



**APPLICATION OF NAÏVE BAYES CLASSIFIER METHOD FOR
PREDICTING CLAIMS IN AUTOMOBILE INSURANCE**

UNDERGRADUATE THESIS

**Submitted as one of the requirements to
obtain
Sarjana Aktuaria**

**By:
MICHELYNN SOLA GRATIA JIRENE
021202000019**

**FACULTY OF BUSINESS
ACTUARIAL SCIENCE STUDY PROGRAM**

CIKARANG

NOVEMBER 2023

PLAGIARISM REPORT

Application of Naive Bayes Classifier Method for Predicting Claims in Automobile Insurance

ORIGINALITY REPORT

13%

SIMILARITY INDEX

7%

INTERNET SOURCES

5%

PUBLICATIONS

8%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Universidade Nova De Lisboa

Student Paper

2%

2

Submitted to Asia Pacific University College of Technology and Innovation (UCTI)

Student Paper

1%

3

Submitted to Kaplan University

Student Paper

<1%

4

Submitted to University of New South Wales

Student Paper

<1%

5

Submitted to University of San Diego

Student Paper

<1%

6

Efstathios Kirkos. "Audit-firm group appointment: an artificial intelligence approach", Intelligent Systems in Accounting Finance & Management, 2009

Publication

<1%

7

Submitted to Mercer University

Student Paper

<1%

ANTI-PLAGIARISM REPORT

AI Scan



This text is most likely to be written by **a human**

There is a **7%** probability this text was entirely written by AI ⓘ

APPLICATION OF NAÏVE BAYES CLASSIFIER METHOD FOR PREDICTING CLAIMS IN AUTOMOBILE INSURANCE

By

Michelynn Sola Gratia Jirene

ID no.

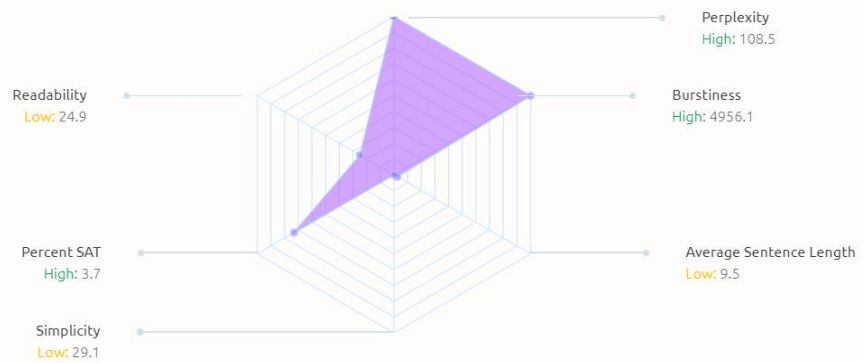
021202000019

A Skripsi presented to the

Faculty of Business President University in partial fulfillment of the requirements for Bachelor Degree in Actuarial Science

0/79 sentences are likely AI generated. ⓘ

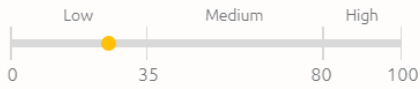
Writing Analysis



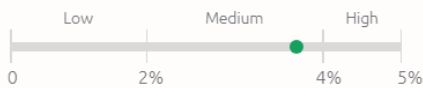
These measurements have been normalized on a scale of 1–100 for display on this chart.

Readability: 24.9

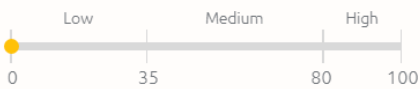
Sentences with short words and low amount of syllables have high readability scores.

**Percent SAT: 3.7%**

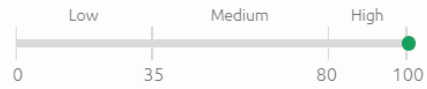
Measures what percentage of words are SAT words, terms from a standardized college admissions exam known for its labyrinthine vocabulary lists.

**Simplicity: 29.1%**

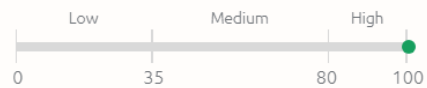
Measures what percentage of words are in the 100 most common words in the English language.

**Perplexity: 108.5**

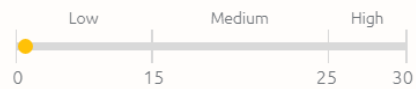
How familiar a piece of text is to large language models like ChatGPT.

**Burstiness: 4956.1**

Unique score developed by GPTZero in 2022 that correlates to variance in writing. Humans generally vary their writing patterns over time.

**Average Sentence Length: 9.5 words**

Unique score that correlates to variance in writing, where humans generally vary writing patterns.



ABSTRACT

Insurance companies often experience difficulties in planning funds. This is caused by the risk of uncertainty in life. Not least, the insurance company went bankrupt because they could not pay their obligations. For this reason, insurance companies need to know the right strategy for setting up reserve funds. One solution that can help insurance companies make decisions and determine strategies is to make claim prediction. In this study, the author will use the Naïve Bayes Classifier method to predict claims in automobile insurance. The Naïve Bayes Classifier itself is a simple probability method where the calculations are based on Bayes' Theorem. The data used is secondary data from Kaggle.com where this data consists of 10,000 samples with 19 features. The prediction results will be divided into two results, namely 0 and 1 where 0 means "no" and 1 means "yes". "yes" or "no" label will inform us whether the customer will claim or not. The data will go through preprocessing in python so that the format is appropriate. The model will be built without feature selection and with feature selection then being compared to determine the best model. Predicted data result will be compared to the actual data and the accuracy of the best model is 82%. Other evaluation method was applied to evaluate how well the model performed by using ROC – AUC score which has a score of 0.87 and 10-fold cross validation which has an average score of 80%. The result of the prediction will help insurance company with underwriting decision and financial approach or planning.

Keywords: *Claim Prediction, Naïve Bayes Classifier, Automobile Insurance*

ABSTRAK

Perusahaan asuransi umumnya mengalami kesulitan dalam perencanaan dana. Hal ini disebabkan oleh resiko ketidakpastian dalam hidup. Tidak sedikit, Perusahaan asuransi mengalami kebangkrutan karena tidak dapat membayar kewajiban mereka. Karena itu, Perusahaan asuransi perlu untuk mengetahui strategi yang tepat dalam mencadangkan dana. Salah satu solusi yang dapat membantu perusahaan asuransi dalam mengambil keputusan dan menetapkan strategi adalah dengan melakukan prediksi klaim. Dalam penelitian ini, penulis akan menggunakan metode Naïve Bayes Classifier untuk memprediksi klaim pada asuransi automobile. Naïve Bayes Classifier sendiri merupakan metode probabilitas sederhana dimana perhitungannya berdasarkan Teorema Bayes. Data yang digunakan adalah data sekunder dari Kaggle.com dengan 10.000 sampel dan 19 fitur. Hasil prediksi akan dibagi menjadi 2 hasil yaitu 0 dan 1 dimana 0 berarti “tidak” dan 1 berarti “ya”. Label “ya” dan “tidak” akan memberikan informasi apakah pelanggan akan melakukan klaim atau tidak. Data akan melalui proses preproses di python agar format sesuai. Model akan dibuat tanpa seleksi fitur dan dengan seleksi fitur lalu dibandingkan untuk menentukan model terbaik. Hasil prediksi data akan dibandingkan dengan data aktual dan akurasi pada model terbaik adalah 82%. Metode evaluasi lain digunakan untuk mengevaluasi seberapa baik performa model dengan menggunakan nilai ROC – AUC dengan nilai 0,87 dan 10-fold cross validation dengan nilai rata-rata 80%. Hasil prediksi akan membantu perusahaan asuransi dalam keputusan underwriting dan rencana pendekatan finansial.

Kata kunci: Prediksi Klaim, Naïve Bayes Classifier, Asuransi Automobile

TABLE OF CONTENTS

COVER	i
PANEL OF EXAMINERS APPROVAL	ii
STATEMENT OF ORIGINALITY	iii
SCIENTIFIC PUBLICATION APPROVAL FOR ACADEMIC INTEREST ..	iv
ADVISOR APPROVAL FOR JOURNAL/INSTITUTION'S REPOSITORY .	v
PLAGIARISM REPORT	vi
ANTI-PLAGIARISM REPORT	vii
ABSTRACT	ix
ACKNOWLEDGEMENT	xi
TABLE OF CONTENTS.....	xii
LIST OF TABLES.....	xv
LIST OF FIGURES	xvii
LIST OF ABBREVIATIONS	xviii
CHAPTER I INTRODUCTION	1
1.1. Research Background.....	1
1.2. Problem of the Research.....	4
1.3. Research Question.....	4
1.4. Research Objective.....	4
1.5. Research Scope and Limitation	5
1.6. Research Benefit	5
1.7. Research Outline	6
CHAPTER II LITERATURE REVIEW	7
2.1. Insurance.....	7
2.2. Data Mining	8
2.3. Machine Learning	9
2.3.1. Machine Learning Definition	9
2.3.2. Supervised and Unsupervised Machine Learning	9
2.3.3. Types of Supervised and Unsupervised Learning Algorithm	11
2.3.4. Data Preprocessing	12

2.3.5. Training and Testing.....	13
2.4. Probability	14
2.5. Conditional Probability	16
2.6. Bayes' Theorem	16
2.6.1. Naïve Bayes Classifier.....	18
2.6.2. Advantage and Disadvantage	22
2.7. Confusion Matrix	22
2.8. ROC - AUC Curve	24
2.9. K-fold Cross Validation	26
2.10. Previous Research and Research Gap	28
CHAPTER III METHODOLOGY	30
3.1. Research Design.....	30
3.2. Sampling Plan	30
3.3. Instrument/Operational Definition	30
3.4. Data Collection Design	31
3.5. Flowchart	32
CHAPTER IV ANALYSIS AND RESULT	34
4.1. Data Preparation.....	34
4.2. Import Data.....	34
4.3. Check and Fill Null Data	37
4.4. Feature Plotting	41
4.5. Transform Data into Numerical	54
4.6. Prediction.....	57
4.7. Evaluation.....	58
4.8. Categorical Feature Model Prediction and Evaluation	60
4.9. Numerical Feature Model Prediction and Evaluation	63
4.10. 5 Feature Selection Model Prediction and Evaluation	66
4.11. 7 Feature Selection Model Prediction and Evaluation	68
4.12. 9 Feature Selection Model Prediction and Evaluation	71
CHAPTER V CONCLUSION	74
5.1. Conclusion	74
5.2. Recommendation	75

REFERENCES	76
APPENDICES	79

LIST OF TABLES

Table 2.1 Sample Table for Naïve Bayes Classifier Example	20
Table 2.2 x and $f(x)$ Approximate Value	25
Table 4.1 Raw Data 1	34
Table 4.2 Raw Data 2	35
Table 4.3 Raw Data 3	36
Table 4.4 Raw Data Info	36
Table 4.5 Sum of Null Data	37
Table 4.6 Columns of Filled Data 1	38
Table 4.7 Columns of Filled Data 2	39
Table 4.8 Columns of Filled Data 3	39
Table 4.9 Recheck Sum of Null Data	40
Table 4.10 New Data Info with Filled Columns	41
Table 4.11 Updated Data 1.....	51
Table 4.12 Updated Data 2.....	52
Table 4.13 Updated Data 3.....	53
Table 4.14 Updated Data Info	53
Table 4.15 Unique Value	54
Table 4.16 Transformed Data 1.....	55
Table 4.17 Transformed Data 2.....	56
Table 4.18 Transformed Data Info	56
Table 4.19 Ten Rows of Probability.....	57
Table 4.20 Classification Report	59
Table 4.21 Categorical Features Data Info	61
Table 4.22 Categorical Feature Model Classification Report.....	62
Table 4.23 Numerical Features Data Info	63
Table 4.24 Numerical Feature Model Classification Report	65
Table 4.25 5 Feature Selection Model Data Info	66
Table 4.26 5 Feature Selection Model Classification Report	67
Table 4.27 7 Feature Selection Model Data Info	68

Table 4.28 7 Feature Selection Model Classification Report	70
Table 4.29 9 Feature Selection Model Data Info	71
Table 4.30 9 Feature Selection Model Classification Report	72

LIST OF FIGURES

Figure 3.1 Flowchart.....	32
Figure 4.1 Count Plot Between Age and Outcome	42
Figure 4.2 Count Plot Between Gender and Outcome	42
Figure 4.3 Count Plot Between Race and Outcome	43
Figure 4.4 Count Plot Between Driving Experience and Outcome.....	43
Figure 4.5 Count Plot Between Education and Outcome	44
Figure 4.6 Count Plot Between Income and Outcome	45
Figure 4.7 Count Plot Between Vehicle Ownership and Outcome	45
Figure 4.8 Count Plot Between Vehicle Year and Outcome	46
Figure 4.9 Count Plot Between Married and Outcome	46
Figure 4.10 Count Plot Between Children and Outcome	47
Figure 4.11 Count Plot Between Vehicle Type and Outcome	48
Figure 4.12 Box Plot Between Speeding Violations and Outcome.....	48
Figure 4.13 Box Plot Between DUIS and Outcome	49
Figure 4.14 Box Plot Between Past Accidents and Outcome	50
Figure 4.15 Box Plot Between Credit Score and Outcome.....	50
Figure 4.16 Box Plot Between Annual Mileage and Outcome	51
Figure 4.17 Confusion Matrix Plot.....	59
Figure 4.18 ROC – AUC Curve	60
Figure 4.19 Categorical Feature Model Confusion Matrix Plot.....	62
Figure 4.20 Categorical Feature Model ROC – AUC Curve	63
Figure 4.21 Numerical Feature Model Confusion Matrix Plot	64
Figure 4.22 Numerical Feature Model ROC – AUC Curve	65
Figure 4.23 5 Feature Selection Model Confusion Matrix Plot	67
Figure 4.24 5 Feature Selection Model ROC – AUC Curve.....	68
Figure 4.25 7 Feature Selection Model Confusion Matrix Plot	69
Figure 4.26 7 Feature Selection Model ROC – AUC Curve.....	70
Figure 4.27 9 Feature Selection Model Confusion Matrix Plot	72
Figure 4.28 9 Feature Selection Model ROC – AUC Curve.....	73

LIST OF ABBREVIATIONS

AI	: Artificial Intelligence
ML	: Machine Learning
NBC	: Naïve Bayes Classifier
ROC	: Receiver Operator Characteristic Curve
TP	: True Positive
TN	: True Negative
FP	: False Positive
FN	: False Negative
TPR	: True Positive Rate
FPR	: False Positive Rate
OJK	: Otoritas Jasa Keuangan
KDD	: Knowledgeable Discovery in Data
AUC	: Area Under the Curve
DUIS	: Driving Under Influences
PDF	: Probability Density Function
CDF	: Cumulative Distribution Function