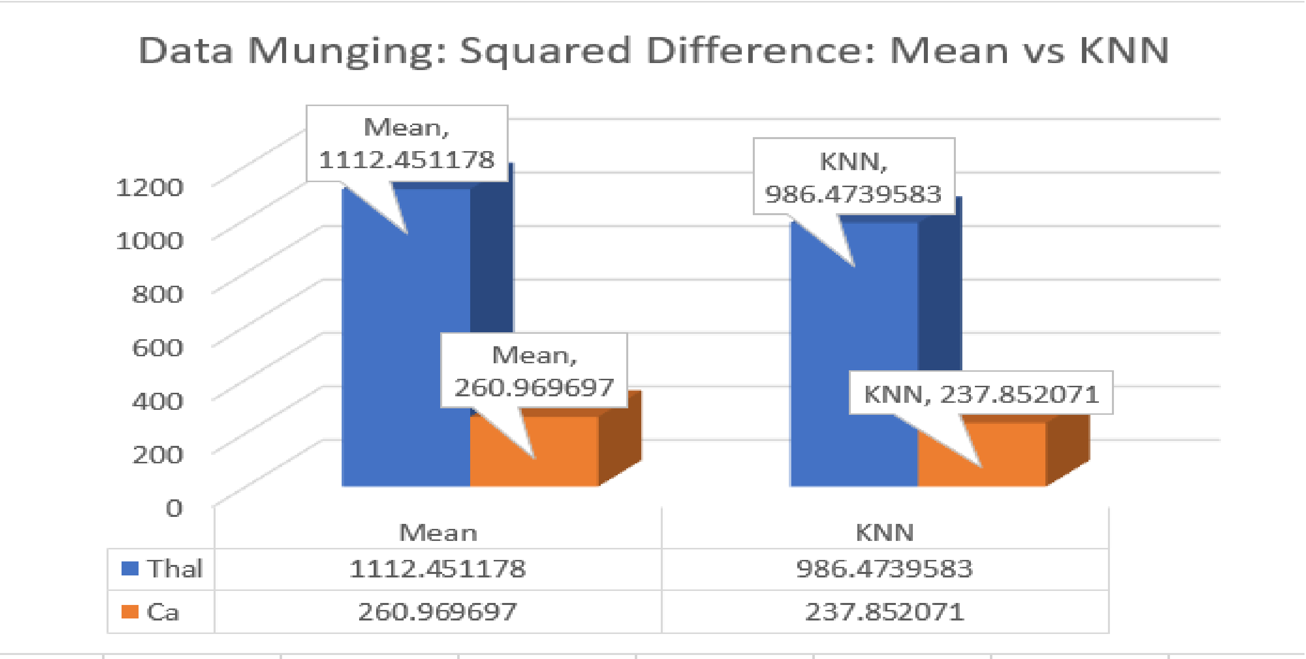
3. Comparison of sum of squared difference & MSE between data munging by mean and data munging by KNN using optimal K



Graph – 5: Sum of squared difference in data munging by replacing with mean vs replacing with k nearest neighbours for bare-nuclei value of breast cancer dataset [using optimal k = 10]

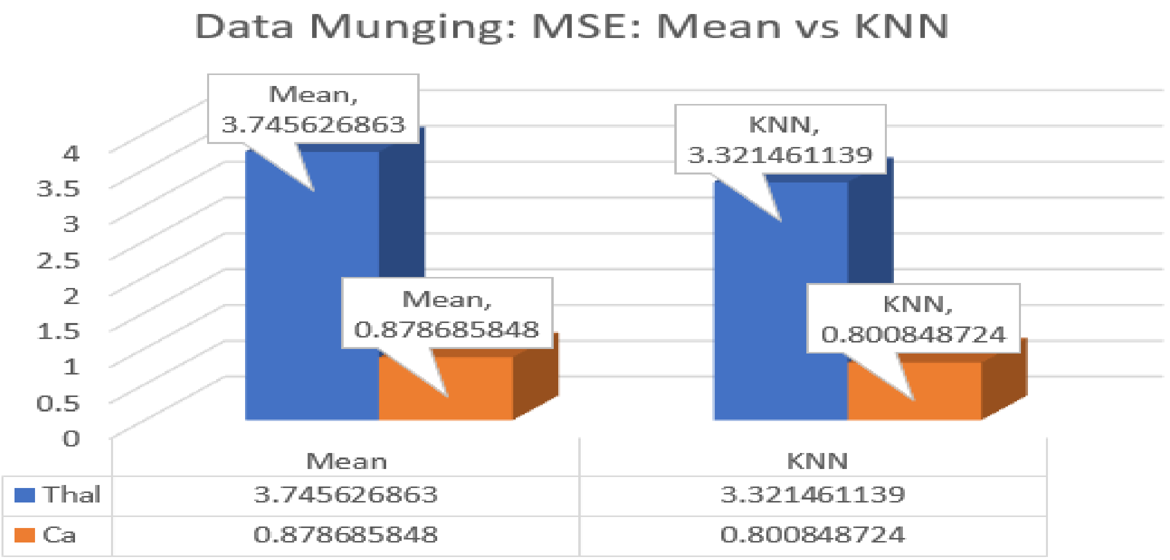


Graph – 6: Mean squared error between predicted and known values using mean vs using KNN for bare-nuclei value of breast cancer dataset[Optimal value of k = 10]



	Mean	KNN
■ Thal	1112.451178	986.4739583
■ Ca	260.969697	237.852071

Graph – 7: Difference in total squared difference between known and predicted values by using data munging by mean vs data munging by k-nearest neighbours in heart disease dataset.

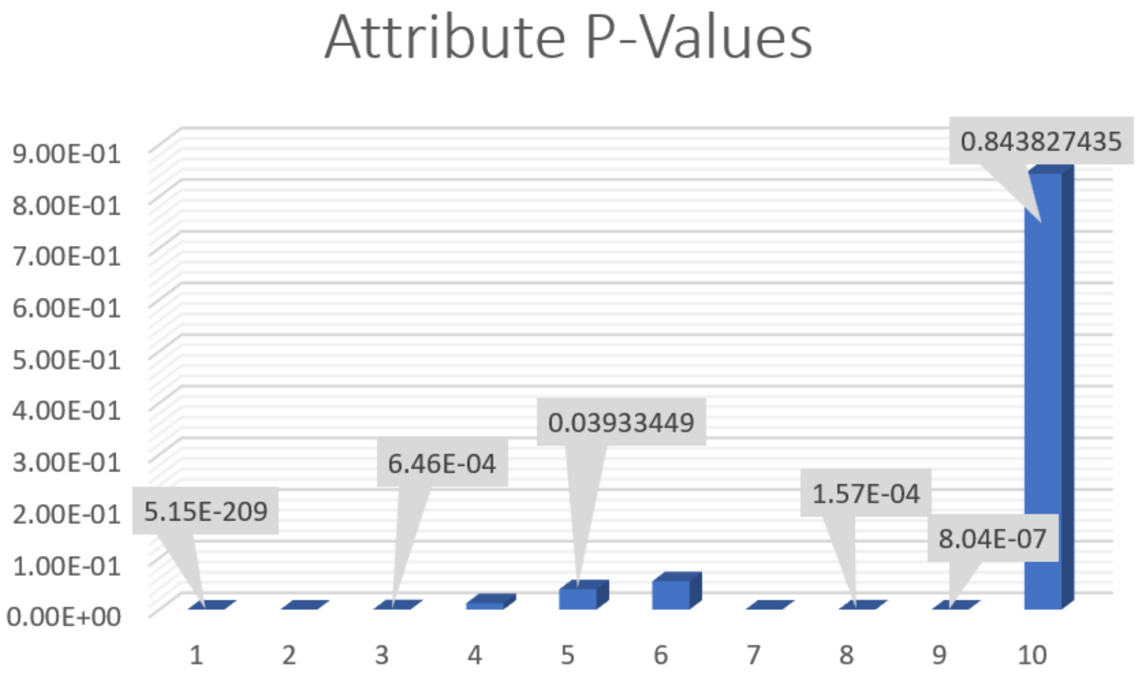


Graph – 8: Difference in mean squared error between known and predicted values by using data munging by mean vs using k nearest neighbours for heart disease dataset.

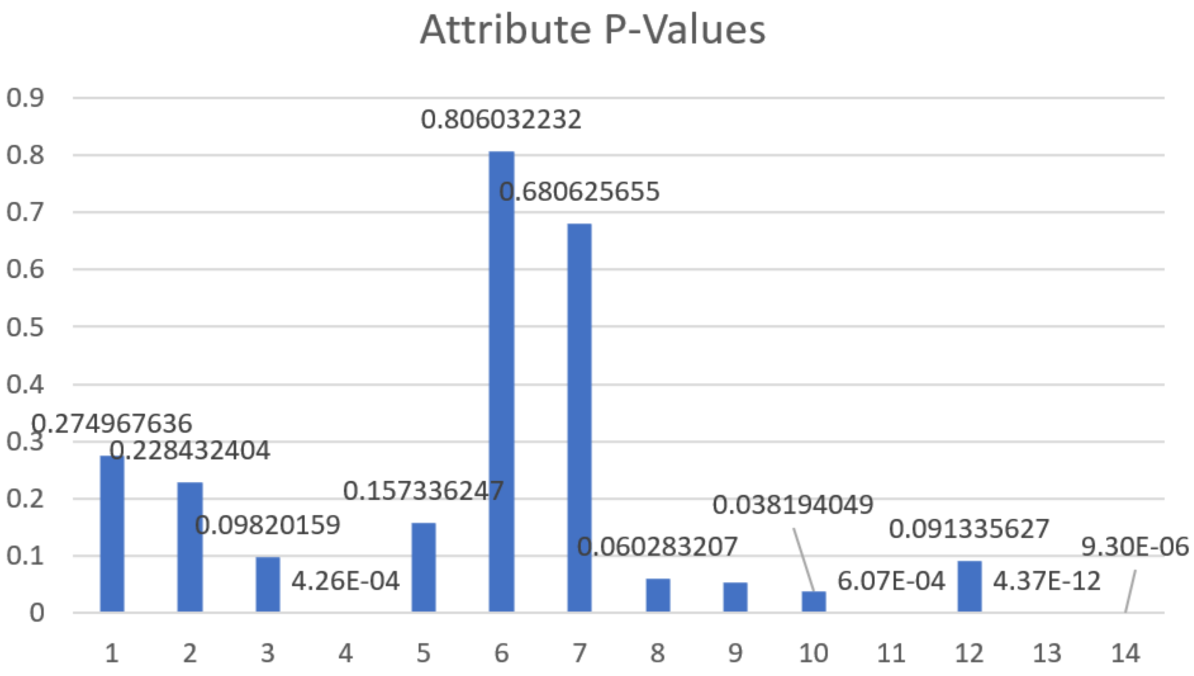
Analysis: For the breast cancer dataset, the sum of squared difference between known and predicted values of the bare nuclei attribute was more than halved when using KNN instead of mean and so was the MSE, suggesting a very significant improvement in accuracy of data munging.

For the heart disease dataset, the difference between mean and KNN was not as great but KNN still improved upon the MSE for both attributes Thal & Ca.

4. Initial P=values of all attributes for the datasets



Graph – 9: P-values for different attributes of breast cancer dataset. -> Ranging from 5.15E-209 to 0.8438



Graph - 10: P-values for different attributes of heart disease dataset. -> Ranging from 9.3R-06 to 0.806

Analysis: As shown, the range of p-values for attributes had an extensive range for the breast cancer dataset from E-209 to 0.8, hence multiple ranges of upperbound-lowerbound were required to get results over all possible combinations of attributes. However, attributes with p-values lower than 0.0001 are important to retain in the diagnosis model, hence we didn't experiment on setting a limit lower than 0.0001

5. Simulated Annealing: P-values vs Accuracies

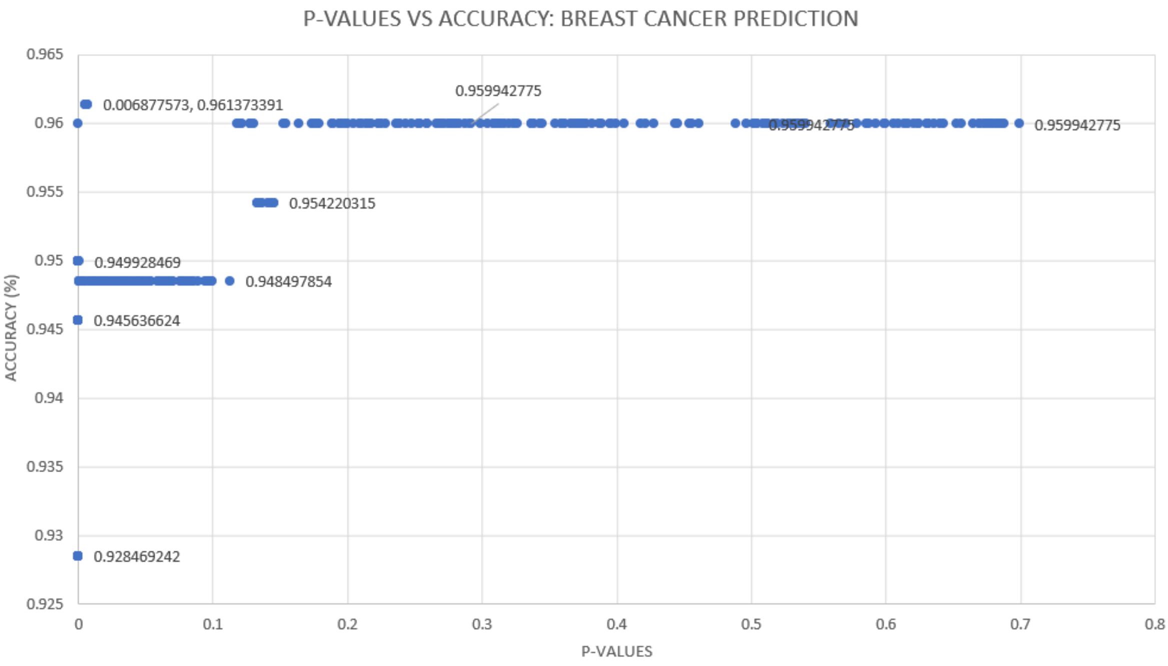
Simulated Annealing Initial conditions:

[BREAST CANCER DATASET]

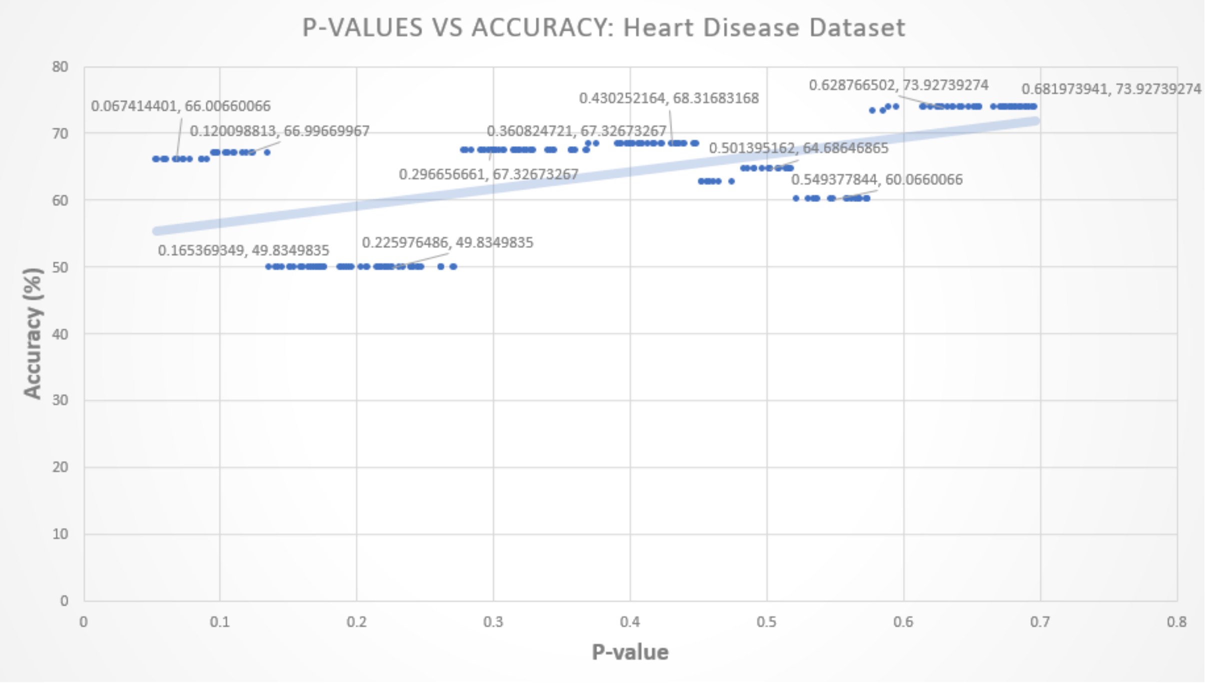
UPPER-BOUND P-VALUE = 0.75
LOWER-BOUND P-VALUE = 0.001 [Approx.]
CONSTANT COOLING RATE = 0.03
INITIAL TEMPERATURE = 1000
INITIAL P-VALUE = 0.75

[HEART DISEASE DATASET]

UPPER-BOUND P-VALUE = 0.70
LOWER BOUND P-VALUE = 0.10
CONSTANT COOLING RATE = 0.3
INITIAL TEMPERATURE = 1000
INITIAL P-VALUE = 0.8



Graph - 11: Accuracy of prediction using linear regression with attribute elimination based on different p-value limits for breast cancer dataset. [Highest at of]



Graph-12: Accuracy of prediction using linear regression with attribute elimination based on different p-limits for heart disease dataset. [Highest at approx 0.682 of 73.927%]

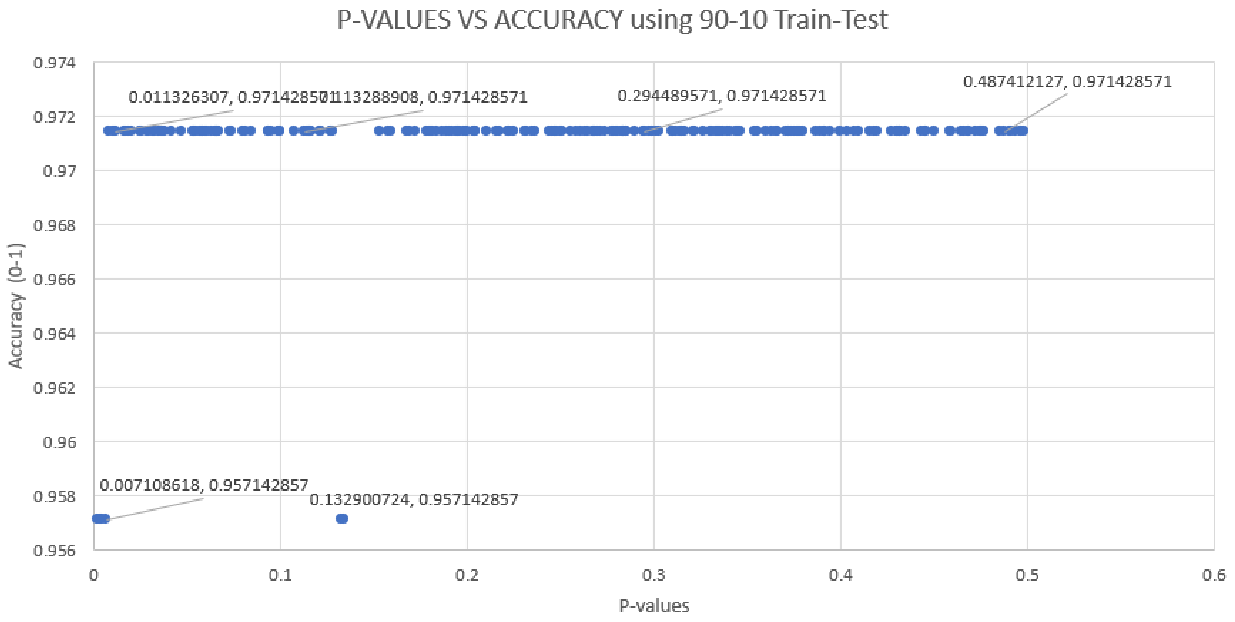
[HEART DISEASE DATASET] We compared the accuracy of linear regression using p-value significance limit of 0.05 versus p-value significance limit of 0.68 for attribute elimination, using 5-fold cross validation and direct accuracy calculation with the following results:

```
P-value = 0.05 Accuracy = 0.6533333333333333
P-value = 0.05 Direct Accuracy = 0.6600660066006601

P-value = 0.68 Accuracy = 0.7466666666666667
P-value = 0.68 Direct Accuracy = 0.7392739273927392
```

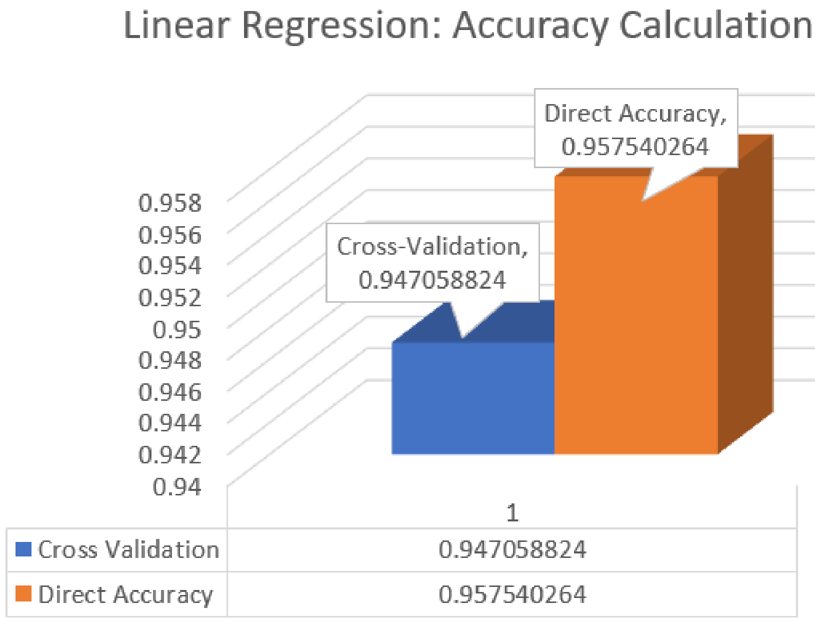
[BREAST CANCER DATASET] Linear regression resulted in an accuracy of prediction of 94.8% for p-value close to 0.05 and higher values of 95.99% at p-value 0.50437 and even higher accuracy of 96.14% ar p-value 0.00512. These results are significant since improved accuracies are found with p-values higher and lower than 0.05, as a higher limit leads to less attributes eliminated and hence a more complex model - but the limit of 0.00512 is much less than 0.05 and provides an accuracy greater than

NOTE: As mentioned at the beginning of the report, the 90-10 train test method has not been used since it resulted in stale accuracy values over a wide range with the following result [A almost constant accuracy of 97.14% can be observed] Hence, this method may limit our evaluation of the ideal p-value.



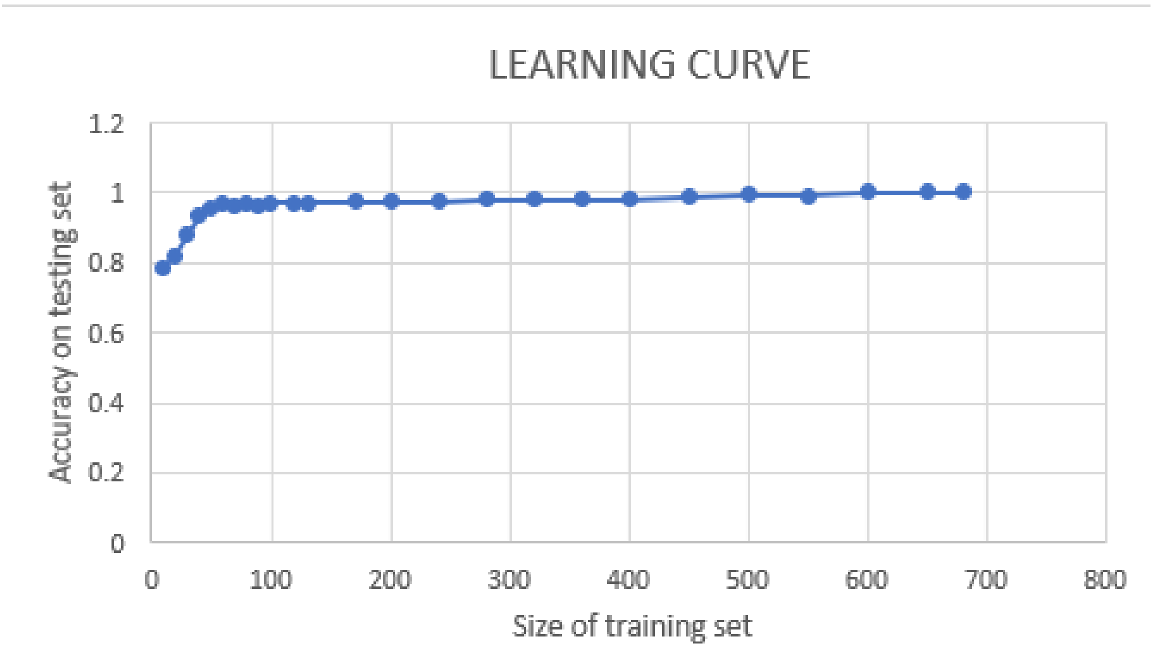
Graph-13: P-value vs accuracy for breast cancer dataset using simulated annealing using 90-10 Train-Test method for evaluation of accuracy.

Analysis: The above graph of p-values versus accuracies with data obtained from simulated annealing feature selection process, using the 90-10 train-test method for calculating accuracy - shows the constant trend of accuracy obtained through this method, with 97.14% being the highest accuracy obtained in the range of p-values from 0 to 0.5, hence not providing a strong evaluation of the p-values and making a decision using this metric clearly doesn’t consider the accuracy of direct calculation -> training on the whole record set and testing on the same.

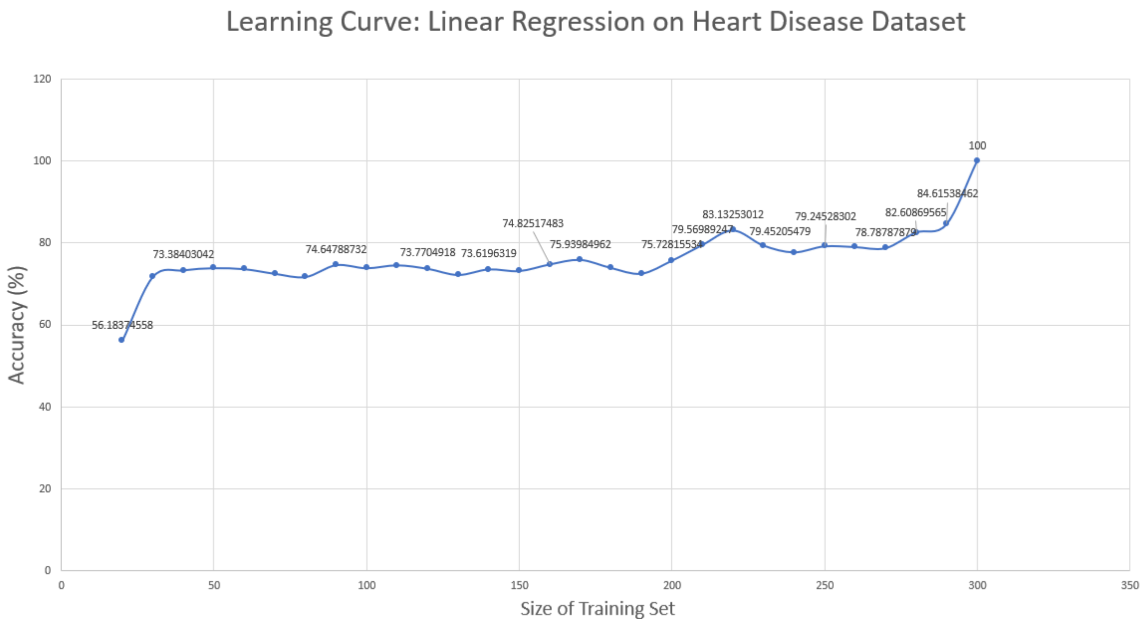


Graph-14: Accuracy of prediction using linear regression done by 5-fold cross validation vs direct calculation with p-value limit of 0.05 and on breast cancer dataset.

6. Learning curves for linear regression



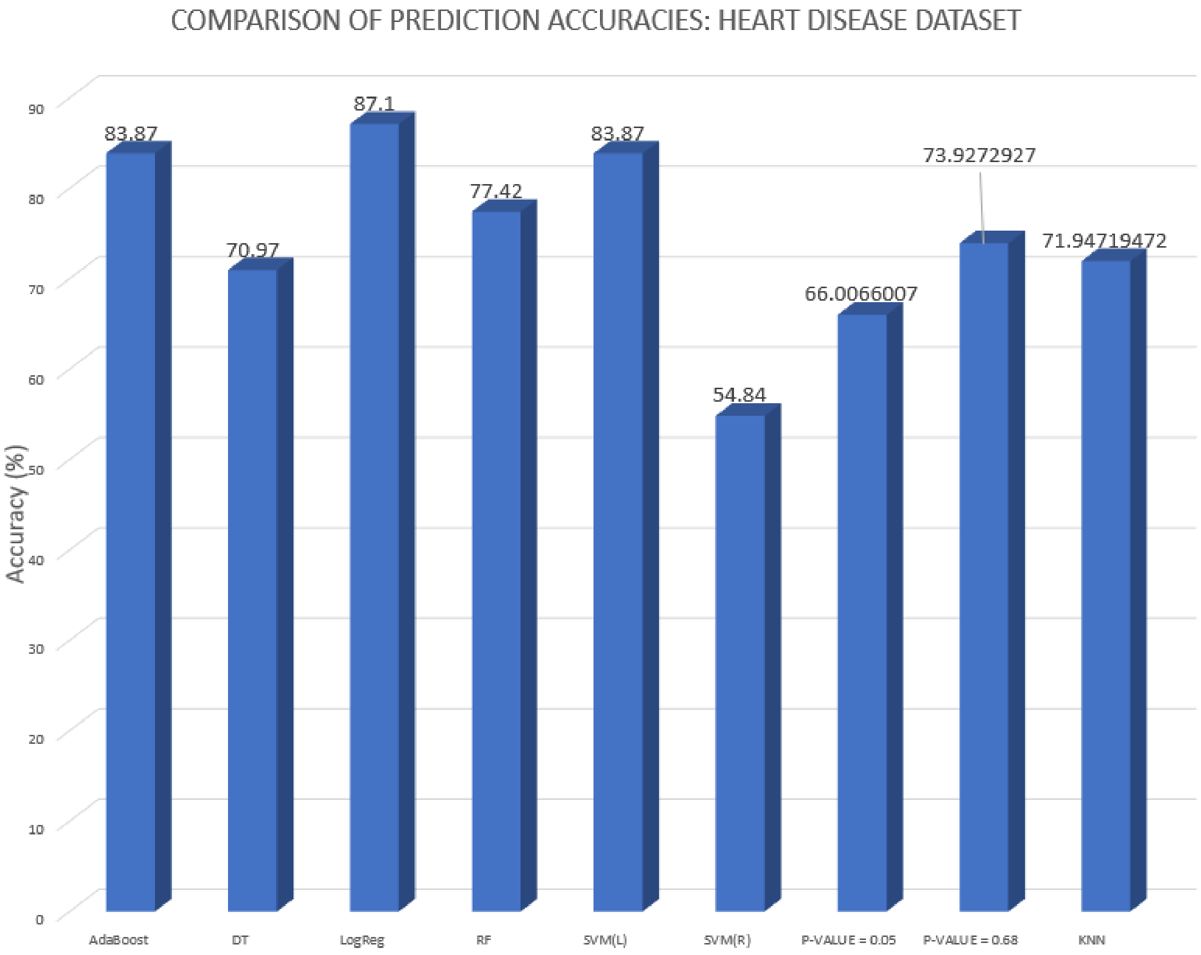
Graph-15: Learning curve for linear regression on breast cancer dataset, showing the accuracy obtained on varying the size of training set from 10 to 680 [Total - 699 records] from the total 699 records [after data munging], we varied the size of the training set of records from 10, 20 - to 680 [The increments are not constant].



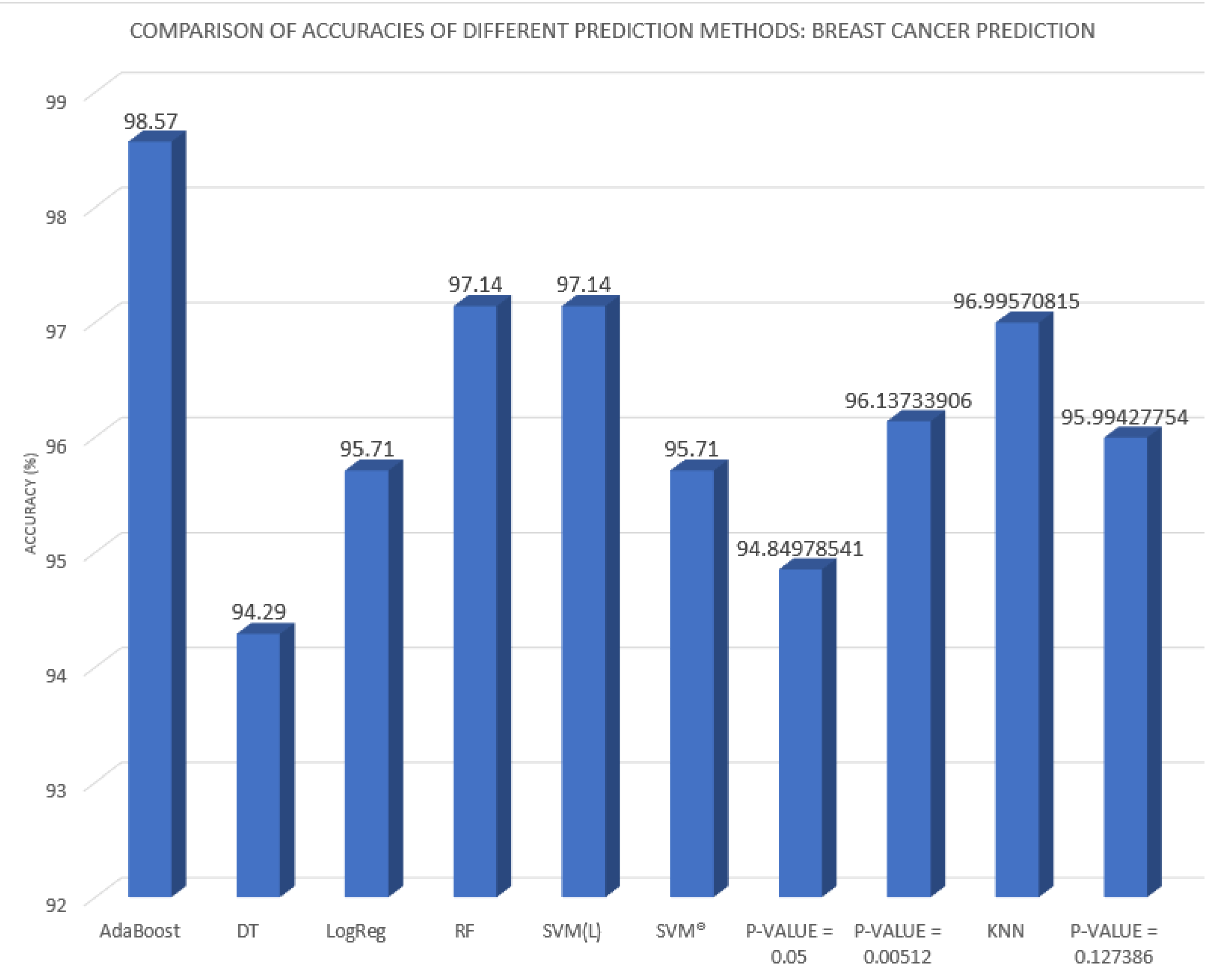
Graph - 16: Learning curve for linear regression on the heart disease dataset, showing the accuracy obtained on varying the training set size from 20 to 300 [Constant increments of 10] [Total - 303 records]

Analysis of learning curves: - For the breast cancer dataset, on varying the training size from 10-680 records with the rest used for testing, we received a relatively regular learning curve with accuracy tending towards 100% as the testing set size reduced and training set size reached the total number of records (699).

For the heart disease dataset, with a total of 303 records in the processed cleveland dataset, we varied the training set size from 10 to 300 and tested the accuracy on the testing set. We obtained a regular shaped graph with an eventual accuracy of 100% with a training set of size of 300 records.



Graph-17: Comparison of accuracies obtained by different machine learning methods used in the research paper vs methods used in the project [Linear Regression: P-value = 0.068, KNN] for heart disease dataset

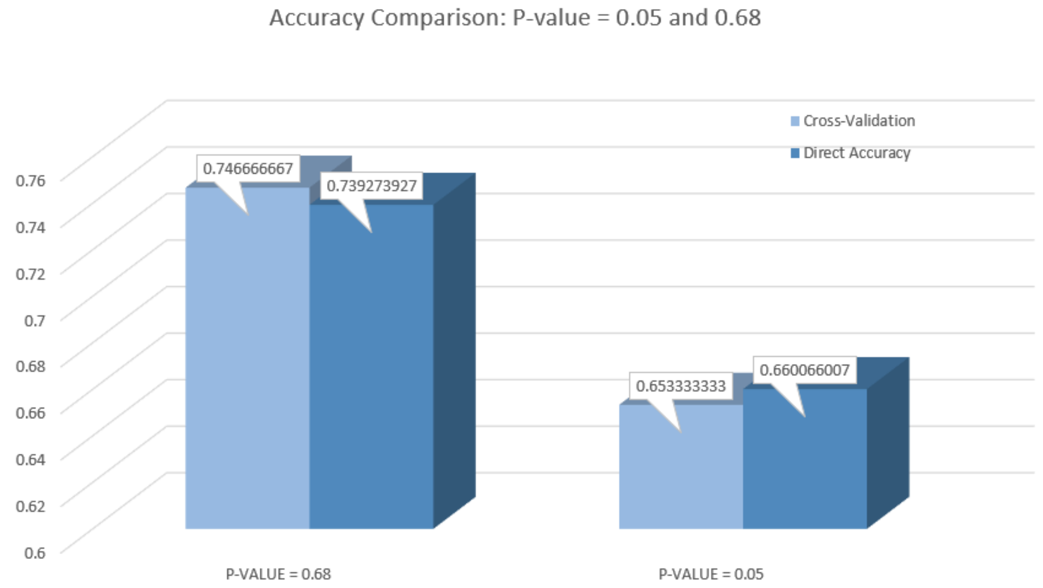


Graph-18: Comparison of accuracies of machine learning methods used in the research paper and linear regression with p-value elimination with 0.05, 0.00512, 0.127386 and by KNN with K = 2 for breast cancer dataset

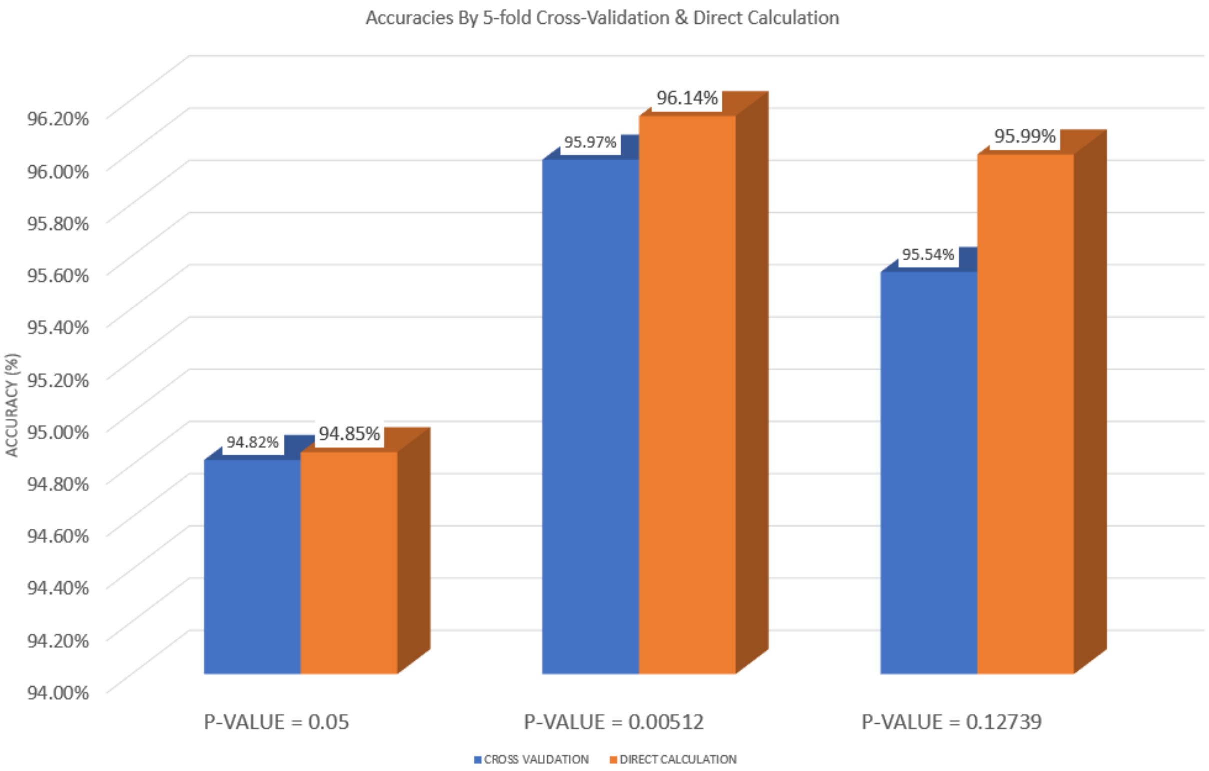
Analysis of overall accuracy comparison: - In the research paper, the machine learning methods of AdaBoost, DT, LogReg, RF, SVM(L) and SVM(R) were evaluated for the breast cancer dataset using the 90-10 Train-Test method. We have compared the accuracy obtained of Linear Regression using P-value = 0.05, P-value = 0.127, P-value = 0.00512 for attribute elimination along with Prediction using KNN evaluated by training with the entire dataset and testing accuracy on the same. The results clearly show KNN & LinReg with the P-values having an accuracy greater than that of DT, LpgReg and SVM(R) -> A considerable improvement and a method that should be considered further. Moreover, KNN with optimal K = 2 having the highest accuracy amongst the proposed methods in this project.

For the heart disease dataset, we compared the accuracies obtained on using p-value significance limit of 0.05 versus p-value significance limit of 0.068 and for KNN using optimal K = 16 with the machine learning methods used in the research paper, note again the difference in the evaluation of the methods in the paper and in this project. P-value = 0.68 with linear regression gave the highest accuracy amongst the three methods of 73.93%, while KNN resulted in an accuracy of 71.94%. The accuracy of 66% obtained by using p-value = 0.05 for attribute elimination clearly highlights a flaw in the p-value and the need to consider alternate p-values to adapt to the dataset & disease.

Accuracy comparison using Cross-Validation & direct calculation



Graph-19: Comparison between linear regression model using attribute elimination with p-value limit of 0.05 and p-value limit of 0.68 on heart disease dataset



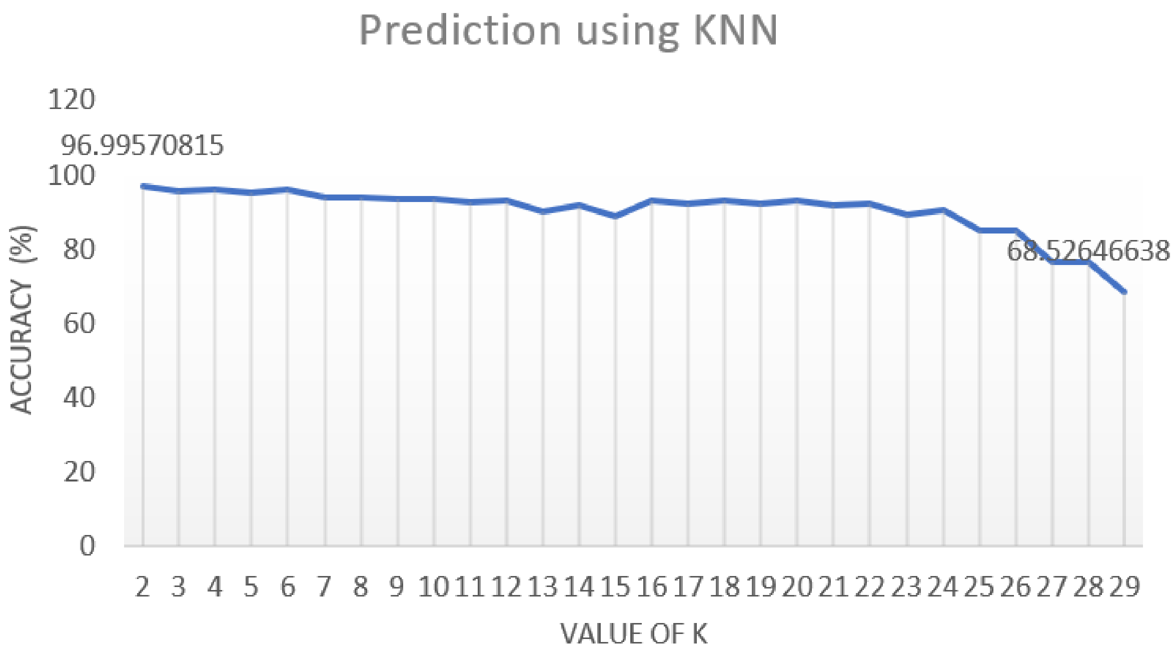
Graph-20: Comparison of accuracies using 5 fold cross validation and using direct calculation with p-values: 0.05, 0.00512, .12739 on breast cancer dataset.

Analysis: Breast cancer dataset -> Higher accuracies obtained using direct calculation

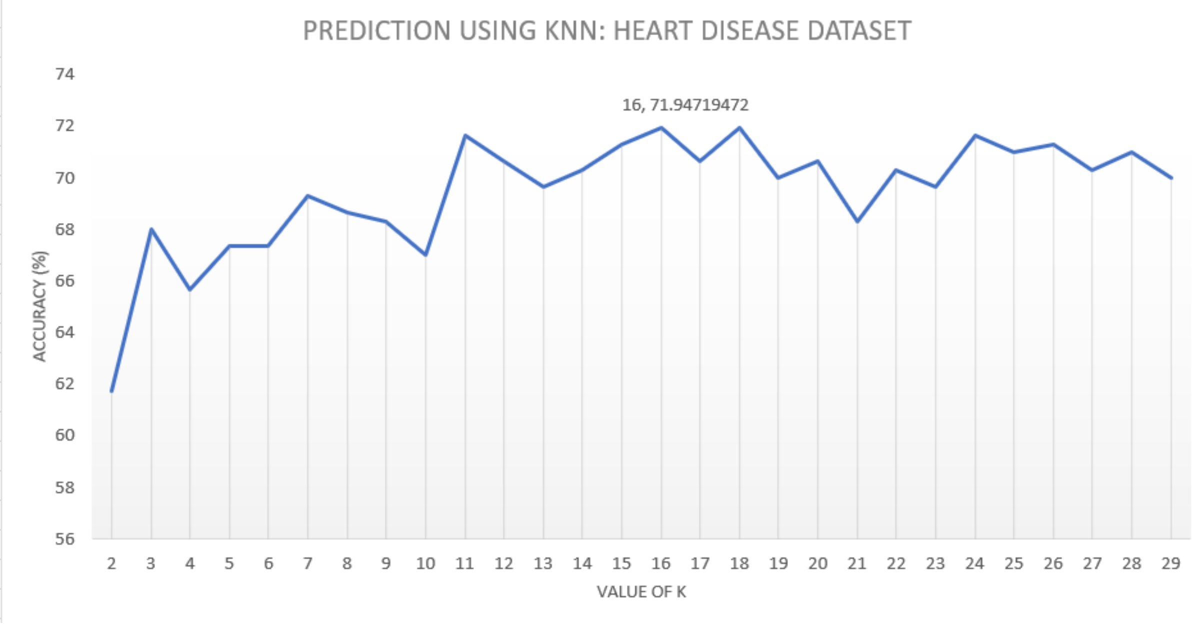
-> Accuracies of prediction for p-value = 0.00512 & 0.12739 using 5 fold cross validation - 95.97% and 95.54% considerably higher than that of p-value = 0.05 - 94.82%.

Heart Disease Dataset -> P-value = 0.68 with linear regression gives considerably higher accuracies (74.667% to 65.33%) than using p-value = 0.05 (73.93% - 66%) with linear regression with both cross validation and direct calculation.

Prediction Using KNN



Graph-21: Illustrating the accuracy obtained for prediction using KNN for different values of k [Optimal value obtained: K = 2 with accuracy = 96.9957%] for the breast cancer dataset



Graph - 22: Accuracy obtained for prediction using KNN for different values of K from 2 to 29 [Optimal value = 16 with accuracy = 71.947%] for the heart disease dataset

For the breast cancer dataset, on evaluating the accuracy of prediction with predicting the diagnosis using plurality test on the k-nearest neighbors' classification, K = 2 gave the highest accuracy of 96.996% using the direct calculation method described above - this accuracy is very significant and higher than methods such as DT, SVM(R) used in the research paper.

For the heart disease dataset, as highlighted on the graph, K = 16 gave the highest accuracy of prediction of 71.947% - again significant as being higher than that of linear regression using p-value limit of 0.05 and DT, SVM(R) methods used in the research paper for the heart disease dataset.

IMPLEMENTATION

Data sets used:

- Wisconsin Breast Cancer dataset
- Processed Cleveland Heart Disease dataset

Obtained from "UCI Machine Learning Repository": https://archive.ics.uci.edu/ml/index.php

A brief explanation of the code

Language used: Java

Java libraries used (from org.apache.commons.math3):

1. distribution.TDistribution - a library for computing probabilities and statistics related to the t-distribution, a probability distribution used in hypothesis testing and confidence interval calculations.
2. linear.Array2DRowRealMatrix - a library for creating and manipulating 2-dimensional matrices of real numbers in Java, based on an array data structure.
3. linear.RealMatrix - a library for creating and manipulating matrices of real numbers in Java, with support for various operations such as addition, multiplication, and inversion.
4. stat.regression.OLSMultipleLinearRegression - a library for performing multiple linear regression analysis, a statistical method for modeling the relationship between a dependent variable and multiple independent variables.

Important methods used:

1. The newSampleData method of the OLSMultipleLinearRegression class in Java is used to set the independent and dependent variables for

a regression analysis. It takes two arguments: a 2-dimensional array representing the independent variables and a 1-dimensional array representing the dependent variable. These data are used to estimate the regression coefficients using the ordinary least squares method.

2. The estimateRegressionParameters method of the OLSMultipleLinearRegression class in Java is used to estimate the regression coefficients using the independent and dependent variables that have been previously set using the newSampleData method.

3. The estimateResiduals method of the OLSMultipleLinearRegression class in Java is used to estimate the residuals of the regression analysis (A residual represents the difference between the actual value of the dependent variable and the predicted value based on the regression model), based on the independent and dependent variables that have been previously set using the newSampleData method, and the regression coefficients that have been estimated using the estimateRegressionParameters method.

4. getPValues: Using a custom OLS model and passing the attribute values for all 9 attributes along with values of predicted attribute, this function finds the p-values for all attributes and prints the same.

We have made two separate programs: Breast Cancer Prediction, and Heart Disease Prediction

Brief description of functions we defined:

We have a class BreastCancerPrediction that encapsulates all the functionalities. We start with defining some constants, such as the initial temperature, cooling rate, and iterations per temperature. These constants are used in the simulated annealing algorithm, which is used for feature selection. As the program starts, the main calls several functions to create and populate a database table with breast cancer data from a file, read and store complete and missing data in arraylists, and determine the optimal number of neighbors to consider in the k-Nearest Neighbors algorithm.

For heart disease prediction, we follow the same steps as with the Breast Cancer Prediction. Here we have two attributes with missing values (Thal and Ca) for several records, however the method of training, prediction and accuracy measurement is no different for both the attributes as compared to the model described above, with the major steps being: getting and processing the input data, performing missing value imputation, feature selection, and prediction and accuracy testing. Techniques used such as selecting best K for KNN, simulated annealing for threshold p-value, K-fold cross validation, fitting a linear regression model, and augmentation methods of adding noise, combining and scaling remain the same.

Function Name	Input	Output	Description
eadAndAddDa	None	None	nction reads data from a file and inse nto a database table named ancerdata/heartdata

adMissingAnd pleteData	almd & alcd ArrayLists to add data to	None	nction reads complete and missing d om the bcancerdata table and stores em in separate arrays.
mpleteByMea r	arrayList alcd of complete records	Prints MSE & SSD	edicts values of bareNuclie by using ean of complete records for complete cords mputes MSE & SSD between the actu predicted values between known & edicted values
stKPredict	arrayList alcd of complete records	teger - optimal lue of k	eates a HashMap to store the accura rcentage (using accuracyofKPredict) ch value of k between 2 and 29 erates through the HashMap to find t y with the highest value. It then retu
stanceMetric	two complete records or one complete & one missing	Integer distance	lculate the euclidean distance betwe o records as a sum of differences tween feature values.
stK	ArrayList of complete records	teger value of timal K	e function iterates through k values om 1-25, calls accuracyOfK for each k d stores the MSE result in a HashMa terwards, the function iterates throu e HashMap to find the k with the nimum MSE value.

uralityTest	ArrayList of bare nuclei values	:eger 2 or 4 wit ghest plurality	counts the occurrences of class 2 anc ıss 4 in an ArrayList of bare nuclei val d returns the class with the highest unt as the predicted classification.
curacyOfK	Integer - K ArrayList alcd of complete records	mean squared error	edicts the value for missing attribute ing input k by average of values of k arest neighbors lculates sum of squared diff betweer own and predicted values & returns SE
earestNeighbo	A complete element, ArrayList of complete records, K	ArrayList of values of missing attribute of the k nearest neighbors	lculates distance between element a ch other element, stores the distanc d attribute value in a HashMap & the ements with least distance to the ement in an arrayList & returns it
tPValues	attribute set X & predicted value set Y	prints the p-values	ith a custom OLS model by passing th tribute values for 9 attributes along v lues of predicted attribute, finds the values for all attributes
lculateAccurac	Array of true values, Array of predicted values	Accuracy - 0 to 1	atches corresponding true and edicted values & returns no of correc edictions/total no of records

ossValidation	1. 2D array of attribute values for records 2. 1D array of values of predicted attribute	Average accuracy of 5 fold cross validation	Shuffles the records & Iterates over t ds Forms training & testing set of recor Predicts values for testing set by ining linear regression model Calculates accuracies over all 5 folds d returns the average accuracy
edict:	1. 2D array of training attribute set 2. 1D array of training values of predicted attribute 3. 2D array of testing attribute set	1D array of predicted values for the testing set	ains a linear regression model on the ining set nd the predicted value using the mod r each record in the testing set and a em to an array turns the array of predicted values
ceptanceProba .	Current p-value, new p-value, temperature	Accuracy of accepting new p-value	irst checks if the new solution is bett e, less]than the current solution. If it accepts the new solution with obability 1.0 (i.e., it always accepts it he new solution is worse (i.e. greate an) than the current solution, the ethod returns the probability of

			cepting the new soln as (CURRENT-NEW)/TEMPERATURE
rmalFeature tion:	ArrayList alcd of complete records limit value for p-value significance limit	Prints the accuracy of linear regression with given p-value limit	eates 2D array of attribute set and 1D ray of predicted values for the compl t. ratively removes attributes until all ve p-value under limit & calculates curacy of model trained on resulting
FeatureSelect	ArrayList alcd of complete records UpperBound & lowerBound for simulated annealing.	Prints the p-values & accuracies and p-value with highest accuracy	es simulated annealing to vary the imit of attribute elimination ains linear regression model on resul tribute set and calculates accuracy of odel for each p-value limit ows p-value with highest accuracy
tLearningCur	ArrayList alcd of complete records	prints training size & accuracy	lculates accuracy of linear regressior ferent sizes of training set by testing e test set by shuffling and dividing th iginal set.
dNoise	one complete record	one complete record	random variable is added to each ature with a given standard deviation dDev) to create a new record

ombine	two complete records	one complete record	aximum value of attributes between cords used to create a new record
ale	one complete record	one complete record	ch attribute value is multiplied by a ven factor to create a new record

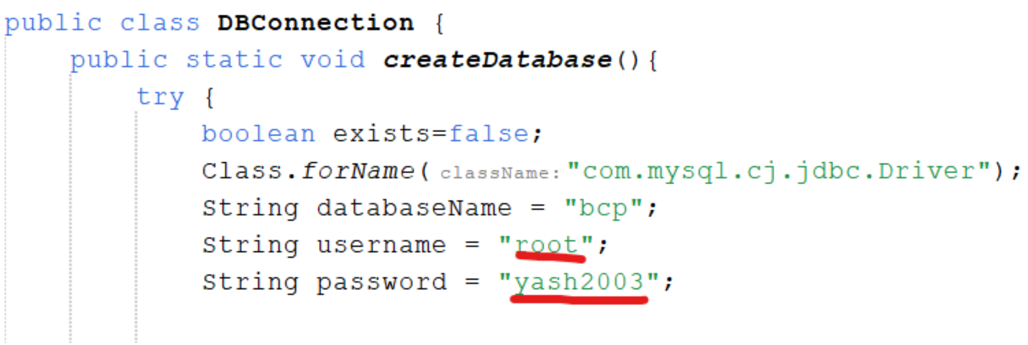
Heart Disease Prediction:

We follow the same steps as with the Breast Cancer Prediction. Here we have two attributes with missing values (Thal and Ca) for several records, however the method of training, prediction and accuracy measurement is no different for both the attributes as compared to the model described above, with the major steps being: getting and processing the input data, performing missing value imputation, feature selection, and prediction and accuracy testing. Techniques used such as selecting best K for KNN, simulated annealing for threshold p-value, K-fold cross validation, fitting a linear regression model, and augmentation methods of adding noise, combining and scaling remain the same.

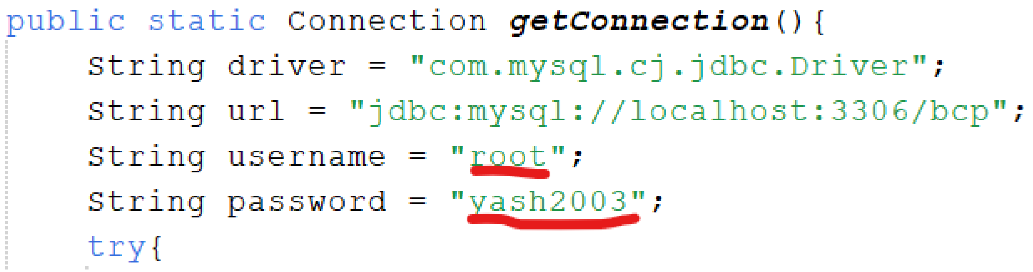
Instructions for running the application:

1. Prerequisite: A local mysql server.
2. Open the DBConnection.java file in each of the two projects (BreastCancerPrediction & HeartDiseasePrediction).

3.



```
public class DBConnection {
    public static void createDatabase(){
        try {
            boolean exists=false;
            Class.forName( className: "com.mysql.cj.jdbc.Driver");
            String databaseName = "bcp";
            String username = "root";
            String password = "yash2003";
```



```
public static Connection getConnection(){
    String driver = "com.mysql.cj.jdbc.Driver";
    String url = "jdbc:mysql://localhost:3306/bcp";
    String username = "root";
    String password = "yash2003";
    try{
```

Change the username and password in these two instances as per the credentials set in your system during your mysql installation. (For most mysql servers, the defaults are username: “root”, password: “”).

4. Run the BreastCancerPrediction.java and the HeartDiseasePrediction.java independently.

Note: The main functions of each of the two programs do not run all the functions in sequence. You may uncomment the function calls to get the outputs of specific functionalities as needed.

CONCLUSION

Conclusion of comparison between using KNN vs using mean for data munging:

For breast cancer dataset, the method of using k-nearest neighbours with optimal value of k = 10 for data munging resulted in more than **halving** the sum of squared differences and more than halving the mean squared error when compared to data munging by replacing with mean value.

For the heart disease dataset, KNN still proved to produce results closer to true value for Thal & Ca fields with lesser sum of squared differences and lesser mean squared error than when replacing with mean

Conclusion of varying p-values for eliminating attributes: For the breast cancer dataset although the accuracies did not show a completely consistent trend, compared to the accuracy of 94.48% for p-value significance level of around 0.05 while using a p-value of around 0.12, a higher accuracy of 96% could be achieved for the linear regression model.

For the heart disease dataset, the accuracies obtained for p-values in the range of 0.6-0.7 were higher than those obtained for the range 0.05 - 0.1. Showing a possibility of a more accurate model on increasing/varying the p-value significance limit, at least for this particular disease and recordset, with highest accuracy of 73.9% being more than that of SVM(R), DT machine learning models used in the research paper. Moreover, the graph showing the accuracies obtained on using the 90-10 Train-Test method used in the research paper indicates a constant trend of 97.14% for the breast cancer dataset, clearly not providing an in-depth evaluation of the different p-values. The evaluation metric used here of direct calculation however, may allow choosing the best p-value for a dataset and disease.

Conclusion of using KNN for prediction

With the great results obtained from using KNN for data munging, the results of KNN for breast cancer using the optimal value of K = 2, extracted amongst K = 2 to K = 30 by comparing accuracies of predicted values against known values on the complete set of records and conducting plurality test upon the classifications of the K nearest neighbors to an item -> the accuracy of 96.99% is significant compared to the other machine learning methods proposed in the paper, with SVM(L), AdaBoost and RF still having better accuracy, but KNN had a greater accuracy than SVM(R) and LogReg, DT methods. However, the different evaluation methods of accuracy [90-10 train test and direct calculation] should be kept in consideration.

REFERENCES

1. Kohli, P. S., & Arora, S. (2018, December). Application of machine learning in disease prediction. In 2018 4th International conference on computing communication and automation (ICCCA) (pp. 1-4). IEEE.

2. UCI Machine Learning Repository: Processed Cleveland Heart Disease Dataset: UCI Machine Learning Repository: Heart Disease Data Set

3. UCI Machine Learning Repository: Wisconsin Breast Cancer Dataset: UCI Machine Learning Repository: Breast Cancer Data Set

4. Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

5. Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1, 1-14.