COMM055

Machine Learning and Data Mining

Coursework 2020

A Comparison Study on
Machine Learning Algorithms and
Data Mining Techniques

Written by:

Aju, Anju (6615183)

Ayyadurai, Agalya (6608114)

Omote, Ai (6613370)

Patil, Ishan (6607828)

Ploysuayngam, Sophita (6604934)

Somchit, Pariyakorn (6606616)

Abstract

The purpose of the study is to evaluate, compare and discuss the results and insights gained from exploring different Machine Learning (ML) algorithms as well as Data Mining (DM) techniques. Therefore, this report is about exploring different algorithms belonging to Classification, Regression and Clustering problem types. The basic design of the study includes Data Preparation and Understanding, Data Pre-Processing, Modelling, Parameter Tuning, Evaluating the models within and against two datasets. One of the major findings is that an algorithm performing well on one dataset need not necessarily perform well on the other dataset. Overall, on two different datasets, the performances of algorithms were evaluated based on their performance metrics, interpretability, interestingness and their ability to generalize on unseen data.

***Keywords**: machine learning, data mining, classification, regression, clustering, algorithms, modelling, evaluation, comparison, parameter tuning, performance metrics, roc, cross-validation, r-squared, mean-squared error, adjusted-rand index, homogeneity, completeness, v-measure, justifications, IBM attrition, mushroom classification, life expectancy, car price prediction.*

Table of Contents

1. Introduction

‘Machine learning[1] is the application of artificial intelligence that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed’. Machine Learning has three categories and two of them are Supervised (labelled) and Unsupervised Learning (unlabelled). Supervised learning is where you are given both the input and output variables and the inputs are mapped to the outputs (Classification and Regression problems). Unsupervised learning is where there are only input variables and no output variables and similar features are categorized together as in Clustering.

The goal of this project is to compare and evaluate various machine learning algorithms on two chosen datasets. The strategy is to explore and evaluate the performance of machine learning algorithms of Classification, Clustering and Regression problem types within and across the chosen datasets.

2. Project Definition

The objectives of the project are:

- To understand and give a problem statement for each dataset and identify hypotheses to be addressed.
- To prepare the data for analysis that includes data loading, data cleaning, dimensionality reduction, data exploration and data visualization.
- To design modelling tasks, select learning algorithms and justifying as required.
- To model algorithms belonging to different problem types as Classification, Regression and Clustering.
- To evaluate the performance of these algorithms using two different performance metrics. ROC and Cross-Validation (for Classification algorithms), Adjusted Rand Index and V-measure (for clustering algorithms), R-Squared and Mean-Squared error (for Regression algorithms).
- To tune the parameters of the algorithm either manually or using automated tuning. Also, a systematic approach for tuning (to tune one specific parameter and using the validation set separated from the train set to understand how the model generalizes well on unseen data) parameters is taken to find the best value where the model would perform well both on train and unseen data.
- To summarize and discuss the results and propose the best performing algorithm on specific dataset.
- Finally, accepting or rejecting any initial assumptions/hypotheses addressed and including any newly discovered patterns or trends in the data.

Dataset	Problem Type(s)
IBM Attrition	Classification and Clustering
Mushroom Classification	
Car Price Prediction	Regression
Life Expectancy	

Table 1

As the main objective is to compare different machine learning techniques, it was decided to explore algorithms from three different problem types. Therefore, for the analysis and modelling, four varied datasets were selected, and the problem type for each dataset is identified as given in Table 1.

2.1 Datasets Overview

An overview of all the chosen datasets, including the data dictionary, data understanding, problem statement for each dataset, hypotheses considered, ML/DM techniques used, performance metrics used are discussed in this section. The summary of the chosen datasets is given in Table 2.

Dataset	**No. of Records**	**No. of Columns**	**Description**
IBM Attrition[7]	1500	35	"Attrition" is the target variable with classes 'yes' or 'no'.
Mushroom Classification[8]	8124	23	The target variable is "class" with values 'edible' or 'poisonous'
Car Price Prediction[9]	205	26	The target variable is the pricing of the car
Life Expectancy[10]	2939	22	The target variable is "life expectancy" for specific countries over the years

Table 2

2.1.1 IBM Attrition

Field name*	**Data type**	**Description**
Age	Numeric – int64	Employee age
Attrition	Object	If the employee decided to stay or leave the company
Business Travel	Object	How often the employee goes for business trips
Daily Rate	Numeric – int64	Their daily income
Department	Object	The department which the employee works in
Education	Numeric – int64	1- Below College, 2- College, 3- Bachelor, 4- Master, 5- Doctor

*Table 3 - Data Dictionary for IBM Attrition (*Lists only few columns - the entire table is in Appendix A)*

Data Understanding: The output variable is "Attrition" which contains values "yes" or "no" meaning if an employee left the company or not. There are 34 input variables that could determine the target variable (Attrition). Therefore, a model can be designed with the data, which could predict if an employee would stay or leave the company based on the specific input features.

Problem Statement: The dataset has a target variable that classifies employees into two categories. Therefore, this belongs to the classification problem type where a classifier (supervised machine learning technique) can be built on this dataset. The dataset could also be clustered into 2 clusters where clustering algorithms (unsupervised machine learning technique) can be used to learn the data. To choose appropriate classification and clustering algorithms along with relevant performance metrics is vital to achieve the objective of the project. This helps in understanding how different algorithms belonging to supervised and unsupervised machine learning techniques perform on the same/different data.

Hypotheses Considered: Before approaching the project tasks, it is vital to address questions related to the data and machine learning techniques to know what results to expect. So, following hypotheses were considered:

- Employees who worked longer hours are likely to leave the company, perhaps due to stress of workload
- Employees leaving the company would be directly proportional to their job satisfaction, where lower the job satisfaction employees are more likely to leave than employees with higher job satisfaction.
- Among the classification algorithms, Logistic Regression (ideal algorithm for binary classification) is assumed to perform well, since the target data is dichotomous.

ML/DM Techniques: The following ML/DM techniques were planned to be used:

- To find the correlation among input features using the correlation matrix
- To visualize the input features by plotting histograms, box plots and bar plots
- One-hot encoding on the categorical columns of the dataset
- To plot a bar graph that shows how many employees for the specific input feature have Attrition values as "Yes" or "No"
- To build models on the data using classification algorithms such as Logistic Regression, Decision Trees, Perceptron, Multi-Layer Perceptron (MLP) and Support-Vector Machines (SVM)
- To also build models on the data using clustering algorithms such as K-Means, K-Modes, Hierarchical Agglomerative Clustering (HAC)
- To tune the parameters of the model manually or using automated tuning (e.g. Grid Search) or a systematic tuning (using the validation set separated from the train set)
- To evaluate and compare the performance of different models.

Performance Metrics:

- For Classification algorithms - Receiver operating characteristic curve (ROC) and Cross-Validation.
- For clustering - Adjusted-Rand Index (ARI) and V-measure.

2.1.2 Mushroom Classification

Data Understanding: The target variable is "class" with values 'p' (poisonous) and 'e' (edible). A model can be designed to classify the mushrooms into edible and poisonous based on the 22 input variables.

Problem Statement: In addition, to the problem statement given for IBM dataset, to compare the performance of the same algorithms on different datasets - IBM and Mushroom.

Field name*	Data type	Description
Class	Object	'poisonous' or 'edible'
Cap Shape	Object	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
Cap Surface	Object	fibrous=f, grooves=g, scaly=y, smooth=s
Cap Colour	Object	brown=n, buff=b, cinnamon=c, grey=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
Bruises	Object	bruises=t, no=f
Odour	Object	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s

*Table 4 - Data Dictionary for Mushroom Classification (*Lists only few columns - the entire table is in Appendix A)*

Hypotheses Considered:

- An assumption is that narrow mushrooms with white gills are more likely to be classified as poisonous[2]
- Another assumption is that any mushroom with a red colour on the cap or stem is likely to be poisonous.
- The best performing algorithm would be decision trees as decision trees work great on categorical values and has a natural "if..then..else" way of making decisions[3].
- Also, the dataset being full of categorical values, decision trees algorithm would perform well on label-encoded data rather than one-hot encoded data because of "Curse of Dimensionality"
- Since the data contains facts it's expected that any model would perform well, like any model would on the well-known IRIS dataset.

ML/DM Techniques and Performance Metrics: Furthermore, to the techniques on IBM dataset,

- To visualize the number of poisonous and edible mushrooms of specific feature using bar plots.

2.1.3 Car Price Prediction

Data Understanding: The target variable is "price" having price values of cars which is continuous. A regression model can be built to predict the price of any car based on the 25 input features.

Problem Statement: The dataset has a target variable that is continuous. Therefore, this belongs to the regression problem type where a model (supervised machine learning) can be built on this dataset. To choose appropriate regression algorithms along with relevant performance metrics is vital to achieve the objective of the project. This helps in understanding how different algorithms perform on the same data.

Hypotheses Considered:

- Assuming cars with brand names as BMW are more likely to be expensive than Nissan.
- Cars with two doors (like sports cars) are more likely to be expensive than four or six door cars.

Field name*	**Data type**	**Description**
Car ID	Numeric – int64	Unique number to identify car
Symboling	Numeric -int64	The value of risk factor symbol associated with car price. Value can go up or down the scale.
Car name	Object	Model of the car
Fuel type	Object	Fuel type of car, either diesel, gas.
Aspiration	Object	Values for aspiration – std and turbo.
Door number	Object	Number of doors; four, two.

*Table 5 - Data Dictionary for Car Price Prediction (*Lists only few columns - the entire table is in Appendix A)*

ML/DM Techniques: Furthermore, to the techniques on IBM dataset, since the target variable is continuous,

- Instead of classification/clustering, regression models such as MLR, Lasso, Ridge, KNeighboursRegressor (KNN) are built.
- Comparing the performance of algorithms on one-hot versus label encoded data.

Performance Metrics: R-Squared, Mean-Squared Error.

2.1.4 Life Expectancy

Field name*	**Data type**	**Description**
Country	Object	The name of the country. Example: Afghanistan
Year	Numeric-int64	Year recorded for life expectancy
Status	Object	Developed or Developing
Life Expectancy	Numeric-float64	Life expectancy of the country
Adult Mortality	Numeric-float64	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant Deaths	Numeric-int64	Number of Infant Deaths per 1000 population

*Table 6 - Data Dictionary for Life Expectancy (*Lists only few columns - the entire table is in Appendix A)*

Data Understanding: The target variable "LifeExpectancy" is continuous which gives the life expectancy of a human from specific country over the years. A regression model can be built to predict the life expectancy based on the 21 input features.

Problem Statement: In addition, to the problem statement for Car Price dataset, to compare the performance of the same algorithms on different datasets – Car Price, Life Expectancy.

Hypotheses Considered:

- A country's development status could have an impact on life expectancy. As people are more likely to have access to good health care in more developed countries than in a developing country.
- People's lifestyle and health conditions could also have an impact on life expectancy. For example, the life expectancy of someone who consumes a lot of alcohol and who has a health condition such HIV/Aids would be lower.
- Best algorithm for both Car Price and Life Expectancy would be Multiple Linear Regression (MLR). As the dataset contains many independent variables, MLR would perform better than more complex algorithms which tend to overfit many features.

ML/DM Techniques and Performance Metrics: Furthermore, to the techniques on Car Price dataset:

- Missing values to be handled either using imputation or interpolation.

2.2 Pipeline Design of the Application

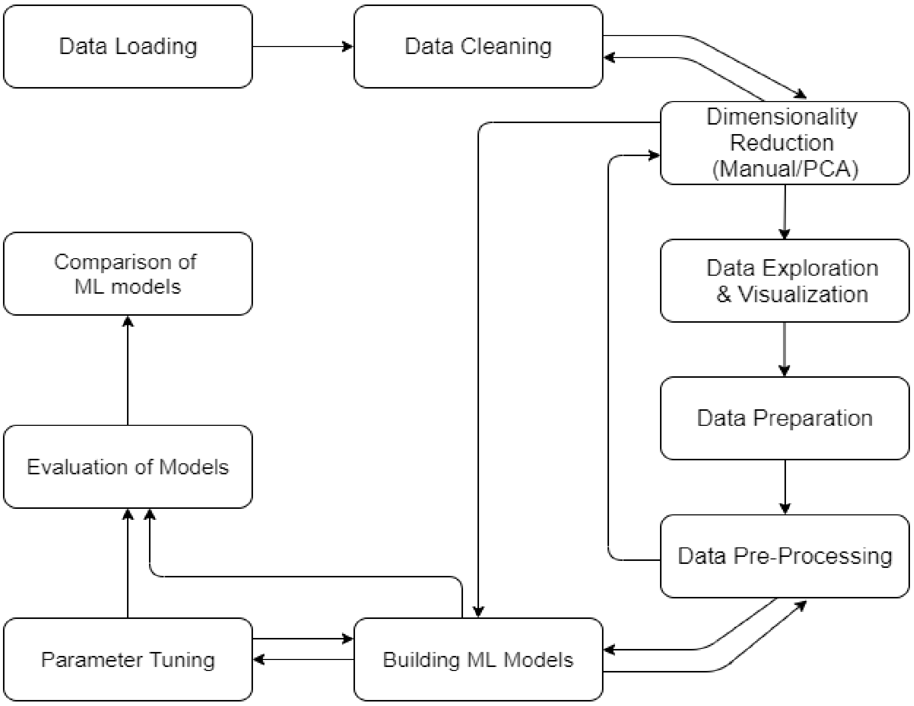


Figure 1 – Application Pipeline Design

The ML/DM python pipeline (Figure 1) follows the CRISP-DM methodology to prepare the data for analysis and build effective models along with evaluation and comparisons.

2.3 Distribution of Workload

Implementation of at least one ML algorithm per member, therefore, to implement 6 Algorithms from different ML approaches. Data preparation and Understanding on four datasets is divided as 3 members working on 2 datasets. Writing report and implementation of further algorithms to be done as a team.

3. Data Preparation

3.1 Data Preparation and Exploration Approach

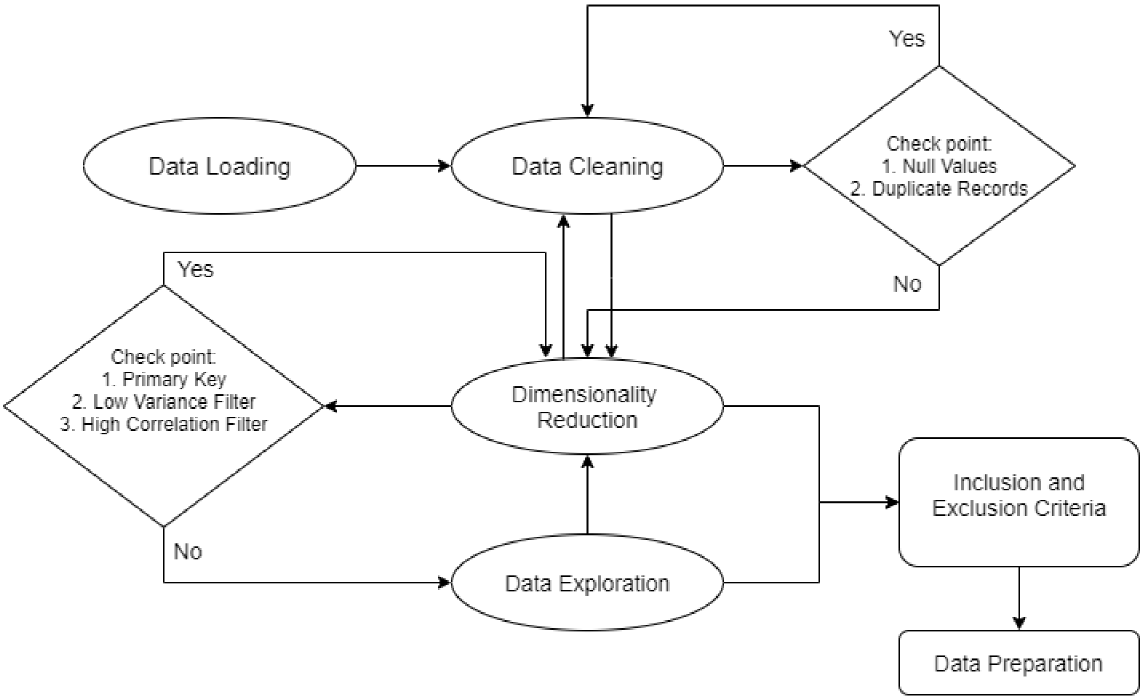


Figure 2 – Data Preparation Approach

3.2 Explanation of the Approach

Figure 2 gives the approach taken for the data preparation and exploration. Each of the given steps are chosen to effectively prepare the data for analysis. The given approach is explained and justified as below:

Data Loading: Loaded the dataset into python using pandas library – as pandas is one of the best tools for exploratory data analysis[4].

Data Cleaning: This step has various checkpoints and it is connected to the other steps as Dimensionality Reduction and Data Exploration to effectively clean the data and determine the inclusion and exclusion criteria. For example, on the dataset Life Expectancy, the following steps were taken:

- Renaming column names as they contain trailing spaces
- Checking for missing values and handling those missing values
- Where there are few missing values, those values were replaced with mean values of the column
- Where there are columns that have more than 15% of its data missing, interpolation technique is used. In this technique a function is created based on fixed data points and missing values can be evaluated using this function which interpolates between known data values[5]. And the remaining values that can't be interpolated are replaced with the mean values
- Checking for duplicate records as they won't contribute to any of the learning

Dimensionality Reduction: This step helps in determining which features to include and exclude for the modelling. Following steps were considered for the exclusion criteria:

- To check and drop columns having unique values (Primary Key columns – e.g. Employee Number). As they don't contribute to any learning by the machine.
- To check and drop columns having only one value in more than 90% of the records ("Low Variance Filter"). The feature taking just one value on almost all its records will not contribute much to learn about the feature.
- High Correlation Filter - to drop one of the columns that have high correlation with another. Therefore, a correlation matrix is plotted to find the highly correlated factors and one of the factors is dropped. The factor to drop is determined by the number of values it has. Factors with many input values are dropped as the model generalizes well with features taking fewer values.

Data Exploration: This step is connected to the previous steps as we explore the data to find inclusion and exclusion criteria. Furthermore, this helps in determining outliers and understanding more about the values a feature can take and the kind of statistical distribution it takes. Outliers are found in the datasets and decided to not remove as they seemed to be actual values the feature can take.

Data Preparation: Finally, the cleaned data is prepared and exported as a new csv file which can be loaded again for specific modelling approach.

3.3 Plan for Measuring Data Quality

Table 7 gives an example on how the data quality is measured throughout the data preparation and exploration approach.

Data	Checks	Before Cleaning	After Cleaning
IBM Attrition	Columns with same value for more 97% of records	‘EmployeeCount’, ‘Over18’, ‘StandardHours’ are all columns that only have one value for 1470 entries	Removed the column
	Checking for columns with unique entries	‘Employee ID’	Removed the column
Mushroom Classification	Missing Values	2480 in ‘stalk-root’ column	Removed the column as the model performed well without the column
	Irrelevant columns	Veil-type only contained one type of veil - no influence in the overall prediction	Removed ‘veil-type’
	Columns with same value for more than 90% of records	‘Gill-attachment’ - 97.4% ‘Veil-color’ - 97.5% ‘Ring-number’ - 92.2% ‘Veil-type’ - 100%	Removed ‘veil-type’
Life Expectancy	Missing Values	There are missing values, 14 columns out of the 22 had missing values.	Filling null values with their mean for few columns Filling null values using interpolation for 3 columns
Car Price Prediction	Checking for columns with unique entries	Only ‘CarID’ has unique values, does not contribute much to car price prediction	Removed the column
	Columns with same value for more 97% of records	The column "enginelocation" has the value “front” occurs more than 98%	Removed the column

Table 7 – Data Quality Check

4. Model Development and Evaluation

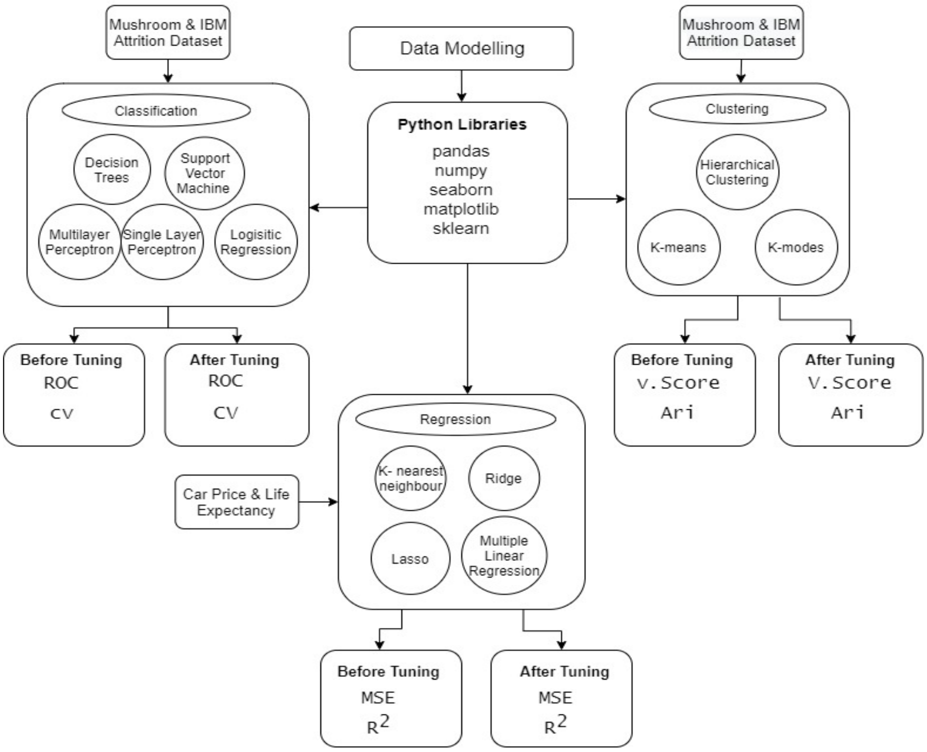


Figure 3 – Design of Modelling Tasks

4.1 Design of Modelling Tasks

Figure 3 gives the modelling tasks designed for the project. The following steps are considered:

Step 1: For each problem type, loading the cleaned data (two datasets for each problem type)

Step 2: Data Pre-Processing which includes Encoding Categorical Variables (one-hot or label) and Train-Validation-Test split of the data for the modelling.

Step3: Writing Python functions for fitting the built models, performance metrics and tuning the parameters.

Step 4: Evaluation and Comparison of the models with default parameter values

Step 5: Optimizing the parameters manually/systematically on the models and evaluating the models based on the performance metrics

Step 6: Optimizing the parameters using automated tuning as Grid Search and evaluating the models

Step 7: Proposing the best and worst performing model on each dataset

4.2 Selection of Learning Algorithms

Classification Algorithms		
Chosen Algorithms	**Advantages**	**Disadvantages**
Decision Trees	Decision trees are ideal for decision making, presenting the data into a tree format. The model requires less effort for data pre preparation as well as pre- processing. Decision trees are used for solving both classification and regression problems. Therefore, decision trees have been applied to both IBM and Mushroom dataset.	Decision Trees suffer from Curse of Dimensionality. Therefore, one-hot encoded categorical values often perform poorly as the dimensions of the data increases, and decision finds it difficult to find the right split
Perceptron	A perceptron model is a single layer neural network, ideal for binary classification (since the datasets are classified into two classes). The perceptron network consists of several inputs, performs computations on the inputs and produces an output. Single layer perceptron work by enabling the neurons to learn and process elements in training set one at time.	Perceptron work best on linearly separable problems and if the data is not separable by a single straight line or plane it could perform poorly. However, Multi-Layer perceptron overcomes this disadvantage
Multi-Layer Perceptron	Multilayer perceptron is formed by the full connections of perceptron in the network. The output of one layer is the input of the next layer. The layers in between the output and the input are called hidden layers. Multi-layer Perceptron was chosen to train the Mushroom and IBM Employee Attrition datasets as it is a powerful model which is able to solve classification problems effectively.	With too many parameters to learn because of the dense layers it results in redundancy and takes a long time to train the network

Chosen Algorithms	Advantages	Disadvantages
Logistic Regression	Logistic Regression is a popular and widely used for classification problems. A statistical method for predicting binary variables which is more suitable for the type of data chosen. Chosen datasets have binary classification problem where the output is either a positive class or a negative class. Also, the algorithm is less prone to overfit the data when there are more records as in the chosen datasets.	Like in Perceptron, Logistic Regression performs well only in linearly separable data. Need more data to make sure the model doesn't overfit.
Support Vector Machines	Support -Vector Machines are effective in high-dimensional spaces. As the data in one-hot encoded, it is assumed that SVM could perform well on the chosen datasets. The algorithm also takes less time to train.	Need right amount of data - neither less nor more - for the model to perform well. When there is noise in data the algorithm wouldn't perform well i.e when the target classes overlap.[6]

Table 8 – Classification Algorithms

Clustering Algorithms		
Algorithms	**Advantages**	**Disadvantages**
Hierarchical Clustering	Hierarchical clustering is an agglomerative algorithm which is easy to understand and use and clusters samples together iteratively with the nearest clusters, until there's no more clusters to merge. It uses distance, mostly Euclidean distance, to find the nearest (most similar) cluster to merge and the point to measure the distance is decided using linkage. There is multiple criterion: single-linkage where it computes between the two most similar parts of a cluster, complete-linkage the two least similar clusters, and mean or average-linkage from the centre of the clusters. In this project, different linkages were evaluated to find the most optimal criterion for each dataset. Hierarchical clustering was used on IBM and mushroom datasets.	The data needs to be cleaned without any missing values for the model to train. Performs poorly with data having various data types
K-means	K-means algorithm is a clustering method that partitions dataset into k-subgroups according to their shared characteristics. Unlike hierarchical clustering, it assigns the samples to a cluster such that the sum of the squared distance between each sample and the cluster's centroid is minimised. Since the algorithm uses distance as a measure of goodness of the model, K-means works better on numerical dataset. Also, once categorical data are transformed into numerical values, k-means tends to cluster around the categorical variables (as they tend to have high variance in a normalised dataset). Therefore, K-means were used on IBM dataset. It was also applied on mushroom dataset to see how well k-means still perform on the categorical dataset and to compare with the result from IBM dataset.	Need to find the optimal k-value. Suffers from Curse of Dimensionality - hence needs Dimensionality Reduction using PCA

Clustering Algorithms		
Algorithms	**Advantages**	**Disadvantages**
K-modes	K-modes is an alternative method of k-means when a dataset is composed only of categorical variables. Where k-means calculate the distance between two points, k-modes aims to minimise the dissimilarity between two clusters using Hamming distance (it is a metric used in information theory to calculate the distance between two binary data strings).	Because it uses different measure, it does not do well when the dataset is mixed or if a given category is particularly prevalent, as the algorithm will not be able to incorporate such information. Moreover, k-modes has to be installed separately as it is not supported by scikit learn.

Table 9 – Clustering Algorithms

Regression Algorithms		
Algorithms	**Advantages**	**Disadvantages**
Nearest Neighbour Regression	Since the output values in the datasets (Car Price Prediction and Life Expectancy) are continuous and not discrete, regression based on the KNN Algorithm is chosen to model for the datasets. The advantage of KNN for regression problems is that based on feature similarity it predicts the output for the given input values based on the nearest neighbours. If there are more than two nearest neighbours found, then it takes the average of these neighbour values. Thus, any KNN model predicts an optimum output value.	It takes a long time to train with more data. It does not learn anything from train data therefore, the algorithm wouldn't generalize well with unseen data
Ridge	It's another algorithm that generalizes well with unseen data as it adds enough bias while modelling which helps in getting approximate output values	Includes all the features in training - could lead to problem with generalizing on unseen data
Lasso	Lasso is also a regression model and as the output variables are continuous for the chosen datasets, the algorithm is used. It is mainly chosen for comparing one regression model with other. Also, it has a regularization term within its function and hence generalizes well compared to other regression models thereby no overfitting the train data	Compresses the features especially the highly correlated ones - thereby the model trains on few features rather than all. This can be both an advantage and a disadvantage
Multiple Linear Regression	Multiple Linear Regression (MLR) is a supervised learning technique. The algorithm is chosen for its efficiency and very simple to understand calculations. When the data is linear the algorithm performs well with less errors in predictions	It performs well only when the input features are linear with the output feature. And the algorithm is sentinel to outliers where both the chosen datasets have outliers (though valuable data) within.

Table 10- Regression Algorithms

4.3 Selection of Algorithm Parameters and Tuning

It is important to select the right parameters (Table 11) and tune the values to build an effective model that doesn’t overfit but generalizes well with unseen data. Tuning the parameters is done:

- Manually – trial and error to find the best value of any parameter based on the model’s performance
- Systematic – where one parameter is taken and tuned based on the performance not just on train set, but a validation set separated from the train set. The different values of algorithm are plotted with the performance score of the model both on train and validation set and finding the optimum value the parameter can take. (Neither overfitting nor underfitting)
- Automated Tuning: using the available functions - Grid Search or Random Search - where the parameters are tuned with the possible pairs and outputs the best estimator a model can take and the performance score of the model with that estimator. In grid search all possible pairs are tuned, in random search random pairs are tuned where former is more effective but the latter is efficient.

Algorithms	Parameters Tuned	Justification
Logistic Regression	max_iter, solver	Needed for the model to converge for different solver
	fit_intercept	for adding bias to avoid overfitting
Decision trees	ccp_alpha	For minimal cost-complexity pruning - to avoid overfitting
	max_depth	Maximum depth of the tree - to avoid overfitting
Peceptron	alpha	To multiply with the regularization term
	eta0	Constant by which updates are multiplied - main parameter for training
	max_iter	For the model to converge
MLP	hidden_layer_sizes	To find the optimum no of hidden layers for the chosen data - for the model to perform well
	activation	Activation function for the hidden layer
	alpha	For regularization
	max_iter	For the model to converge
SVM	C	For regularization
	Gamma	Trial and error - to check the performance of model with different values
K-Means	init	Method for initialization where the cluster centres need to be - random or kmeans++
	n_init	to run the algorithm a specific number of time - until its clustered
	n_clusters	important parameter - no of clusters to form with the data
Birch	threshold	To set the optimum distance between two different clusters
	n_clusters	important parameter - no of clusters to form with the data
K-Modes	init	Method for initialization where the cluster centres need to be
	n_clusters	important parameter - no of clusters to form with the data
HAC	linkage	Which type of linkage is optimum for the chosen data
	n_clusters	important parameter - no of clusters to form with the data
MLR	normalize	For normalization
	fit_intercept	with and without intercept
Ridge	alpha	important parameter - regularization strength - to find the optimum value for training the best model
	solver, fit_intercept	solver, with or without intercept
Lasso	alpha	important parameter - to find the optimum alpha for a best model
KNN	n_neighbours	no of neighbours to use
	algorithm, weights	Which algorithm to compute the nearest neighbours and also the weight function for prediction

Table 11 – Selection of Parameters

4.4 Evaluation of the Learned Models

Classification

The AUC (Area Under the Curve) scores inform the performance in distinguishing between classes in classification problems. The higher the AUC score, the better the model is at predicting. It can be seen in the picture below that in the default models the Support Vector Machine outperforms other models on the IBM dataset with the AUC score of 0.809 whereas the perceptron performs poorly yielding the AUC score of 0.5.

On Default Models:

Figure 4 shows that overfitting might occur in the decision tree, the multi-layer perceptron, and the support vector machine models since their AUC and accuracy scores are 1.000.

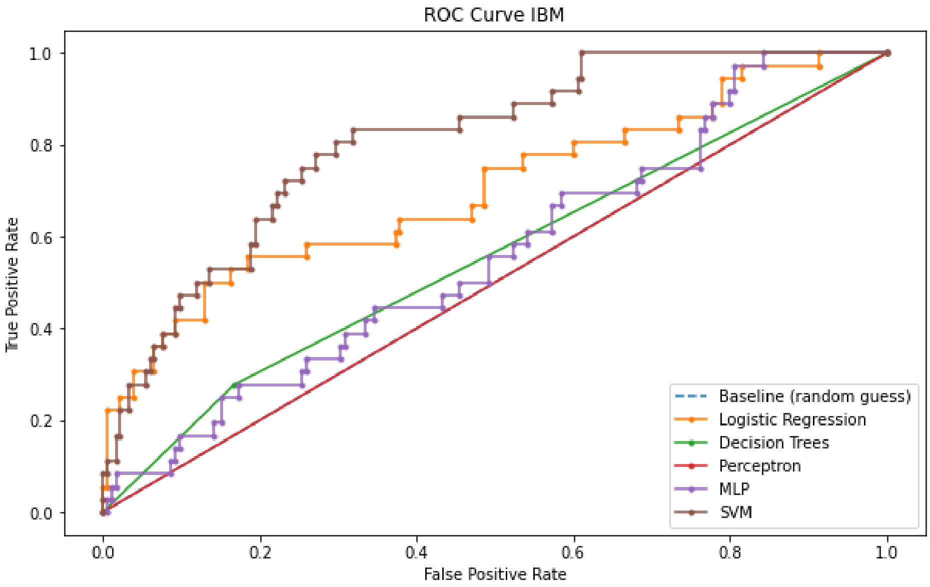


Figure 4

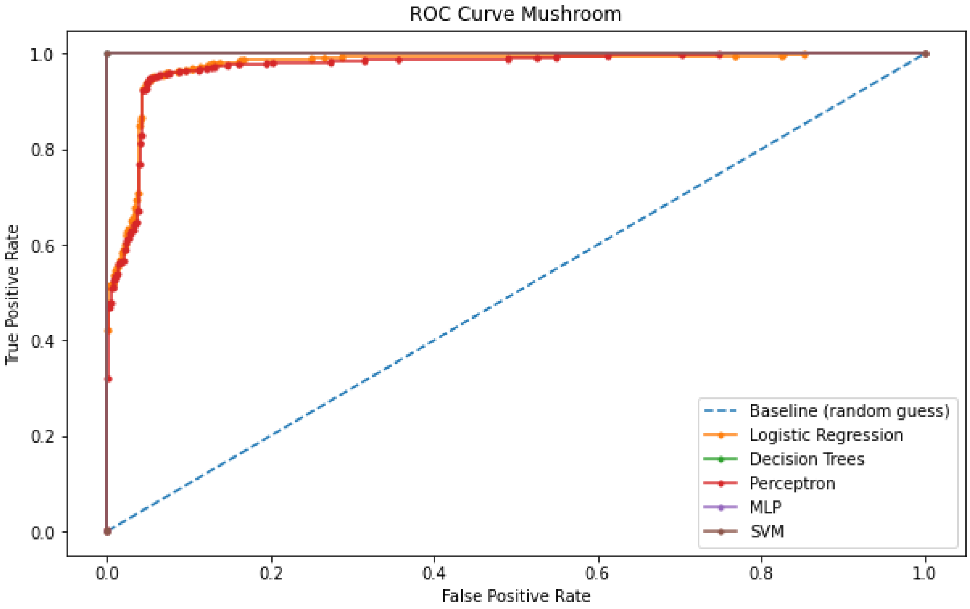


Figure 5

Overall, the prediction from the manually tuned models (Figure 6, 7) achieve almost the same performance as the default models. On IBM dataset, the logistic regression model gives the highest AUC score of 0.815 and the perceptron is still the worst classifier. On the Mushroom dataset, overfitting is suspected in the decision tree, the multi-layer perceptron, and the support vector machine models. The logistic regression and the perceptron have high levels of efficiency with AUC scores 0.975 and 0.972 respectively.

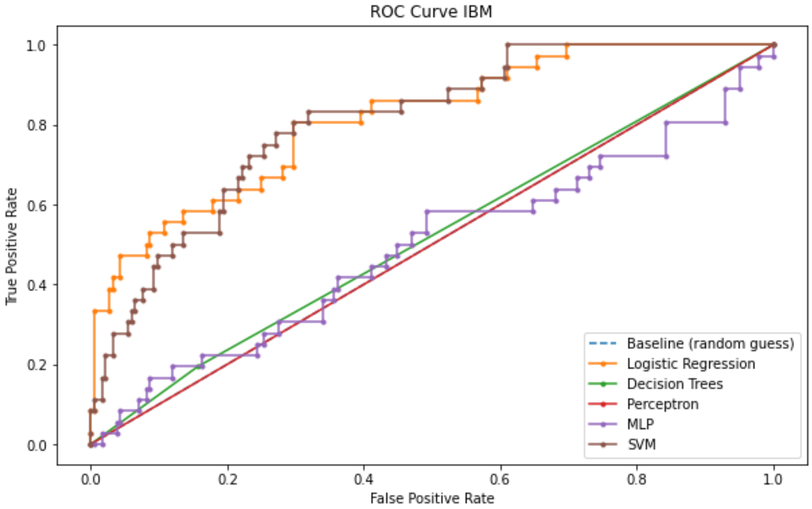


Figure 6 – Manually Tuned

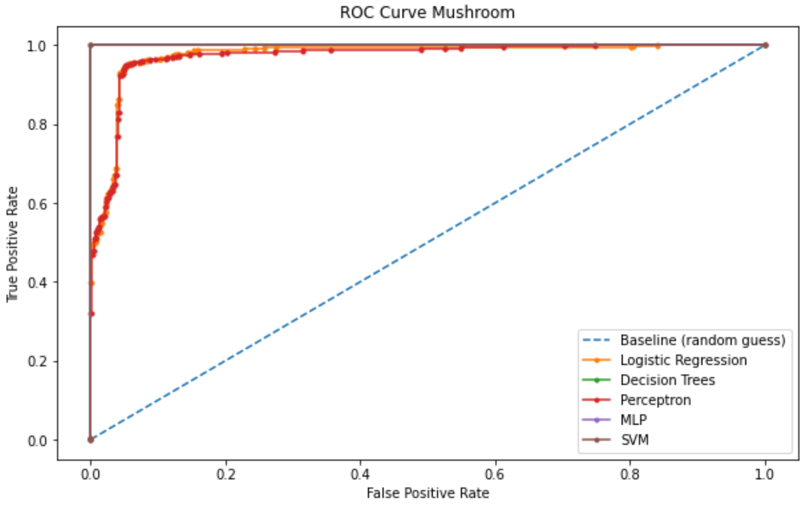


Figure 7 – Manually Tuned

When the models are tuned automatically, the performance of the perceptron is relatively poor compared to other models while the support vector machine obtains the highest AUC score of 0.809 on the IBM dataset. The obtained AUC scores for the Mushroom dataset are equal to those from manually tuned models.

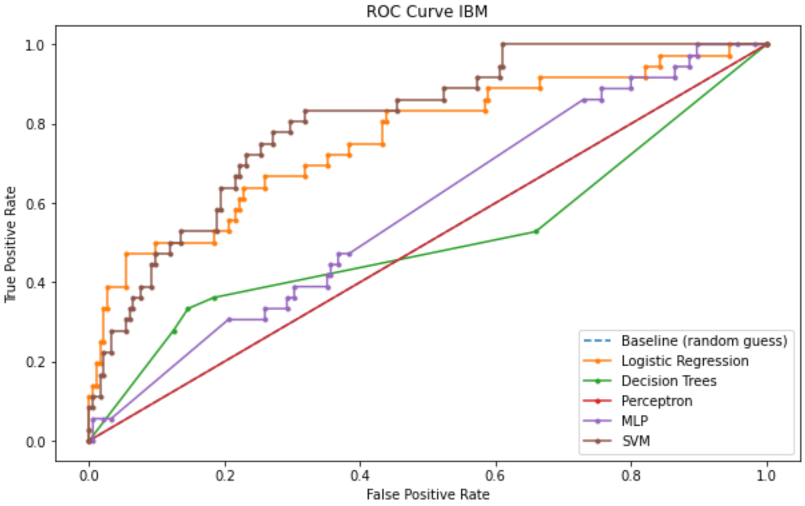


Figure 8 – Auto Tuned

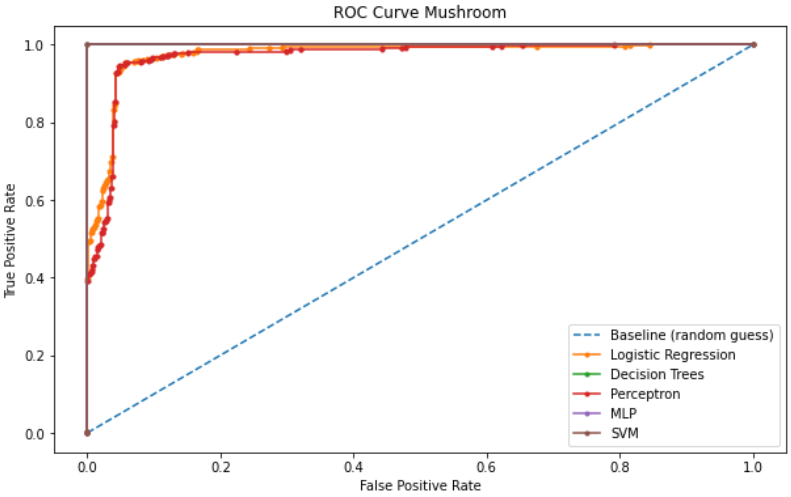


Figure 9-Auto Tuned

The cross-validation (CV) score is an average performance of the cross-validation approaches. The support vector machine achieves the highest CV score of 0.878 for the default models followed by the logistic regression, the perceptron and the decision tree model respectively on the IBM dataset. The multi-layer perceptron obtains the poorest performance. In the Mushroom dataset, the decision tree, MLP and SVM yield the CV score of 1.000 which we suspect that the models are well trained.

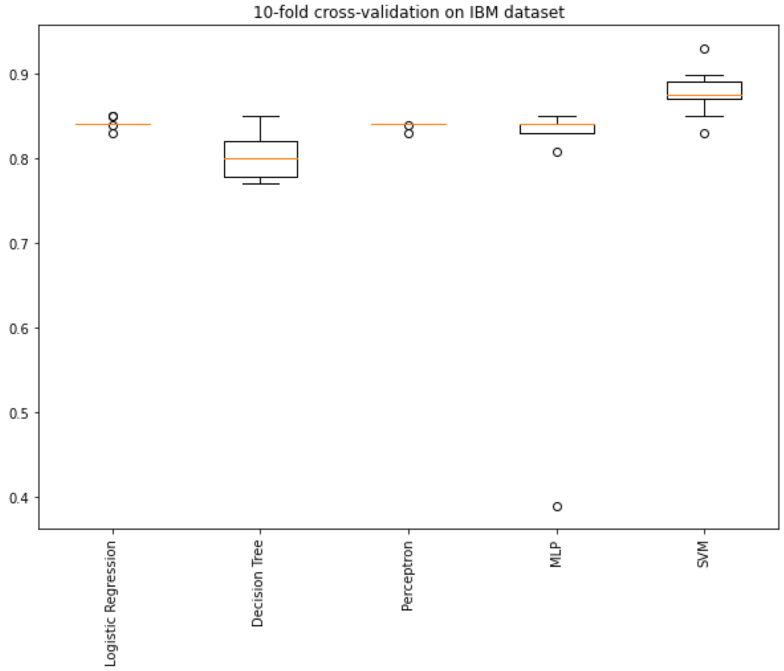


Figure 10-Default models

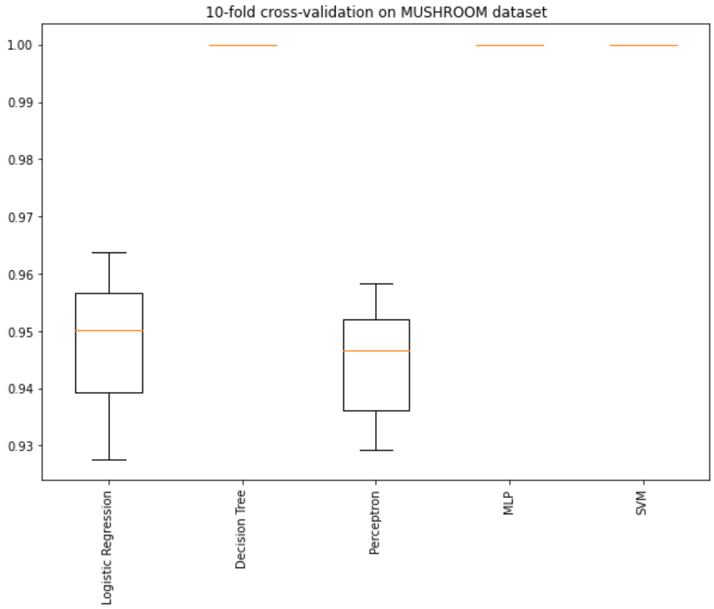


Figure 11-Default models

Logistic Regression is the most suitable model for predicting the IBM data when the models are manually tuned whereas MLP obtains the lowest CV score. In the Mushroom dataset, the decision tree, the MLP and the SVM are trained perfectly with a 1.000 CV score. The logistic regression and the perceptron have an equal performance with 0.947 CV score.

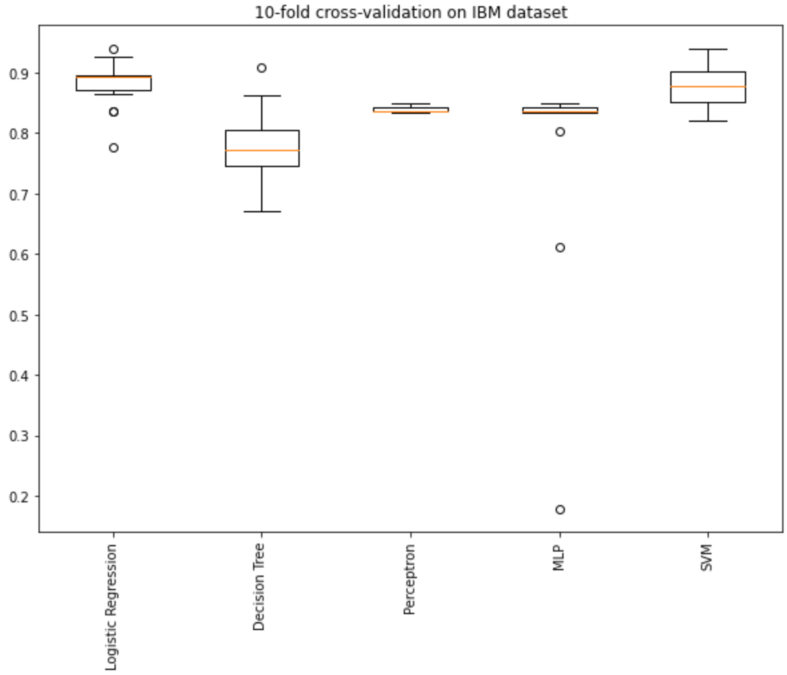


Figure 12-Manual Tuning

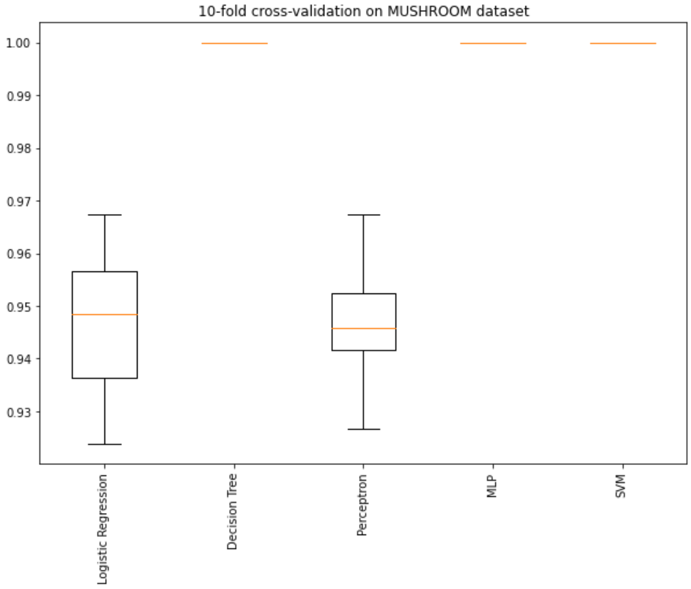


Figure 13-Manual Tuning

The support vector machine obtains the highest CV score among the automatically tuned models on the IBM dataset. The CV score of the MLP and decision tree models improve from the manually tuned models. In the Mushroom dataset, the decision tree, MLP and SVM models achieve the CV score of 1.000 followed by the logistic regression model which performs slightly better than the perceptron.

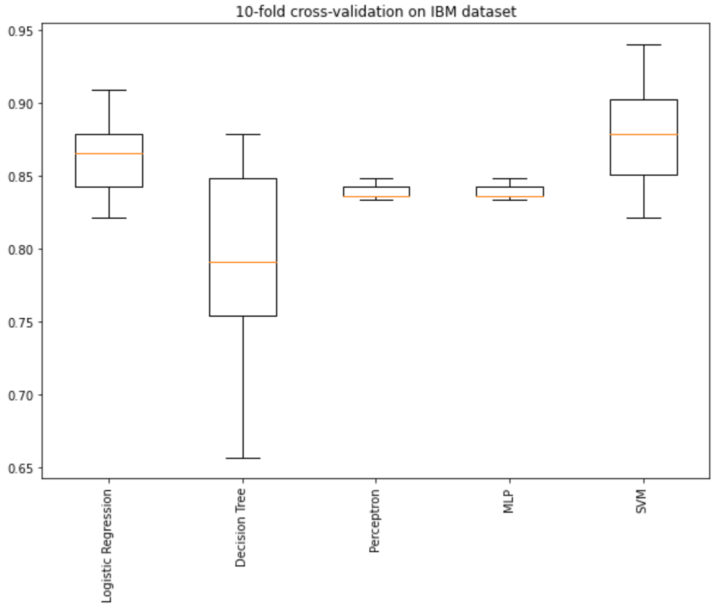


Figure 14-Auto Tuned

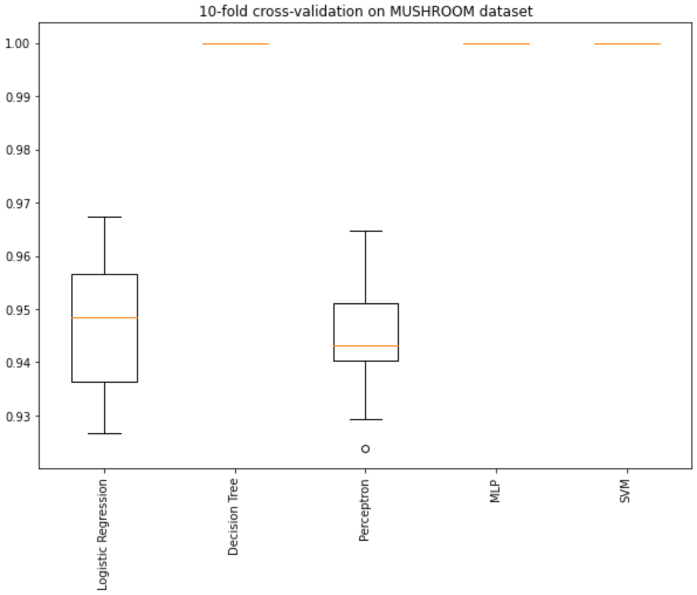


Figure 15-Auto tuned

Clustering

Evaluation: From the following table it is seen that, on HAC and Birch performed well compared with other algorithms in terms of homogeneity score and completeness both on IBM and Mushroom Data.

Clustering

On IBM				
Algorithms	Adjusted Rand Index Score	Homogeneity Score	Completeness Score	V Measure Score
K- Means	0.00038	0.00442	0.00094	0.00155
Birch	0.00118	0.00658	0.00142	0.00233
HAC	0.00118	0.00658	0.00142	0.00233
K-Modes	-0.00086	0.00444	0.04962	0.00814
On Mushroom				
K-Means	0.15587	0.43779	0.15263	0.22635
Birch	0.19296	0.62881	0.22248	0.32867
HAC	0.19674	0.6389	0.2252	0.33302
K-Modes	-0.00008	0.00083	0.06682	0.00164

Regression

Life Expectancy				
Model - After Tuning	R^2		MSE	
	Train	Test	Train	Test
Multiple-Linear Regression	0.96621	**0.94984**	**3.14768**	**4.56009**
Ridge Regression	0.95662	0.94097	4.04079	5.36732
Lasso Regression	0.681	0.66093	29.71351	30.82812
KNNRegressor	**1**	-0.16017	0	105.48202
Car Prediction				
Model - After Tuning	R^2		MSE	
	Train	Test	Train	Test
Multiple-Linear Regression	**0.98357**	**0.88048**	**1073925.98**	**3982220.33**
Ridge Regression	0.97234	0.82177	1807889.41	5938436.95
Lasso Regression	0.97262	0.76398	1789375.33	7826749.55
KNNRegressor	**0.99726**	0.72951	179323.122	9012507.18

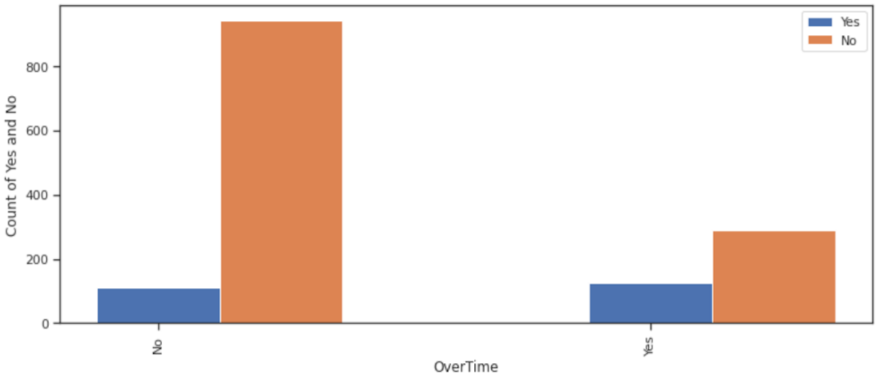
Evaluation: The models are optimized with right parameter values and the performance metrics are got. On both datasets Multiple-Linear Regression is the best performing model based on the R squared and MSE scores. Although KKRegressor gives 100% accuracy on train set, it clearly shows the models didn't learn and generalize well with unseen data like it was assumed earlier.

5. Results

5.1 Accept or Reject Hypothesis

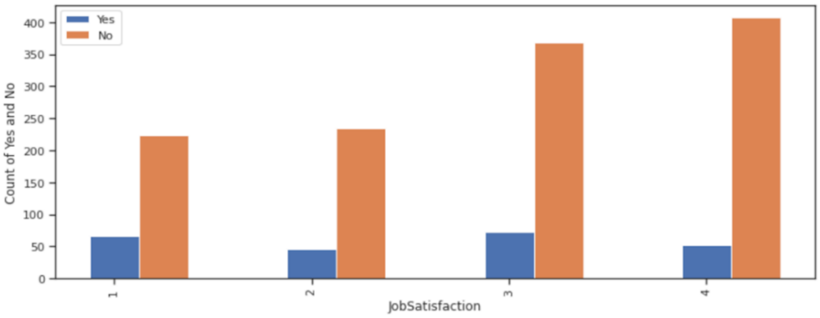
IBM dataset

Hypothesis 1: Employees who work longer hours are likely to leave the company, perhaps due to stress of workload.



Overall, the graph displays that employees who didn't work time stayed within the company. Also, the graph shows from the 400+ employees who did work overtime, around 100 of those employees had left the company. This could be due to heavy workload and unbearable stress. Around 900 employees who had stayed within the company did not work overtime. In this case, we can accept the hypothesis that employees who worked longer hours are more likely to leave the company.

Hypothesis 2: Employees leaving the company would be directly proportional to their job satisfaction, where lower the job satisfaction employees are more likely to leave than employees with higher job satisfaction.



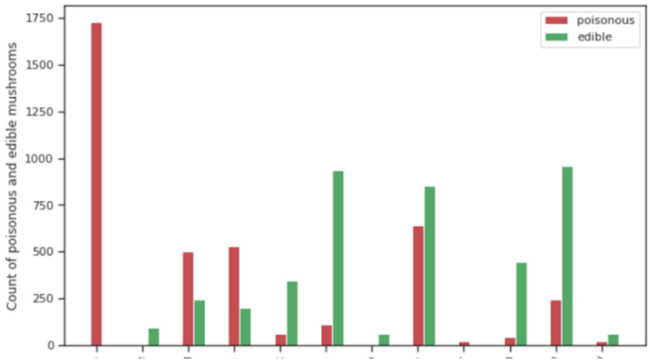
The graph above clearly illustrates that those who most satisfied with the company and gave a satisfaction score of 3 or 4 , stayed within the company. However, employees who gave a score of 3, had the most number of employees leave the company. Therefore, this hypothesis has been rejected.

Hypothesis 3 : Among the classification algorithms, Logistic Regression (ideal algorithm for binary classification) is assumed to perform well, since the target data is dichotomous.

Even though, the prediction was made that Logistic Regression will be the best algorithm for this dataset, however we obtained highest accuracy from SVM across train, test, validation data.

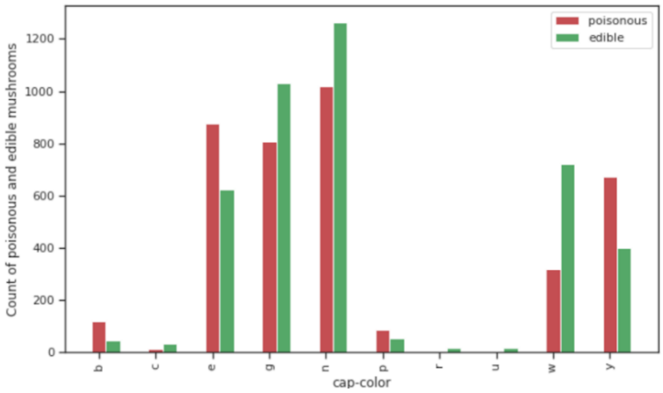
Mushroom Dataset

Hypothesis 1: An assumption is that narrow mushrooms with white gills are more likely to be classified as poisonous, obtained from research.



The graph shows that mushrooms with buff colored gills are the most poisonous. Around 1700 species of mushrooms with b- buff colour being identified as poisonous. Therefore, we can accept the hypothesis.

Hypothesis 2: Another assumption is that any mushroom with a red colour on the cap or stem is likely to be poisonous.



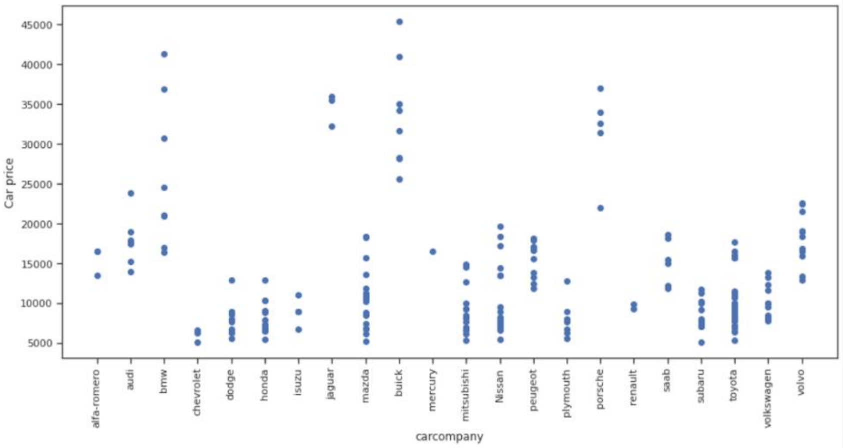
Graph shown above illustrates that mushrooms with brown (n) and red (e) cap color were identified as the most poisonous, therefore this hypothesis can be accepted.

Hypothesis 3: The best performing algorithm would be decision trees as decision trees work great on categorical values and have a natural "if..then..else" way of making decisions

Result: Our results tell us that decision trees gave a high accuracy score across train, test and validation data for default, manually tuned and automatically tuned models.

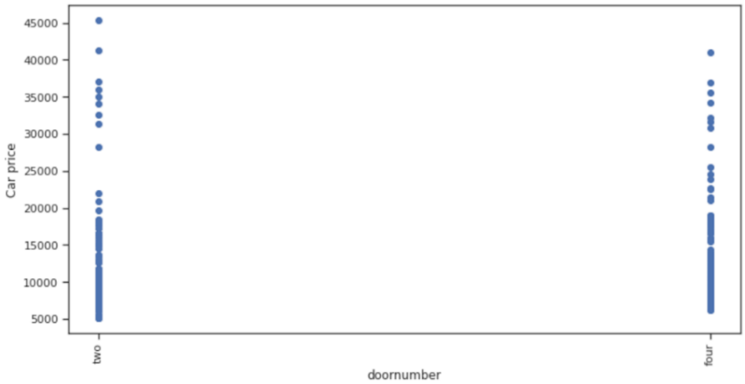
Life Expectancy dataset

Hypothesis 1: Assuming cars with brand names as BMW are more likely to be expensive than Nissan.



The scatter plot shows that both buick and BMW are the more expensive car brands. BMW has car models with a value of more than £40,000, whereas Nissan's most expensive model is only around £20,000. Therefore, we can accept the hypothesis that brand name has impact on price of car.

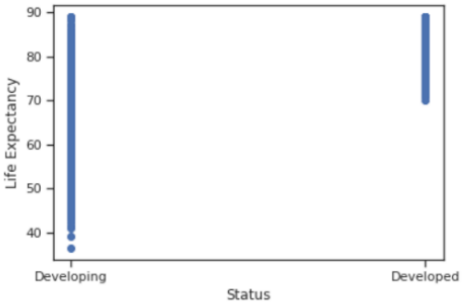
Hypothesis 2: Cars with two doors (like sports cars) are more likely to be expensive than four or six door cars.



The above plot shows that cars with two doors are more expensive in comparison to four door cars. Cars with two doors. Cars with two doors have cars with prices over 45,000, whereas cars with four doors have prices around 40,000. Even Though the values are very close each other, we can accept the hypothesis.

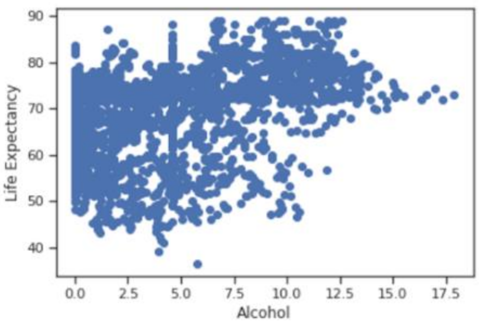
Life Expectancy

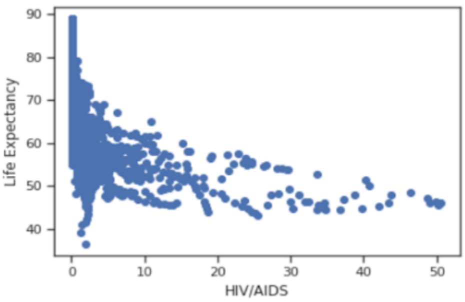
Hypothesis 1: A country's development status could have an impact on life expectancy. As people are more likely to have access to good health care in more developed countries than in a developing country.



The graph shows that in developing countries the life expectancy is very varied, from ages even below 20 . Whereas in developed countries, the life expectancy is between 68-90. Therefore the country's development status does have an impact on life expectancy and we can accept the hypothesis.

Hypothesis 3: People's lifestyle and health conditions could also have an impact on life expectancy. For example, the life expectancy of someone who consumes a lot of alcohol and who has a health condition such HIV/Aids would be lower.





First scatter plot shows the as the consumption of alcohol increases among the population (1000 people), the life expectancy slowly decreases. The second scatterplot shows that as HIV/AIDs increase across the population, the life expectancy decreases. Therefore, we can accept the hypothesis that lifestyle and health does have an impact on Life Expectancy.

Contributions

Name	Tasks Done
Anju	Ridge Regression on Life and Car Price, Parameter Tuning, Evaluation
Agalya	KNN on Life and Car price, Parameter Tuning, Evaluation
Ai	K-Modes on IBM and Mushroom, Parameter Tuning, Evaluation
Ishan	K-Means on IBM and Mushroom , Parameter Tuning, Evaluation
Sophita	HAC on IBM and Mushroom, Parameter Tuning, Evaluation
Earnie	Lasso Regression on Life and Car Price, Parameter Tuning, Evaluation
Anju, Agalya, Earnie	Data Preparation and Understanding on IBM, Life Expectancy
Sophita, Ishan, Ai	Data Preparation and Understanding on Mushroom, Car Price prediction
Everyone	Choosing Appropriate Algorithms and Performance Metrics for Classification
Ishan, Ai, Sophita	Choosing Appropriate Algorithms and Performance Metrics for Clustering
Earnie, Agalya, Anju	Choosing Appropriate Algorithms and Performance Metrics for Regression
Everyone	Classification Algorithms, Linear Regression, Birch, Parameter Tuning, Evaluation
Everyone	Report Documentation
Equal contribution was given by everyone in the team to the coursework and it was an incredible experience working as a team with diverse talents.	
E - Signatures :	
Aju, Anju	Anju
Ayyadurai, Agalya	M. Agalya
Omote, Ai	
Patil, Ishan	Patil

Ploysuayngam, Sophita	โสภิตา พลอยสวยงาม
Somchit, Pariyakorn	ปริยากร สมจิตต์

References

[1] AI, D. and Learning, M., 2020. Introduction To Artificial Intelligence And Machine Learning Tutorial | Simplilearn. [online] Simplilearn.com. Available at: <https://www.simplilearn.com/machine-learning-tutorial> [Accessed 6 June 2020].

[2] Sciencing. 2020. How To Identify Poisonous Mushrooms. [online] Available at: <https://sciencing.com/identify-poisonous-mushrooms-2057768.html> [Accessed 8 June 2020]

[3] Medium. 2020. *Decision Trees—A Simple Way To Visualize A Decision*. [online] Available at: <https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb#:~:text=A%20decision%20tree%20is%20a%20flowchart%2Dlike%20structure%20in%20which,taken%20after%20computing%20all%20attributes).> [Accessed 16 June 2020].

[4] Medium. 2020. *Are You Still Using Pandas For Big Data?*. [online] Available at: <https://towardsdatascience.com/are-you-still-using-pandas-for-big-data-12788018ba1a> [Accessed 16 June 2020].

[5] Docs.scipy.org. 2020. *Interpolation (Scipy.Interpolate) — Scipy V1.4.1 Reference Guide*. [online] Available at: <https://docs.scipy.org/doc/scipy/reference/tutorial/interpolate.html> [Accessed 16 June 2020].

[6] Medium. 2020. *Top 4 Advantages And Disadvantages Of Support Vector Machine Or SVM*. [online] Available at: <https://medium.com/@dhiraj8899/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107#:~:text=Disadvantages%3A,the%20SVM%20will%20under%20perform.> [Accessed 16 June 2020].

[7] Kaggle.com. 2020. *IBM HR Analytics Employee Attrition & Performance*. [online] Available at: <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset> [Accessed 4 June 2020].

[8] Kaggle.com. 2020. *Mushroom Classification*. [online] Available at: <https://www.kaggle.com/uciml/mushroom-classification> [Accessed 4 June 2020].

[9] Kaggle.com. 2020. *Car Price Prediction (Linear Regression - RFE)*. [online] Available at: <https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe> [Accessed 4 June 2020].

[10]Kaggle.com. 2020. *Life Expectancy (WHO)*. [online] Available at: <https://www.kaggle.com/kumarajarshi/life-expectancy-who> [Accessed 4 June 2020].

Appendix

1. IBM Data Dictionary (entire table)

Field name	**Data type**	**Description**
Age	Numeric	Employee age
Attrition	Symbolic	If the employee decided to stay or leave the company
Business Travel	Symbolic	How often the employee goes for business trips
Daily Rate	Numeric	Their daily income
Department	Symbolic	The department which the employee works in
Distance from Home	Numeric	How far the employee is from their place of stay during work times
Education	Numeric	1- Below College, 2- College, 3- Bachelor, 4- Master, 5 - Doctor

Education Field	Symbolic	Their education background, such as ‘Life Sciences’ or ‘Medical’.
Employee Count	Numeric	Specifies that a record relates to only employee
Employee Number	Numeric	Employee ID number
Environment Satisfaction	Numeric	A scale of 1-4 on how satisfied they are with the environment they work in. 1- Low, 2- Medium, 3 - High and 4 - Very High
Gender	Symbolic	1- Female. 2- Male
Hourly Rate	Numeric	Salary for the employee on an hourly basis
Job Involvement	Numeric	Involvement of employees within their job. 1- Low, 2 - Medium , 3 - High and 4 - Very High.
Job Level	Numeric	The level of their job is represented in numeric value between 1-4.
Job Role	Numeric	The role of the employee. 1=Hc Rep, 2=HR, 3=Lab Technician, 4=Manager, 5= Managing Director, 6= Research Director, 7= Research Scientist, 8=Sales Executive, 9= Sales Representative
Job Satisfaction	Numeric	Numeric value relating to the employee job satisfaction. 1- ‘Low’, 2- ‘Medium’, 3 -’High’ and 4 - ‘ Very High’.
Marital Status	Symbolic	1- Divorced, 2- Married, 3 - Single
Monthly Income	Numeric	Monthly salary for each employee
Monthly Rate	Numeric	Monthly rate
NumCompaniesWorked	Numeric	Numeric value of all the number of companies the employee has worked at.
Over 18	Symbolic	1- yes, 2 - no
Over Time	Symbolic	If the employee works overtime. ‘ Yes’ or ‘No’
PercentSalaryHike	Numeric	Percentage increase in Salary for the employee.
PerformanceRating	Numeric	1 - ‘low’, 2- ‘Good’ , 3- ‘Excellent’ 4 - ‘Outstanding’
RelationshipSatisfaction	Numeric	1 - ‘low’ 2- ‘medium’ 3- ‘high’ 4- ‘very high’
StandardHours	Numeric	Standard working hours of employee
StockOptionLevel	Numeric	The stock option levels for employee
TotalWorkingYears	Numeric	The total number of years the employee has worked
TrainingTimeLastYear	Numeric	The number of hours the employee has dedicated for training in the past year
WorkLifeBalance	Numeric	The time the employee has spent between work and outside
YearsAtCompany	Numeric	The total number of years the employee has worked within the company
YearsInCurrentRole	Numeric	The number of years the employee has worked in the particular role.
YearsSinceLastPromotion	Numeric	Last promotion of the employee

YearsWithCurrManager	Numeric	Years spent with current manager

2. Mushroom Classification Data Dictionary (entire table)

Field name	**Data type**	**Description**
Class	Symbolic	‘ poisonous or ‘edible’
Cap Shape	Symbolic	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
Cap Surface	Symbolic	fibrous=f, grooves=g, scaly=y, smooth=s
Cap Colour	Symbolic	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
Bruises	Symbolic	bruises=t,no=f
Odor	Symbolic	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
Gill Attachment	Symbolic	attached=a, descending=d, free=f, notched=n
Gill Spacing	Symbolic	close=c, crowded=w, distant=d
Gill Size	Symbolic	broad=b, narrow=n
Gill Colour	Symbolic	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r , orange=o, pink=p, purple=u, red=e, white=w, yellow=y
Stalk Shape	Symbolic	enlarging=e, tapering=t
Stalk root	Symbolic	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
Stalk surface above ring	Symbolic	fibrous=f, scaly=y, silky=k, smooth=s
Stalk surface below ring	Symbolic	fibrous=f, scaly=y, silky=k, smooth=s
Veil type	Symbolic	brown=n, buff=b ,cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
Veil color	Symbolic	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
Ring number	Symbolic	none=n, one=o, two=t
Ring type	Symbolic	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
Spore print color	Symbolic	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o, purple=u,white=w,yellow=y
Population	Symbolic	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
Habitat	Symbolic	g,leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

3. Life Expectancy Data Dictionary (entire table)

Field name	**Data type**	**Description**
Country	Symbolic	The name of the country. Example: Afghanistan
Year	Numeric	Year recorded for life expectancy
Status	Categorical	Developed or Developing
Life Expectancy	Numeric	Life expectancy of the country
Adult Mortality	Numeric	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant Deaths	Numeric	Number of Infant Deaths per 1000 population
Alcohol	Numeric	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
Percentage expenditure	Numeric	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis B	Numeric	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	Numeric	Measles - number of reported cases per 1000 population
BMI	Numeric	Average Body Mass Index of entire population
Under five deaths	Numeric	Number of under-five deaths per 1000 population
Polio	Numeric	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total expenditure	Numeric	General government expenditure on health as a percentage of total government expenditure (%)
diphtheria	Numeric	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
Hiv/Aids	Numeric	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Numeric	Gross Domestic Product per capita (in USD)
Population	Numeric	Population of the country
Thinness 10-19 yrs	Numeric	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness 5-9yrs	Numeric	Prevalence of thinness among children for Age 5 to 9(%)
Income composition	Numeric	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Numeric	Number of years of Schooling(years)

4. Car Price Prediction Data Dictionary (entire table)

Field name	Data type	Description
Car ID	Numeric	Unique number to identify car
symboling	Numeric	The value of risk factor symbol associated with car price. Value can go up or down the scale.
Car name	Categorical	Model of the car
Fuel type	Categorical	Fuel type of car, either diesel, gas.
Aspiration	Categorical	std, turbo.
Door number	Categorical	Number of doors; four, two.
Car body	Categorical	Car body style. Eg: hardtop, wagon, sedan, hatchback, convertible.
Drive wheel	Categorical	4wd, fwd, rwd
Engine location	Categorical	front, rear.
Wheel base	Numeric	continuous from 86.6 120.9.
Car length	Numeric	continuous from 141.1 to 208.1.
Car width	Numeric	continuous from 60.3 to 72.3.
Car height	Numeric	continuous from 47.8 to 59.8.
Curb weight	Numeric	continuous from 1488 to 4066.
Engine type	Categorical	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
Cylinder number	Categorical	eight, five, four, six, three, twelve, two.
Engine size	Numeric	continuous from 61 to 326.
Fuel system	Categorical	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
Bore ratio	Numeric	continuous from 2.54 to 3.94.
Stroke	Numeric	continuous from 2.07 to 4.17
Compression ratio	Numeric	continuous from 7 to 23.
Horse power	Numeric	continuous from 48 to 288.
Peakrpm	Numeric	continuous from 4150 to 6600.
Citympg	Numeric	continuous from 13 to 49.
highwaympg	Numeric	continuous from 13 to 49.
Price	Numeric	Price value of the car