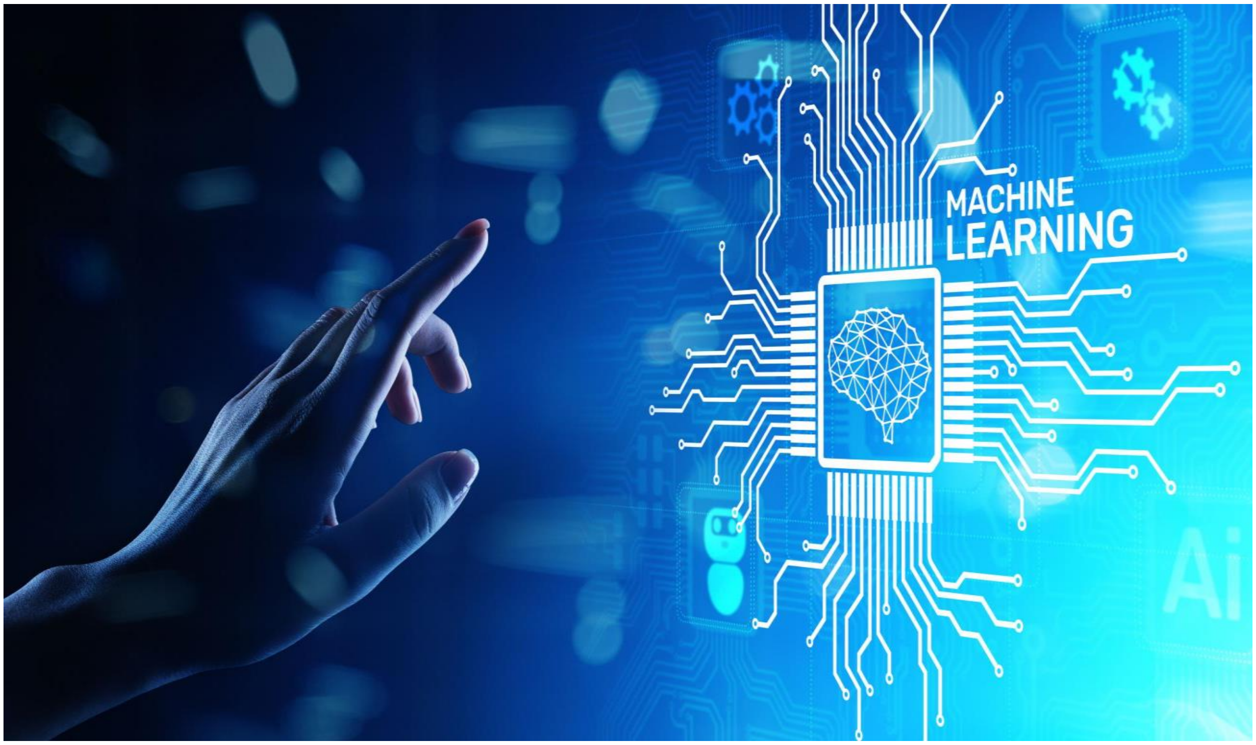
**Module:** Machine Learning (ST3189)

**UoL Student Number:** 200644307

**Page Count: 10 pages (**Excluding Cover Page, Table of Contents and Bibliography)



## Table of Contents

# Introduction

The project consists of three tasks to be completed on different datasets using supervised and unsupervised learning techniques to evaluate and present the data.

Task 1 required the use of unsupervised learning techniques to analyze a dataset. The technique used for this data was K-Means Clustering.

Task 2 involved the use of supervised learning techniques in the form of regression. Several regression models were tested, and the best fitting models were then further explored.

Task 3 also required the use of supervised learning, and called for a dataset suited towards a classification analysis, where the best fitting models were similarly analysed.

# Task 1 – Unsupervised Learning

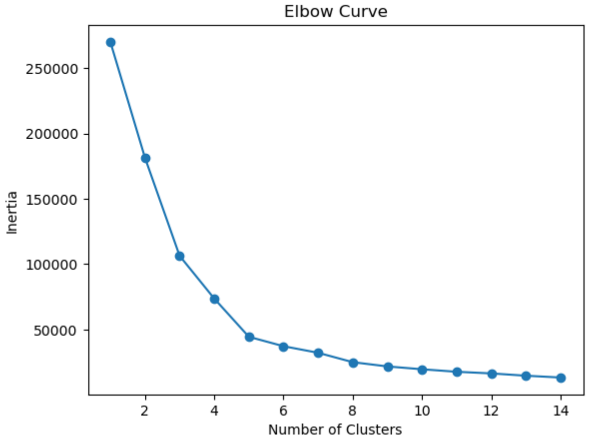
Unsupervised learning involves the use of "machine learning algorithms to analyze and cluster unlabeled datasets" by identifying patterns that can be used to effectively group data (Delua, 2021).

The dataset selected for the first task was a mall customer segmentation dataset, consisting of five variables, including annual income and spending score (based on customer purchasing data). The **K-Means Clustering** algorithm was used to determine the most effective way to segment the customer base based on their income and purchasing behavior, in order to improve marketing strategies to the respective customer segments in the future.

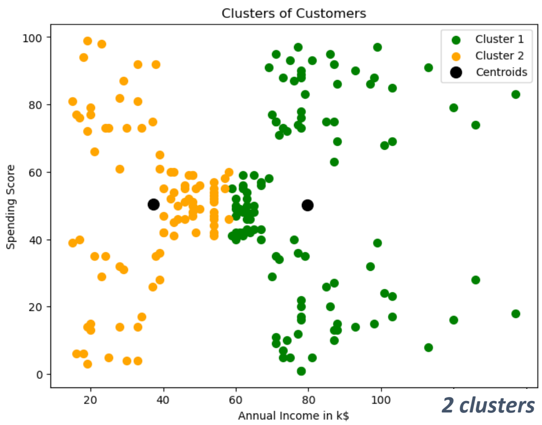
In order to carry out the K-Means algorithm, the optimal number of clusters 'k' is selected. K points are then designated as centroids at random points of the data, and each data point is then assigned to a centroid based on its proximity. The centroids are then recalculated to be the center of all observations in that area (cluster).

The optimal number of clusters 'k' were determined through the 'Elbow Method', which is where the value of k is chosen based on the point of the curve at which the inertia (within-cluster sum of squares) stops decreasing at a high rate (resembles the shape of an elbow).

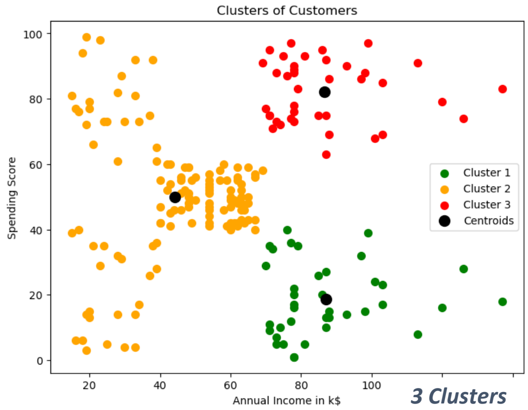
As seen by the Elbow Curve, the inertia starts decreasing linearly at about 5 clusters, thus this was taken as the optimal k value.



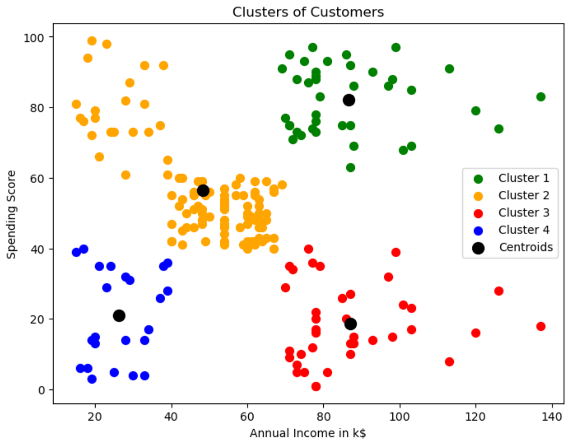
The clusters were then visualized, comparing the results with 2,3,4 and 5 clusters.



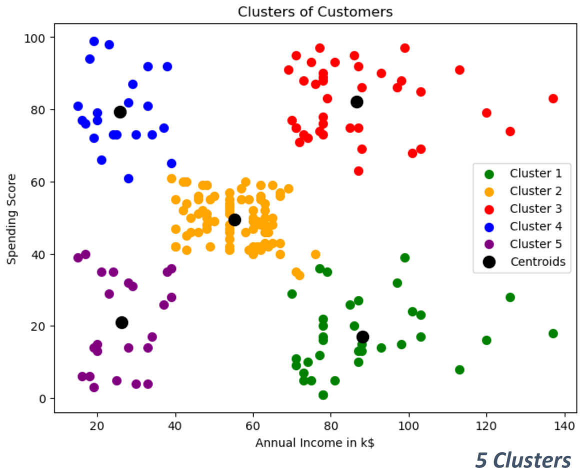
*2 clusters*



*3 Clusters*



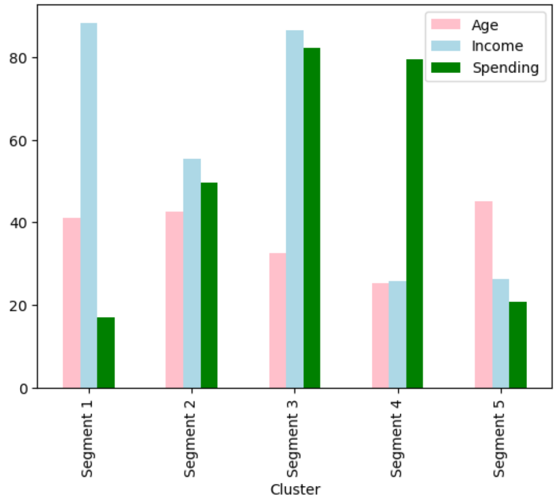
*4 Clusters*

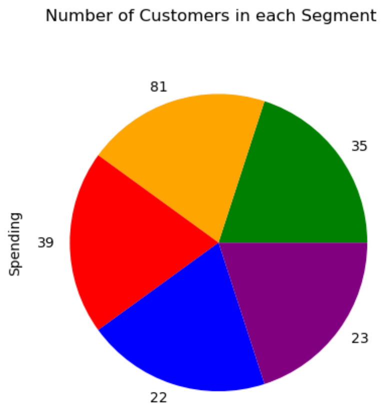


*5 Clusters*

As seen, the segmentation with 5 clusters seems to be the most successful, therefore it is confirmed as the optimal number.

**How can the shopping mall advertise to each of the customer segments to improve future sales?**





It is observed that customers in Cluster 1 have a high annual income but a low spending score, therefore these customers should be notified more about promotions and sales, as they have a high purchasing power.

Customers in Cluster 2 have both moderate income and spending levels, and also represent the largest segment with 81 customers, so these customers should be sent frequent promotion emails to keep them engaged.

Cluster 3 represents customers that have a high income as well as spending score, so this segment represents the most significant group of customers for the mall, and this should be the target group. These customers should be alerted of promotions and new releases frequently.

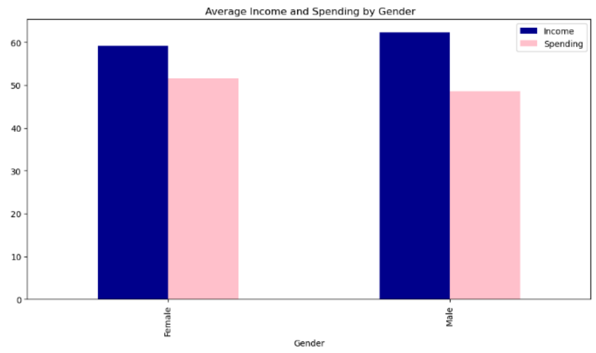
Cluster 4 consists of customers who have a high spending score despite a relatively low annual income, as well as the lowest average age. These customers could be maintained by providing them with more discount codes and coupons.

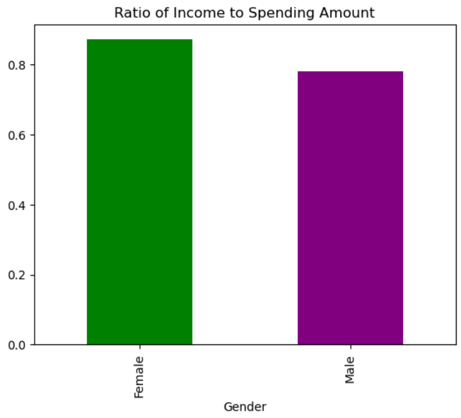
The customers in Cluster 5 have both a low annual income and spending score and this cluster only consists of 23 customers (with a higher average age). Therefore, this segment should not be a big focus for the mall, as there is low potential for improvement in sales.

Similar solutions were constructed by (Parsons, 2003), who stated that “price-based promotions are the prime attractors for increasing spending at a shopping mall”.

### Do women spend more at shopping malls than men?

It can be seen that women do spend more than men relative to their respective incomes, and despite males having a slightly higher income, females still have a higher level of spending.





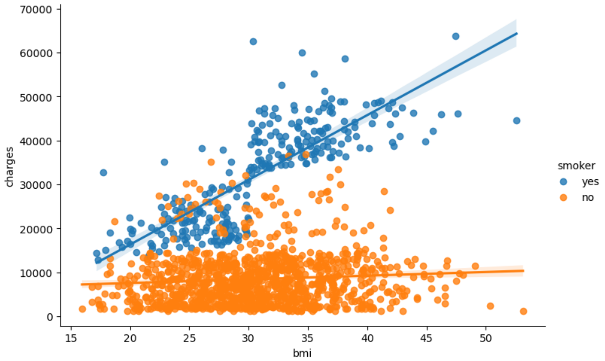
However, in a 2009 study, (Kuruvilla et al., 2009) states that typically “men are found to spend less time shopping than women, but tend to spend more money than women when they do go shopping”, which contradicts the results established from this research.

## Task 2 – Regression

Supervised learning is an approach to machine learning that involves the use of datasets that are labelled, which are then trained to predict certain outcomes, as stated by (Delua, 2021).

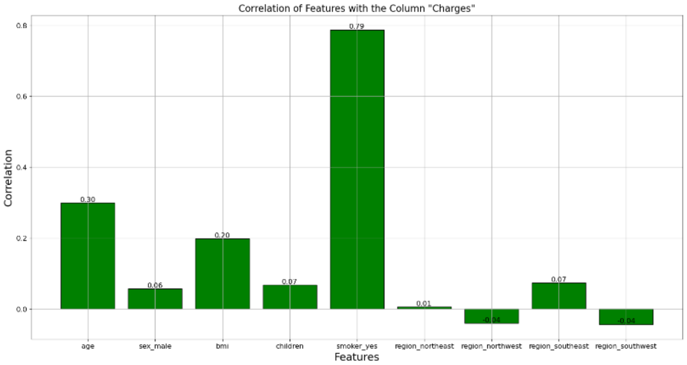
Regression is a supervised learning technique where the relationship between a numerical dependent (target) variable and independent (predictor) variables are explored. "Regression models are helpful for predicting numerical values based on different data points" (Delua, 2021).

The dataset chosen for regression was a medical cost insurance dataset consisting of seven variables including age, BMI, whether the individual is a smoker or not, as well as the continuous target variable, the amount of insurance charges.



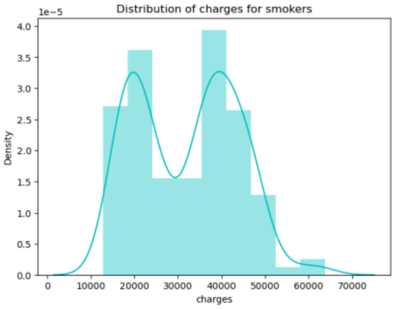
This graph represents the plot of BMI vs. charges, with respect to the individual being a smoker or not. It is observed that individuals who were smokers had an overall higher insurance charge compared to non-smokers, at every level of BMI.

The variables representing sex, smoker and region were all categorical variables, so were therefore converted into metric variables using a label encoder.

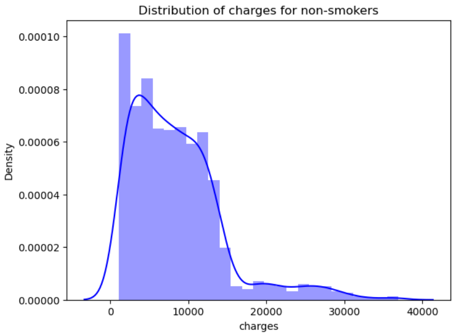


The correlation of each of the other variables with charges (target variable) were also recorded. It can be seen that there is a high positive correlation value of 0.79 between individuals that smoke and charges. Notable positive correlations were also recorded for age, BMI and children with charges.

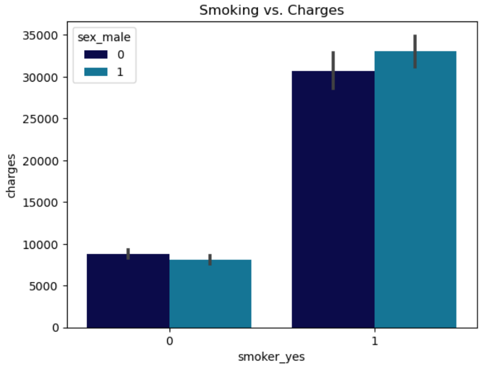
**How does smoking impact the level of medical insurance charges for an individual?**



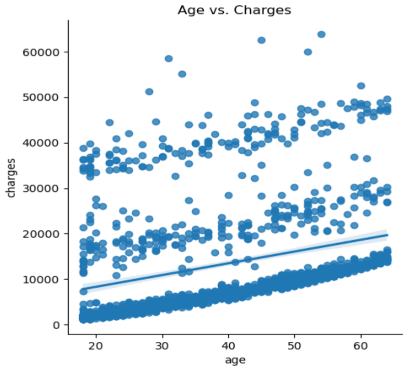
It is clearly observed that the insurance charges for a smoker are overall higher than that of a non-smoker, and there is a higher number of smokers with high charges than lower charges.

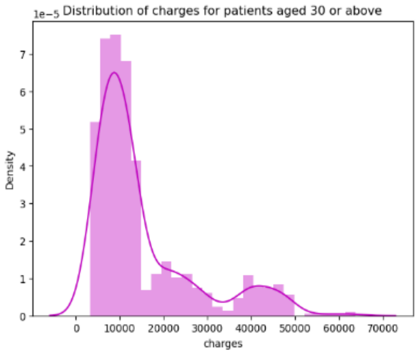


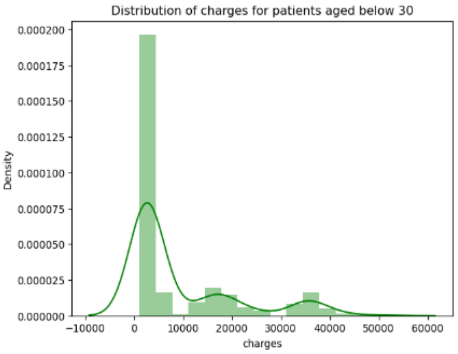
The plot to the right shows that male smokers have a higher charge than female smokers do while the opposite is observed for non-smokers. This indicates that the impact of smoking on insurance charge is heavier on males in comparison to females. Similar correlations were tested in a Japanese study, where (Izumi et al., 2001) stated that "smokers incurred more medical costs than never smokers, by 11% in males, but costs were almost the same in females." However, the results deciphered from this model differ slightly, as there is a clear increase in insurance charge for female smokers as well.



**What effect does age have on the level of medical insurance charge?**

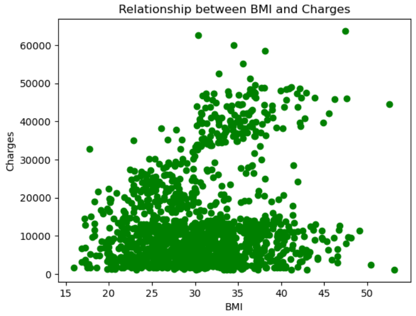


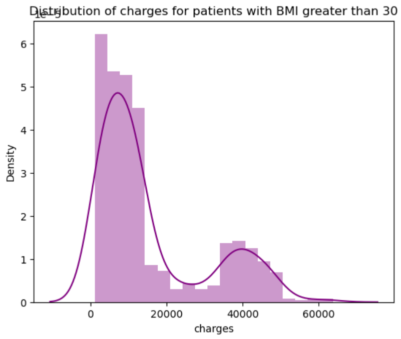


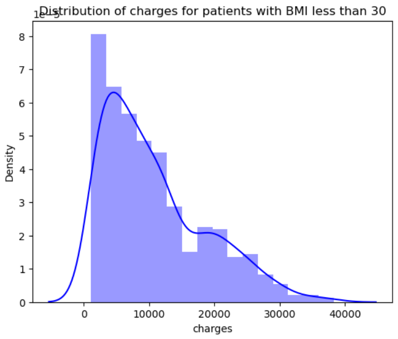


As seen by the figures above, the level of insurance charge does increase with age, as confirmed by the afore mentioned positive correlation of 0.30. In addition, when comparing charges for individuals who are 30 and above with those below 30, there is a clear indication that insurance charges are higher for the 30 and above category.

**How does BMI affect the insurance charge for an individual?**







It is evident that medical insurance charge is higher as the BMI of an individual increases. The distribution of charges for BMI values above 30 were also compared with that of values under 30. A BMI of over 30 was chosen as the criterion as this value indicates obesity, which is a major health risk that could impact medical insurance charges. The diagram above depicts that insurance charges are indeed higher for individuals with a BMI over 30, possibly due to more illnesses and therefore more costly treatments. (Thorpe et al., 2004) also had similar results stating that "costs incurred by the obese were 37 percent higher than costs for those with normal weight in 2001”. This is due to the higher risk of diseases such as hypertension and diabetes, which would require more extensive medical diagnosis and treatment, as mentioned by (Hammond & Levine, 2010).

## Regression Models

The dataset was fitted using multiple different regression models in order to predict future medical insurance costs. The dataset was split into training and testing data in a ratio of 80:20 and the models were then tested.  When comparing the accuracy of the results, it was evident that Gradient Boosting Regressor, Random Forest Regressor, XGB Regressor and Linear Regression were the best fit models for this data, so these were further explored. After performing hyperparameter tuning on the relevant models, the results were as below.

| Model | R^2 score | MAE | RMSE | Explained variance score |
|---|---|---|---|---|
| Gradient Boosting Regressor | 0.89112 | 1767.991172 | 4162.466699 | 0.897951 |
| Random Forest Regressor | 0.880288 | 2615.050963 | 4364.59691 | 0.884533 |
| XGB Regressor | 0.847856 | 2960.020327 | 4920.432937 | 0.850552 |
| Linear Regression | 0.799988 | 3933.272649 | 5641.626559 | 0.800266 |

Descriptions of the parameters above are as follows:

**R^2 score:** This statistic takes a value between 0 and 1, and explains the extent of variance in a response variable explained by a fitted model in comparison to the mean of the actual variable (Saunders et al., 2012). A value close to 1 is more preferrable.

**Mean absolute error (MAE):** This is the average of all errors collected within the model.

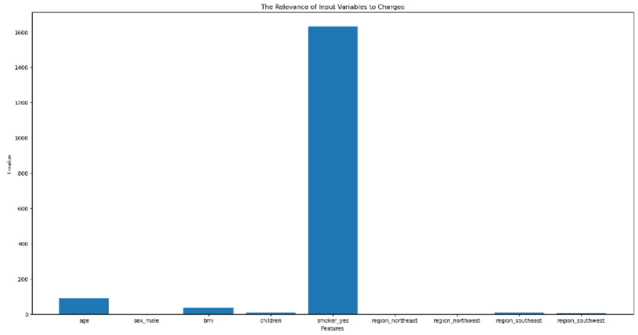
**Root mean squared error (RMSE):** This is considered as a measure of the lack of fit of a model. A large value would indicate that a model does not fit the data very well (Witten et al., 2022).

**Explained variance score:** This measures the variance between a model and the actual data (similar to R^2).

It is clear that the Gradient Boosting Regressor model is the best fit, with the highest R^2 statistic and the lowest MAE AND RMSE scores. Therefore, this model is the best fit for predicting insurance charges, with 89.8% of variance being explained.

### Feature selection

Feature selection, which is the “process of identifying and selecting a subset of input variables that are most relevant to the target variable” (Brownlee, 2020), was carried out to determine the most valuable features that influence medical insurance charges.



It is clearly seen that the most relevant features to charges are smoker_yes (whether the individual is a smoker or not), age, BMI and children. Therefore, linear regression was carried out once again using only the four features selected above.

| Model | R^2 score | MAE | RMSE | Explained variance score |
|---|---|---|---|---|
| Linear Regression | 0.781115 | 4213.798595 | 5829.378522 | 0.781454 |

Despite the feature selection identifying the above variables as the most valuable, the linear regression parameters have actually worsened in value, with a lower R^2 score and higher MAE and RSME values. This indicates that feature selection is not very useful for this dataset, and that all variables should be taken instead when modelling.

**Possible alternative**

It can be noted that having the charges variable (target variable) undergo a logarithmic transformation and then re-running the Linear Regression model returns the following result.

| Model | R^2 score | MAE | RMSE | Explained variance score |
|---|---|---|---|---|
| Linear Regression | 0.804731 | 0.2696916 | 0.4190157 | 0.806612 |

The R^2 and explained variance scores have both increased slightly, showing an improvement in the model fit. In addition, the MAE and RMSE values have drastically decreased, indicating that there far less errors within the model. Therefore, this is an alternate approach to the regression modelling that can be considered.

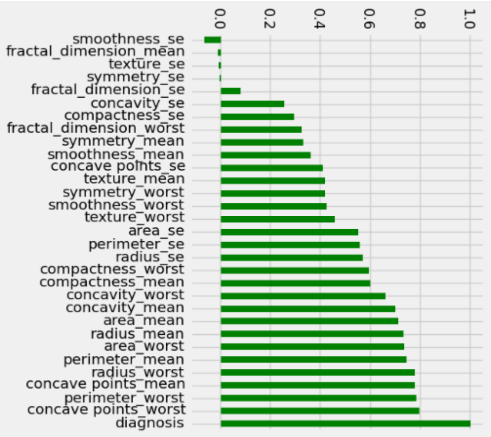
## Task 3- Classification

Classification is also a supervised learning method that aims to classify and predict how data will be categorized based on recognizing patterns within the data, with (Delua, 2021) stating that “Classification problems use an algorithm to accurately assign test data into specific categories”.

The dataset selected for this task was a breast cancer dataset and a model was formed to classify tumors into malignant (cancerous) and benign (non-cancerous) types, and to predict the cancer type of future tumors.

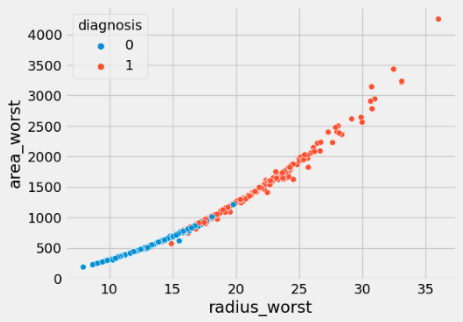
The data consisted of 569 samples and 32 variables, including ID, several features of the tumor, and the categorical target variable, diagnosis, which took two forms; malignant or benign. The target variable was converted into a numerical form, with 1 denoting a malignant tumor and 0 denoting a benign one.

### What are the best features that correlate to the diagnosis of a tumor?

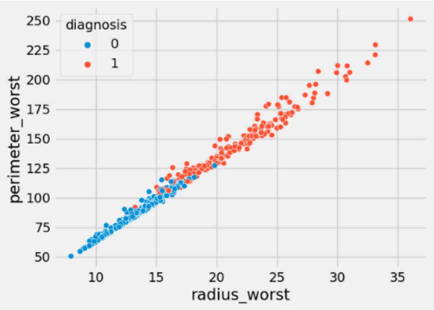


From the correlation diagram, it can be seen that the most positively correlated variables with diagnosis are the worst concave points, worst (largest) perimeter and mean concave points, so it can be deduced that these are the most valuable features. However, (El-Sebakhy et al., 2006) found that “the best three attributes are mean texture, worst mean area, and worst mean smoothness”, as well as (Street et al., 1993) stating similarly that “mean texture, worst area and worst smoothness” were the variables “which best separate benign from malignant samples”, showing that the results of this research differ.

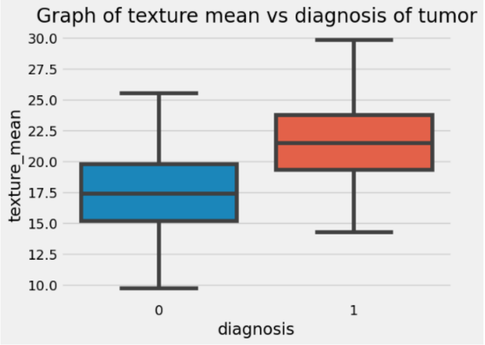
### Is there a significant relationship between the radius worst and area/perimeter worst of a tumor?



It can be seen that there is a strong positive correlation between the worst radius and worst area of a tumor, and that malignant tumors have a high value for both variables. The same is observed for the worst radius plotted with the worst perimeter.



**Does the mean texture have a significant relationship with the diagnosis of a tumor?**



It is clearly seen that malignant tumors have a higher mean texture than benign tumors, as the distribution of the two boxplots are noticeably different. A study done by (Kolios et al., 2021) agreed that “texture could hypothetically be used to predict patient prognosis”, confirming that the texture of tumor does have an effect on the ultimate diagnosis and subsequent treatment.

## Classification Models

Four classification models were tested, which are as follows:

**Support Vector Classifier:** This model “constructs hyperplanes to learn a decision boundary in order to separate data points that belong to different classes” (Enver, 2021).

**Logistic Regression:** This is a model used for binary classifications, where the aim is to predict whether a certain point belongs to one class or another.

**Decision Tree Classifier:** This model used features of the dataset to create nodes based on yes/no questions and the data is split continuously until all data points are assigned to classes (Bento, 2021).

**Random Forest Classifier:** This model “consists of multiple decision trees each of which outputs a prediction” (Molina, 2021).

The dataset was split into training and testing data in a ratio of 80:20 and the models were then tested.

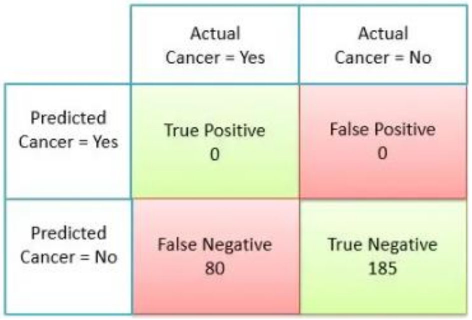
The fit of the model can be examined through the use of an accuracy score, which returns the number of correct predictions out of the total predictions made by the model’s testing data (Pierre, 2021).

The accuracy scores for each model were as follows:

| **Model** | **Accuracy Score** |
|---|---|
| Support Vector Classifier | 94.736842 |
| Logistic Regression | 95.614035 |
| Decision Tree Classifier | 93.859649 |
| Random Forest Classifier | 96.491228 |

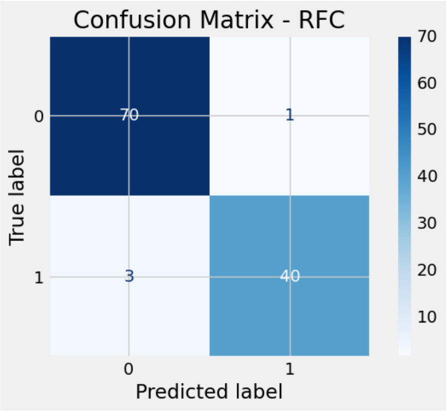
As observed by the results above, Random Forest Classifier and Logistic Regression have the highest accuracy scores, showing that the predictions of the models are accurate to the true value, and are therefore the best fit models for predicting the cancer type of a tumor for this data.

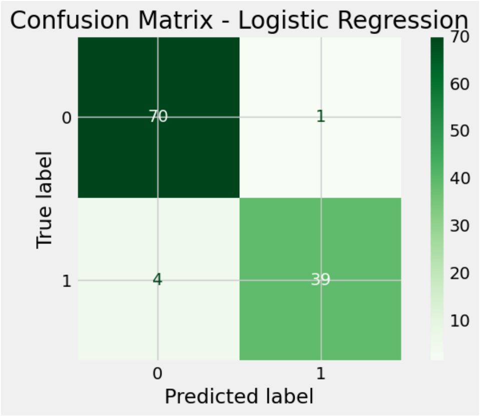
A confusion matrix can also be used to assess the performance of a classification model, which forms a matrix of actual values against the predicted values of the test data, showing how accurate the model is.



| | Actual<br>Cancer = Yes | Actual<br>Cancer = No |
|---|---|---|
| Predicted<br>Cancer = Yes | True Positive<br>0 | False Positive<br>0 |
| Predicted<br>Cancer = No | False Negative<br>80 | True Negative<br>185 |

The confusion matrices for the two models, which depict the accuracy of the model's predictions, are represented below:





Out of 114 data points, both models were able to correctly predict 70 malignant tumors, and the RFC model correctly predicted 40 benign tumors, while the Logistic Regression model correctly predicted 39 benign tumors.

## Bibliography

Delua, J. (2021) *Supervised vs. unsupervised learning: What's the difference?*, *IBM*. Available at: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning (Accessed: March 25, 2023).

Brownlee, J. (2020) *How to perform feature selection for regression data*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/feature-selection-for-regression-data/#:~:text=Feature%20selection%20is%20the%20process%20of%20identifying%20and,and%20a%20numerical%20target%20for%20regression%20predictive%20modeling. (Accessed: March 27, 2023).

Parsons, A.G. (2003) “Assessing the effectiveness of shopping mall promotions: Customer Analysis,” *International Journal of Retail & Distribution Management*, 31(2), pp. 74–79. Available at: https://doi.org/10.1108/09590550310461976.

Izumi, Y. *et al.* (2001) “Impact of smoking habit on medical care use and its costs: A prospective observation of National Health Insurance Beneficiaries in Japan,” *International Journal of Epidemiology*, 30(3), pp. 616–621. Available at: https://doi.org/10.1093/ije/30.3.616.

Thorpe, K.E. *et al.* (2004) “The impact of obesity on rising medical spending,” *Health Affairs*, 23(Suppl1). Available at: https://doi.org/10.1377/hlthaff.w4.480.

Hammond, R. and Levine, R. (2010) “The economic impact of obesity in the United States,” *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, p. 285. Available at: https://doi.org/10.2147/dmsott.s7384.

Witten, D., Hastie, T. and Tibshirani, R. (2022) *An introduction to statistical learning: With applications in R*. Boston, Massachusetts: Springer.

Saunders, L.J., Russell, R.A. and Crabb, D.P. (2012) *The coefficient of determination: What determines a useful R2 statistic?*, *Investigative Ophthalmology & Visual Science*. The Association for Research in Vision and Ophthalmology. Available at: https://iovs.arvojournals.org/article.aspx?articleid=2127111 (Accessed: April 1, 2023).

Enver, A. (2021) *Support vector classifier simply explained [with code]*, *For Predictions in Minutes. No Code Required.* PI.EXCHANGE. Available at: https://www.pi.exchange/blog/support-vector-classifier (Accessed: March 27, 2023).

Bento, C. (2021) *Decision tree classifier explained in real-life: Picking a vacation destination*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575 (Accessed: March 27, 2023).

Molina, E. (2021) *A practical guide to implementing a random forest classifier in Python*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/a-practical-guide-to-implementing-a-random-forest-classifier-in-python-979988d8a263 (Accessed: March 27, 2023).

Pierre, S. (2021) *How to evaluate classification models in python: A beginner's guide*, *Built In*. Available at: https://builtin.com/data-science/evaluating-classification-models (Accessed: March 28, 2023).

El-Sebakhy, E.A. *et al.* (2006) “Evaluation of breast cancer tumor classification with unconstrained functional networks classifier,” *IEEE International Conference on Computer Systems and Applications, 2006.* [Preprint]. Available at: https://doi.org/10.1109/aiccsa.2006.205102.

Street, W.N., Wolberg, W.H. and Mangasarian, O.L. (1993) “Nuclear feature extraction for breast tumor diagnosis,” *SPIE Proceedings* [Preprint]. Available at: https://doi.org/10.1117/12.148698.

Kolios, C. *et al.* (2021) “MRI texture features from tumor core and margin in the prediction of response to neoadjuvant chemotherapy in patients with locally advanced breast cancer,” *Oncotarget*, 12(14), pp. 1354–1365. Available at: https://doi.org/10.18632/oncotarget.28002.

Kuruvilla, S.J., Joshi, N. and Shah, N. (2009) “Do men and women really shop differently? an exploration of gender differences in mall shopping in India,” *International Journal of Consumer Studies*, 33(6), pp. 715–723. Available at: https://doi.org/10.1111/j.1470-6431.2009.00794.x.