Physiotherapy Duration Prediction

Sai Manognya[1*], Terala Hemanth[2], Shayaan Hussain[3]

[1]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.
[2]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.
[3]Department of Electronics and Communication Engineering SR University, Warangal, Telangana, India

*Email: **shayaan.hussain2@gmail.com**

Abstract: : Physiotherapy is a kind of treatment through physical methods and various exercises. A single session of physiotherapy lasts thirty minutes to one hour a day. The complete treatment duration may range from a couple of days to even a few months depending on the condition of the patient. However, the physiotherapists can not accurately predict how long it would take to treat a patient. We have proposed a solution using machine learning to predict the duration based on how long the physiotherapists had taken to treat the patients in history.

Keywords: Machine Learning, Physiotherapy Duration Prediction, Linear Regression Algorithm, Decision Tree Algorithm, Random Forest Algorithm.

1. INTRODUCTION

Physiotherapy has lately become a basic need in the lives of many people across the globe. It is a way to treat the wounds and diseases through physical methods. For example, a person with a fracture or a person who has undergone a surgery, they cant properly move their muscles and bones. Physiotherapy deals with special types of exercises that would gradually bit by bit improve the movement until the person can finally move the bodypart without any difficulty

1. What is physiotherapy?

- Physiotherapists help people affected by injury, illness or disability through movement and exercise, manual therapy, education and advice.
  - They maintain health for people of all ages, helping patients to manage pain and prevent disease.
  - The profession helps to encourage development and facilitate recovery, enabling people to stay in work while helping them remain independent for as long as possible.

2. What physiotherapists do?

- Physiotherapy is a science-based profession and takes a ‘whole person’ approach to health and well being, which includes the patient’s general lifestyle.
- At the core is the patient’s involvement in their own care, through education, awareness, empowerment and participation in their treatment.
- You can benefit from physiotherapy at any time in your life. Physiotherapy helps with back pain or sudden injury, managing long-term medical condition such as asthma, and in preparing for childbirth or a sporting event.

Figure 1. Physiotherapy

2. PROBLEM DEFINITION

This is a study of physiotherapy to predict the number of days needed for treatment using physiotherapy. This can be done using regression models.

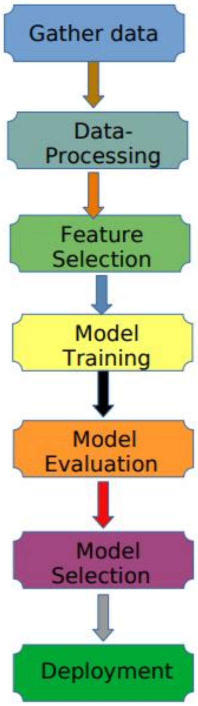


Figure 2. Processing Steps for Machine Learning for Physiotherapy Duration Prediction

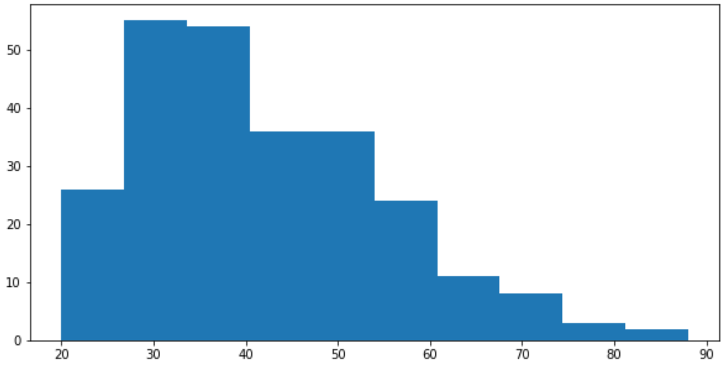
3. DATASET AND ATTRIBUTES

We collected real time data from a physiotherapist. The collected dataset has the following attributes :

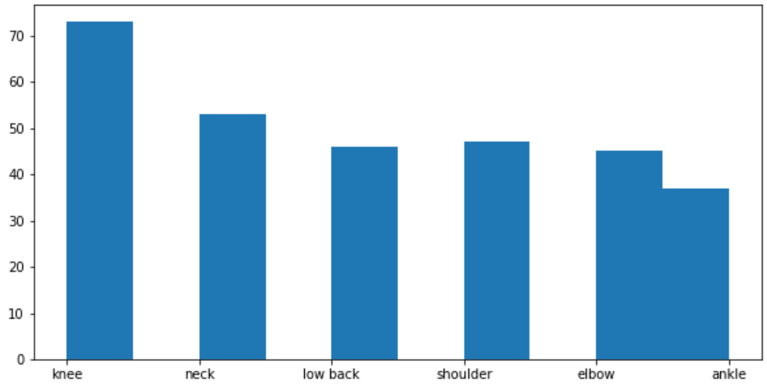
Frozen Shoulder	It is a condition in which person suffers with pain and stiffness in shoulder
Cervical Radiculopathy	When the nerve root in cervical spine is damaged, the person feels numbness, weakness and irritation in the neck area
Patella Fracture	The knee cap is broken or damaged
Lumbar Radiculopathy	When the nerve root in lumbar spine is damaged, the person feels numbness, weakness and irritation in the low back area
Tennis Elbow	Weakening of the tendons that join your forearm muscles to your bones due to overuse of elbow
Osteoarthiris	The damage of the cushion at the end of a bone
Plantar Fasciitis	Inflammation of the fibrous tissue at the ankle

- Age: Age in years
- Affected Body Part
    - Value 1: Knee
    - Value 2: Neck
    - Value 3: Low Back
    - Value 4: Shoulder
    - Value 5: Elbow
    - Value 6: Ankle
- Diagnosis
    - Value 1: Patella Fracture
    - Value 2: Cervical Radiculopathy
    - Value 3: Lumbar Radiculopathy
    - Value 4: Frozen Shoulder
    - Value 5: Tennis Elbow
    - Value 6: Osteoarthritis
    - Value 7: Plantar Fasciitis
- Duration of pain
    - Value 1: Acute (0-6 weeks)
    - Value 2: Subacute (6-12 weeks)
    - Value 3: Chronic (More than 12 weeks)
- Intensity of pain on NPRS
  NPRS stands for Numeric Pain Rating Scale

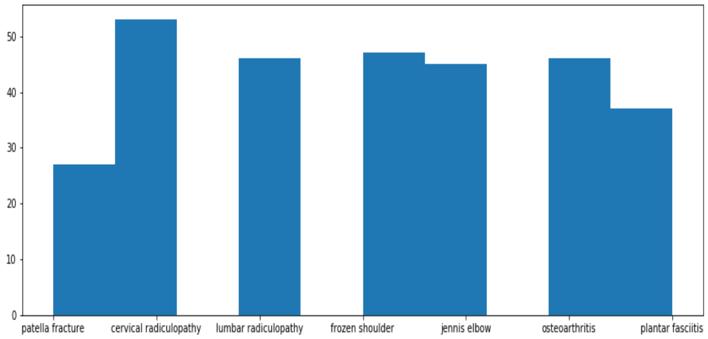
  It ranges between 1 to 10
- Treatment Approach
    - Value 1: Manual
    - Value 2: Mechanical
    - Value 3: Manual and Mechanical
- Duration of treatment
  Number of days the patient was treated



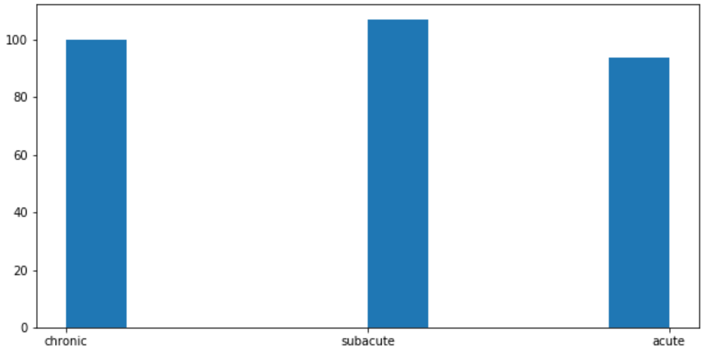
Age



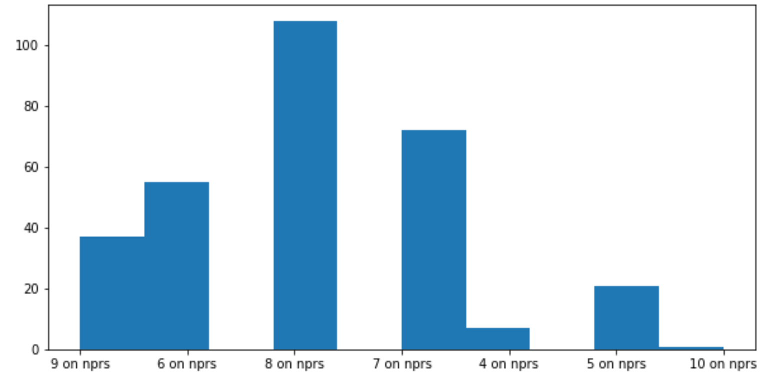
Affected Body Part



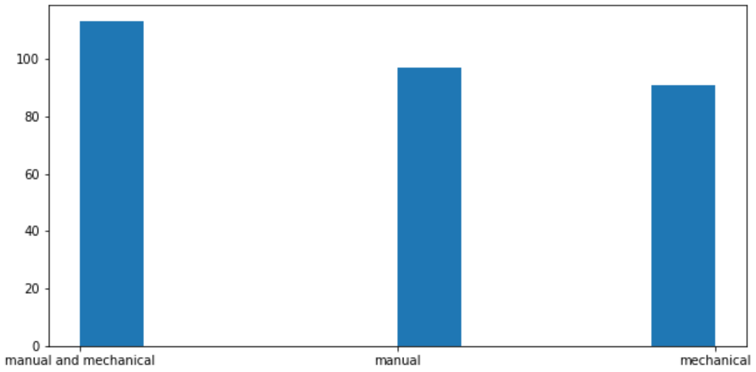
Diagnosis



Duration of pain



Intensity of pain



Treatment approach

Figure 3. Visualizing attributes of the dataset

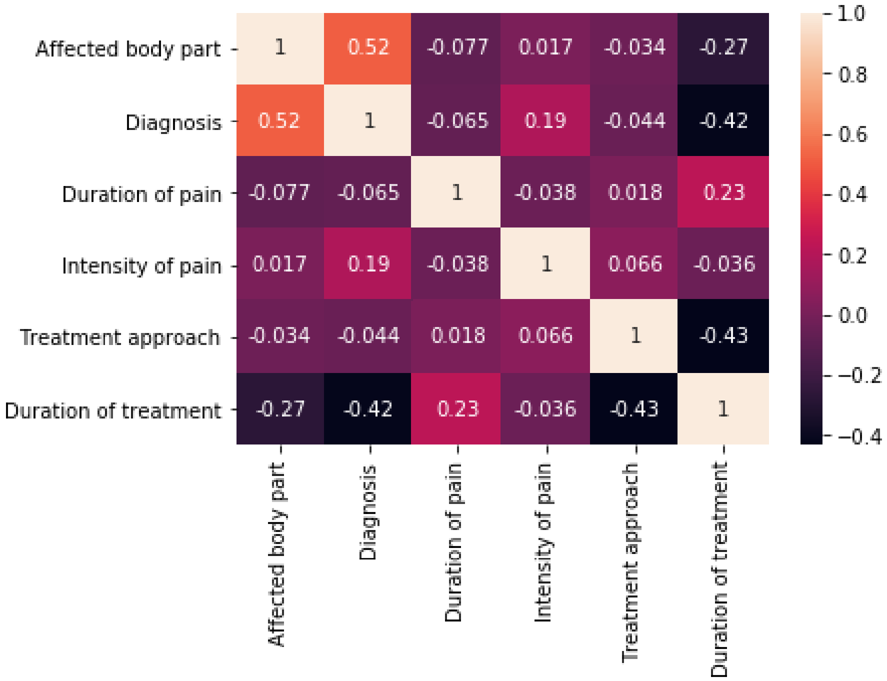
4. DATA PRE-PROCESSING

Correlation Matrix

A correlation matrix is simply a table that displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. coefficients for different variables.

How it is calculated?

A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (X_i) in the table is correlated with each of the other values in the table (X_j)... The diagonal of the table is always a set of ones because the correlation between a variable and itself is always 1. Let's perform the Correlation matrix to understand the relation between the dependent variable and the independent variable and within the independent variable.



	Affected body part	Diagnosis	Duration of pain	Intensity of pain	Treatment approach	Duration of treatment
Affected body part	1	0.52	-0.077	0.017	-0.034	-0.27
Diagnosis	0.52	1	-0.065	0.19	-0.044	-0.42
Duration of pain	-0.077	-0.065	1	-0.038	0.018	0.23
Intensity of pain	0.017	0.19	-0.038	1	0.066	-0.036
Treatment approach	-0.034	-0.044	0.018	0.066	1	-0.43
Duration of treatment	-0.27	-0.42	0.23	-0.036	-0.43	1

Figure 4. Correlation Matrix

The dataset consists of records of 301 patients out of which, the age of 46 patients is missing. The treatment duration of the diagnosis that exist in our dataset do not vary with age. The duration ranges between 5 to 45 days for patients of all ages so we can drop the whole column to reduce the complexity of model. The affected body part is also correlated to the diagnosis so we can drop that column too. The dataset has a lot of string values so we need to deal with various methods to convert them to numeric values.

Diagnosis	It is a string value of the diagnosis. For linear regression, we can use dummy coding and for decision tree and random forest algorithms, we can assign a unique integer value to each string because each value is treated separately in these algorithms
Duration of Pain	Acute, Subacure or Chronic. It describes since how long the patient has been suffering from the diagnosed disease
Intensity of pain	A numeric rating on a scale of 1 to 10
Treatment Approach	Some patients opt for manual treatment without any machinery where as some prefer machinery. There are treatment methods for both. Treatment may also be manual and mechanical simultaneously
Treatment Duration	The number of days patient took for recovery

Table 2. Attributes Validation

Age	Numeric [20 to 88;unique=56;mean=41.88;median=40]
Affected body part	String [unique=6]
Diagnosis	String [unique=7]
Duration of Pain	String [unique=3]
Intensity of Pain	Numeric [4 to 10;unique=7;mean=7.22;median=7]
Treatment Approach	String [unique=3]
Treatment Duration	Numeric [5 to 45;unique=26;mean=12.14;median=11

Table 3. Dataset range and datatype[9].

Enough EDA performs on the data to evaluate the dataset and gather knowledge about the data. Let's perform some Machine Learning model and Experimentation to create a model that helps us to achieve our goal we state in the problem definition

5. Algorithms

We use different machine learning model to solve our regression problem:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor

1. Linear Regression:

Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values.

Given a data set $\{y_i, x_{i1},....,x_{ip}\}_{i=1 \text{ to } n}$ of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable *y* and the *p*-vector of regressors x is linear. This relationship is modeled through a *error variable ε* — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$y = m_1x_1 + m_2x_2 + \dots\dots\dots + m_px_p + c$

The cost function is defined as:

$E = (y_i - (m_1x_{i1} + m_2x_{i2} + \dots.. + m_px_{ip} + c)^2 / 2$

2. Decision Tree:

A decision tree is like a flowchart structure which consists of nodes, branches and leafs. Each internal node represents a ‘test’ on an attribute. Outcome of a test would be directed onto a specific path or a branch which leads to another node or a leaf. The leafs represent the predicted value.

A decision tree requires an end criteria which can be the count of values, the coefficient of variance and a max depth. If any of them is reached, the next node has to be the leaf node. The output in the leaf node is the average of the target variables corresponding to the specified value in the sub-dataset at the given node.

The next feature or node to be checked is decided based on standard deviation reduction of each feature with respect to target variable. The standard deviation is calculated as follows:

$$\text{SD}=Square\ Root\ of\ \frac{\sum(x-mean)^\wedge 2}{n}$$

A decision tree consists of three elements:

1. Decision nodes – it tests a feature
2. Branches – they represent values of result
3. Leaf nodes – they represent the predicted value

3. **Random forest:**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision tree at training time.

Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set
X = x_1, x_2, . . . ,x_n responses with Y = y_1, y_2 , . . . , y_n, bagging repeatedly (*B* times) selects a random sample with their replacement of the training set and fits trees to these samples:

For *b* = 1, ..., *B*:

1. Sample, with replacement, *n* training examples from *X*, *Y*; call these *Xb*, *Yb*.
2. Train a classification or regression tree *fb* on *Xb*, *Yb*.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

6. RESULTS

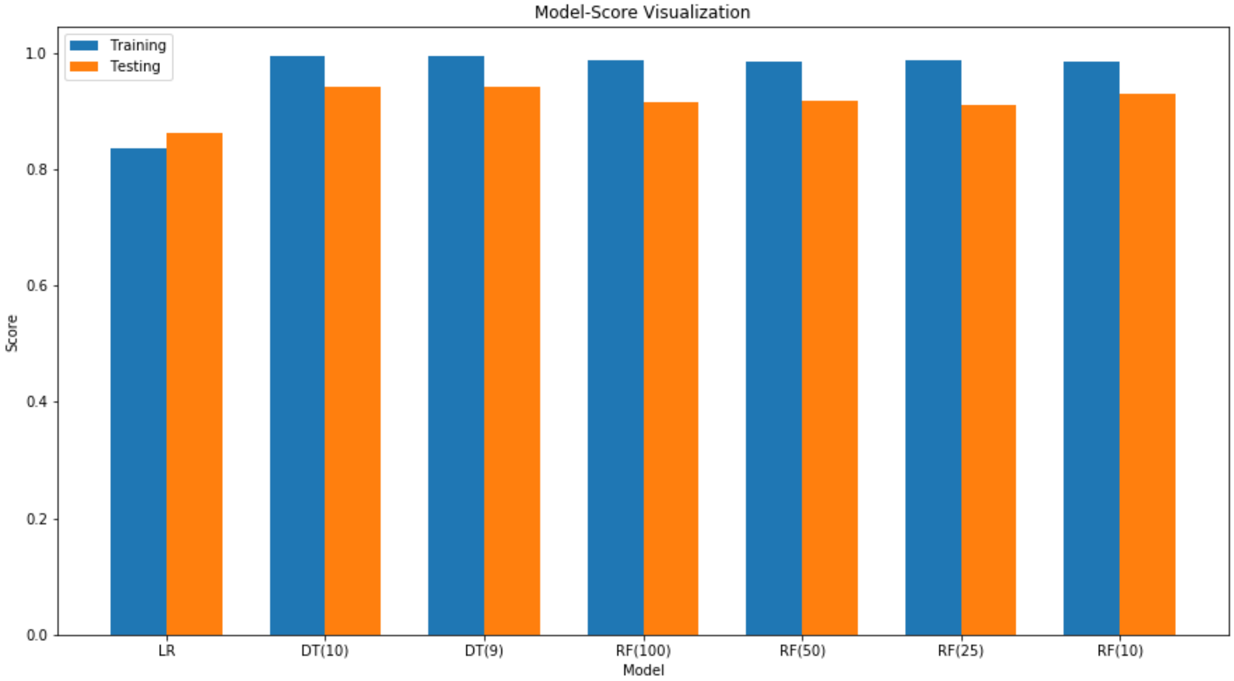


Figure 5. Result of various models with the proposed model

		Random Forest				**Decision Tree**				**Linear Regression**
		Estimators				Depth				
		100	50	25	10	10	9	8	7	
Training	MAE	0.37	0.37	0.37	0.34	0.13	0.16	0.3	0.42	1.67
	MSE	0.53	0.5	0.52	0.55	0.2	0.22	0.52	0.65	6.6
	RMSE	0.73	0.71	0.72	0.74	0.45	0.47	0.72	0.8	2.56
	SCORE	0.98	0.98	0.98	0.98	0.99	0.99	0.98	0.98	0.83
Testing	MAE	0.76	0.73	0.77	0.68	0.61	0.54	0.67	0.71	1.57
	MSE	2.42	2.47	2.64	2.25	1.72	1.54	1.7	1.63	4.03
	RMSE	1.55	1.57	1.62	1.5	1.31	1.24	1.3	1.28	2
	SCORE	0.91	0.91	0.9	0.82	0.93	0.95	0.94	0.94	0.86

Table 4. Model Analysis

The algorithms that we used Linear Regression, Decision Tree and Random Forest, all provide a good result with very less error. However, the decision tree regressor with a depth of 9 provides the best result in training as well as testing as compared to others so this model is used for deployment

7. CONCLUSION

From the above table of scores and errors, we can conclude that decision tree regressor shows the best results with a training score 0.99 and testing score 0.95.

8. REFERENCES

[1] https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

[2] https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96

[3] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[4] https://en.wikipedia.org/wiki/Physical_therapy

[5] https://en.wikipedia.org/wiki/Patella_fracture

[6] https://en.wikipedia.org/wiki/Tennis_elbow

[7] https://www.spine-health.com/conditions/neck-pain/what-cervical-radiculopathy

[8] https://www.physio-pedia.com/Lumbar_Radiculopathy

[9] https://www.mayoclinic.org/diseases-conditions/osteoarthritis/symptoms-causes/syc-20351925

[10] https://orthoinfo.aaos.org/en/diseases--conditions/frozen-shoulder/

[11] https://www.healthline.com/health/plantar-fasciitis

[12] https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

[13] https://www.csp.org.uk/careers-jobs/what-physiotherapy

[14] https://en.wikipedia.org/wiki/Random_forest

[15] https://en.wikipedia.org/wiki/Linear_regression

[16] https://en.wikipedia.org/wiki/Decision_tree

[17] **Essentials of Orthopaedics & Applied Physiotherapy** **By Jayant Joshi, Prakash P Kotwal**