OPTIMIZING THE PROVISION OF STORAGE RESOURCES IN CLOUD COMPUTING USING MACHINE LEARNING

A Project Work
submitted in partial fulfillment of the
requirements for the degree of

Bachelor of Technology
in
Computer Science and Engineering

Submitted By
Mohit Gupta
1513105096

Under the supervision of
Mr. A.Daniel
Assistant Professor

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA – 201306

GALGOTIAS UNIVERSITY

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **"OPTIMIZING THE PROVISION OF STORAGE RESOURCES IN CLOUD COMPUTING USING MACHINE LEARNING"** is the bonafide work of "**MOHIT GUPTA(1513105096)"** who carried out the project work under my supervision.

SIGNATURE OF SUPERVISOR
Mr. A.Daniel
Assistant Professor
School of Computing Science & Engineering

SIGNATURE OF HEAD
Dr. MUNISH SHABARWAL,
PhD (Management), PhD (CS)
Professor & Dean,
School of Computing Science & Engineering

ABSTRACT

Cloud computing is a very popular field at present which is growing very fast and the future of the field seems really wide. With progressive spotlight on cloud computing as a possible solution for a flexible, on-demand computing infrastructure for lots of applications, many companies and unions have started using it. Obviously, cloud computing has been recognized as a model for supporting infrastructure, platform and software services. Within cloud systems, massive distributed data center infrastructure, virtualized physical resources, virtualized middleware platform such as VMware as well as applications are all being provided and consumed as services. The cloud clients should get good and reliable services from a provider and the provider should allocate the resources in a proper way so as to render good services to a client. This brings about the problem of optimization where clients request for more services than they actually require leading to wastage of the cloud storage resource. This demands for optimization both on the part of the client and the cloud service provider. This has led to increased research in the various techniques that can be used for resource allocation within cloud services. This project focuses on the analysis of machine learning as a technique that can be used to predict the cloud storage service request patterns from the clients and therefore optimize the user storage resource demand and usage for the case of cloud computing storage IaaS. Data on cloud storage resource usage was subjected to experiments using machine learning techniques so as to determine which give the most accurate prediction.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

OVERALL DESCRIPTION-

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources such as networks, servers, storage, applications, and services that can be rapidly provisioned and released with minimal management effort or service provider interaction (NIST).

Cloud computing systems provide environments to enables resource provision in terms of scalable infrastructures, middleware, application development platforms and value-added business applications. Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) are three basic service layers.

SaaS: The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings (NIST).

PaaS: The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application- hosting environment (NIST).

IaaS: The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (NIST).

Various resources are available within the cloud computing environment. Resource consumption by an application in the form of CPU time, disk space, amount of memory and network bandwidth is a useful set of information when available before allocating resources. The need to know resource consumption has the benefit of helping cloud service providers in resource optimization.

In cloud computing, Resource Allocation is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module. An optimal Resource Allocation Strategy should avoid the following criteria i.e. Resource Contention in which demand exceeds supply for a shared resource, such as memory, CPU, network or storage. In modern IT, where cost cuts are the norm, addressing resource contention is a top priority. The main concern with resource contention is the performance degradation that occurs as a result. Second Criteria is Scarcity of Resource which happens when there are limited resources and the demand for resources is high. In such situation user cannot avail facility of resource. Third criteria are Resource Fragmentation –In these criteria resources are isolated. There would be enough resources but cannot allocate it to the needed application due to fragmentation into small entities. If fragmentation is done into big entities then we can use it optimum. Forth criteria is Over Provisioning – Over provisioning arises when the application gets surplus resources than the demanded one. Due to this, the investment is high and revenue is low. The fifth criterion is Under Provisioning, which occurs when the application is assigned with fewer numbers of resources than it demanded.

PURPOSE-

Over provisioning arises when the application gets surplus resources than the demanded ones. As more companies put workloads on Amazon Web Services or other public cloud platforms, many are paying for more cloud than they need. That over provisioning is the problem. Over provisioning arises when a client requests for much more resources (processor, RAM and storage) than what they actually use. This is a problem for organizations offering IaaS in the sense that some resources go unused. It is also a problem for the clients because they unknowingly end up paying for more processor and storage resources than they use.

MOTIVATIONS AND SCOPE

This project is only limited to optimize the provisioning of only one of the IAAS resources – storage using Machine Learning. The main scope of this project is to optimize the provisioning of the storage resources so that less resources get unused. This will also enable savings on the part of the customer since they can be accurately advised on their specific resource requirements hence more informed resource requests on their part, and on the side of the cloud service provider, they can be able to optimally plan for the resources available. The machine learning algorithm which is being used is The Support Vector Machine as it is one of the most powerful machine learning algorithms that can perform well on data sets that have many attributes, even if there are very few cases on which to train the model.

LITERATURE SURVEY

INTRODUCTION

Cloud computing offers three main services which are PaaS, SaaS and IaaS. There is lack of an appropriate machine learning technique that can be used to predict storage resource consumption for the case of IaaS cloud computing service. Machine learning can be used to predict these storage resource usages based on previous client usage patterns. Cloud computing can be offered using various deployment models and there are various players involved in offering cloud computing service. Various resource allocation models are reviewed so as to get an overview of the various resources and the methods that are used to allocate these resources.

The benefits of this include a better awareness on storage resource usage for future planning on the part of the organizations Cloud computing services and hence saving on costs as well as better client advisory on future requirements when they request for storage service requests thereby saving clients from unnecessary additional costs. The poor performance results produced by statistical estimation models have flooded the estimation area for over the last decade. Their inability to handle categorical data, cope with missing data points, spread of data points and most importantly lack of reasoning capabilities has triggered an increase in the number of studies using non-traditional methods like machine learning techniques. The area of machine learning draws on concepts from diverse fields such as statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity and control theory. In this case the researcher picked on Safaricom limited as a local cloud computing service provider and corporate organizations it provides these services to as the clients for the case of IaaS. If not well managed, a cloud service provider may end up under-utilizing the available resources on his part and the cloud client may end up paying for more services than they actually require. This can be both detrimental to the client in terms of unnecessary cost and to the service provider in terms of resourcewastages.

CLOUD COMPUTING

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics (On-demand self-service, Broad network access, Resource pooling, Rapid elasticity, Measured Service); three service models (Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS), Cloud Infrastructure as a Service (IaaS)); and, four deployment models (Private cloud, Community cloud, Public cloud, Hybrid cloud). Key enabling technologies include: (1) fast wide-area networks, (2) powerful, inexpensive server computers, and (3) high-performance virtualization for commodity hardware (NIST).

Cloud Computing Service Models

Software as a Service (Saas)

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings (NIST).

Platform as a Service (PaaS)

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment (NIST).

Infrastructure as a Service (IaaS)

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (NIST).

Benefits of Cloud Computing

The advantages of - “renting” these - “virtual” resources over traditional on-premise IT includes:

• On demand and elastic services—quickly scale up or down.

• Self-service, automated provisioning and de-provisioning.

• Reduced costs from economies of scale and resource pooling.

• Pay-for-use—costs based on metered service usage.

Cloud Deployment Models

Public Cloud

The deployment of a public cloud computing system is characterized on the one hand by the public availability of the cloud service offering and on the other hand by the public network that is used to communicate with the cloud service. The cloud services and cloud resources are procured from very large resource pools that are shared by all end users. These IT factories, which tend to be specifically built for running cloud computing systems, provision the resources precisely according to required quantities. Byoptimizing operation, support, and maintenance, the cloud provider can achieve significant economies of scale, leading to low prices for cloud resources. In addition, public cloud portfolios employ techniques for resource optimization; however, these are transparent for end users and represent a potential threat to the security of the system

If a cloud provider runs several datacenters, for instance, resources can be assigned in such a way that the load is uniformly distributed between all centers. Some of the best-known examples of public cloud systems are Amazon Web Services (AWS) containing the Elastic Compute Cloud (EC2) and the Simple Storage Service (S3) which form an IaaS cloud offering and the Google App Engine with provides a PaaS to its customers. The customer relationship management (CRM) solution Salesforce.com is the best-known example in the area of SaaS cloud offerings.

Private Cloud

Private cloud computing systems emulate public cloud service offerings within an organization's boundaries to make services accessible for one designated organization. Private cloud computing systems make use of virtualization solutions and focus on consolidating distributed IT services often within data centers belonging to the company. The chief advantage of these systems is that the enterprise retains full control over corporate data, security guidelines, and system performance. In contrast, private cloud offerings are usually not as large-scale as public cloud offerings resulting in worse economies of scale.

Community Cloud

In a community cloud, organizations with similar requirements share a cloud infrastructure. It may be understood as a generalization of a private cloud, a private cloud being an infrastructure which is only accessible by one certain organization.

Hybrid Cloud

A hybrid cloud service deployment model implements the required processes by combining the\ cloud services of different cloud computing systems, e.g. private and public cloud services. The hybrid model is also suitable for enterprises in which the transition to full outsourcing has already been completed, for instance, to combine community cloud services with public cloud services.

Cloud Environment Roles

In cloud environments, individual roles can be identified similar to the typical role distribution in Service Oriented Architectures and in particular in (business oriented) Virtual Organizations. As the roles relate strongly to the individual business models it is imperative to have a clear definition of the types of roles involved in order to ensure common understanding.

Cloud Providers offer clouds to the customer – either via dedicated APIs (PaaS), virtual machines and / or direct access to the resources (IaaS). Hosts of cloud enhanced services (SaaS) are typically referred to as Service Providers, though there may be ambiguity between the terms Service Provider and Cloud Provider.

Cloud Resellers aggregate cloud platforms from cloud providers to either provide a larger resource infrastructure to their customers or to provide enhanced features. This relates to community clouds in so far as the cloud aggregators may expose a single interface to a merged cloud infrastructure. They will match the economic benefits of global cloud infrastructures with the understanding of local customer needs by providing highly customized, enhanced offerings to local companies (especially SME's) and world-class applications in important European industry sectors. Similar to the software and consulting industry, the creation of European cloud partner ecosystems will provide significant economic opportunities in the application domain – first, by mapping emerging industry requests into innovative solutions and second by utilizing these innovative solutions by European companies in the global marketplace.

Cloud Adopters or (Software / Services) Vendors enhance their own services and capabilities by exploiting cloud platforms from cloud providers or cloud resellers. This enables them to e.g. provide services that scale to dynamic demands – in particular new business entries who cannot estimate the uptake / demand of their services as yet. The cloud enhanced services thus effectively become software as a service.

Cloud Consumers or Users make direct use of the cloud capabilities – as opposed to cloud resellers and cloud adopters, however, not to improve the services and capabilities they offer, but to make use of the direct results, i.e. either to execute complex computations or to host a flexible data set. Note that this involves in particular larger enterprises which outsource their in-house infrastructure to reduce cost and efforts.

Cloud Tool Providers do not actually provide cloud capabilities, but supporting tools such as programming environments, virtual machine management etc.

Cloud Auditor - A third-party (often accredited) that conducts independent assessments of cloud environments assumes the role of the cloud auditor. The typical responsibilities associated with this role include the evaluation of security controls, privacy impacts, and performance. The main purpose of the cloud auditor role is to provide an unbiased assessment (and possible endorsement) of a cloud environment to help strengthen the trust relationship between cloud consumers and cloud providers.

Cloud Broker - This role is assumed by a party that assumes the responsibility of managing and negotiating the usage of cloud services between cloud consumers and cloud providers. Mediation services provided by cloud brokers include service intermediation, aggregation, and arbitrage.

Cloud Carrier - The party responsible for providing the wire-level connectivity between cloud consumers and cloud providers assumes the role of the cloud carrier. This role is often assumed by network and telecommunication providers (NIST).

Cloud Characteristics

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. The essential characteristics for cloud computing include the ones highlighted below.

On-demand self-service – A consumer can unilaterally provision computingcapabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Broad network access – Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

Resource pooling – The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid elasticity – Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Machine Learning

Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests.

Machine Learning is concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A major focus of Machine Learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too complex to describe generally in programming languages, so that in effect programs must automatically describe programs.

Applications of machine learning

In recent years many successful machine learning applications have been developed, ranging from data-mining programs that learn to detect fraudulent credit card transactions, to information-filtering systems that learn users' reading preferences, to autonomous vehicles that learn to drive on public highways. At the same time, there have been important advances in the theory and algorithms that form the foundations of this field.

The poor performance results produced by statistical estimation models have flooded the estimation area for over the last decade. Their inability to handle categorical data, cope with missing data points, spread of data points and most importantly lack of reasoning capabilities has triggered an increase in the number of studies using non-traditional methods like machine learning techniques.

Types of Machine Learning

There are two main types of Machine Learning algorithms. In this project, supervised learning is adopted here to build models from raw data and perform regression and classification.

Supervised learning: Supervised Learning is a machine learning paradigm for acquiring the input-output relationship information of a system based on a given set of paired input-output training samples. As the output is regarded as the label of the input data or the supervision, an input-output training sample is also called labeled training data, or supervised data. Learning from Labeled Data, or Inductive Machine Learning. The goal of supervised learning is to build an artificial system that can learn the mapping between the input and the output, and can predict the output of the system given new inputs. If the output takes a finite set of discrete values that indicate the class labels of the input, the learned mapping leads to the classification of the input data. If the output takes continuous values, it leads to a regression of the input. It deduces a function from training data that maps inputs to the expected outcomes. The output of the function can be a predicted continuous value (called regression), or a predicted class label from a discrete set for the input object (called classification). The goal of the supervised learner is to predict the value of the function for any valid input object from a number of training examples. The most widely used classifiers are the Neural Network (Multilayer perceptron), Support Vector Machines, k-nearest neighbor algorithm, Regression Analysis, Artificial neural networks and time series analysis.

Unsupervised learning: Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with supervised learning or reinforcement learning, there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output.

Random Forest Algorithm

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.

How it works:

Random Forest is a supervised learning algorithm. Like you can already see from it's name, it creates a forest and makes it somehow random. The „forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Why Random forest algorithm

- The same random forest algorithm or the random forest classifier can use for both classification and the regression task.
- Random forest classifier will handle the missing values.
- When we have more trees in the forest, random forest classifier won't overfit the model.
- Can model the random forest classifier for categorical values also.

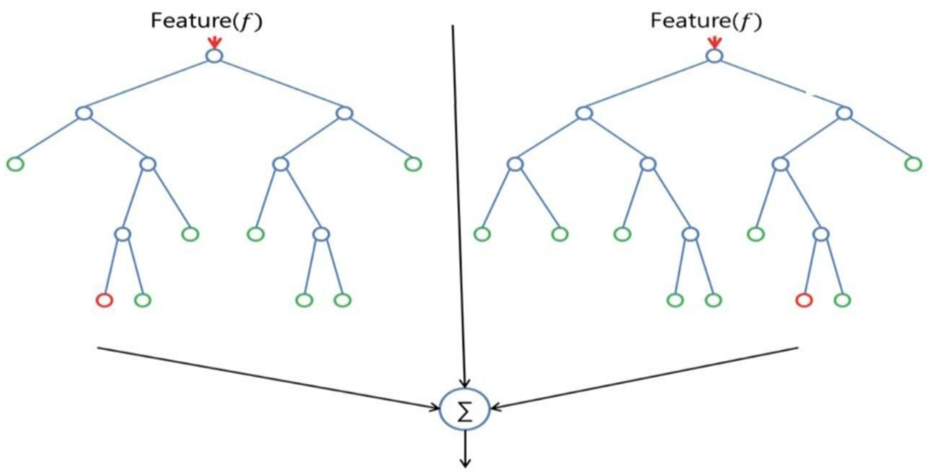


Fig.1 Random Forest With Two Trees

Random forest algorithm real life example

Suppose Mady somehow got 2 weeks leave from his office. He wants to spend his 2 weeks by traveling to the different place. He also wants to go to the place he may like.

So he decided to ask his best friend about the places he may like. Then his friend started asking about his past trips. It’s just like his best friend will ask, You have been visited the X place did you like it?

Based on the answers which are given by Mady, his best start recommending the place Mady may like. Here his best formed the decision tree with the answer given by Mady.

As his best friend may recommend his best place to Mady as a friend. The model will be biased with the closeness of their friendship. So he decided to ask few more friends to recommend the best place he may like.

Now his friends asked some random questions and each one recommended one place to Mady. Now Mady considered the place which is high votes from his friends as the final place to visit.

How Random forest algorithm works

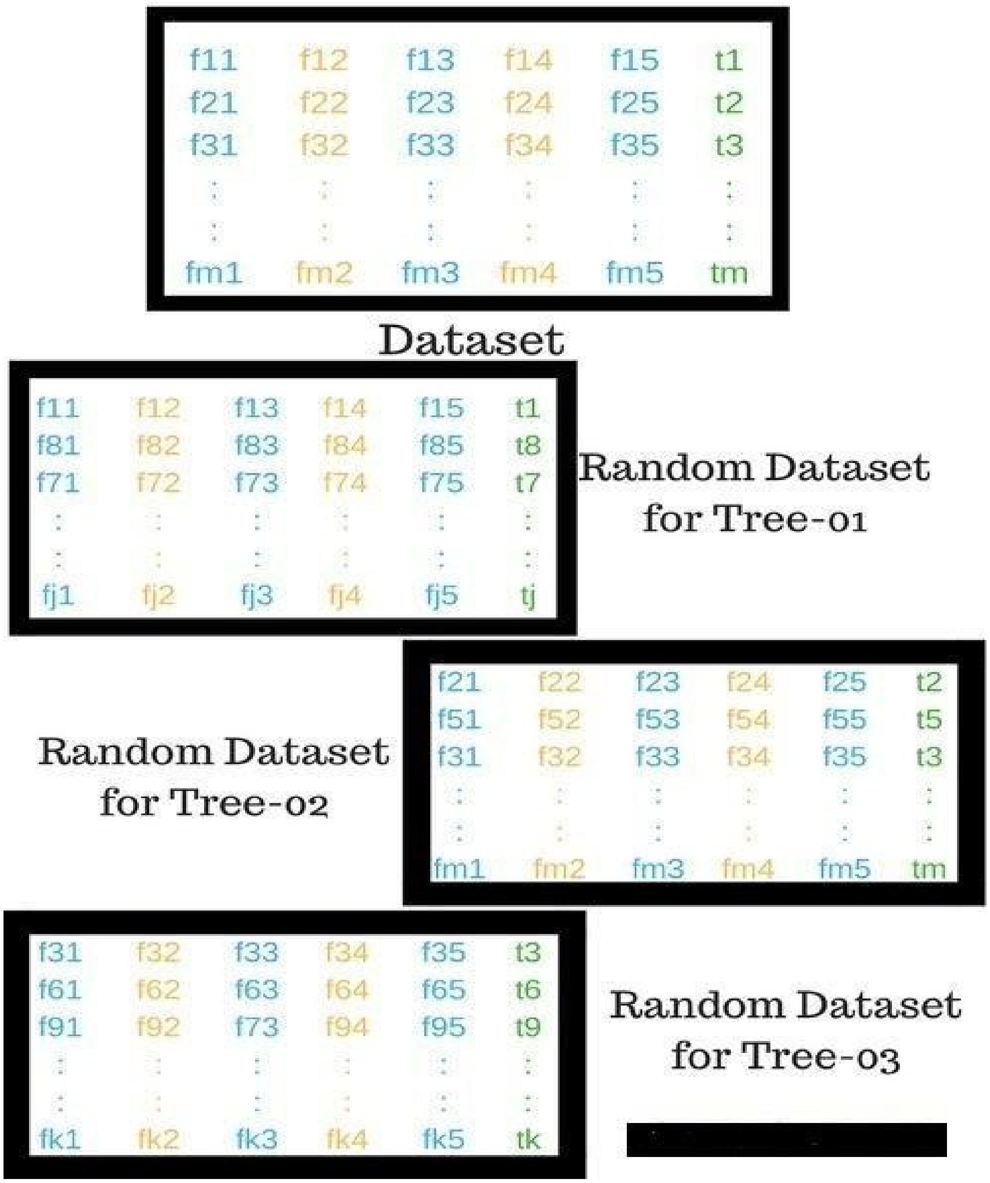


Fig.2

Let’s look at the pseudocode for random forest algorithm and later we can walk through each step in the random forest algorithm.

The pseudocode for random forest algorithm can split into two stages.

- Random forest creation pseudocode.
- Pseudocode to perform prediction from the created random forest classifier.

Random Forest pseudocode:

1. Randomly select “k” features from total “m” features.
    1. Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

In the next stage, we are using the randomly selected “k” features to find the root node by using the best splitapproach.

The next stage, We will be calculating the daughter nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node.

Finally, we repeat 1 to 4 stages to create “n” randomly created trees. This randomly created trees forms the random forest.

Random forest prediction pseudocode:

To perform prediction using the trained random forest algorithm uses the below pseudocode.

1. Takes the test features and use the rules of each randomly created decision tree to predict the oucome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm we need to pass the test features through the rules of each randomly created trees. Suppose let's say we formed 100 random decision trees to from the random forest.

Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target votes will be calculated. Suppose the 100 random decision trees are prediction some 3 unique targets x, y, z then the votes of x is nothing but out of 100 random decision tree how many trees prediction is x.

Likewise for other 2 targets (y, z). If x is getting high votes. Let's say out of 100 random decision tree 60 trees are predicting the target will be x. Then the final random forest returns the x as the predicted target.

Random forest algorithm applications

Fig.3

Below are some the application where random forest algorithm is widely used.

1. Banking
2. Medicine
3. Stock Market
4. E-commerce

Proposed model

System Architecture

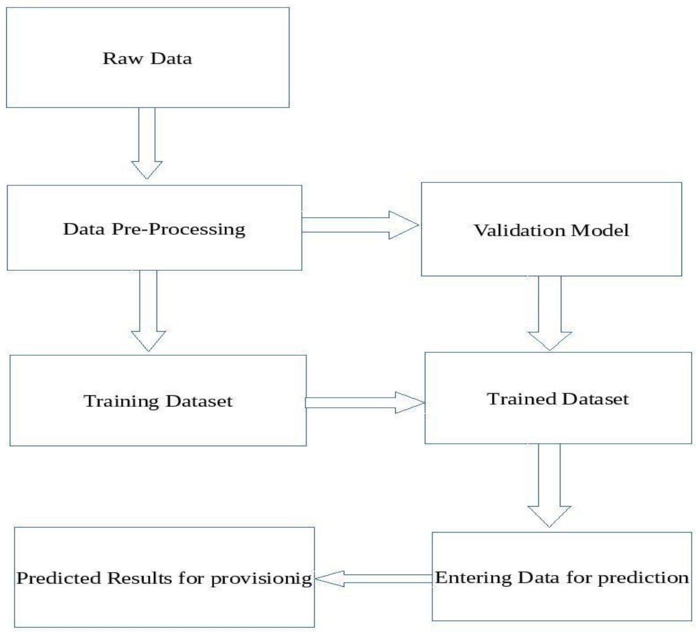


Fig.4

Selection of variables/Type of data required

Dependent variable: A dependent variable is one whose values are to be modeled and predicted by other variables. The dependent variables or target variables in this study were usage of the storage resource on a monthly basis, and the total amount of resource consumed after six months.

Independent variables: An independent variable also known as predictor variable is one whose value will be used to predict the value of the dependent or target variables.

The independent variables in this study are the client requests for resources, and number of users in the organization while the dependent variable was the resource usage after a period of six months, number of users in the organization and the type of cloud IaaS storage service the client is offered.

Data Collection of training data

The data was collected through the use of questionnaires which were presented to the Safaricom systems engineers.

Classification of data

Once the data was collected, it was then classified based on magnitude of client requests for the cloud computing storage service on a six-month cumulative basis measured in gigabytes. The reason for the classification was to group the various clients' requests so as to identify those whose requests are above 1000 GB. The classification index for both the dependent and independent variables is given by the table below.

Training Data

The training data was comprised of 200 data elements and it was presented in Microsoft excel's .csv format. It was made up of parameters which included, the amount of storage requested, amount of storage used and the number of users within a particular organization.

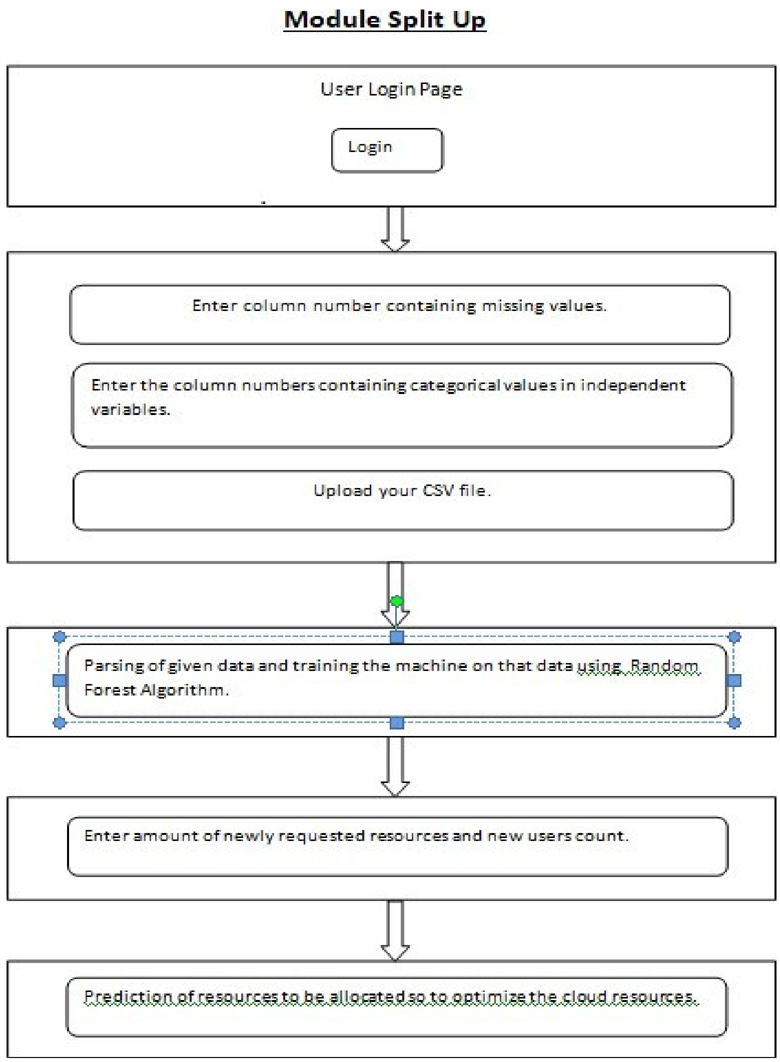

Module Split Up
User Login Page
Login
Enter column number containing missing values.
Enter the column numbers containing categorical values in independent variables.
Upload your CSV file.
Parsing of given data and training the machine on that data using Random Forest Algorithm.
Enter amount of newly requested resources and new users count.
Prediction of resources to be allocated so to optimize the cloud resources.

Fig.5

Implementation

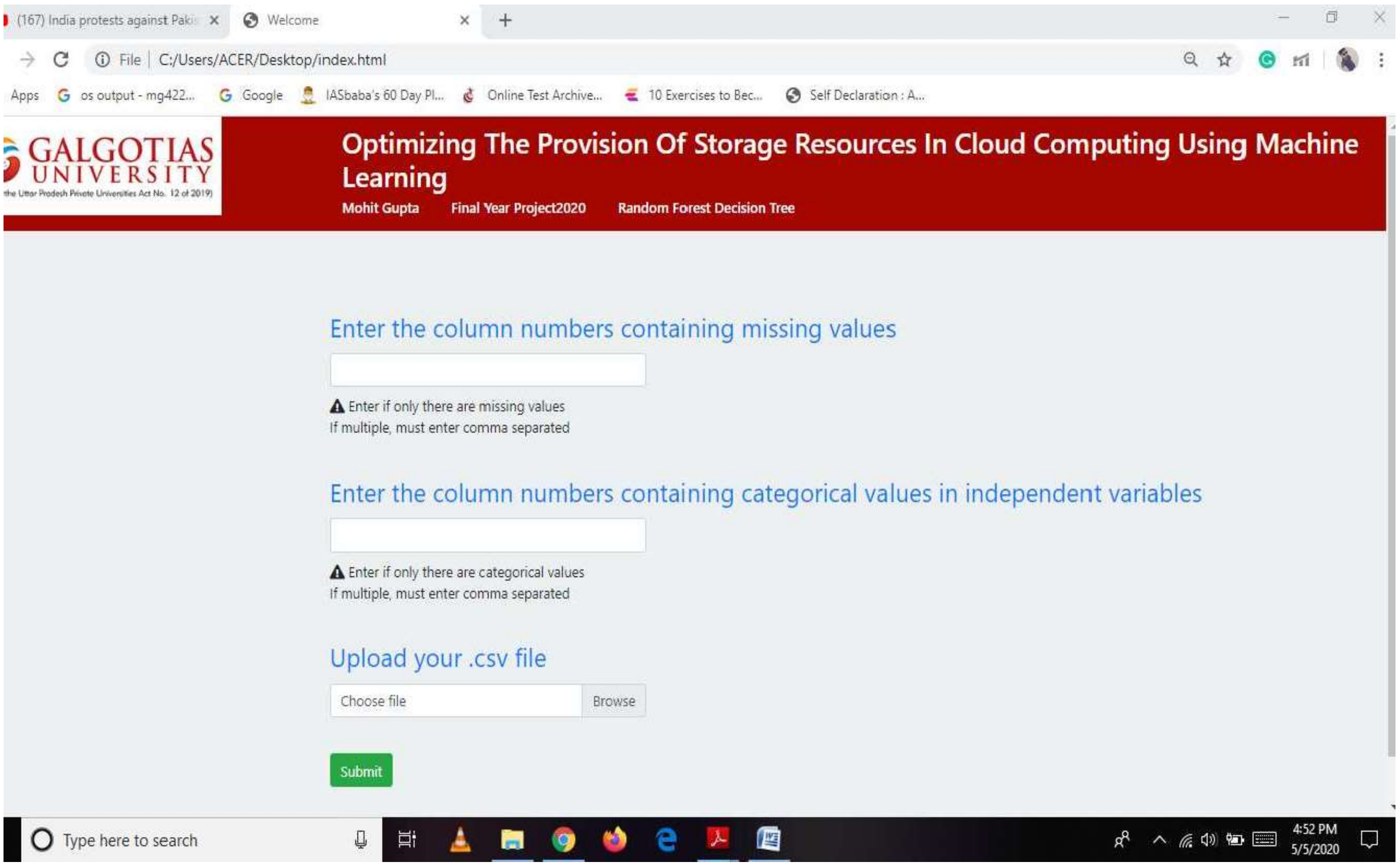


Fig.6 Webpage to enter data and upload csv file.

Sample data collected about request of resources by the user

A1 TYPE OF SERVICE

	A	B	C	D
1	TYPE OF S	RESOURCE	NO. OF U:	TOTAL USAGE AFTER SIX MONTHS(GB)
2	STORAGE	1000	248	206.3
3	STORAGE	2000	249	412.7
4	STORAGE	3500	146	412.7
5	STORAGE	1000	294	330.1
6	STORAGE	10000	107	2475.9
7	STORAGE	2000	227	907.8
8	STORAGE	3000	66	1031.6
9	STORAGE	5000	260	1650.6
10	STORAGE	1000	264	247.6
11	STORAGE	4000	228	1238
12	STORAGE	5000	152	1650.6
13	STORAGE	10000	77	3713.9
14	STORAGE	3000	186	1238
15	STORAGE	4500	64	1650.6
16	STORAGE	4000	127	1238
17	STORAGE	5000	104	1650.6
18	STORAGE	3000	202	1031.6
19	STORAGE	4000	232	1650.6
20	STORAGE	10000	120	2475.9
21	STORAGE	8000	129	1650.6
22	STORAGE	3000	68	825.3
23	STORAGE	4000	157	1444.3
24	STORAGE	5000	154	2063.3
25	STORAGE	8000	88	2888.6
26	STORAGE	2500	141	1031.6
27	STORAGE	3000	50	1031.6
28	STORAGE	5000	117	825.3
29	STORAGE	9000	83	2888.6
30	STORAGE	1000	221	330.1
31	STORAGE	4000	227	1568.1
32	STORAGE	5000	180	1584.1
33	STORAGE	10000	388	3755.6

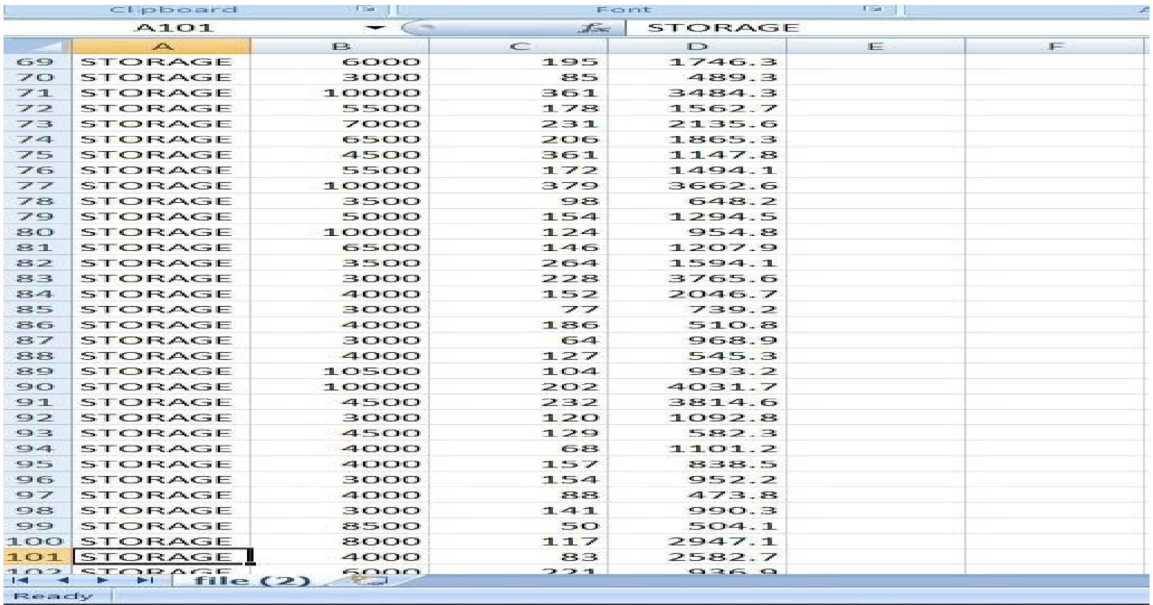
Fig.7

A67 STORAGE

	A	B	C	D
35	STORAGE	3500	105	729.2
36	STORAGE	3000	86	500.8
37	STORAGE	4000	124	958.9
38	STORAGE	3000	89	535.3
39	STORAGE	4000	127	983.2
40	STORAGE	10500	415	4021.7
41	STORAGE	10000	393	3804.6
42	STORAGE	4500	135	1082.8
43	STORAGE	3000	92	572.3
44	STORAGE	4500	136	1091.2
45	STORAGE	4000	113	828.5
46	STORAGE	4000	123	942.2
47	STORAGE	3000	83	463.8
48	STORAGE	4000	126	980.3
49	STORAGE	3000	86	494.1
50	STORAGE	8500	307	2937.1
51	STORAGE	8000	272	2572.7
52	STORAGE	4000	122	926.9
53	STORAGE	6000	193	1732.4
54	STORAGE	2000	64	211.3
55	STORAGE	14000	570	5539.7
56	STORAGE	3000	84	470.7
57	STORAGE	5000	158	1340.5
58	STORAGE	3500	98	645.4
59	STORAGE	4500	130	1027.3
60	STORAGE	3000	92	569.7
61	STORAGE	6500	212	1935.8
62	STORAGE	5000	150	1249.7
63	STORAGE	4500	143	1166.8
64	STORAGE	6500	219	2011.3
65	STORAGE	4000	121	924.1
66	STORAGE	6500	219	2011.4
67	STORAGE	4000	121	918.3
68	STORAGE	5000	160	1358.1

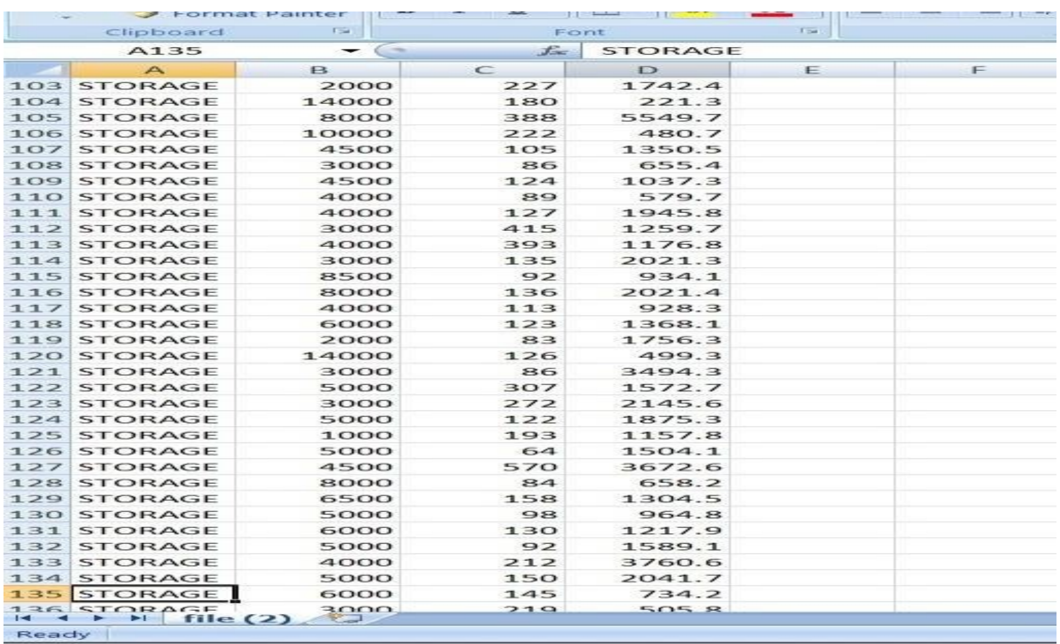
file (2)

Fig.8



	A	B	C	D	E	F
69	STORAGE	6000	195	1746.3		
70	STORAGE	3000	85	489.3		
71	STORAGE	10000	361	3484.3		
72	STORAGE	5500	178	1562.7		
73	STORAGE	7000	231	2135.6		
74	STORAGE	6500	206	1865.3		
75	STORAGE	4500	361	1147.8		
76	STORAGE	5500	172	1494.1		
77	STORAGE	10000	379	3662.6		
78	STORAGE	3500	98	648.2		
79	STORAGE	5000	154	1294.5		
80	STORAGE	10000	124	954.8		
81	STORAGE	6500	146	1207.9		
82	STORAGE	3500	264	1594.1		
83	STORAGE	3000	228	3765.6		
84	STORAGE	4000	152	2046.7		
85	STORAGE	3000	77	739.2		
86	STORAGE	4000	186	510.8		
87	STORAGE	3000	64	968.9		
88	STORAGE	4000	127	545.3		
89	STORAGE	10500	104	993.2		
90	STORAGE	10000	202	4031.7		
91	STORAGE	4500	232	3814.6		
92	STORAGE	3000	120	1092.8		
93	STORAGE	4500	129	582.3		
94	STORAGE	4000	68	1101.2		
95	STORAGE	4000	157	838.5		
96	STORAGE	3000	154	952.2		
97	STORAGE	4000	88	473.8		
98	STORAGE	3000	141	990.3		
99	STORAGE	8500	50	504.1		
100	STORAGE	8000	117	2947.1		
101	STORAGE	4000	83	2582.7		

Fig.9



	A	B	C	D	E	F
103	STORAGE	2000	227	1742.4		
104	STORAGE	14000	180	221.3		
105	STORAGE	8000	388	5549.7		
106	STORAGE	10000	222	480.7		
107	STORAGE	4500	105	1350.5		
108	STORAGE	3000	86	655.4		
109	STORAGE	4500	124	1037.3		
110	STORAGE	4000	89	579.7		
111	STORAGE	4000	127	1945.8		
112	STORAGE	3000	415	1259.7		
113	STORAGE	4000	393	1176.8		
114	STORAGE	3000	135	2021.3		
115	STORAGE	8500	92	934.1		
116	STORAGE	8000	136	2021.4		
117	STORAGE	4000	113	928.3		
118	STORAGE	6000	123	1368.1		
119	STORAGE	2000	83	1756.3		
120	STORAGE	14000	126	499.3		
121	STORAGE	3000	86	3494.3		
122	STORAGE	5000	307	1572.7		
123	STORAGE	3000	272	2145.6		
124	STORAGE	5000	122	1875.3		
125	STORAGE	1000	193	1157.8		
126	STORAGE	5000	64	1504.1		
127	STORAGE	4500	570	3672.6		
128	STORAGE	8000	84	658.2		
129	STORAGE	6500	158	1304.5		
130	STORAGE	5000	98	964.8		
131	STORAGE	6000	130	1217.9		
132	STORAGE	5000	92	1589.1		
133	STORAGE	4000	212	3760.6		
134	STORAGE	5000	150	2041.7		
135	STORAGE	6000	145	734.2		

Fig.10

A169 STORAGE

	A	B	C	D	E	F
137	STORAGE	10000	121	963.9		
138	STORAGE	5500	219	540.3		
139	STORAGE	7000	121	988.2		
140	STORAGE	6500	160	4026.7		
141	STORAGE	4500	64	3809.6		
142	STORAGE	5500	570	1087.8		
143	STORAGE	10000	84	577.3		
144	STORAGE	3500	158	1096.2		
145	STORAGE	1000	98	833.5		
146	STORAGE	4000	130	947.2		
147	STORAGE	5000	92	468.8		
148	STORAGE	10000	212	985.3		
149	STORAGE	6500	150	499.1		
150	STORAGE	3500	143	2942.1		
151	STORAGE	3000	219	2577.1		
152	STORAGE	4000	121	931.9		
153	STORAGE	3000	219	1737.4		
154	STORAGE	4000	121	216.3		
155	STORAGE	10500	160	5544.7		
156	STORAGE	10000	195	475.7		
157	STORAGE	4500	85	1345.5		
158	STORAGE	3000	361	650.4		
159	STORAGE	4500	178	1032.3		
160	STORAGE	5000	231	574.7		
161	STORAGE	8500	206	1940.8		
162	STORAGE	5000	141	1254.7		
163	STORAGE	6000	172	1171.8		
164	STORAGE	9000	379	2016.3		
165	STORAGE	4500	98	929.1		
166	STORAGE	8000	154	2016.4		
167	STORAGE	4000	124	923.3		
168	STORAGE	10500	146	1363.1		
169	STORAGE	6000	264	1751.3		
170	STORAGE	5000	228	494.3		

file (2)

Ready

Fig.11

Paste Format Painter

Clipboard Font

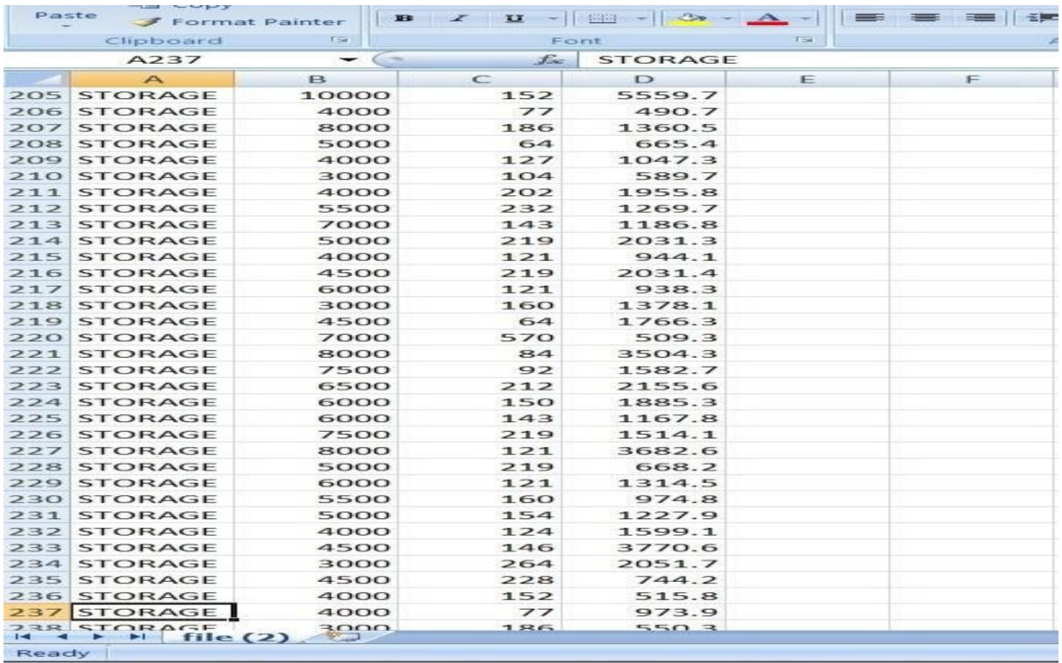
A203 STORAGE

	A	B	C	D	E	F
171	STORAGE	7000	152	3489.3		
172	STORAGE	3000	77	1567.7		
173	STORAGE	4500	186	2140.6		
174	STORAGE	6000	64	1870.3		
175	STORAGE	4000	127	1152.8		
176	STORAGE	6000	104	1499.1		
177	STORAGE	8000	202	3667.6		
178	STORAGE	5000	232	653.2		
179	STORAGE	6000	143	1299.5		
180	STORAGE	9000	219	959.8		
181	STORAGE	8000	121	1212.9		
182	STORAGE	6000	219	1604.1		
183	STORAGE	4000	121	3775.6		
184	STORAGE	10000	160	2056.7		
185	STORAGE	4000	64	749.2		
186	STORAGE	8000	570	520.8		
187	STORAGE	5000	84	978.9		
188	STORAGE	4000	158	555.3		
189	STORAGE	3000	98	1003.2		
190	STORAGE	4000	130	4041.7		
191	STORAGE	5500	92	3824.6		
192	STORAGE	7000	212	1102.8		
193	STORAGE	5000	150	592.3		
194	STORAGE	5000	143	1111.2		
195	STORAGE	6000	219	848.5		
196	STORAGE	5500	121	962.2		
197	STORAGE	4000	219	483.8		
198	STORAGE	8000	121	1000.3		
199	STORAGE	7500	160	514.1		
200	STORAGE	8000	154	2957.1		
201	STORAGE	5000	124	2592.7		
202	STORAGE	3000	146	946.9		
203	STORAGE	6000	264	1752.4		
204	STORAGE	4000	228	231.3		

file (2)

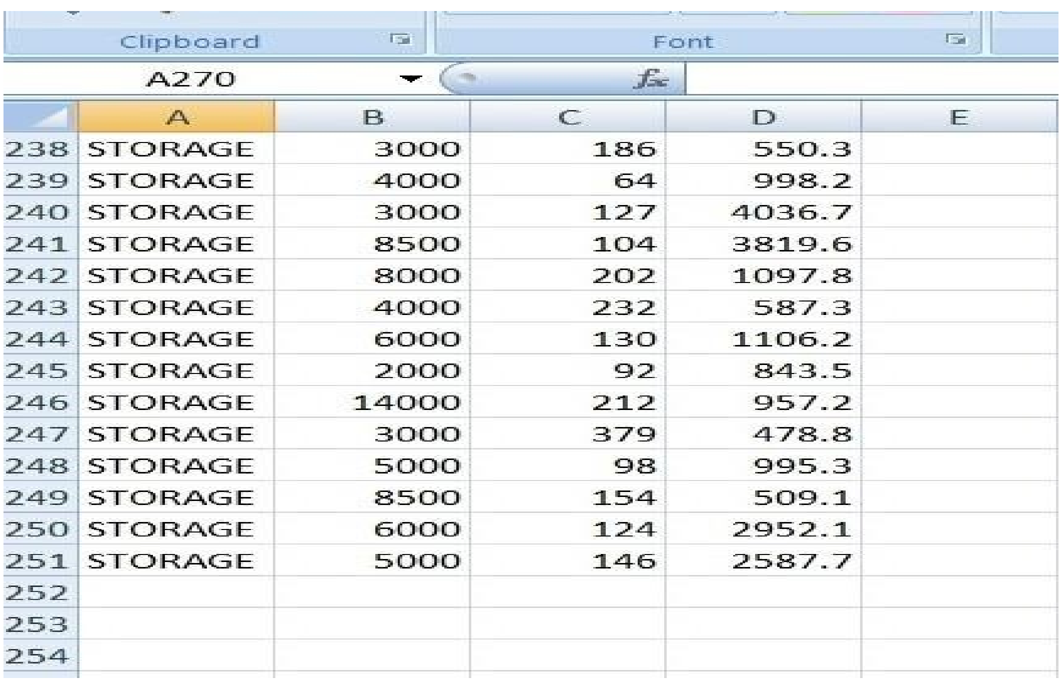
Ready

Fig.12



	A	B	C	D	E	F
205	STORAGE	10000	152	5559.7		
206	STORAGE	4000	77	490.7		
207	STORAGE	8000	186	1360.5		
208	STORAGE	5000	64	665.4		
209	STORAGE	4000	127	1047.3		
210	STORAGE	3000	104	589.7		
211	STORAGE	4000	202	1955.8		
212	STORAGE	5500	232	1269.7		
213	STORAGE	7000	143	1186.8		
214	STORAGE	5000	219	2031.3		
215	STORAGE	4000	121	944.1		
216	STORAGE	4500	219	2031.4		
217	STORAGE	6000	121	938.3		
218	STORAGE	3000	160	1378.1		
219	STORAGE	4500	64	1766.3		
220	STORAGE	7000	570	509.3		
221	STORAGE	8000	84	3504.3		
222	STORAGE	7500	92	1582.7		
223	STORAGE	6500	212	2155.6		
224	STORAGE	6000	150	1885.3		
225	STORAGE	6000	143	1167.8		
226	STORAGE	7500	219	1514.1		
227	STORAGE	8000	121	3682.6		
228	STORAGE	5000	219	668.2		
229	STORAGE	6000	121	1314.5		
230	STORAGE	5500	160	974.8		
231	STORAGE	5000	154	1227.9		
232	STORAGE	4000	124	1599.1		
233	STORAGE	4500	146	3770.6		
234	STORAGE	3000	264	2051.7		
235	STORAGE	4500	228	744.2		
236	STORAGE	4000	152	515.8		
237	STORAGE	4000	77	973.9		

Fig.13



	A	B	C	D	E
238	STORAGE	3000	186	550.3	
239	STORAGE	4000	64	998.2	
240	STORAGE	3000	127	4036.7	
241	STORAGE	8500	104	3819.6	
242	STORAGE	8000	202	1097.8	
243	STORAGE	4000	232	587.3	
244	STORAGE	6000	130	1106.2	
245	STORAGE	2000	92	843.5	
246	STORAGE	14000	212	957.2	
247	STORAGE	3000	379	478.8	
248	STORAGE	5000	98	995.3	
249	STORAGE	8500	154	509.1	
250	STORAGE	6000	124	2952.1	
251	STORAGE	5000	146	2587.7	
252					
253					
254					

Fig.14

Results and Discussions

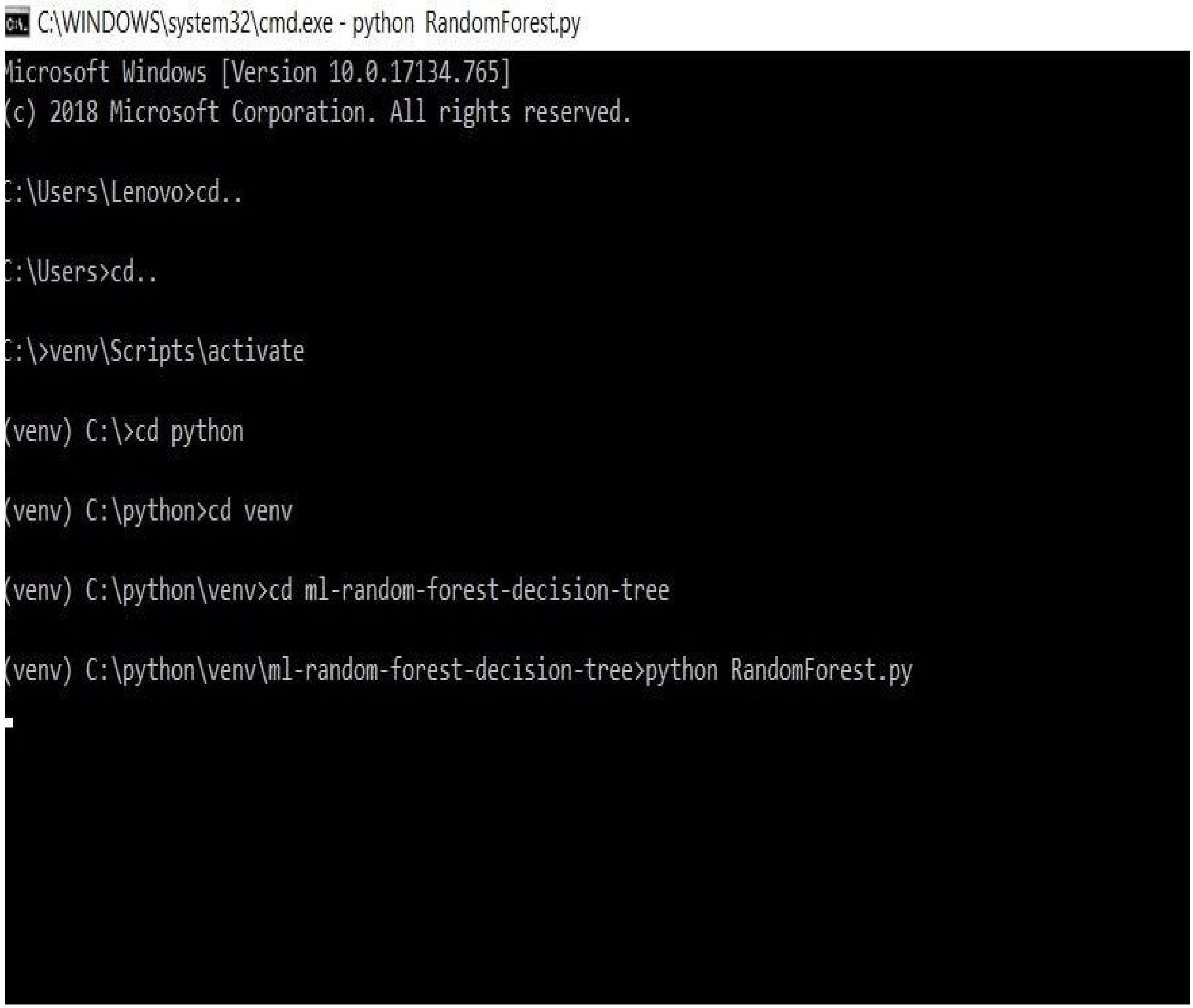


Fig.15

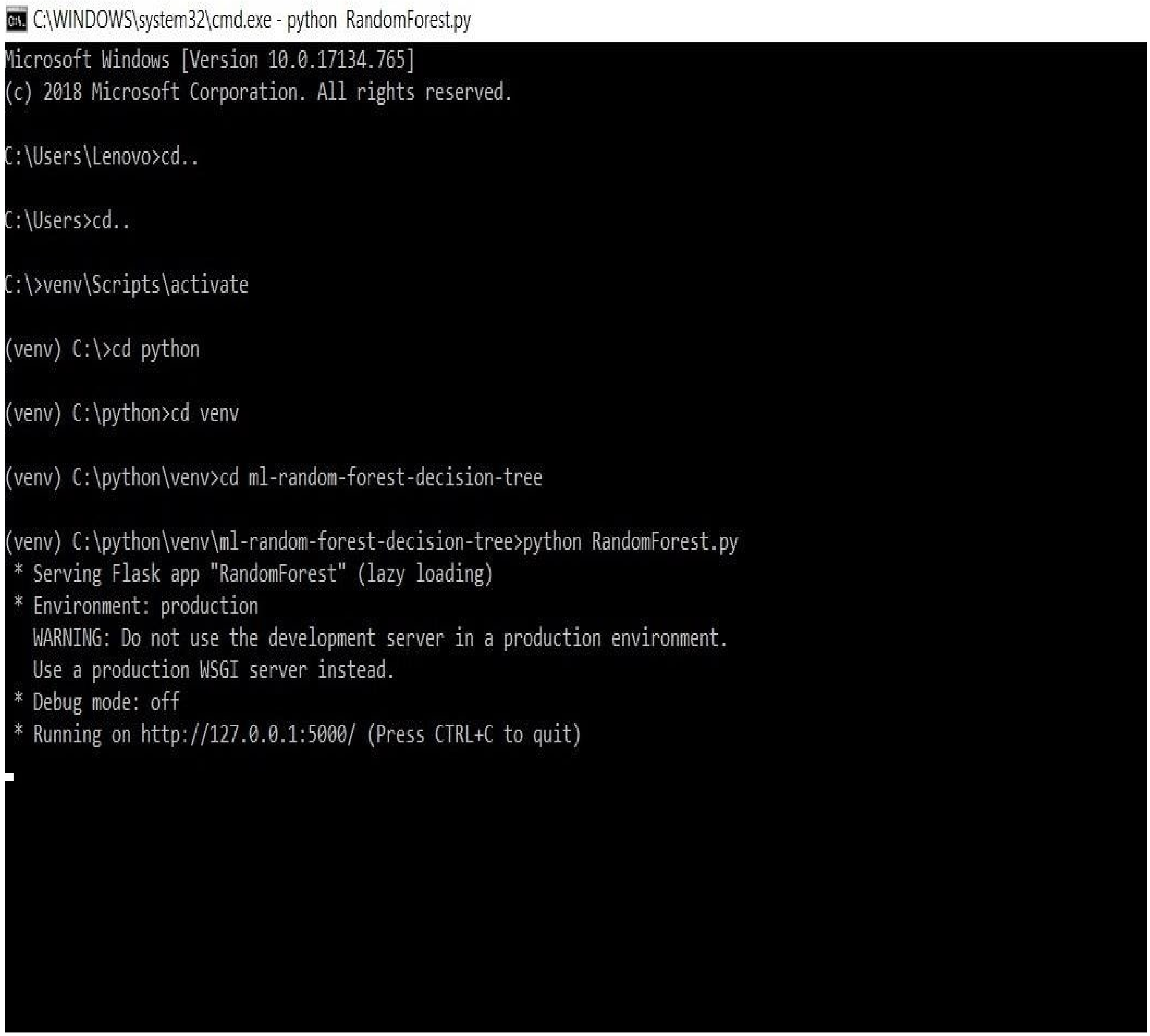
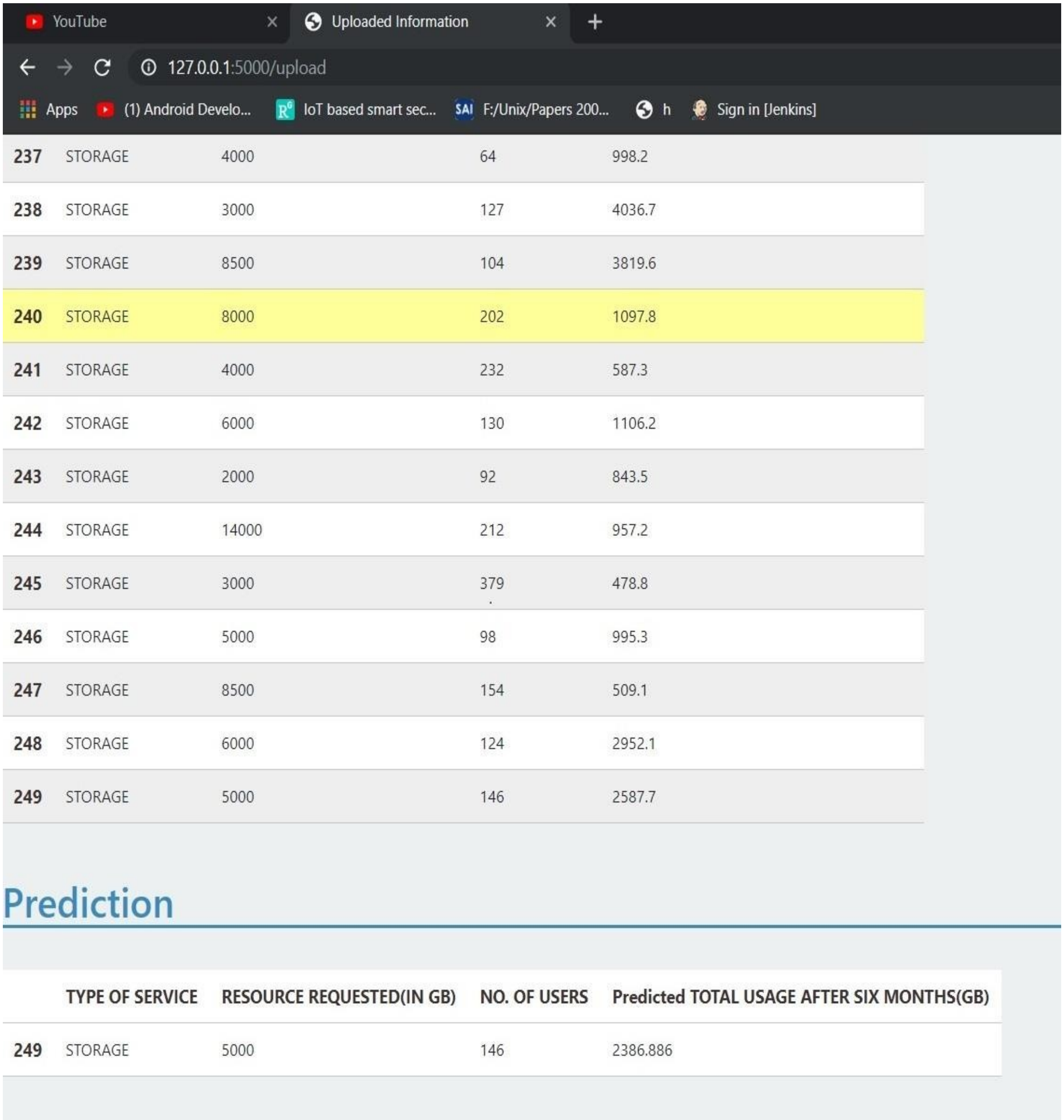

C:\WINDOWS\system32\cmd.exe - python RandomForest.py
Microsoft Windows [Version 10.0.17134.765]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\Lenovo>cd..
C:\Users>cd..
C:\>venv\Scripts\activate
(venv) C:\>cd python
(venv) C:\python>cd venv
(venv) C:\python\venv>cd ml-random-forest-decision-tree
(venv) C:\python\venv\ml-random-forest-decision-tree>python RandomForest.py
* Serving Flask app "RandomForest" (lazy loading)
* Environment: production
WARNING: Do not use the development server in a production environment.
Use a production WSGI server instead.
* Debug mode: off
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)

Fig.16



237	STORAGE	4000	64	998.2
238	STORAGE	3000	127	4036.7
239	STORAGE	8500	104	3819.6
240	STORAGE	8000	202	1097.8
241	STORAGE	4000	232	587.3
242	STORAGE	6000	130	1106.2
243	STORAGE	2000	92	843.5
244	STORAGE	14000	212	957.2
245	STORAGE	3000	379	478.8
246	STORAGE	5000	98	995.3
247	STORAGE	8500	154	509.1
248	STORAGE	6000	124	2952.1
249	STORAGE	5000	146	2587.7

Prediction

	TYPE OF SERVICE	RESOURCE REQUESTED(IN GB)	NO. OF USERS	Predicted TOTAL USAGE AFTER SIX MONTHS(GB)
249	STORAGE	5000	146	2386.886

Fig.17

Conclusions and Future Works

CONCLUSION

In this project, we not only built and used a random forest in Python, but we also developed an understanding of the model by starting with the basics. We have developed basic concept of machine learning and knowledge of various packages and tools that can be used in python for machine learning , basic of html and javascript are used in this project of ours to make the user interface of ours. We first looked at an individual decision tree, the building block of a random forest, and then saw how we can overcome the high variance of a single decision tree by combining hundreds of them in an ensemble model known as a random forest. The random forest uses the concepts of random sampling of observations, random sampling of features, and averaging predictions. Hopefully this project has given us the confidence and understanding needed to start using the random forest on various projects. The random forest is a powerful machine learning model, but that should not prevent us from knowing how it works. The more we know about a model, the better equipped we will be to use it effectively and explain how it makes predictions.

FUTURE SCOPE

In this project we have used machine learning algorithm: Random forest decision tree, to analyse and predict our result for optimization of resource storage provisioning, so that the problem of under and over provisioning can be solved and the money of the clients and storage of the vendor can be easily saved. For the future use of the algorithm are.

- Enhanced understanding of machine learning algorithms to analyse , predict and save future resorces in different situations.

Automatic allotment of provisioning for the users without the user being troubled for various formalities and data.

References

1) A. Matsunaga and J. A. B. Fortes, ―On the use of machine learning to predict the time and resources consumed by applications, in Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010, p. 495-504.

2) Arkaitz Ruiz-Alvarez, Marty Humphrey, A Model and Decision Procedure for Data Storage in Cloud Computing, in Proceedings of the IEEE/ACM International Symposium on Cluster, Ottawa Canada, 2012.

3) Machine Learning Techniques for optimizing the provision of storage resources in Cloud Computing Infrastructure as a service (IaaS): A Comparative Study by Edgar Otieno Addero

4) https://www.researchgate.net/

5) https://wikipedia.org/

6) https://stackoverflow.com/