SKRIPSI

PERBANDINGAN PREDIKSI *FRAUD* KLAIM ASURANSI KENDARAAN BERMOTOR DENGAN MENGGUNAKAN MODEL *LIGHT GRADIENT BOOSTING MACHINE* DAN *EXTREME GRADIENT BOOST*

DIEGO AREND SIANIPAR

NPM: 6161901089

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN

2023

FINAL PROJECT

COMPARISON OF INSURANCE CLAIM FRAUD PREDICTION USING LIGHT GRADIENT BOOSTING MACHINE AND EXTREME GRADIENT BOOST MODEL

DIEGO AREND SIANIPAR

NPM: 6161901089

DEPARTMENT OF MATHEMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2023

LEMBAR PENGESAHAN

PERBANDINGAN PREDIKSI *FRAUD* KLAIM ASURANSI KENDARAAN BERMOTOR DENGAN MENGGUNAKAN MODEL *LIGHT GRADIENT BOOSTING MACHINE* DAN *EXTREME GRADIENT BOOST*

Diego Arend Sianipar

NPM: 6161901089

Bandung, 26 Mei 2023

Menyetujui,

Pembimbing 1	Pembimbing 2
Benny Yong, Ph.D.	**Robyn Irawan, M.Sc.**

Ketua Penguji	Anggota Penguji
Maria Anestasia, M.Si., M.Act.Sc.	**Felivia, MActSc, ASAI**

Mengetahui,

Ketua Program Studi

Dr. Livia Owen

PERNYATAAN

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

PERBANDINGAN PREDIKSI *FRAUD* KLAIM ASURANSI KENDARAAN BERMOTOR DENGAN MENGGUNAKAN MODEL *LIGHT GRADIENT BOOSTING MACHINE* DAN *EXTREME GRADIENT BOOST*

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,
26 Mei 2023

Diego Arend Sianipar
NPM: 6161901089

ABSTRAK

Asuransi sangat dibutuhkan belakangan ini karena masih maraknya kemungkinan risiko yang di luar dugaan kita sebagai manusia. Namun, masih ada beberapa orang yang melakukan klaim asuransi atas kejadian yang telah diatur sedemikian rupa sesuai dengan kontrak polis asuransi, yang padahal tidak terjadi demikian. Maka dari itu, perusahaan asuransi harus berhati-hati untuk menjaga dari penipuan atau *fraud* klaim asuransi. *Machine learning* adalah alat yang menggunakan data dan melatih dirinya untuk mendapatkan hasil yang diinginkan dengan membagi data menjadi dua bagian yaitu, data latih dan data uji. *Decision tree* adalah algoritma yang berbentuk sebuah pohon yang pembangunan modelnya ditentukan oleh nilai *information gain*. *Gradient boosting* adalah algoritma dengan tujuan utama untuk mencari fungsi aproksimasi yang memetakan hasil sehingga nilai dari kesalahan perhitungan menjadi minimum. *Gradient tree boosting* merupakan salah satu bentuk dari *gradient boosting* yang menggunakan pohon dan dihitung bobot setiap *terminal node*. XGBoost adalah salah satu algoritma *gradient tree boosting* yang paling banyak digunakan dalam berbagai penelitian dan menggunakan perhitungan *gain of variance* pada pembentukan pohonnya. Skripsi akan menggunakan algoritma LightGBM yang serupa dengan XGBoost tetapi berbeda pada perhitungan *gain of variance*nya. Evaluasi model menggunakan *5-fold cross validation* dengan memerhatikan ukuran seperti *Confusion matrix*, *Area Under Curve* (AUC), *Accuracy*, *Precision*, *Recall*, dan *F1-score* untuk membandingkan performa antara LightGBM dan XGBoost. Ditemukan bahwa model LightGBM secara umum dapat mengklasifikasi *fraud* klaim asuransi dengan lebih baik dibandingkan model XGBoost.

Kata-kata kunci: *fraud*; *machine learning*; XGBoost; LightGBM; *gain of variance*.

ABSTRACT

Insurance is much needed lately due to the risk that is out of our control as humans. However, there are still people that do their insurance claims on what was arranged on the contract that didn't actually happen. Therefore, insurance companies have to be careful to protect themselves from claim fraud. Machine learning is a tool to train itself to obtain the results with splitting the data into training data and testing data. Decision tree is a machine learning algorithm that is determined by information gain. Gradient boosting is an algorithm with the main purpose to find the approximation function that map results so the value of miscalculation becomes minimum. Gradient tree boosting is a specialization of gradient boosting that use trees and the weight of each terminal nodes. XGBoost is one of the gradient tree boosting algorithm that use gain of variance to determined how trees will be built. This paper will use LightGBM algorithm that is similar to XGBoost but is different on calculating the gain of variance. Model evaluation will use 5-fold cross validation focusing on measures like: Confusion matrix, Area Under Curve (AUC), Accuracy, Precision, Recall, and F1-score to compare the performance between LightGBM and XGBoost. It is found that in general, LightGBM model can classify fraud insurance claim compare to XGBoost.

Keywords: fraud; machine learning; XGBoost; LightGBM; gain of variance.

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas berkat dan rahmat-Nya dalam penyelesaian penulisan skripsi dengan judul "Perbandingan Prediksi *Fraud* Klaim Asuransi Kendaraan Bermotor Dengan Menggunakan Model *Light Gradient Boosting Machine* dan *Extreme Gradient Boost*". Skripsi ini dibangun sebagai salah satu syarat yang harus dipenuhi untuk menyelesaikan Program Studi Matematika. Pada kesempatan ini, penulis akan mengungkapkan terima kasih atas dukungan dalam bentuk apapun dari berbahai belah pihak, yakni:

- Orang tua dan keluarga penulis yang sudah memberi dukungan, doa, dan semangat selama proses penyelesaian skripsi ini.
- Bapak Benny Yong, Ph.D. dan Bapak Robyn Irawan, M.Sc. selaku dosen pembimbing yang sudah meluangkan waktu dan selalu sabar memberikan kritik serta saran selama proses penyelesaian skripsi ini.
- Ibu Felivia, MActSc, ASAI dan Ibu Maria Anestasia, M.Si., MActSc selaku dosen penguji yang sudah memberikan kritik, saran, dan kontribusinya dalam penyelesaian serta penyempurnaan skripsi ini.
- Seluruh dosen Universitas Katolik Parahyangan dalam Program Studi Matematika dan di luarnya, serta dosen-dosen dari luar Universitas Katolik Parahyangan, yang sudah mengasah ilmu penulis sehingga penulisan skripsi ini menjadi rapi dan penyelesaiannya lancar.
- Teman-teman yang saya temui di *arcade* Bandung yang sudah menemani dan menghibur penulis di luar perkuliahan.
- Teman-teman lainnya yang sudah memberi dukungan secara langsung maupun tidak langsung sebelum ujian sidang skripsi dan selama penulisan skripsi ini.
- Ibu Catharina Neysia selaku pendamping penulis yang sudah menemani dan memberikan dukungan, doa, dan semangat selama proses penyelesaian skripsi ini.
- Seluruh mahasiswa jurusan Matematika angkatan 2019 atas kebersamaannya selama studi penulis di Universitas Katolik Parahyangan.

Penulis bersyukur sudah melewati pembangunan skripsi ini yang memberikan banyak kenangan dan pelajaran yang akan teringat untuk masa depan. Penulis juga mendoakan semua pihak yang sudah sabar menanggapi kesalahan penulis dari awal sampai akhir pembangunan skripsi ini. Penulis berharap semua pihak tersebut akan jadikan proses pembangunan skripsi penulis sebagai kenangan.

Penulis menyadari bahwa masih banyak kekurangan baik dari isi skripsi dan sikap penulis selama penyelesaian skripsi ini. Oleh karena itu, untuk isi skripsi ini, penulis menerima kritik dan saran yang membangun dalam bentuk apapun agar skripsi ini menjadi lebih baik, lengkap dan terus berkembang, serta berguna dan berkesan bagi pembaca. Untuk sikap penulis selama penyelesaian skripsi ini, penulis akan jadikan proses ini sebagai pelajaran di masa yang akan datang untuk mengembangkan sikap penulis.

Bandung, 26 Mei 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Dalam lingkungan kehidupan saat ini, terdapat sangat banyak risiko atas kejadian-kejadian yang di luar prediksi kita. Karena itu, masyarakat mulai sadar akan pentingnya asuransi sebagai proteksi keuangan dalam bahaya. Proteksi keuangan ini digunakan sebagai jaminan dari kejadian tidak terduga yang melibatkan jumlah uang yang besar. Maka dari itu, perusahaan asuransi akan melakukan banyak transaksi mengenai klaim. Tentunya dari ribuan klaim yang dilakukan akan ada orang yang mencoba melakukan pemalsuan atas suatu kejadian untuk mendapatkan klaim meskipun tidak dalam bahaya. Perusahaan asuransi harus antisipasi pemalsuan ini terjadi agar uang yang dikeluarkan tidak jatuh ke orang yang salah. Pembohongan seperti ini yang biasa disebut dengan *fraud* [1]. Maka dari itu, cara-cara untuk memprediksi terjadinya *fraud* dari pemegang-pemegang polis dengan karakteristik tertentu sangat dibutuhkan. Prediksi atas fraud dapat dilakukan dengan beberapa model prediksi yang performanya dapat dibandingkan untuk mencari model terbaik.

Salah satu model prediksi yang populer dalam statistika adalah regresi linear. Akan tetapi, regresi linear tidak cocok untuk masalah klasifikasi seperti masalah memprediksi sebuah klaim asuransi adalah *fraud* atau tidak karena banyaknya asumsi-asumsi yang harus dipenuhi [2]. Maka dari itu, dalam skripsi ini digunakan pembelajaran mesin untuk melakukan klasifikasi pada masalah *fraud* klaim asuransi.

Pembelajaran mesin adalah sub-bagian dari *artificial intelligence* yang mengandalkan iterasi-iterasi dalam prosesnya agar semakin baik dalam penyelesaian masalah. Algoritma diberi masukan berupa sebagian dari data untuk pelatihan (*traning*) yang menjadi bagian utama dari pembelajaran mesin. Lalu untuk pembangunan model, sisa dari data digunakan untuk diuji (*testing*) ketepatan luaran algoritma dan dibandingkan terhadap luaran pada data sebenarnya. Pada metode pembelajaran mesin terdapat bermacam algoritma yang dapat digunakan dari sederhana sampai kompleks [3]. Beberapa metode yang dibahas dalam skripsi ini berlandaskan dari sebuah metode yang sederhana, yaitu *decision tree*.

Decision tree adalah algoritma yang berbentuk seperti pohon terbalik yang terdiri dari *root node* di atas, *internal nodes* berisi penentu pembagian data dengan *leaf nodes* atau *terminal nodes* menjadi *node* berisi label kelas atau kelompok. *Decision tree* sangat mudah untuk diinterpretasi karena bentuknya yang sederhana serta cara membagi dalam pohonnya mirip dengan cara manusia membuat keputusan. Akan tetapi, *decision tree* mempunyai kelemahan dalam memprediksi karena sangat sensitif terhadap perubahan dalam data dan sangat mudah untuk terjadinya *overfit* [3]. Maka dari itu, terdapat beberapa metode-metode yang dikembangkan dari *decision tree* untuk

menghindari masalah tersebut.

Boosting adalah algoritma yang memiliki banyak pohon dengan setiap pohon berbentuk *decision tree*. Akan tetapi, pohon dibentuk dengan memperhitungkan kesalahan dari pohon-pohon sebelumnya yang meningkatkan kekuatan pada setiap pohon berikutnya. Algoritma ini memperbaiki kelemahan *decision tree* dengan menggunakan banyak pohon dengan tingkat kompleksitas yang kecil. Karena itu, *boosting* mengembangkan kelemahan dari *decision tree* yang sensitif akan perubahan pada data dengan membagi *dataset* berbeda-beda untuk setiap pohon [3]. Pengembangan dari *boosting* yang banyak digunakan adalah *gradient boosting*. Terdapat dua algoritma utama dalam *gradient boosting* yang akan dibahas dalam skripsi ini, yaitu LightGBM dan XGBoost.

Pada skripsi ini, akan digunakan algoritma LightGBM dan XGBoost untuk menyelesaikan masalah penentuan *fraud* klaim asuransi. Berdasarkan Quinto [3], LightGBM secara umum berjalan lebih baik dibandingkan dengan algoritma yang lebih sering dipakai bernama XGBoost. Akan tetapi, jurnal Zhang dan Gong [4] menunjukkan performa XGBoost lebih bagus dibandingkan dengan LightGBM. Maka dari itu, skripsi ini akan membandingkan LightGBM dengan XGBoost untuk permasalahan prediksi *fraud* klaim asuransi.

1.2 Rumusan Masalah

Untuk menjabarkan masalah yang akan dibahas pada skripsi ini, diberikan rumusan masalah seperti berikut:

1. Bagaimana cara menyelesaikan masalah penentuan *fraud* klaim asuransi?
2. Bagaimana cara menguji performa model yang dihasilkan dari algoritma pembelajaran mesin dalam penentuan *fraud* klaim asuransi?
3. Bagaimana cara menentukan algoritma yang lebih baik dalam mengklasifikasi *fraud* klaim asuransi?

1.3 Tujuan

Tujuan yang ingin dicapai pada skripsi ini adalah sebagai berikut:

1. Menggunakan algoritma LightGBM dan XGBoost untuk memodelkan masalah penentuan *fraud* klaim asuransi.
2. Menggunakan *5-fold cross validation* untuk menguji performa model.
3. Menentukan algoritma yang lebih baik dengan ukuran-ukuran seperti AUC, *accuracy*, *precision*, *recall*, dan *F1-score*.

1.4 *State of the Art*

Dalam skripsi ini, akan dikembangkan ide penggunaan algoritma LightGBM pada data *fraud* klaim asuransi dan membandingkannya dengan algoritma XGBoost yang serupa dan telah dikembangkan lebih dahulu. Pada jurnal Taha dan Malebary [5], algoritma LightGBM diuji dari beberapa ukuran

seperti akurasi dan lainnya. Untuk skripsi ini, aspek-aspek yang sama akan diuji dan dibandingkan dengan algoritma XGBoost untuk menentukan algoritma yang lebih baik pada sebuah *dataset* klaim asuransi nasabah yang bersumber dari Github.