

A Mini-Project Report
on
MODCLOTH FIT PREDICTION USING MACHINE
LEARNING

submitted in partial fulfillment of the requirements

for the award of degree of
BACHELOR OF TECHNOLOGY

in
Information Technology

by
U . Shivani (19WH1A1279)
T . Pooja (19WH1A1284)
P . Girija(19WH1A1286)
G . Vaishnavi (19WH1A12B9)

Under the esteemed guidance of
Dr. P. S Latha Kalyampudi
Associate Professor



Department of Information Technology

BVRIT HYDERABAD College of Engineering for Women

Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyderabad – 500090

(Affiliated to Jawaharlal Nehru Technological University, Hyderabad)

(NAAC 'A' Grade & NBA Accredited- ECE, EEE, CSE IT)

January, 2023

DECLARATION

We hereby declare that the work presented in this project entitled “ **MODCLOTH FIT PREDICTION USING MACHINE LEARNING**” submitted towards completion of Minor Project in IV year I sem of B.Tech IT at “BVRIT HYDERABAD College of Engineering for Women”, Hyderabad is an authentic record of our original work carried out under the esteem guidance of **Dr. P. S Latha Kalyampudi , Associate Professor**, Department of Information Technology.

U. Shivani (19WH1A1279)

T. Pooja (19WH1A1284)

P. Girija (19WH1A1286)

G. Vaishnavi (19WH1A12B9)



BVRIT HYDERABAD

College of Engineering for Women

Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyderabad – 500090

(Affiliated to Jawaharlal Nehru Technological University, Hyderabad)

(NAAC 'A' Grade & NBA Accredited- ECE, EEE, CSE IT)

CERTIFICATE

This is to certify that the mini-project report on “ **MODCLOTH FIT PREDICTION USING MACHINE LEARNING**” is a bonafide work carried out by **U. Shivani (19WH1A1279), T. Pooja (19WH1A1284), P. Girija (19WH1A1286)** and **G. Vaishnavi (19WH1A12B9)** in the fulfillment for the award of B.Tech degree in **Information Technology, BVRIT HYDERABAD College of Engineering for Women, Bachupally, Hyderabad** affiliated to Jawaharlal Nehru Technological University, Hyderabad under my guidance and supervision.

The results embodied in the project work have not been submitted to any other university or institute for the award of any degree or diploma.

Internal Guide

Dr. P. S Latha Kalyampudi

Associate Professor

Department of IT

Head of the Department

Dr. Aruna Rao S L

Professor & HoD

Department of IT

External Examiner

ACKNOWLEDGEMENT

We would like to express our profound gratitude and thanks to **Dr. K. V. N. Sunitha, Principal, BVRIT HYDERABAD College of Engineering for Women** for providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. Aruna Rao S L, Professor & Head, Department of IT, BVRIT HYDERABAD College of Engineering for Women** for all the timely support, constant guidance and valuable suggestions during the period of our project.

We are extremely thankful and indebted to our internal guide, **Dr. P. S. Latha Kalyampudi, Associate Professor, Department of IT, BVRIT HYDERABAD College of Engineering for Women** for her constant guidance, encouragement and moral support throughout the project.

Finally, we would also like to thank our Project Coordinators **Ms. K.S Niraja, Assistant Professor, Mr. Ch. Anil Kumar, Assistant Professor**, all the faculty and staff of Department of IT who helped us directly or indirectly, parents and friends for their cooperation in completing the project work.

U . Shivani (19WH1A1279)

T . Pooja (19WH1A1284)

P . Giriya (19WH1A1286)

G . Vaishnavi (19WH1A12B9)

ABSTRACT

MoD Cloth fit Prediction is used to predict the size of a particular Mod Cloth product fit for the customer. Fit prediction is critical in order to improve customers' shopping experiences and to reduce product return rates. Modelling customers' fit feedback is challenging due to its subtle semantics, arising from the subjective evaluation of products, and imbalanced label distribution. This project proposes a new predictive framework to tackle the product fit problem, which captures the customers' feedback on the product fit, and employs a metric learning technique to resolve the issues. It also contribute the public dataset collected from online clothing retailer (MoD Cloth). The whole dataset is split into training and testing sets. This dataset is useful in identifying key features that determine the fit of a clothing product on a customer. Random Forest algorithm is applied to predict the Size of the product.

LIST OF FIGURES

Figure No.	Figure Name	Page No.
3.2.1	Architecture Design for size prediction model	6
3.3.1	Use case diagram to predict size	7
4.2.1	Raw Data of MoDCloth Dataset	8
4.2.2	Data Preprocessing of MoDCloth Data	10
4.2.3	Data Visualization of MoDCloth Dataset	10
4.2.4	Splitting the Data into training & testing data	11
4.2.5	Accuracy of the Random Forest model	12
4.2.6.1	Random Forest	12
4.3	Predicting the Size with user inputs	14

CONTENTS

TOPIC	PAGE NO.
Abstract	V
List of Figures	VI
1. Introduction	1
1.1 Objective	2
1.2 Problem Definition	2
1.3 Aim of the Project	2
2. Literature Survey	3
2.1 Major Issues	4
3. System Analysis and Design	5
3.1 Proposed System	5
3.2 System Architecture	6
3.3 Use Case Diagram	7
4.Implementation	8
4.1 Modules	8
4.2 Proposed Algorithms	8
4.2.1 Raw Data	9
4.2.2 Data PreProcessing	9
4.2.3 Data Visualization	10
4.2.4 Testing Data	11
4.2.5 Predicting the Size using Algorithm	11
4.2.6 Random Forest Algorithm	12
4.3 Results	14
5. Conclusion and Future Scope	15
References	16

1. INTRODUCTION

With the growth of the online fashion industry and the wide size variations across different clothing products, automatically providing accurate and personalized fit guidance is worthy of interest. As retailers often allow customers to provide fit feedback (e.g. “small”, “fit”, “large”) during the product return process or when leaving reviews, predictive models have been recently developed based on this kind of data [1, 8, 9]. A few recent approaches to this problem use two sets of latent variables to recover products and customers’ “true” sizes, and model the fit as a function of the difference between the two variables [8, 9].

However, we notice that customers’ fit feedback reflects not only the objective match/mismatch between a product’s true size and a customer’s measurements, but also depends on other subjective characteristics of a product’s style and properties. We see two customers expressing concerns that a jacket seems ‘baggy,’ but both have different feedback regarding this fit shortcoming. Additionally, such fit feedback is ordinal in nature and is unevenly distributed (most transactions are reported as “fit”), which differs from general item recommendation tasks and requires domain-specific techniques.

In this we pose product size recommendation problem as fit prediction problem and tackle the aforementioned challenges in the following ways: First, unlike previous work which focuses on recovering “true” sizes, we develop a new model to factorize the semantics of customers’ fit feedback, so that representations can capture customers’ fit preferences on various product aspects (like shoulders, waist etc.).

We apply an ordinal regression procedure to learn these representations such that the order of labels is preserved. Second, using a heuristic we sample good representations from each class and project them to a metric space to address label imbalance issues. We collect customers' fit feedback from two different clothing websites and contribute two public datasets. Through experiments on these datasets, we show the effectiveness of uncovering fine grained aspects of fit feedback and highlight the ability of metric learning approaches with prototyping in handling label imbalance issues.

1.1 OBJECTIVE

The objective of Mod cloth Fit Prediction Using Machine Learning is that the size and fit problem as a classification problem where given a product and a customer we want to predict if the product will 'fit' to the customer using the Random Forest algorithm, which is an supervised learning algorithm.

1.2 PROBLEM DEFINITION

Given a customer C , a product P , and a set of reviews for the product R , we want to predict if the product will be small, fit or large for the customer. Both product and customer are defined by their respective features $P = p_i$ and $C = c_i$ where each feature can be either continuous or categorical. $R = r_i$ consists of the reviews left by customers on the product page. We define the output space as $F = (\text{small}, \text{fit}, \text{large})$.

1.3 AIM OF THE PROJECT

The main aim of this project is to predict the size of a particular Mod Cloth product fit for the customer. The Fit prediction is critical in order to improve customers shopping experiences. This makes it easier for both the Producers and the consumers to reduce product return rates. The model helps the customers to know their required product size for a particular mod-cloth.

2. LITERATURE SURVEY

2.1 RELATED WORK

- According to Shreya Singh, G Mohammed Abdulla, Sumit Borar and Sagar Arora. Mainly focuses on providing a production ready size recommendation system for shoes and address the challenge of providing recommendation for users. The product size recommendation problem is fairly recent with a only few studies proposed so far [1, 8, 9]. One approach recovers products and customers’ “true” sizes and uses these as features in a standard classifier for fit prediction [8]. In parallel to our work, another approach extends the above method and proposes Bayesian logit and probit regression models with ordinal categories to model fit [9]. Our approach differs from these studies in that we focus on capturing fit semantics and handle label imbalance issues using metric learning approaches with prototyping. Another recent approach uses skip-gram models to learn customers’ and products’ latent features [1]. However, this approach assumes the availability of platform-specific features whereas our model works on more limited (and thus more readily available) transaction data.

- According to R Misra, M Wan and J. McAuley Proposed a new predictive framework to tackle the product fit problem, which captures the semantics behind customers’ fit feedback, and employs a metric learning technique to resolve label imbalance issues. Metric learning has previously been applied to several recommendation problems. For music recommendation, one study encodes songs by their implicit feedback and employs metric learning to retrieve songs similar to a query song [6]. Another study uses users’ meta-data to build representations and employs metric learning to retrieve suitable partners for online dating recommendation [7]. In contrast, we learn representations of transactions that capture the ordinal nature of fit. Recently, collaborative filtering was combined with metric learning and its effectiveness was shown on various recommendation tasks [3]. However, the approach (that relies on binary implicit feedback data) does not translate directly to the ordinal nature of the product size recommendation problem.

- According to Romain Guigourès, Yuen King Ho, Evgenii Koriagin, Abdul-Saboor Sheikh, Urs Bergmann, and Reza Shirvany . Proposed a hierarchical Bayesian approach to tackle the challenging problem of size recommendation in e-commerce fashion. Our approach jointly models a size purchased by a customer, and its possible return event: 1. no return, 2. returned too small 3. returned too big. In this work, we focus on using customer reviews, purchases and returns data of the customers, product information in predicting the right fit for a customer. Customer reviews might contain crucial size and fit information which might help in predicting the right fit to a customer over using product and customer information alone. To extract the information from customer reviews we have embedded the reviews using a pre-trained language model [2].

- According to G Mohammed Abdulla and Sumit Borar, Proposed a size recommendation system to automatically pre-select consumer's size based on past purchase and content data without explicitly asking for users measurements. Many studies have used prototyping techniques with Nearest Neighbor based classification methods. Köstinger et al. [5] propose to jointly identify good prototypes and learn a distance metric, and also show that this leads to better generalization compared to k-Nearest Neighbor (k- NN) classification. Another study proposes a novel algorithm for deriving optimal prototypes, specifically for the 1-NN setting [11]. Following these lines, we develop a simple, fast and effective heuristic to choose relevant prototypes for Large Margin Nearest Neighbor (LMNN).

2.1 Major Issues

- The challenge for the MoD Cloth dataset is that there are many null values.
- Converting height from inches to centimeters (Object to float).

3. SYSTEM ANALYSIS AND DESIGN

3.1 PROPOSED SYSTEM

In the proposed system, the project will predict the size of a particular Mod Cloth product. It will take the input of different parameters like product id, height, waist, bust size, foot size etc.

Based on the issues with forecasting mistakes and the danger of overfitting brought on by huge datasets. The results of the data analysis and submission to the company are inefficient and ineffectual. In order to solve the issue, we can predict the size of particular product using the applicable Random Forest technique.

The Fit prediction is critical in order to improve customers shopping experiences. This makes it easier for both the Producers and the consumers to reduce product return rates. The model helps the customers to know their required product size / fit for a particular mod-cloth.

3.2 ARCHITECTURE DESIGN

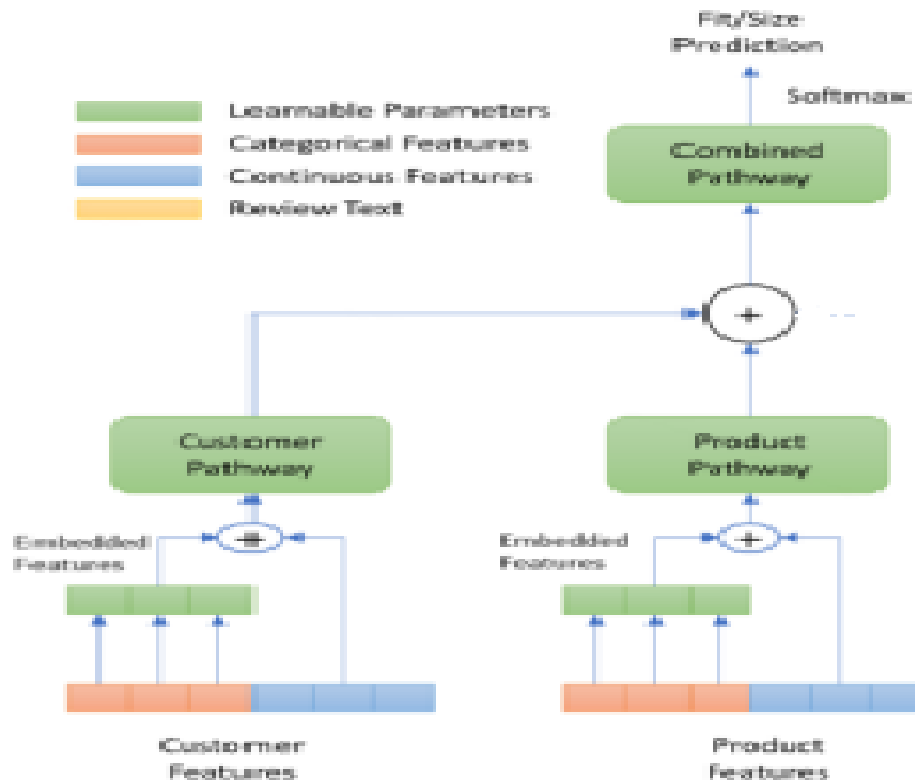


Figure 3.2.1: Architecture Design for size prediction model

3.3 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UM) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



Figure 3.3: Use case diagram to predict size.

4. IMPLEMENTATION

4.1 MODULES

1. Data Collection & Exploration
2. Data Preprocessing
3. Training the Model
4. Evaluating the Model
5. Results

4.2 PROPOSED ALGORITHM :

Based on the issues with overfitting risk brought on by huge datasets. The company's analysis of the data revealed that it was inefficient and unsuccessful. Thus, in order to solve the issue, we will predict the size of an MoD Cloth product for a particular customer using the Random Forest technique. The Fit prediction is critical in order to improve customers' shopping experiences. This makes it easier for both the Producers and the consumers to reduce product return rates. The model helps the customers to know their required product size/fit for a particular mod cloth. The system makes advantage of the hips, cup size, bust size, bust, height, fit, and shoe size. With the use of this dataset, an supervised learning model is built, and it is then used to predict the size by taking the different parameters as input. The project's suggested methodology is broken down into six steps:

1. Raw Data
2. Data PreProcessing
3. Data visualization
4. Testing Data
5. Predicted the size/fit using the Algorithm

6. Algorithm

4.2.1 RAW DATA:

The term raw data refers to unprocessed data that is taken from a dataset or, in some cases, created as an output of data processing. Extraction, organizing, and occasionally analysis are the procedures needed for raw data. The data set consists of 82790 records.

	item_id	waist	quality	cup size	hips	bra size	category	bust	height	user_name	length	fit	user_id	shoe size	shoe width	size
0	123373	29.0	5.0	d	38.0	34.0	new	36.0	5ft 6in	Emily	just right	small	991571	NaN	NaN	7
1	123373	31.0	3.0	b	30.0	36.0	new	NaN	5ft 2in	sydneybraden2001	just right	small	587883	NaN	NaN	13
2	123373	30.0	2.0	b	NaN	32.0	new	NaN	5ft 7in	Ugggh	slightly long	small	395665	9.0	NaN	7
3	123373	NaN	5.0	dd/e	NaN	NaN	new	NaN	NaN	alexmeyer626	just right	fit	875643	NaN	NaN	21
4	123373	NaN	5.0	b	NaN	36.0	new	NaN	5ft 2in	dberrones1	slightly long	small	944840	NaN	NaN	18

Figure 4.2.1: Raw Data of MoDCloth Dataset

4.2.2 DATA PREPROCESSING:

MoD Cloth will collect a significant amount of data for testing and training purposes. The acquired dataset will be aggregated and categorized based on the review of the customers. These datasets will undergo data pre-processing when these criteria are taken into account. The project used data from the Kaggle named ModCloth dataset and imported using python libraries by selecting the required fields in the required format. Then converted some of the fields to the required format. Here height is converted from foot inches to centimeters using regular expressions.


```
[ ] 1 df[['foot','inches']] = df['height'].str.split(expand = True)
    2 df['foot'] = df['foot'].str.extract('(\\d*)').astype(float) #Match 0 or more digits(\\d*)
    3 df['inches'] = df['inches'].str.extract('(\\d*)').astype(float)
    4 #Converting height into cm
    5 df['height'] = df['foot'].astype(float) * 30.48 + df['inches'].astype(float)*2.45
    6 del df['foot']
    7 del df['inches']
```

```
[ ] 1 df.head()
```

	item_id	waist	quality	cup size	hips	bra size	category	bust	height	user_name	length	fit	user_id	shoe size	shoe width	size
0	123373	29.0	5.0	d	38.0	34.0	new	36.0	167.10	Emily	just right	small	991571	NaN	NaN	7
1	123373	31.0	3.0	b	30.0	36.0	new	NaN	157.30	sydneybraden2001	just right	small	587883	NaN	NaN	13
2	123373	30.0	2.0	b	NaN	32.0	new	NaN	169.55	Ugggh	slightly long	small	395665	9.0	NaN	7
3	123373	NaN	5.0	dd/e	NaN	NaN	new	NaN	NaN	alexmeyer626	just right	fit	875643	NaN	NaN	21
4	123373	NaN	5.0	b	NaN	36.0	new	NaN	157.30	dberrones1	slightly long	small	944840	NaN	NaN	18

Figure 4.2.2: Data Preprocessing of ModCloth Data

4.2.3 DATA VISUALIZATION:

Data visualization is the process of using graphs and performance metrics to assess a model's performance. The major use of data visualization is to reclassify the data into new categories so that the algorithm used can be expanded to include an observation of each output variable that results from an observed input variable. And data cleaning is performed based on the visualization.

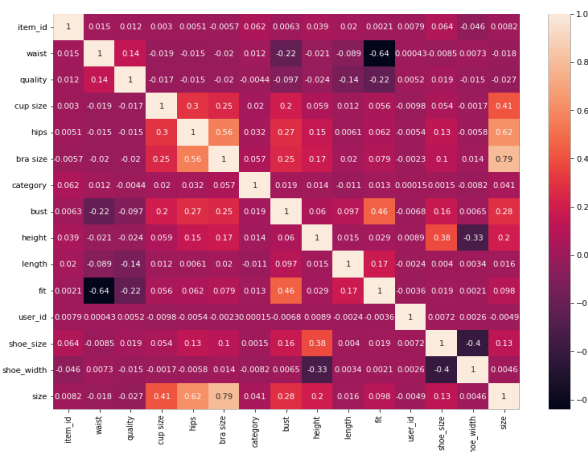


Figure 4.2.3: Data Visualization of Mod-Cloth Dataset

4.2.4 TESTING DATA:

Testing the data is the next important step after visualising the data in an algorithm; the test set may be defined as a set of used to assess the performance of a model using performance metrics. The programme that uses the test set must be able to generalise and perform well with the data set in order to accurately provide the expected data and produce a programme that is useful in nature. Furthermore, overfitting, which is when the computer memorises the data set, occurs when this happens. To balance overfitting, regularisation is used to apply to the model and reduce it.

```
[51] 1 x = df.values[:,0:9]
      2 y = df.values[:,9]
      3 x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25,random_state=42)

[52] 1 x_train.shape, x_test.shape, y_train.shape, y_test.shape

((54755, 9), (18252, 9), (54755,), (18252,))
```

Figure 4.2.4: Splitting the Data into training & testing data

4.2.5 PREDICTING THE SIZE/FIT USING ALGORITHM:

The proposed method predicts the size/fit of a particular MoD Cloth product by taking the parameters such as height, waist, bust, hip size, show size, etc. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. Based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

```
1 rf = RandomForestClassifier()
2 rf.fit(x_train,y_train)
3 rf_train_pred = rf.predict(x_train)
4 rf_test_pred = rf.predict(x_test)
5 print("Training accuracy : ",accuracy_score(y_train,rf_train_pred)*100)
6 print("Testing accuracy : ",accuracy_score(y_test, rf_test_pred)*100)
```

Training accuracy : 98.91516756460598
Testing accuracy : 68.91847468770546

Figure 4.2.5: Accuracy of the Random Forest Model

4.2.6 RANDOM FOREST:

The algorithm used in this study is based on supervised learning. In contrast to unsupervised learning, there will be labeled data for this grouping. Based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

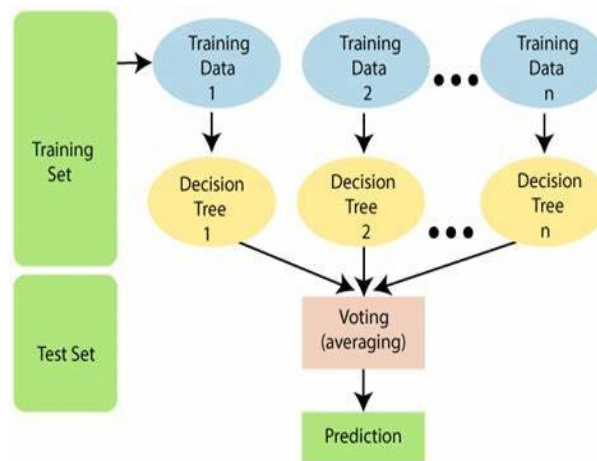


Figure 4.2.6.1: Random Forest

The algorithm is categorized into five steps:

- Select random K data points from the training set.
- Build the decision trees associated with the selected data points (Subsets).
- Choose the number N for the decision trees that you want to build.
- Repeat Steps 1 and 2.
- For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

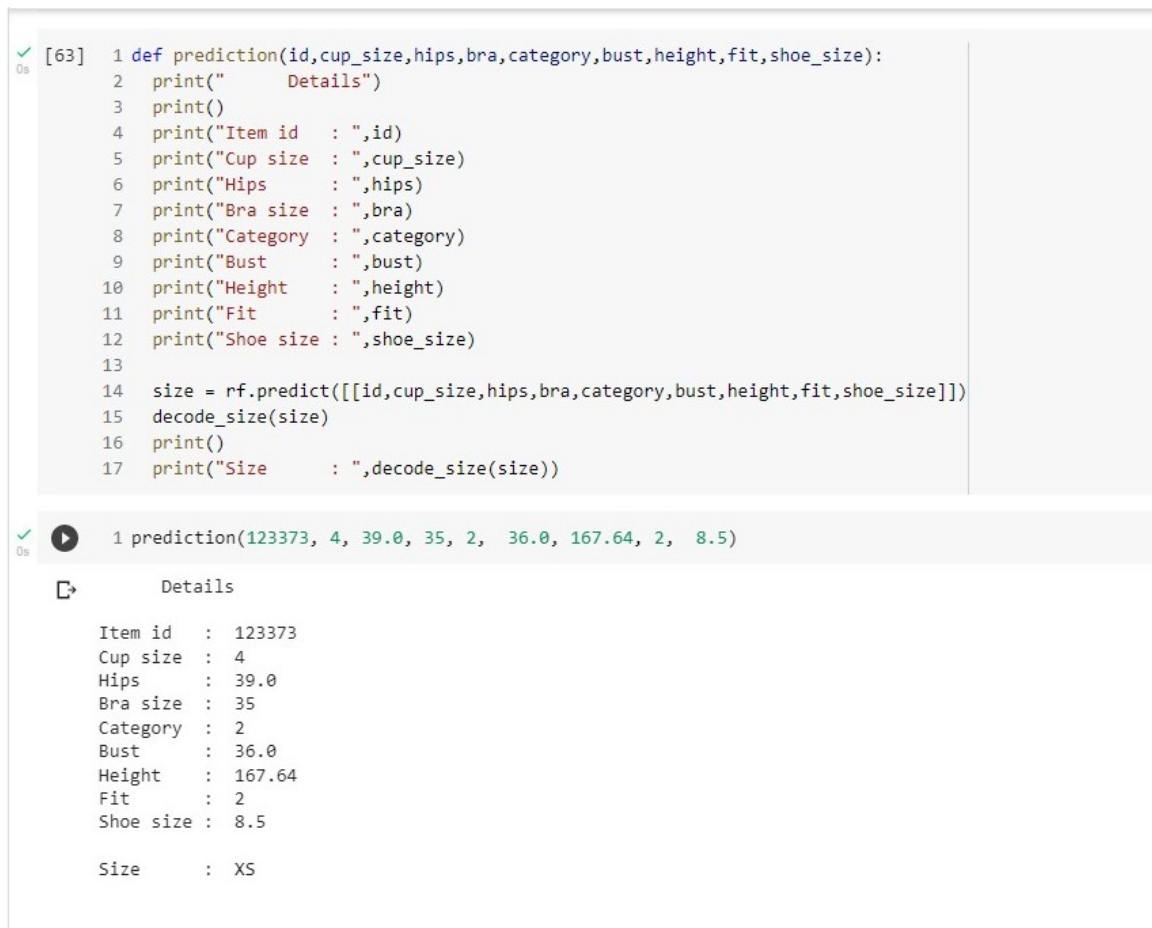
Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and predicts the final output.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output.

4.3 RESULTS:

The application makes a prediction of the size by taking the parameters such as height, waist, bust, hip size, show size, etc., using Random Forest to find the best size for the customer. The outcomes discussed are supported by the figures shown below.

Random Forest gives us the highest accuracy when compared to other algorithms. The model gives 98 % of training accuracy and 70 % testing accuracy. The model will predict the size by taking the parameters such as height, waist, bust, hip size, show size, etc.



```
[63] 1 def prediction(id,cup_size,hips,bra,category,bust,height,fit,shoe_size):
      2     print("      Details")
      3     print()
      4     print("Item id   : ",id)
      5     print("Cup size  : ",cup_size)
      6     print("Hips     : ",hips)
      7     print("Bra size  : ",bra)
      8     print("Category : ",category)
      9     print("Bust     : ",bust)
     10     print("Height   : ",height)
     11     print("Fit      : ",fit)
     12     print("Shoe size : ",shoe_size)
     13
     14     size = rf.predict([[id,cup_size,hips,bra,category,bust,height,fit,shoe_size]])
     15     decode_size(size)
     16     print()
     17     print("Size      : ",decode_size(size))
```

```
1 prediction(123373, 4, 39.0, 35, 2, 36.0, 167.64, 2, 8.5)
```

```
Details
Item id   : 123373
Cup size  : 4
Hips     : 39.0
Bra size  : 35
Category : 2
Bust     : 36.0
Height   : 167.64
Fit      : 2
Shoe size : 8.5

Size      : XS
```

Figure 4.3: Predicting the Size with user inputs

5. CONCLUSION AND FUTURE SCOPE

The project's conclusion is Size and Fit recommendation is an important problem in e-commerce because it helps customers in choosing the right fit and thereby reduces size and fit-related returns. Most of the earlier works embedded user and product information to predict the right fit. In this paper, we propose a machine learning-based approach to predict the fit based on customer reviews along with customer and product information. Through extensive experimentation on different datasets curated from data of one of the largest e-commerce platforms, we show the effectiveness of our approach.

We plan to extend this work by including other customer feedback like customer uploaded images and customer QnA to predict the right fit. Customers sometimes ask questions related to size and fit which get answered by other customers who have already bought the product. Also using customer-uploaded images along with product information might help the model to learn the interactions between the user and product to predict the right fit.

REFERENCES

- [1] V Sembium, R Rastogi, L Tekumalla et al., "Bayesian models for product size recommendations[C]", Proceedings of the 2020 World Wide Web Conference. International World Wide Web Conferences Steering Committee, pp. 679-687, 2020.
- [2] R Misra, M Wan and J. McAuley, "Decomposing fit semantics for product size recommendation in metric spaces[C]", Proceedings of the 12th ACM Conference on Recommender Systems, pp. 422-426, 2019.
- [3] Romain Guigourès, Yuen King Ho, Evgenii Koriagin, Abdul-Saboor Sheikh, Urs Bergmann, and Reza Shirvany. 2018. A hierarchical bayesian model for size recommendation in fashion. In Proceedings of the 12th ACM Conference on Recommender Systems. ACM, 392–396.
- [4] A. Chan, J. Fan, and W. Yu, "Prediction of men's shirt pattern based on 3d body measurements," International Journal of Clothing Science and Technology, vol. 17, no. 2, pp. 100–108, 2018.
- [5] G Mohammed Abdulla and Sumit Borar. 2017. Size recommendation system for fashion e-commerce. In KDD Workshop on Machine Learning Meets Fashion.