

**SKRIPSI**

**KLASIFIKASI BIG DATA DENGAN ALGORITMA NAIVE  
BAYES PADA SISTEM TERDISTRIBUSI HADOOP**



**Mohamad Fahrizal Septrianto**

**NPM: 2013730017**

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS  
UNIVERSITAS KATOLIK PARAHYANGAN  
2017**



**UNDERGRADUATE THESIS**

**BIG DATA CLASSIFICATION WITH NAIVE BAYES  
ALGORITHM ON HADOOP DISTRIBUTED SYSTEM**



**Mohamad Fahrizal Septrianto**

**NPM: 2013730017**

**DEPARTMENT OF INFORMATICS  
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES  
PARAHYANGAN CATHOLIC UNIVERSITY  
2017**



# LEMBAR PENGESAHAN

## KLASIFIKASI BIG DATA DENGAN ALGORITMA NAIVE BAYES PADA SISTEM TERDISTRIBUSI HADOOP

Mohamad Fahrizal Septrianto

NPM: 2013730017

Bandung, 30 Mei 2017

Menyetujui,

Pembimbing



Dr. Veronica Sri Moertini



Ketua Tim Penguji



Dott. Thomas Anung Basuki

Anggota Tim Penguji



Aditya Bagoes Saputra, M.T.

Mengetahui,

Ketua Program Studi



Mariskha Tri Adithia, P.D.Eng





## PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

### **KLASIFIKASI BIG DATA DENGAN ALGORITMA NAIVE BAYES PADA SISTEM TERDISTRIBUSI HADOOP**

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,  
Tanggal 30 Mei 2017



Mohamad Fahrizal Septrianto  
NPM: 2013730017





## ABSTRAK

Internet saat ini telah menjadi salah satu sarana utama untuk melakukan komunikasi, pencarian informasi, bahkan untuk melakukan transaksi jual beli barang. Hal ini menyebabkan data yang diproses menjadi sangat banyak dan tidak beraturan. Karena banyaknya pengguna dan data yang dioper di dalam internet maka muncullah kebutuhan untuk mengolah data yang sangat bervariasi dalam jumlah yang sangat besar dan dengan kebutuhan akan kecepatan pemrosesan yang tinggi menjadi pokok permasalahan yang dihadapi saat ini.

Sebuah sistem terdistribusi adalah salah satu dampak kemajuan teknologi yang dapat mengolah big data dengan komputasi secara paralel terdistribusi. Hadoop adalah contoh kerangka kerja yang dapat menjalankan perangkat lunak dengan sistem yang terdistribusi. Hadoop adalah salah satu kerangka kerja yang dirancang untuk memproses dan menganalisis data yang sangat banyak. Hadoop memiliki komponen-komponen yang dapat mengolah data menggunakan banyak komputer dalam satu waktu secara paralel, sehingga waktu yang dieksekusi selama pemrosesan data menjadi minimum.

Untuk dapat mengetahui dan mengambil informasi berharga yang tersembunyi dalam *big data*, dibutuhkan teknik *data mining/machine-learning* untuk melakukan proses pembuatan model spesifik yang nantinya bisa digunakan untuk memprediksi kemunculan data berikutnya. Hal ini dapat berguna untuk suatu perusahaan dalam melakukan *predictive analysis* untuk menaikkan nilai perusahaan tersebut.

Pada skripsi ini, telah berhasil dikembangkan sebuah perangkat lunak yang berjalan di dalam lingkungan Hadoop berbasis *MapReduce* yang mampu melakukan pembuatan model klasifikasi *naive bayes* sekaligus melakukan pengujian terhadap model tersebut menggunakan *big data*. Pada eksperimen yang telah dilakukan menggunakan *cluster* hadoop yang ada pada laboratorium milik FTIS (Fakultas Teknologi Informasi dan Sains), ditetapkan 2 variabel eksperimen, yaitu (1) ukuran *big data* dan (2) ukuran blok. Dari hasil eksperimen yang telah dilakukan diketahui bahwa ukuran *big data* yang diumpankan pada perangkat lunak sangat bergantung pada spesifikasi perangkat keras milik *cluster* Hadoop, sehingga dapat mempengaruhi waktu eksekusi. Lalu, ukuran blok pada HDFS (*Hadoop Distributed File System*) juga mempengaruhi waktu eksekusi perangkat lunak yang berjalan pada sistem terdistribusi Hadoop.

**Kata-kata kunci:** Hadoop, Map Reduce, Big Data, Naive Bayes



## ABSTRACT

Nowadays, internet has become one of the most used tool for communicating, searching for information, even for online transaction like buying and selling some products. Also, it has grows to become a digital place for companies to sell and market their products. Because internet users continue to grow quickly and diverse, this could cause all the data that are processed is getting more and more big and also had unstructured pattern. It also made the needs for processing huge data at an efficient amount of time to be the major common problem faced by most companies/individual.

A distributed system is an impact from the present technology that can meet human expectations in huge data processing. Hadoop is an example of a framework that can run software in a distributed system. Hadoop is a framework that build for huge data processing and analysis. Hadoop have many component that support it to process a numerous data in many computers at the same time with parallel method. Because of the ability to concurrently run a job at the same time, Hadoop can minimize the process time from the job.

To be able to know and retrieve hidden valuable information inside the big data, it required data mining technique/machine learning algorithm to perform the process of making a specific model that can later be used to predict a fact in the future so that company could make a better decision. This could be useful for company in doing predictive analysis to raise the value of the company.

In this research, a MapReduce based software that runs in Hadoop environment has been successfully developed which is able to make a naive bayes classifier model for classification as well as testing the models that use large data. In experiments that has been done using existing hadoop clusters in the laboratory owned by FTIS (Faculty of Information Technology and Sciences), 2 experimental variables were defined, which is: (1) big data size and (2) block size. The results told us that the size of big data fed on the software is very dependent on the hardware specifications owned by hadoop cluster, so it can affect the execution time. Also, block size on HDFS (Hadoop Distributed File System) also affects software execution time running on the hadoop distributed system according to the size of the data itself.

**Keywords:** Hadoop, Map Reduce, Big Data, Naive Bayes



*Teruntuk Agustine Bornita dan Samodro Rudy Santoso, yang telah dengan setia mengasuh, mendidik, memperhatikan, dan memberi asupan gizi yang cukup, sehingga skripsi ini dapat terselesaikan.*



## KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas selesainya skripsi yang berjudul "Klasifikasi Big Data dengan Algoritma Naive Bayes pada Sistem Terdistribusi Hadoop". Penulisan skripsi ini diajukan untuk memenuhi salah satu syarat memperoleh gelar sarjana pada program studi Teknik Informatika, Fakultas Teknologi Informasi dan Sains, Universitas Katolik Parahyangan.

Atas dukungan materil dan moril yang diberikan dalam penyusunan makalah ini, maka penulis mengucapkan banyak terima kasih kepada :

1. Kedua orangtua, Samodro Rudy Santoso dan Agustine Bornita yang telah banyak memberi motivasi untuk selalu semangat dalam mengerjakan skripsi dan dukungan berupa materil dalam bentuk *cash* maupun kredit
2. Dr. Veronica Sri Moertini selaku dosen pembimbing yang telah memberikan banyak motivasi dan masukan yang sangat bermanfaat kepada penulis ditengah kebingungan dalam pengambilan suatu keputusan
3. Professor Nouredin Sadawi (*Research Associate and Software Engineer at the Faculty of Medicine, Department of Surgery and Cancer, Imperial College London*) yang telah sangat membantu penulis dalam memahami pengujian yang dilakukan pada model NBC (*Naive Bayes Classifier*) yang menggunakan *multi-class classification* dalam videonya yang berjudul "Confusion Matrix for Multiple Classes"
4. Aditya Bagoes Saputra, M.T. selaku dosen PSC (Pengenalan Sistem Cerdas) yang sempat diajak berdiskusi dalam memahami *term - term* yang ada di dalam buku referensi pendukung skripsi ini
5. Kedua kakak kandung, Deryl Wicaksono dan Aulia Rizky Wicaksono yang telah banyak membantu meredakan ketegangan otak dengan berdiskusi, berdialog, merencanakan masa depan, bahkan hingga bermain game bersama, sehingga membuat penulis selalu kembali termotivasi untuk mengerjakan skripsi ini
6. Mentor - mentor *internship* Blibli yang telah mengenalkan dan mengajarkan fundamental - fundamental dari *Spring framework Java* yang digunakan di beberapa modul program pada skripsi ini
7. *Udemy Online Courses*, platform pembelajaran online dengan menggunakan video yang telah banyak membantu penulis dalam memahami konsep Java 8, Java lambda function, Hadoop, HDFS, MapReduce, dan Spring framework
8. Liptia Venica(IT 2012) yang telah membantu melakukan *setting* lab untuk cluster Hadoop dan memberikan pemahaman - pemahaman fundamental pada sistem terdistribusi Hadoop yang sesungguhnya pada perangkat yang ada di Lab FTIS UNPAR
9. Rekan - rekan KMKLab, Radiontech, dan ISFnet batch 2, khususnya Wilianto Indrawan(IT 2012), Distra Vantari(IT 2012), dan Rachael Awuy (IT 2013) yang selalu siap sedia untuk diajak berdiskusi dan tanya jawab demi mendapatkan pandangan lain dalam mengerjakan

skripsi ini, sehingga penulis mendapatkan pandangan - pandangan baru untuk dijadikan pertimbangan dalam menghadapi permasalahan yang muncul pada saat penyusunan skripsi

10. Rekan - rekan seperjuangan IT UNPAR 2013, 2012, 2011, dan 2010 yang juga turut andil secara langsung maupun tidak langsung dalam pengerjaan skripsi sehingga dapat terselesaikan
11. Semua pihak yang telah banyak memberikan bantuan kepada penulis dalam menyelesaikan skripsi

Akhirnya penulis menyadari masih banyak kekurangan dan kelemahan pada skripsi ini. Untuk itu saran dan kritik yang konstruktif akan sangat membantu agar skripsi ini dapat menjadi lebih baik.

Bandung, Mei 2017

Penulis



# DAFTAR ISI

KATA PENGANTAR	xv
DAFTAR ISI	xvii
DAFTAR GAMBAR	xxi
DAFTAR TABEL	xxv
<b>1 PENDAHULUAN</b>	<b>1</b>
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
1.4 Batasan Masalah	3
1.5 Metodologi Penelitian	3
1.6 Sistematika Pembahasan	3
<b>2 LANDASAN TEORI</b>	<b>5</b>
2.1 Penambangan Data	5
2.1.1 Definisi Penambangan Data	5
2.1.2 Tugas Yang Dapat Dilakukan Oleh <i>Data Mining</i>	5
2.1.3 Langkah - langkah Penambangan Data	7
2.2 <i>Big Data</i>	8
2.2.1 Definisi <i>Big Data</i>	8
2.2.2 Karakteristik <i>Big Data</i>	8
2.3 Sistem Terdistribusi Hadoop	9
2.3.1 Definisi Hadoop	9
2.3.2 Fitur - fitur dari <i>Hadoop</i>	10
2.3.3 Cara Kerja Hadoop	10
2.3.4 Elemen dari Hadoop	11
2.3.4.1 HDFS (Hadoop Distributed File System)	11
2.3.4.2 MapReduce	15
2.4 Teorema Bayes	20
2.5 <i>Naive Bayes Classifier</i>	21
2.5.1 Algoritma	21
2.5.1.1 Pembuatan Model	22
2.5.1.2 Klasifikasi	24
2.5.2 <i>Zero-Frequency Problem</i>	26
2.5.3 Prediktor Numerik	26
2.5.4 Pengukuran <i>Naive Bayes Classifier</i>	26
2.6 Framework Yang Digunakan Dalam Membangun Perangkat Lunak	29
2.6.1 <i>Spring Framework</i>	29
2.6.2 <i>Maven</i>	31
2.6.3 <i>Thymeleaf</i>	31

<b>3</b>	<b>ANALISIS</b>	<b>33</b>
3.1	Deskripsi Masalah dan Solusi Umum	33
3.1.1	Deskripsi Masalah	33
3.1.2	Solusi Umum	33
3.2	Analisis Perangkat Lunak	34
3.2.1	Analisis Skema Algoritma <i>Naive Bayes Classifier</i> Berbasis <i>Map Reduce</i>	34
3.2.1.1	Modul Kelola <i>Input</i>	35
3.2.1.2	Modul <i>Train Naive Bayes M-R Based</i>	38
3.2.1.3	Modul <i>Testing Naive Bayes M-R Based</i>	50
3.2.1.4	Modul Klasifikasi <i>Naive Bayes</i>	61
3.2.2	Diagram Kelas	62
3.2.2.1	Modul Kelola <i>Input</i>	63
3.2.2.2	Modul <i>Training dan Testing Naive Bayes M-R Based</i>	64
3.2.2.3	Modul Klasifikasi <i>Naive Bayes</i>	65
<b>4</b>	<b>PERANCANGAN</b>	<b>67</b>
4.1	Perancangan Antarmuka	67
4.1.1	<i>Dashboard</i>	67
4.1.2	<i>Input Set Manager</i>	68
4.1.3	<i>Renew NBC Model Manager</i>	69
4.1.4	<i>Testing Manager</i>	69
4.1.5	<i>Classification Manager</i>	70
4.1.6	<i>Error Rate Dashboard</i>	71
4.2	Diagram Kelas Lengkap	71
4.2.1	Diagram Kelas Modul <i>Train Naive Bayes M-R Based</i>	71
4.2.2	Diagram Kelas Modul <i>Testing Naive Bayes M-R Based</i>	79
4.2.2.1	Kelas <i>Main</i>	79
4.2.2.2	<i>Package Base</i>	80
4.2.2.3	<i>Package Mapper</i>	82
4.2.2.4	<i>Package Reducer</i>	87
4.2.3	Diagram Kelas Modul Kelola Input	91
4.2.3.1	<i>InputController</i>	92
4.2.3.2	<i>HdfsService</i>	93
4.2.3.3	<i>InputSetDtoRequest</i>	94
4.2.4	Diagram Kelas Modul Klasifikasi <i>Naive Bayes</i>	94
4.2.4.1	<i>Controller</i>	96
4.2.4.2	<i>Service</i>	99
4.2.4.3	<i>Model</i>	102
<b>5</b>	<b>IMPLEMENTASI, PENGUJIAN, DAN EKSPERIMEN</b>	<b>107</b>
5.1	Deskripsi Perangkat Keras dan Lunak yang Digunakan	107
5.2	Implementasi Antarmuka	107
5.2.1	<i>Shell</i>	107
5.2.2	Berbasis Web HTML	108
5.2.2.1	Layout Menu	108
5.2.2.2	Dashboard	109
5.2.2.3	Input Set Manager	109
5.2.2.4	Renew Model Manager	110
5.2.2.5	Testing Manager	111
5.2.2.6	Classification Manager	112
5.2.2.7	Error Rate Dashboard	112
5.3	Implementasi Package, Kelas, dan Method dengan Java	114

5.4	Pengujian Kebenaran . . . . .	114
5.4.1	Perhitungan Manual Dengan Data Studi Kasus . . . . .	114
5.4.1.1	Pembuatan Model NBC . . . . .	114
5.4.1.2	Klasifikasi Menggunakan Model NBC . . . . .	115
5.4.2	Perhitungan Menggunakan Perangkat Lunak yang Dibangun Dengan Data Studi Kasus . . . . .	118
5.4.2.1	Pembuatan model NBC . . . . .	118
5.4.2.2	Klasifikasi Menggunakan Model NBC . . . . .	119
5.4.3	Perbandingan dan Kesimpulan . . . . .	121
5.4.3.1	Model . . . . .	121
5.4.3.2	Klasifikasi . . . . .	122
5.5	Eksperimen . . . . .	122
5.5.1	Pembuatan model . . . . .	122
5.5.1.1	Dataset-1 . . . . .	122
5.5.1.2	Dataset-2 . . . . .	125
5.5.1.3	Dataset-3 . . . . .	126
5.5.1.4	Kesimpulan . . . . .	128
5.5.2	Performansi Big Data . . . . .	129
5.5.2.1	Uji Pengaruh Ukuran Blok Terhadap Kecepatan . . . . .	129
5.5.2.2	Uji Pengaruh Ukuran Data, Jumlah Atribut Prediktor, dan Jenis Atribut . . . . .	130
5.5.2.3	Kesimpulan . . . . .	133
<b>6</b>	<b>KESIMPULAN DAN SARAN</b>	<b>135</b>
6.1	Kesimpulan . . . . .	135
6.2	Saran Penelitian Lanjutan . . . . .	136
	<b>DAFTAR REFERENSI</b>	<b>139</b>
<b>A</b>	<b>STRUKTUR DAN KODE PROGRAM MODUL TRAIN NAIVE BAYES M-R BASED</b>	<b>141</b>
A.1	Struktur Program . . . . .	141
A.2	Kode Program . . . . .	141
<b>B</b>	<b>STRUKTUR DAN KODE PROGRAM MODUL TESTING NAIVE BAYES M-R BASED</b>	<b>147</b>
B.1	Struktur Program . . . . .	147
B.2	Kode Program . . . . .	147
B.2.1	Package Base . . . . .	148
B.2.2	Package Mapper . . . . .	149
B.2.3	Package Reducer . . . . .	157
<b>C</b>	<b>STRUKTUR DAN KODE PROGRAM MODUL KELOLA INPUT DAN KLASIFIKASI</b>	<b>163</b>
C.1	Struktur Program . . . . .	163
C.2	Kode Program . . . . .	164
C.2.1	Kode Program Kebutuhan Modul Kelola Input . . . . .	166
C.2.2	Kode Program Kebutuhan Modul Klasifikasi . . . . .	171
C.2.2.1	Controller . . . . .	171
C.2.2.2	Service . . . . .	177
C.2.2.3	Model . . . . .	185
<b>D</b>	<b>PENGUJIAN KEBENARAN</b>	<b>195</b>
D.1	Pembuatan Model NBC . . . . .	195
D.2	Testing Model NBC . . . . .	196

<b>E</b>	<b>EKSPERIMEN PEMBUATAN MODEL DAN TESTING</b>	<b>197</b>
E.1	Hasil Pembuatan Model dan Testing Dataset-1 . . . . .	197
E.1.1	Contoh Dataset-1 . . . . .	197
E.1.2	Hasil Pembuatan Model Dataset-1 . . . . .	197
E.1.3	Hasil Testing Dataset-1 . . . . .	199
E.2	Hasil Pembuatan Model dan Testing Dataset-2 . . . . .	200
E.2.1	Contoh Dataset-2 . . . . .	200
E.2.2	Hasil Pembuatan Model Dataset-2 . . . . .	200
E.2.3	Hasil Testing Dataset-2 . . . . .	201
E.3	Hasil Pembuatan Model dan Testing Dataset-3 . . . . .	201
E.3.1	Contoh Dataset-3 . . . . .	201
E.3.2	Hasil Pembuatan Model Dataset-3 . . . . .	202
E.3.3	Hasil Testing Dataset-3 . . . . .	203

## DAFTAR GAMBAR

2.1	Langkah Penambahan Data	7
2.2	Arsitektur master-slave	9
2.3	Alur kerja Secondary NameNode dan NameNode	12
2.4	FileSystem Metadata pada NameNode dan Secondary NameNode	12
2.5	Client membaca data dari HDFS	13
2.6	Failure takeover 1	13
2.7	Failure takeover 1	14
2.8	Client menulis data ke HDFS	15
2.9	Ilustrasi framework MapReduce	16
2.10	Arsitektur YARN	17
2.11	Arsitektur YARN	17
2.12	Proses MapReduce	18
2.13	Proses Map	19
2.14	Proses Reduce	20
2.15	<i>Bayesian Network</i> dengan Kedalaman 1 level	21
2.16	<i>Precision dan Recall</i>	28
2.17	Request Processing Workflow Spring MVC	30
3.1	Rancangan Keseluruhan Modul Program	35
3.2	Modul-Specification	35
3.3	Flow Chart Modul Input	36
3.4	Missing-values	37
3.5	Flow Chart Modul Training	38
3.6	<i>Context diagram</i> modul Training	39
3.7	DFD level 1 modul Training	40
3.8	DFD level 2: proses 1.0	42
3.9	P-Spec proses 1.0.0	44
3.10	P-Spec proses 1.0.1	44
3.11	P-Spec proses 1.0.2	44
3.12	P-Spec proses 1.0.3	45
3.13	DFD level 2: proses 1.1	45
3.14	P-Spec training reduce: pada proses 1.1.0	46
3.15	DFD level 2: proses 1.2	47
3.16	P-Spec training reduce: pada proses 1.2.0	49
3.17	P-Spec training reduce: pada proses 1.2.1	49
3.18	P-Spec training reduce: pada proses 1.2.2	50
3.19	P-Spec training: pada proses 1.2.3	50
3.20	Flow Chart Modul Testing	51
3.21	<i>Context diagram</i> modul Testing	51
3.22	DFD level 1 modul Testing	53
3.23	DFD level 2: proses 1.0	55
3.24	P-Spec training reduce: pada proses 1.0.0	57

3.25	P-Spec training reduce: pada proses 1.0.1	57
3.26	P-Spec training reduce: pada proses 1.0.2	58
3.27	P-Spec training reduce: pada proses 1.0.3	58
3.28	P-Spec training reduce: pada proses 1.0.4	58
3.29	DFD level 2: proses 1.1	59
3.30	DFD level 2: proses 1.2	59
3.31	P-Spec testing reduce: pada proses 1.2.0	60
3.32	P-Spec testing reduce: pada proses 1.2.1	61
3.33	Flow Chart Modul Klasifikasi	62
3.34	Diagram kelas modul kelola input	63
3.35	Diagram kelas modul <i>training</i> dan <i>testing</i>	64
3.36	Diagram kelas modul klasifikasi <i>naive bayes</i>	65
4.1	<i>Dashboard</i>	67
4.2	<i>Input Set Manager</i>	68
4.3	<i>Renew NBC Model Manager</i>	69
4.4	<i>Testing Manager</i>	70
4.5	<i>Classification Manager</i>	70
4.6	<i>Error Rate Dashboard</i>	71
4.7	Diagram Kelas <i>Train Naive Bayes M-R Based</i>	72
4.8	Diagram Kelas Modul <i>Testing: Main</i>	79
4.9	Diagram Kelas Modul <i>Testing: Package Base</i>	80
4.10	Diagram Kelas Modul <i>Testing: Package Mapper</i>	82
4.11	Diagram Kelas Modul <i>Testing: Package Reducer</i>	87
4.12	Struktur MVC Pada Modul Kelola Input	91
4.13	Diagram Kelas Modul Kelola Input	92
4.14	Struktur MVC Pada Modul Klasifikasi	95
4.15	Diagram Kelas Modul Kelola Input	96
4.16	Diagram Kelas Modul Kelola Input	99
4.17	Diagram Kelas Modul Kelola Input	102
5.1	Implementasi Antarmuka Layout Menu	108
5.2	Implementasi Antarmuka Dashboard	109
5.3	Implementasi Antarmuka Input Set Manager	110
5.4	Implementasi Antarmuka Renew Model Manager	111
5.5	Implementasi Antarmuka Testing Manager	111
5.6	Implementasi Antarmuka Classification Manager	112
5.7	Implementasi Antarmuka Error Rate Dashboard	113
5.8	Import Model NBC - Pengujian	120
5.9	Hasil Klasifikasi Dataset-1	120
5.10	Hasil Klasifikasi Dataset-1	121
5.11	Perbandingan Hasil Model Manual dan Program	121
5.12	Perbandingan Hasil Klasifikasi Manual dan Program	122
5.13	Memasukkan dataset-1 ke dalam HDFS	124
5.14	Memasukkan dataset-2 ke dalam HDFS	125
5.15	Memasukkan dataset-3 ke dalam HDFS	128
5.16	Grafik pengaruh ukuran blok terhadap rata-rata waktu eksekusi	130
5.17	Grafik pengaruh ukuran data terhadap rata-rata waktu eksekusi dengan atribut prediktor sebanyak 6	131
5.18	Grafik pengaruh ukuran data terhadap rata-rata waktu eksekusi dengan atribut prediktor sebanyak 4	132

5.19 Grafik pengaruh ukuran data terhadap rata-rata waktu eksekusi dengan atribut prediktor sebanyak 2 . . . . .	133
5.20 Grafik pengaruh ukuran data terhadap rata-rata waktu eksekusi dengan jumlah dan tipe atribut prediktor berbeda . . . . .	134





## DAFTAR TABEL

2.1	Contoh Dataset . . . . .	22
2.2	Tabel frekuensi atribut Outlook . . . . .	23
2.3	Tabel frekuensi atribut Outlook (modifikasi) . . . . .	23
2.4	Tabel kemungkinan atribut Outlook . . . . .	23
2.5	Tabel Distribusi . . . . .	23
2.6	Tabel rata - rata dan standar deviasi atribut Humidity . . . . .	23
2.7	Contoh <i>Confusion Matrix</i> . . . . .	27
2.8	<i>Confusion Matrix - Accuracy</i> . . . . .	27
2.9	<i>Confusion Matrix</i> Kelas Play . . . . .	29
3.1	Pemetaan Proses <i>MapReduce</i> pada DFD ke Method pada Kelas . . . . .	64
5.1	Contoh Dataset . . . . .	114
5.2	Table frekuensi atribut Outlook . . . . .	115
5.3	Table frekuensi atribut Outlook . . . . .	115
5.4	Table kemungkinan atribut Outlook . . . . .	116
5.5	Table Distribusi . . . . .	116
5.6	Table rata - rata dan standar deviasi atribut Humidity . . . . .	117
5.7	Contoh Dataset . . . . .	118
5.8	<i>Confusion Matrix - mushroom classification</i> . . . . .	124
5.9	Hasil Evaluasi Model NBC - <i>mushroom classification</i> . . . . .	124
5.10	<i>Confusion Matrix - car evaluation</i> . . . . .	126
5.11	Hasil Evaluasi Model NBC - <i>car evaluation</i> . . . . .	126
5.12	<i>Confusion Matrix - homicide reports</i> . . . . .	128
5.13	Hasil Evaluasi Model NBC - <i>homicide reports</i> . . . . .	128
5.14	Hasil uji pengaruh ukuran blok terhadap kecepatan program dalam detik . . . . .	129
5.15	Hasil uji pengaruh ukuran data terhadap kecepatan program dengan atribut prediktor sebanyak 6 dan atribut kelas sebanyak 1 dalam detik . . . . .	130
5.16	Hasil uji pengaruh ukuran data terhadap kecepatan program dengan atribut prediktor sebanyak 4 dan atribut kelas sebanyak 1 dalam detik . . . . .	131
5.17	Hasil uji pengaruh ukuran data terhadap kecepatan program dengan atribut prediktor sebanyak 2 dan atribut kelas sebanyak 1 dalam detik . . . . .	132



# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang

Seiring dengan berkembangnya teknologi, kebutuhan akan penggunaan Internet melaju sangat pesat. Tidak dapat disangkal, bahwa kebanyakan dari manusia masa kini sudah melihat Internet sebagai kebutuhan primer-nya, karena kemudahan dan manfaat yang ditawarkan oleh Internet yang sangat banyak. Hal tersebut menyebabkan frekuensi penggunaan Internet yang semakin tinggi pula, sehingga secara tidak langsung Internet telah menjadi sarana utama dalam mendapatkan informasi dan telah berhasil meniadakan batasan informasi yang bisa diakses dari mana saja.

Dengan begitu, banyak perusahaan yang menjadikan Internet sebagai salah satu sarana utama untuk mengembangkan produk/jasa yang mereka miliki, melihat pengguna-nya yang sangat banyak dan bermacam - macam dari segala penjuru dunia yang juga menggunakan Internet. Karena banyaknya pengguna dan data yang dioper di dalam Internet maka kebutuhan untuk mengolah data yang sangat bervariasi dan jumlah yang sangat besar dengan kecepatan yang tinggi menjadi pokok permasalahan yang dihadapi saat ini (*Big Data*).

*Big Data* merupakan suatu terminologi modern untuk sekumpulan data yang memiliki kesulitan tersendiri untuk diproses dengan cara tradisional (menggunakan satu buah komputer). Tiga hal terpenting yang menjadi pokok permasalahan dalam *Big Data* adalah : (1) mengolah data yang berjumlah sangat besar, (2) mengolah data yang memiliki tipe sangat bermacam - macam / variatif, (3) mengolah data dengan performa yang optimal. *Big Data* tidak melulu berasal dari Internet, di dalam kehidupan kita sehari - hari sering kali kita berurusan dengan data, seperti data pada sensor sidik jari ketika absensi, data pembelian pada supermarket, data sensor kelembaban udara pada 10 tahun terakhir untuk memprediksi cuaca, kenaikan dan penurunan harga saham, *bitcoin*, dsb. *Big Data* menjadi topik yang diminati karena dengan data yang begitu banyak, dapat diteliti pola yang terjadi pada data tersebut selama beberapa kurun waktu tertentu untuk digunakan dalam menganalisis data dan membuat keputusan serta memberikan prediksi kemunculan data berikutnya dengan tingkat akurasi yang tinggi berdasarkan data yang dipelajari. Perusahaan - perusahaan saat ini tengah memulai untuk mengumpulkan setiap data yang dapat mereka peroleh dari *customer* untuk melihat pola perilaku dari *customer* mereka dan membuat keputusan yang dapat menguntungkan perusahaan berdasarkan hal tersebut. Tentu saja hal ini tidak dapat dilakukan menggunakan teknik komputasi yang tradisional (menggunakan satu buah komputer berteknologi tinggi), karena biaya dan waktu yang terlalu mahal dan lama.

*Apache Hadoop* merupakan platform yang dibuat untuk menangani permasalahan yang muncul pada *Big Data* dan melakukan analisis pada *Big Data*. Hadoop merupakan sebuah platform *open-source* yang terdiri dari beberapa *cluster* yang saling bekerja sama untuk mengolah data berdasarkan sistem yang terdistribusi dan mampu melibatkan ratusan bahkan ribuan *cluster* yang dapat menjadi *node worker*-nya. Hadoop memiliki dua komponen utama yaitu HDFS (*Hadoop Distributed File System*) dan MapReduce. MapReduce adalah sebuah model pemrograman untuk memproses data yang sangat besar. MapReduce menggunakan algoritma paralel dan terdistribusi. Fungsi MapReduce tersebut akan menyaring, memperkecil, dan melakukan agregasi terhadap data sehingga data yang tidak diperlukan akan dihilangkan. HDFS adalah sebuah sistem file yang

terdistribusi yang didesain untuk beroperasi di dalam suatu kumpulan hardware (*a set of commodity hardware*). Jika dibandingkan dengan *file-system* lainnya, HDFS dirancang untuk menyimpan data set yang besar dan memiliki bandwidth yang tinggi untuk melakukan streaming data tersebut.

Di samping itu, Data Mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Melihat perkembangan dan pertumbuhan data yang kian semakin tinggi, teknik data mining sangat cocok untuk diimplementasikan pada *Big Data*. Karena, diharapkan teknik data mining memiliki tingkat akurasi yang tinggi sebanding dengan volume data yang kian meninggi. Kakas *data mining* meramalkan tren dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. *Data Mining* dapat menjawab pertanyaan - pertanyaan bisnis yang dengan cara tradisional memerlukan banyak waktu untuk menjawabnya. *Data Mining* mengeksplorasi data - data yang ada untuk menemukan pola - pola yang tersembunyi dan mencari informasi pemrediksi yang mungkin saja terlupakan oleh para pelaku bisnis karena tidak terpikirkan sebelumnya oleh mereka. Pada *data mining* teknik tersebut dinamakan sebagai teknik klasifikasi. Salah satu algoritma yang cukup populer digunakan pada teknik klasifikasi adalah algoritma NBC (*Naive Bayes Classifier*). Berbeda dengan beberapa algoritma yang menerapkan teknik klasifikasi lainnya, pada algoritma NBC tidak diperlukan proses yang berjalan secara iteratif. Sehingga, algoritma NBC ini cocok untuk diimplementasikan pada Hadoop dengan berbasiskan MapReduce.

Fokus penelitian skripsi ini adalah untuk menggunakan sistem terdistribusi Hadoop dalam memecahkan 3 masalah utama yang dimiliki oleh *Big Data* dalam menerapkan algoritma teknik data mining menggunakan NBC dalam melakukan klasifikasi berdasarkan data yang diberikan.

## 1.2 Rumusan Masalah

Dari latar belakang tersebut, rumusan masalah yang dibahas pada skripsi ini adalah :

1. Bagaimana merancang algoritma *Naive Bayes Classifier* berbasis *MapReduce* pada lingkungan sistem terdistribusi Hadoop ?
2. Bagaimana mengimplementasikan algoritma *Naive Bayes Classifier* berbasis *MapReduce* pada lingkungan sistem terdistribusi Hadoop ?
3. Bagaimana melakukan pengujian pada algoritma *Naive Bayes Classifier* berbasis *MapReduce* dengan *Big Data* ?
4. Bagaimana melakukan eksperimen terhadap algoritma *Naive Bayes Classifier* pada lingkungan terdistribusi Hadoop menggunakan *Big Data* ?

## 1.3 Tujuan

Berdasarkan identifikasi masalah, tujuan penelitian sebagai berikut:

1. Merancang algoritma *Naive Bayes Classifier* berbasis *MapReduce* pada lingkungan terdistribusi Hadoop.
2. Mengimplementasikan algoritma *Naive Bayes Classifier* berbasis *MapReduce* pada lingkungan terdistribusi Hadoop.
3. Menguji hasil implementasi algoritma *Naive Bayes Classifier* dengan *Big Data*.
4. Melakukan eksperimen pada algoritma *Naive Bayes Classifier* pada lingkungan terdistribusi Hadoop menggunakan Big Data

## 1.4 Batasan Masalah

Batasan masalah dari penelitian ini antara lain :

1. Karena keterbatasan sumber daya yang dimiliki, digunakan 1 komputer sebagai *master node* dan 4 komputer sebagai *slave node* selama pengujian dan eksperimen berlangsung.
2. Pada eksperimen dengan big data, digunakan data nyata (diperoleh dari <https://www.kaggle.com/murderaccountability/homicide-reports>) yang digandakan.

## 1.5 Metodologi Penelitian

Langkah-langkah yang dilakukan dalam penelitian ini adalah :

1. Melakukan studi literatur tentang sistem terdistribusi Hadoop dan *tools* lainnya yang dapat membantu
2. Melakukan studi literatur tentang klasifikasi menggunakan algoritma NBC (*Naive Bayes Classifier*)
3. Mempelajari Hadoop MapReduce dan membuat program - program kecil yang dapat mendukung implementasi dari algoritma NBC berbasis MapReduce
4. Merancang algoritma NBC berbasis *MapReduce*
5. Mengumpulkan data yang dianalisis/diuji (input)
6. Melakukan implementasi klasifikasi menggunakan algoritma NBC pada sistem terdistribusi Hadoop dengan *Big Data*
7. Menganalisis studi kasus untuk data yang berukuran kecil, menengah, dan sangat besar (*Big Data*)
8. Merancang teknik analisis hasil data dari output pada hasil dari pekerjaan *Hadoop MapReduce*
9. Melakukan pengujian dan eksperimen untuk menguji performa sistem

## 1.6 Sistematika Pembahasan

Sistematika pembahasan pada skripsi ini, yaitu :

1. Bab 1 Pendahuluan, berisi tentang latar belakang dan permasalahan utama yang dibahas pada penelitian ini yang kemudian akan dipecahkan menjadi beberapa poin penting, tujuan dari penelitian, batasan masalah, metodologi penelitian yang digunakan, dan sistematika pembahasan pada penelitian ini.
2. Bab 2 Landasan Teori, berisi tentang teori dasar dan pengetahuan mengenai Sistem Terdistribusi Hadoop dan Algoritma *Naive Bayes Classifier*. Pada bab ini dijelaskan juga mengenai beberapa *framework* yang digunakan dalam membangun perangkat lunak.
3. Bab 3 Analisis, berisi tentang analisis masalah yang telah dideskripsikan pada Bab 1 dan menentukan seluruh kebutuhan dari perangkat lunak yang dibangun.
4. Bab 4 Perancangan Perangkat Lunak, berisi tentang rancangan perangkat lunak yang dibangun. Perancangan perangkat lunak meliputi perancangan antarmuka, diagram kelas lengkap, dan rincian metode - metode yang ada pada kelas.

5. Bab 5 Implementasi, Pengujian, dan Eksperimen Perangkat Lunak, berisi tentang hasil dari implementasi, pengujian, dan eksperimen yang dilakukan pada perangkat lunak pada lingkungan terdistribusi Hadoop.
6. Bab 6 Kesimpulan dan Saran, berisi tentang kesimpulan atas hasil penelitian yang telah dilakukan, apakah semua masalah pada rumusan masalah dapat terselesaikan atau tidak, serta saran untuk penelitian yang masih bisa dikembangkan dari penelitian ini.