SKRIPSI

ANALISIS DAN PERBANDINGAN PERFORMA REGRESI LINEAR, KUANTIL, POLINOMIAL, DAN *SPLINE* PADA DATA BESAR KLAIM ASURANSI

TIMITHY

NPM: 6161901105

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN

2023

FINAL PROJECT

ANALYSIS AND COMPARISON OF LINEAR, QUANTILE, POLYNOMIAL, AND *SPLINE* REGRESSION PERFORMANCES ON INSURANCE CLAIM DATA

TIMITHY

NPM: 6161901105

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

2023

LEMBAR PENGESAHAN

ANALISIS DAN PERBANDINGAN PERFORMA REGRESI LINEAR, KUANTIL, POLINOMIAL, DAN *SPLINE* PADA DATA BESAR KLAIM ASURANSI

Timithy

NPM: 6161901105

Bandung, 3 Agustus 2023

Menyetujui,

Pembimbing 1	Pembimbing 2
Dr. Livia Owen	Robyn Irawan, M.Sc.

Ketua Penguji	Anggota Penguji
Prof. Dr. Julius Dharma Lesmono	Dr. Andreas Parama Wijaya

Mengetahui,

Ketua Program Studi

Dr. Livia Owen

PERNYATAAN

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

ANALISIS DAN PERBANDINGAN PERFORMA REGRESI LINEAR, KUANTIL, POLINOMIAL, DAN *SPLINE* PADA DATA BESAR KLAIM ASURANSI

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,
3 Agustus 2023

Timithy
NPM: 6161901105

ABSTRAK

Perusahaan asuransi harus memiliki pengertian yang mendalam akan kepentingan arus kas, seperti mempertimbangkan besarnya nominal uang yang perlu dibayarkan apabila tertanggung mengajukan klaim. Pemodelan menggunakan data masa lalu yang diperoleh perusahaan adalah cara yang ampuh untuk mencegah kejadian yang tidak diinginkan, seperti keharusan melakukan peminjaman atau, sampai batas tertentu, kebangkrutan. Salah satu model statistik yang paling populer adalah model regresi, yang meliputi regresi linear dan regresi kuantil. Regresi kuantil dikenal karena ketangguhannya terhadap pencilan, yang seringkali muncul pada data dalam jumlah observasi yang besar. Di sisi lain, model-model linear tidak bisa bekerja dengan optimal dalam mengevaluasi data nonlinear, sehingga diperkenalkan regresi polinomial. Dalam praktiknya, sebaran data dalam rentang tertentu belum tentu sama dengan sebaran dalam rentang lainnya sehingga suatu model regresi juga belum tentu dapat mewakili semua bagian data yang dimodelkan. Untuk mengatasi masalah ini, salah satu model regresi yang lebih fleksibel adalah regresi *spline* yang memampukan pengguna membangun model yang mempertimbangkan persebaran data dengan pola yang berbeda-beda antar bagiannya. Dengan model-model yang tersedia, untuk mengatasi permasalahan-permasalahan tertentu dari sebuah himpunan data, dapat digunakan model yang sesuai, alhasil memberikan hasil yang lebih kredibel dan akurat. Dalam skripsi ini, perbandingan model akan didasarkan dengan tiga metrik, yakni RMSE, *Adjusted R-Squared*, dan AIC. Perlu ditekankan bahwa dalam skripsi ini, regresi linear telah dimodifikasi dengan mengimplementasikan interaksi. Berdasarkan hasil tersebut, model terbaik yang diperoleh adalah model regresi linear yang melibatkan interaksi dan regresi *spline*.

Kata-kata kunci: Asuransi; Regresi Linear; Regresi Kuantil; Regresi Polinomial; Regresi *Spline*.

ABSTRACT

Insurance companies must have a good understanding about the importance of having a solid cash flow such as considering how much money they might have to spend in an instant by the time the insured files a claim. Modelling using past datasets acquired by the company itself is a powerful tool to prevent undesirable events, such as the need to take a loan, or, to an extent, bankruptcy. One of the most popular statistical method to model are regressions. Quantile regression is known for it's robustness against outliers, which oftenly occur in datasets with huge observations. On the other hand, linear models will most definitely have a hard time evaluating nonlinear data, hence polynomial regression is introduced. In practice, the distribution of data within a certain range is not necessarily the same as the distribution within a range other so that a regression model also may not necessarily represent all parts of the data modeled. One of the more flexible regressions available is spline regression which allows users to split datas into several areas to then be further analyzed. Having these and other powerful tools available allows analysts to have a more diversified techniques to fight againts certain problems occured in a dataset with the ideal model, hence giving a more credible and accurate results. Three metrics will be used to determine the best models, which are RMSE, Adjusted R-Squared, and AIC. Note that linear regression will be modified by implementing interaction. Based on the results obtained, the best models are linear regression with interaction and spline regression.

Keywords: Insurance; Linear Regression; Quantile Regression; Polynomial Regression; Spline Regression.

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas berkat dan rahmat-Nya yang memampukan penulis untuk menyelesaikan penulisan skripsi ini dengan lancar. Skripsi "Analisis dan Perbandingan Performa Regresi Linear, Kuantil, Polinomial, dan *Spline* pada Data Besar Klaim Asurans" dibangun sebagai salah satu syarat yang harus dipenuhi guna menyelesaikan Program Studi Matematika. Pada kesempatan ini, penulis rindu untuk mengucapkan terima kasih atas dukungan dalam bentuk apapun dari berbagai belah pihak, yakni:

- Orang tua dan keluarga penulis yang selalu memberi dukungan, kasih, dan doa yang tulus selama proses penyelesaian skripsi ini.
- Ibu Dr. Livia Owen dan Bapak Robyn Irawan, M.Sc. selaku dosen pembimbing yang selalu meluangkan waktu dan dengan sabar memberikan kritik serta saran yang membangun sehingga penulis mampu menyelesaikan skripsi ini dengan baik.
- Bapak Prof. Dr. Julius Dharma Lesmono dan Bapak Dr. Andreas Parama Wijaya selaku dosen penguji yang telah memberikan kritik, saran, dan kontribusinya dalam penyempurnaan skripsi ini.
- Seluruh dosen baik dari dalam Universitas Katolik Parahyangan, yakni dari Program Studi Matematika dan di luarnya, serta yang dari luar Universitas Katolik Parahyangan, yang mengasah ilmu penulis sehingga penulisan skripsi ini dapat diselesaikan dengan lancar.
- Evan Felix, Felix Alpha Winarto, Jovansen Hiustar, Raymond Susanto, dan Willy Zoe Ardiya Tanako selaku sahabat-sahabat penulis yang selalu memberikan dorongan, dukungan, dan hiburan dalam suka maupun duka selama keberlangsungan proses perkuliahan.
- Billy Setiawan, Janice Kusuma Djiwantara, Joice Ivana, dan Vania Rosalie Hadiono yang telah menjadi mitra penulis baik di dalam maupun di luar dunia perkuliahan.
- Aditya Pradipta, Biqytofa, Egha Hafidzal Wirakusuma, Kevin Kartijaya, dan lain-lain yang telah menghibur dan menemani penulis dalam suka dan duka penulis.
- Seluruh mahasiswa angkatan 2019 atas kebersamaannya dalam bentuk apapun selama studi penulis di Universitas Katolik Parahyangan.

Penulis berharap setiap kenangan yang telah dibangun terukir di dalam hati setiap pribadi dan penulis mendoakan kesuksesan untuk setiap jalan yang sedang, atau akan ditempuh.

Penulis menyadari bahwa masih banyak kekurangan dan kesalahan dalam skripsi ini. Oleh karena itu, penulis menerima kritik dan saran yang membangun dalam bentuk apapun supaya menjadi skripsi ini menjadi lebih baik, berkembang, dan berguna bagi pembaca.

Bandung, 3 Agustus 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Manusia merupakan makhluk hidup yang memiliki kehendak bebas, di mana mereka bebas untuk membuat keputusan, dan mengatur masa depannya sampai batas tertentu. Namun, kehendak bebas tersebut tidak berarti bahwa segala sesuatu akan terjadi sesuai kehendaknya. Intinya, manusia tidak luput dari kejadian yang tidak mengenakkan, menakutkan, bahkan mengancam nyawanya [1]. Kejadian-kejadian tersebut bisa saja mengancam keberlangsungan hidup manusia karena dapat memunculkan risiko finansial yang mengakibatkan masalah lanjutan. Maka dari itu, salah satu pendekatan yang berhasil diciptakan manusia untuk meminimalisir risiko finansial dalam kehidupan adalah asuransi.

Pada dasarnya, perusahaan asuransi merupakan perusahaan yang fokus utamanya adalah memperoleh laba dari premi nasabah sehingga perusahaan asuransi memerlukan pengelolaan dana yang baik demi mencegah berbagai masalah finansial perusahaan [2]. Perusahaan asuransi harus membayar manfaat yang relatif besar apabila nasabah mengalami kerugian-kerugian tertentu dan bergantung pada hak yang tercantum pada polis asuransi. Penyeleksian nasabah secara ketat sangat penting bagi perusahaan asuransi, sehingga riwayat kesehatan nasabah sangat krusial demi mencegah pembayaran klaim yang berada di luar jangkauan kemampuan perusahaan. Akibatnya, dibutuhkan keakuratan perhitungan yang dapat menggambarkan kebutuhan perusahaan asuransi dengan tercukupi agar tidak menutup kemungkinan pengalokasian dana ke bidang-bidang yang dapat berkontribusi dalam pengembangan perusahaan tersebut.

Perusahaan asuransi tidak bisa menerima nasabah dengan sembarangan, salah satu cara menyeleksi nasabah adalah dengan memberikan berbagai pertanyaan. Data diperoleh dari jawaban pertanyaan-pertanyaan tersebut akan dikelola menjadi variabel independen. Secara realistis, data dari sebuah perusahaan tidak selalu sesuai dengan yang diharapkan, seperti dengan adanya besar klaim yang abnormal, yang dikenal dengan istilah pencilan [3]. Adanya pencilan dapat memengaruhi hasil analisis data dengan sangat signifikan, sehingga diperlukan pendekatan untuk mengatasi pengaruh buruk yang disebabkan oleh data pencilan.

Regresi dapat digunakan sebagai alat untuk memprediksi nilai suatu variabel terikat menggunakan hubungan antara variabel terikat tersebut dengan variabel independen. Asumsi dasar dari model regresi seperti linearitas, homoskedastisitas, normalitas, dan asumsi-asumsi lainnya harus terpenuhi agar regresi tersebut dapat digunakan dengan optimal [4]. Staffa, Kohane, dan Zurakowski [5] menyatakan bahwa regresi kuantil dapat digunakan apabila asumsi dasar dari regresi linear tidak terpenuhi. Selain itu, keunggulan utama dari regresi kuantil adalah fleksibilitasnya

dalam menganalisis distribusi dari variabel terikat, terutama apabila variabelnya tidak berdistribusi normal. Berbagai penelitian mengenai perkembangan regresi kuantil telah dilakukan, salah satunya dari Alicja Wolny-Dominiak, Agnieszka Ornat-Acedańska, dan Grażyna Trzpiot [6], yang membahas mengenai perbandingan berbagai jenis regresi terhadap regresi kuantil. Berdasarkan riset tersebut, regresi kuantil merupakan regresi yang kokoh terhadap pencilan yang kemungkinan besar ada dalam sebuah himpunan data. Fleksibilitas dan kekokohan tersebutlah yang memperkuat keinginan penulis untuk mendalami model tersebut.

Faktanya, hubungan antara variabel bebas dan terikat dari data-data yang tersedia di dunia nyata tidak selalu linear. Alhasil, diperlukan model-model lain yang lebih fleksibel dan yang lebih mudah menangkap hubungan nonlinear antar variabel. Berbeda dengan model-model regresi yang telah disinggung, sesuai dengan namanya, regresi polinomial mampu mengakomodasi hubungan yang kuadratik dengan menghadirkan bentuk kuadrat dari variabel bebas yang digunakan [7]. Hal ini membuat regresi polinomial menangkap dan memproses pola-pola cekung dari hubungan variabel bebas terhadap variabel terikat. Selain itu, regresi polinomial dapat mencakup berbagai pola data dengan meningkatkan derajat dari polinom agar diperoleh model yang lebih sesuai. Walaupun demikian, perlu dipertimbangkan bahwa dengan meningkatkan derajat dari polinomial yang digunakan, model cenderung *overfit* dan menyebabkan model tidak relevan apabila diaplikasikan ke data lain. Serupa dengan regresi polinomial, regresi *spline* juga mampu mengakomodasi hubungan kuadratik antar variabel. Namun, porsi tertentu pada sebuah data belum tentu memiliki hubungan antara variabel independen dengan variabel terikat yang sama dengan porsi yang lain. Regresi *spline* mampu membagi data menjadi beberapa segmen pada titik-titik yang di mana terjadi perubahan pola persebaran data secara drastis sehingga memberikan hasil yang lebih stabil [8].

Model-model yang telah disinggung memiliki kelebihan dan kekurangannya masing-masing. Skripsi ini diharapkan dapat menjadi membuka pintu-pintu baru untuk penelitian-penelitian selanjutnya dengan menambahkan model-model yang kurang umum untuk mengatasi permasalahan-permasalahan tertentu yang hadir dalam data di dunia nyata, terutama dalam bidang asuransi. Dengan seringnya terjadi permasalahan ekonomi dalam perusahaan asuransi, agar diperoleh perhitungan yang akurat, diharapkan bahwa perusahaan dapat memilih model yang tepat untuk mengatasi berbagai macam permasalahan, salah satunya seperti besar klaim.

1.2 Rumusan Masalah

Rumusan masalah yang terbentuk, berdasarkan latar belakang, yang akan diselesaikan di dalam skripsi ini, antara lain

1. Apakah regresi linear, kuantil, polinomial, dan *spline* dapat digunakan untuk memodelkan besar klaim nasabah sebuah perusahaan asuransi?

2. Berdasarkan data yang digunakan, variabel apa saja yang paling memengaruhi besar klaim nasabah?

3. Bagaimana perbandingan performa masing-masing model regresi terhadap besar klaim nasabah?

1.3 Tujuan

Berdasarkan berbagai masalah yang telah dipaparkan sebelumnya, bagian ini akan memberi pemahaman yang lebih mendalam pada penelitian ini.

1. Memperoleh hasil pemodelan harga klaim menggunakan regresi linear, kuantil, polinomial, dan *spline*, serta menganalisis hasil yang diperoleh menggunakan model-model tersebut.
2. Menentukan signifikansi variabel berdasarkan model-model yang dibangun.
3. Membandingkan performa keempat model regresi menggunakan kesimpulan yang diperoleh terhadap besar klaim nasabah menggunakan ukuran-ukuran tertentu dan dilanjutkan dengan menentukan model terbaiknya.

1.4 Batasan Masalah

Berikut adalah hal-hal yang akan dibatasi dalam skripsi ini:

1. Himpunan data yang akan digunakan adalah berasal dari Kaggle[1].
2. Fokus dari skripsi ini adalah pada model-model regresi yang digunakan sehingga asumsi-asumsi dasar yang digunakan dalam model-model regresi tidak diperiksa.
3. Derajat tertinggi dari regresi polinomial dan *spline* yang digunakan adalah tiga.
4. Dari empat model regresi yang digunakan, interaksi hanya akan diimplementasikan pada model regresi linear.

1.5 *State of the Art*

Dalam skripsi ini, digunakan empat model, yaitu regresi linear, kuantil, polinomial, dan *spline*. Pemilihan model-model ini didasarkan oleh poin-poin berikut:

1. Secara umum, model-model seperti regresi kuantil, polinomial, dan *spline* jarang digunakan sehingga model-model tersebut meningkatkan ketertarikan penulis untuk membahasnya di dalam skripsi ini.
2. Performa sebuah model dapat dipengaruhi oleh terpenuhi atau tidaknya asumsi-asumsi dasar model tersebut [4]. Distribusi dari sebuah data tidak bisa diduga sehingga dibutuhkan model yang fleksibel dan tidak dipengaruhi oleh distribusi data tersebut yang seringkali disebabkan oleh pencilan [3]. Sebagai model yang dapat mengatasi pencilan dengan baik, diharapkan regresi kuantil dapat lebih dikenal, diaplikasikan, dan diperdalam.
3. Hubungan antar atribut data tidak selalu berbanding lurus. Dengan kehadiran regresi polinomial, hubungan kuadrat antar atribut data dapat diakomodasi secara efektif dengan menambahkan suku kuadrat atribut lainnya [7].

[1] https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

4. Selain menghindari *overfit* yang rawan terjadi dalam polinomial berderajat tinggi, regresi *spline* cenderung menghasilkan luaran yang lebih stabil [8].