**CS552 – Data Science with Python Assignment-2**

**Cumhuriye Tüluğ Küçüköğüt**
**S025791**

## Abstract

Clothes are a part of people's lives and much of the marketing depends on them. Shopping sites use natural language processing, machine learning, cloud computing, etc. uses computer science. Classification in machine learning is widely used in real-life product categorization problems. In this classification method, product names and images are generally used.

This project is a study to find the label of the images, to find out that the clothes belong to their real labels. SVM, KNN and RandomForestGenerator algorithms are used to provide this classification.For this study Fashion-MNIST data is used.

## Introduction

E-commerce sites are getting bigger day by day so used technologies are improved too.According to researchers[5] recognizing product categories become more important nowadays.This project aims to categorize clothes according to labels using different types of classification algorihms and Fashion-MNIST data.

Firstly data are collected and divided into train and test, categorization is done via KNN,SVM and RandomForestGenerator.At the final part of the report, comparision was provided with statistical methods.

In section 2, algorihms are summarized which has been used in the work. In section 3 experimental setup is explained. In section 4 dataset and predictions are told finally conclusions are represented in section 5.

## Background

Fashion-MNIST is used to classify clothes. Fashion-MNIST consists of a total of 70000 samples divided into 60000 training and 10000 test data. Images are in 10 types and their scale is 28x28. Classification algorithms were used to provide this study and are described below.

## Classification Algorithms

### SVM

This algorithm is used both regression and classification tasks. Given a set of training examples, each marked as belonging to one or the other of both categories, a SVM training algorithm creates a model that assigns new examples to one category or the other, turning it into a nonprobabilistic binary linear classifier. For seperating two classes, hyperplanes are used. This algorithm’s aim is to find this plane and this plane should have the maximum margin for more confidential results[1]

### KNN

It is based on the estimation of the class of the vector formed by the independent variables of the value to be estimated based on the information in which class the nearest neighbors are dense. The KNN (K-Nearest Neighbors) Algorithm makes predictions based on two basic values; Distance and number of neighbors.[2]

### Random Forrest Classifier

Random Forest algorithm is used both classification and regression.Random forests build decision trees on randomly selected data samples, take predictions from each tree, and choose the best solution by voting.[3]

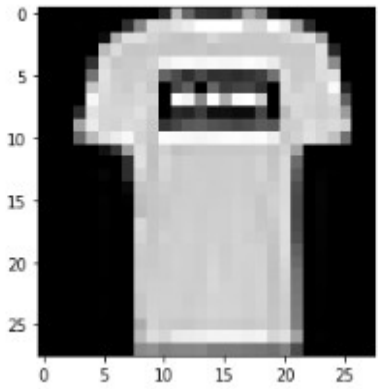
## Experimental Details and Design

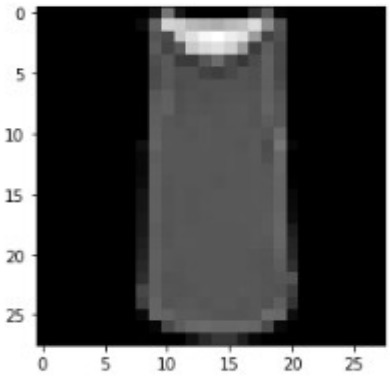
Firstly Fashion-MNIST dataset is used in this study and they are divided into train and test. Fashion-MNIST includes some clothes images and labels. Since the labels contain numerical data, these numbers were first matched with the clothing names.These are explained in Table 1.

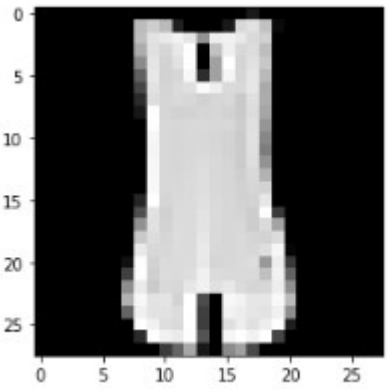
| | |
|---|---|
| 0 | tshirt |
| 1 | pants |
| 2 | sweater |
| 3 | dress |
| 4 | coat |
| 5 | sandals |
| 6 | shirt |
| 7 | sneaker |
| 8 | bag |
| 9 | ankle boat |

Table 1

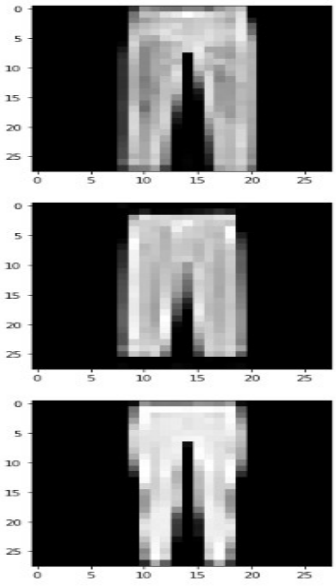
Some fashion mnist data are shown below.
Thirt



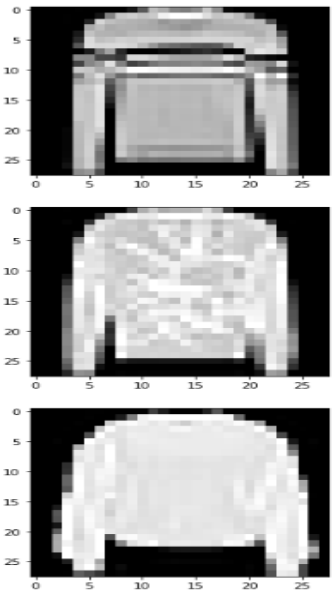




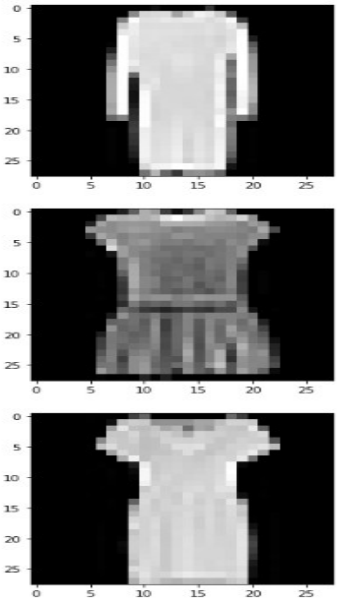
Pants



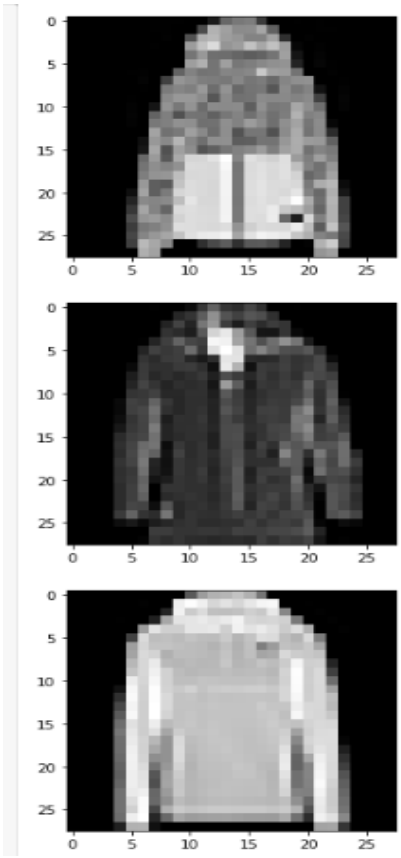
Sweater



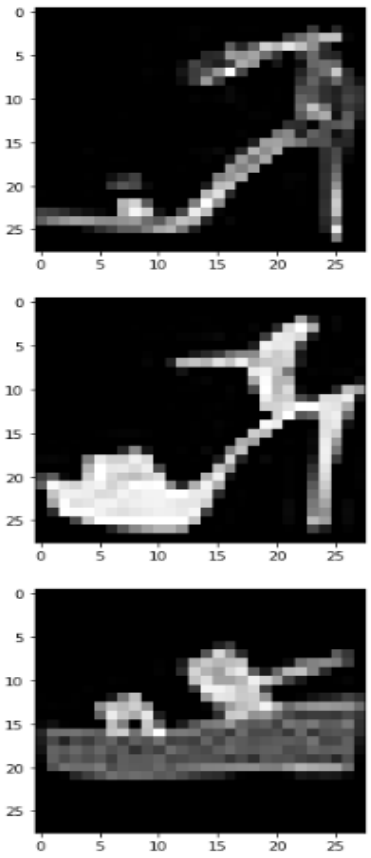
Dress



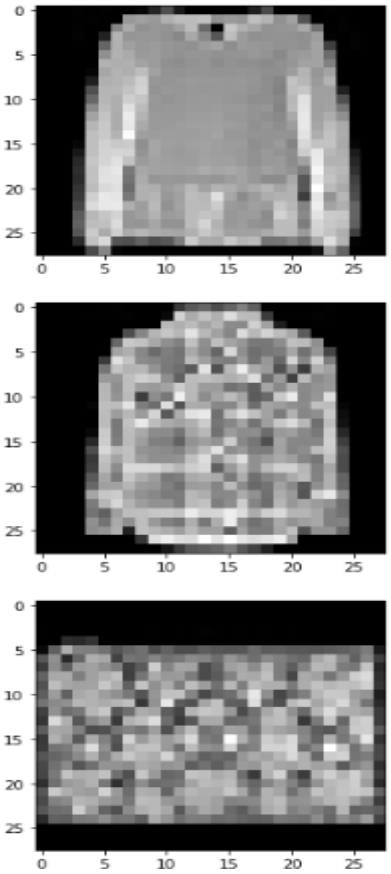
Coat



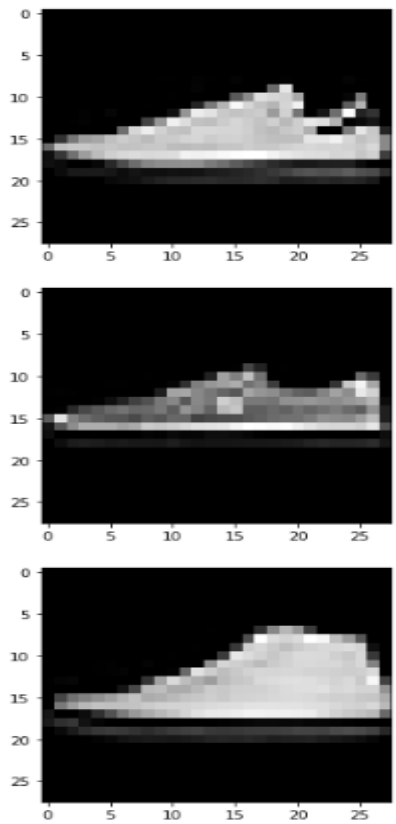
Sandals



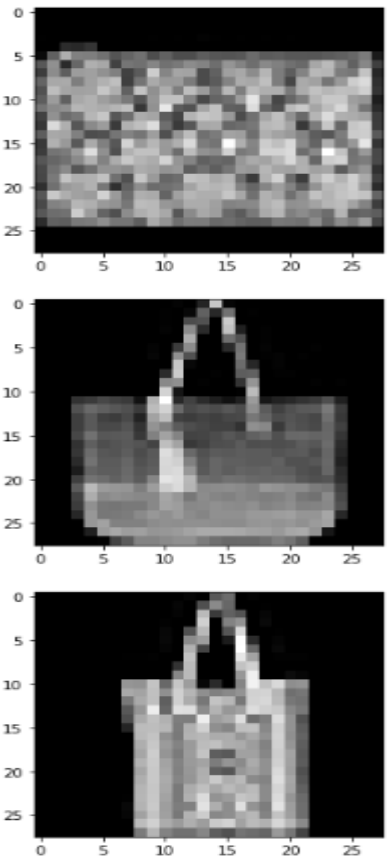
Shirts



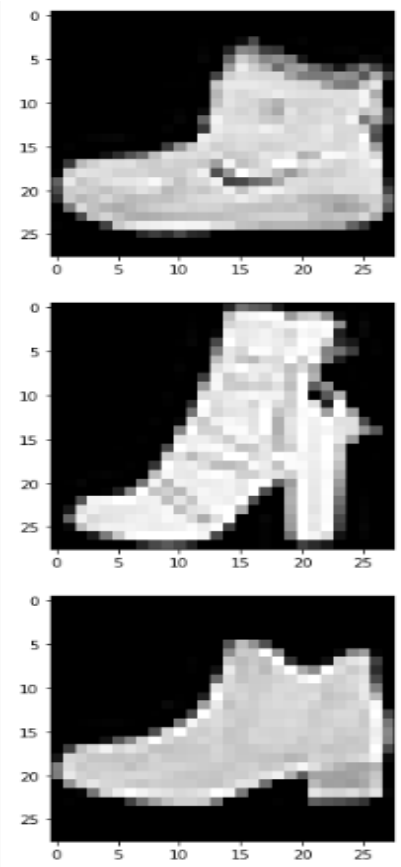
Sneakers



Bag



Ankle Boat



After visualizing some clothes, data is scaling with standard scaler and firstly knn algorithm is implemented. For this algorithm different neighbour numbers are used and according to scores best option is 5. Results are shown in Table 2.

| n_neighbour | score |
|---|---|
| 5 | 0.8536 |
| 7 | 0.8525 |
| 3 | 0.8499 |

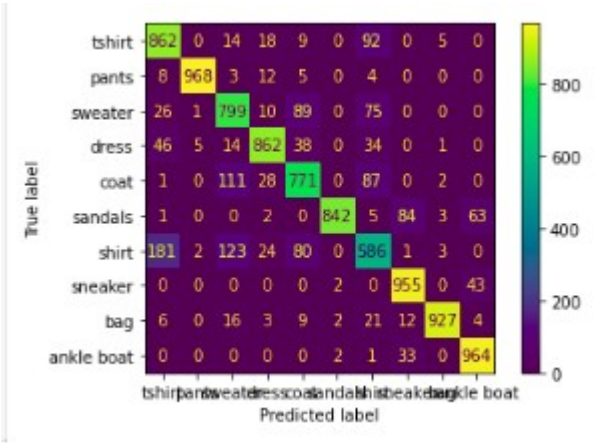
Table 2

Table 3 shows the algorihm's performance.

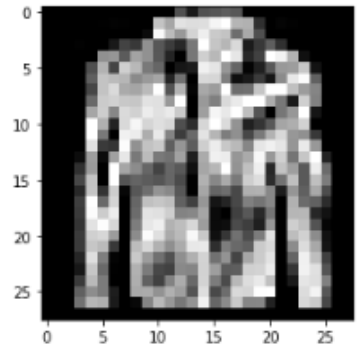
| | |
|---|---|
| Accuracy | 0.8536 |
| Recall | 0.8536 |
| Precision | 0.8536 |
| F1 | 0.8536 |

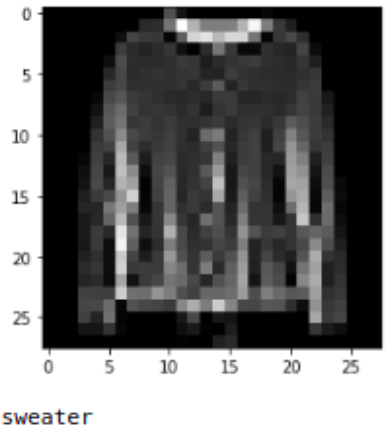
Table 3

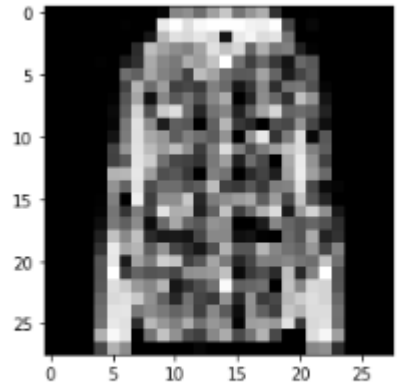
Confusion matrix is shown below.



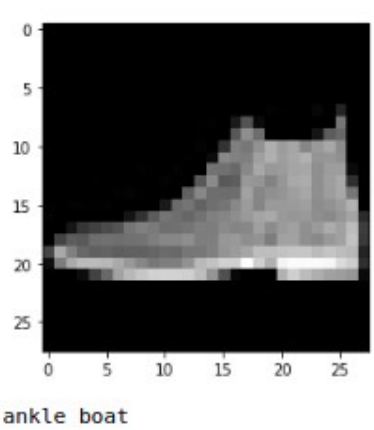

According the results 1464 unmatched results are found. Some unmatched results are shown below.

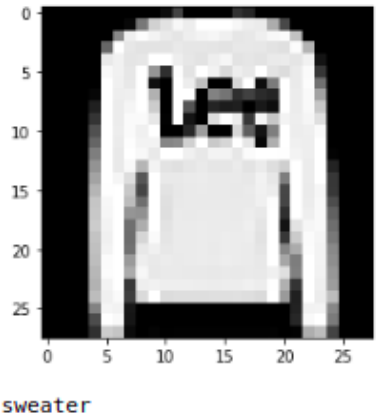





According the results matched results are found. Some matched results are shown below.

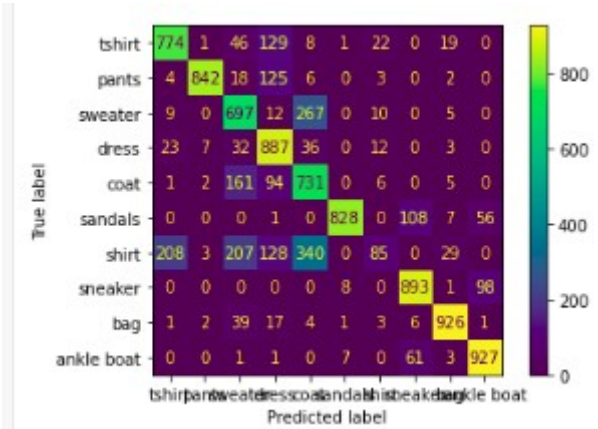


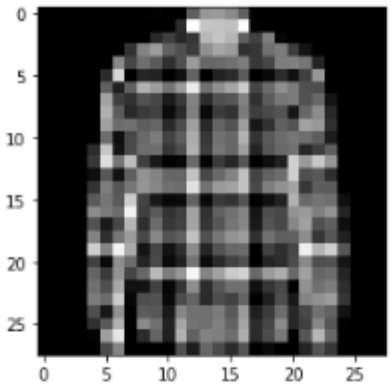

Secondly RandomForestGenerator classifier algorithm is used and results are shown below.Score is 0.759.

| | |
|---|---|
| Accuracy | 0.759 |
| Recall | 0.759 |
| Precision | 0.759 |
| F1 | 0.759 |

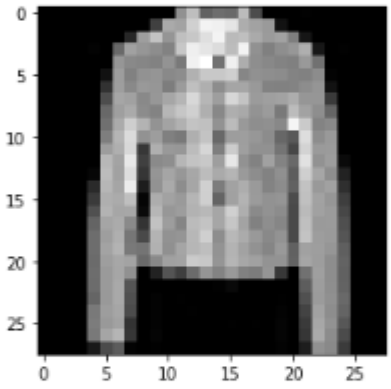
Confusion Matrix:



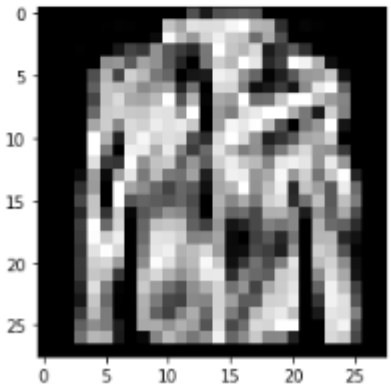
According the results 2410 unmatched results are found. Some unmatched results are shown below.



coat

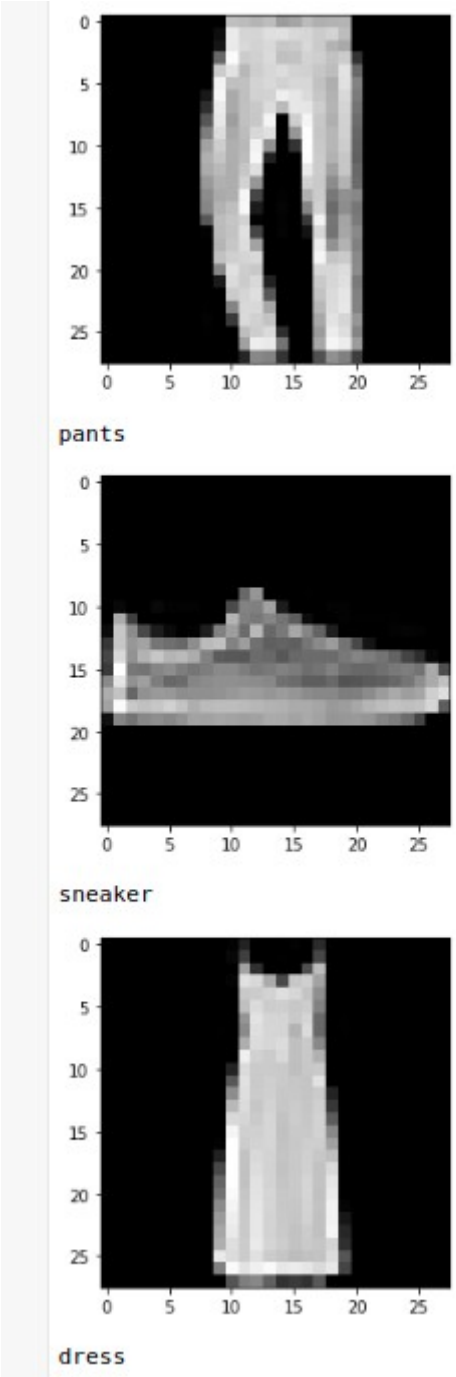


sweater



sweater

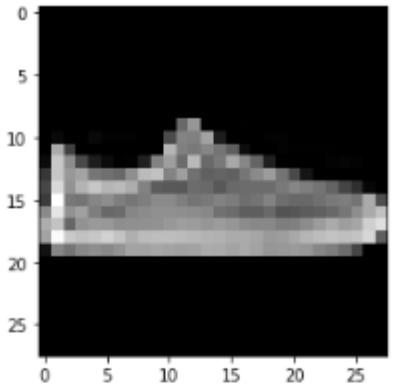
According the results matched results are found. Some matched results are shown below.



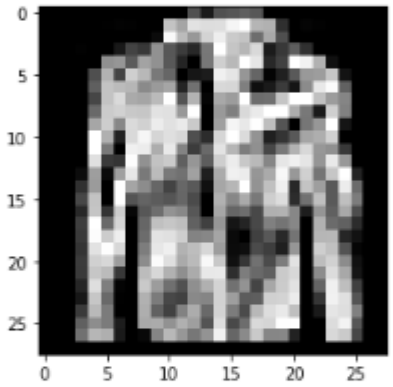

Finally SVM algorithm is used and results are shown below. Score is found 0.8454

| | |
|---|---|
| Accuracy | 0.8454 |
| Recall | 0.8454 |
| Precision | 0.8454 |
| F1 | 0.8454 |

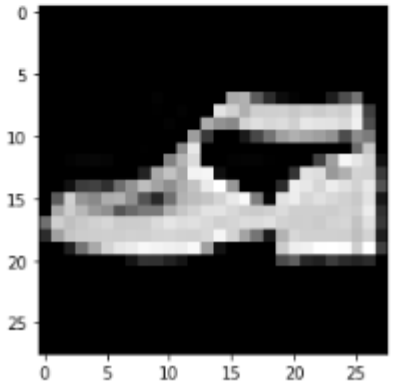
According the results 1546 unmatched results are found. Some unmatched results are shown below.



```
sandals
```

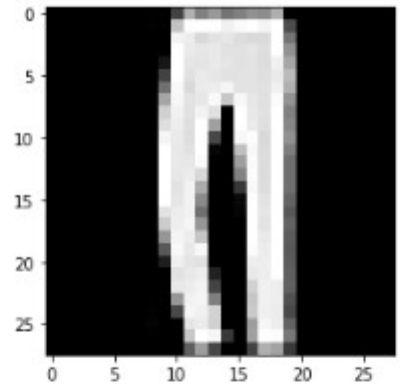


```
sweater
```

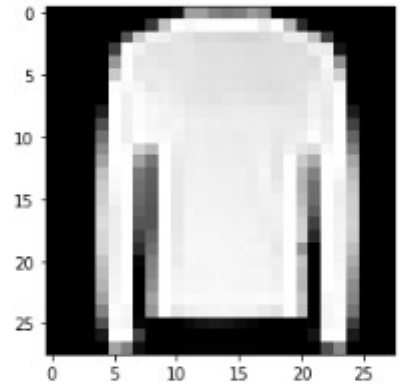


```
sandals
```

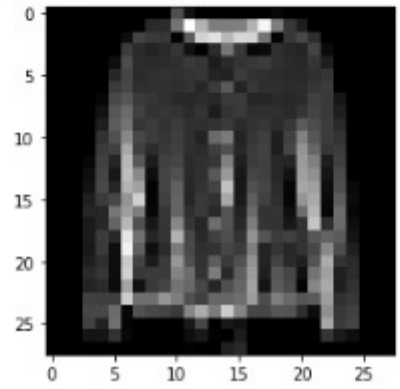
Some of correct results are shown.



pants



sweater



sweater


1)https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
2)https://medium.com/@arslanev/makine-%C3%B6%C4%9Frenmesi-knn-k-nearest-neighbors-algoritmas%C4%B1-bdfb688d7c5f
3)https://www.datacamp.com/community/tutorials/random-forests-classifier-python

5)https://sigir-ecom.github.io/ecom18DCPapers/ecom18DC_paper_8.pdf