Aisyah Journal of Informatics and Electrical Engineering
Universitas Aisyah Pringsewu

Journal Homepage
http://jti.aisyahuniversity.ac.id/index.php/AJIEE

IMPLEMENTASI ORANGE DATA MINING UNTUK PREDIKSI BIAYA ASURANSI

Cahyono Budy Santoso

[1] Program Studi Sistem Informasi
Universitas Binaniaga, Bogor, Indonesia
cahyono@stikombinaniaga.ac.id

ABSTRACT

Insurance is a policy that eliminates or reduces the cost of losses incurred by various risks. Various factors affect the cost of insurance. These considerations contribute to the formulation of insurance policies. Machine learning (ML) for the insurance industry sector can make insurance policy formulation more efficient. This study shows how different regression models can forecast insurance costs. And this study compares the results of models such as Neural Network, Gradient Boosting, Random Forest, k-Nearest Neighbors, Decision tree, Multiple Linear Regression, and Support Vector Machine. This paper offers the best approach to the Gradient Boosting model with RMSE value of 4527.749, MAE value of 2460.358, MSE value of 20500507.210865 and R2 value of 0.858.

Keywords: regression, machine learning, insurance

ABSTRACT

Asuransi adalah kebijakan yang menghilangkan atau mengurangi biaya kerugian yang terjadi oleh berbagai risiko. Berbagai faktor mempengaruhi biaya asuransi. Pertimbangan-pertimbangan ini berkontribusi pada formulasi kebijakan asuransi. Machine learning (ML) untuk sektor industri asuransi dapat membuat perumusan polis asuransi menjadi lebih efisien. Studi ini menunjukkan bagaimana model regresi yang berbeda dapat meramalkan biaya asuransi. Dan penelitian ini membandingkan hasil model misalnya *Neural Network, Gradient Boosting, Random Forest, k-Nearest Neighbors, Decision tree, Multiple Linear Regression*, dan *Support Vector Machine*. Paper ini menawarkan pendekatan terbaik pada model *Gradient Boosting* dengan nilai RMSE sebesar 4527.749, nilai MAE sebesar 2460.358, nilai MSE sebesar 20500507.210865 dan nilai R2 sebesar 0.858.

***Keywords**: regresi, machine learning, asuransi*

I. PENDAHULUAN

Pada era sekarang ini banyak masalah ancaman dan ketidakpastian. Orang, rumah tangga, perusahaan, dan properti terpapar berbagai macam bentuk risiko. Resiko tersebut dapat berupa resiko kematian, kesehatan, dan kerugian harta benda atau aset. Hidup dan kesejahteraan adalah bagian terbesar dari kehidupan manusia. Tetapi resiko biasanya tidak dapat dihindari, sehingga dunia keuangan telah mengembangkan berbagai produk untuk melindungi individu dan organisasi dari risiko ini dengan menggunakan modal keuangan untuk menggantinya. Oleh karena itu, asuransi adalah polis yang mengurangi atau menghilangkan biaya kerugian yang ditimbulkan oleh berbagai risiko[1].

Mengenai nilai asuransi dalam kehidupan individu, menjadi penting bagi perusahaan asuransi untuk mengukur jumlah yang ditanggung oleh polis ini dan biaya asuransi yang harus dibayar untuk itu. Berbagai variabel memperkirakan biaya ini. Setiap faktor ini penting, jika ada faktor yang dihilangkan saat jumlahnya dihitung, kebijakan berubah secara keseluruhan. Oleh karena itu penting bahwa tugas-tugas ini dilakukan dengan akurasi yang tinggi, karena kesalahan manusia dapat terjadi, sehingga perusahaan asuransi menggunakan orang yang berpengalaman di bidang ini. Mereka juga menggunakan alat yang berbeda untuk menghitung premi asuransi. *Machine Learning* (ML) bermanfaat di sini. ML dapat menggeneralisasikan upaya atau metode untuk merumuskan kebijakan. Model ML dilatih berdasarkan data asuransi dari masa lalu. Faktor-faktor yang diperlukan untuk mengukur pembayaran kemudian dapat didefinisikan sebagai input model, kemudian model dapat mengantisipasi biaya polis asuransi dengan benar. Hal tersebut mengurangi upaya manusia dan sumber daya dan meningkatkan profitabilitas perusahaan. Dengan demikian akurasi dapat ditingkatkan dengan *ML*. Tujuan penelitian ini adalah memperkirakan biaya asuransi. Nilai dari biaya asuransi didasarkan pada variabel yang berbeda. Sehingga biaya asuransi adalah nilai berkelanjutan. Regresi adalah pilihan terbaik yang tersedia untuk memenuhi kebutuhan penelitian ini. Penelitian ini menggunakan regresi linier berganda dalam analisis karena ada banyak variabel independen yang digunakan untuk menghitung variabel dependen (target). Untuk penelitian ini, menggunakan dataset untuk biaya asuransi kesehatan [2]. *Preprocessing dataset* dilakukan terlebih dahulu. Kemudian dilakukan pelatihan model regresi dengan data pelatihan dan akhirnya mengevaluasi model ini berdasarkan data pengujian. Dalam penelitian ini menggunakan beberapa model regresi, misalnya, *Neural Network, Gradient Boosting, Random Forest, k-Nearest Neighbors, Decision tree, Multiple Linear Regression,* dan *Support Vector Machine*. Sebagai hasil penelitian ditemukan bahwa *Gradient Boosting* memberikan akurasi tertinggi dengan nilai r-kuadrat 0.858.

II. TINJAUAN PUSTAKA

Regresi

Analisis regresi adalah metode prediktif yang mengeksplorasi hubungan antara dependen (target) dan variabel independen (s) (prediktor). Teknologi ini digunakan untuk melakukan peramalan, mengestimasi model time series, dan mencari hubungan sebab akibat antar variabel. Dalam analisis ini, misalnya, menganalisis hubungan antara biaya asuransi (variabel target) dan enam variabel independen berdasarkan umur, BMI, jumlah anak, tempat tinggal individu, atau jenis kelamin dan apakah pelanggan adalah orang yang merokok.

Analisis regresi memperkirakan hubungan antara dua variabel atau lebih, seperti yang dinyatakan sebelumnya. Menggunakan model regresi yang berbeda untuk memperkirakan biaya asuransi kesehatan berdasarkan enam variabel independen, dan dengan menggunakan regresi ini, kita dapat memperkirakan biaya asuransi kesehatan masa depan berdasarkan data saat ini dan masa lalu. Ada beberapa keuntungan menggunakan analisis regresi sebagai berikut:

- Menunjukkan hubungan penting antara variabel dependen dan independen.
- Menunjukkan intensitas efek pada variabel dependen dari beberapa variabel independen.

Analisis regresi juga membantu seseorang untuk membandingkan hasil pengukuran variabel pada berbagai skala, seperti efek variabel independen dan variabel dependen. Keunggulan ini memungkinkan

peneliti pasar, analis data, dan ilmuwan data untuk menentukan rentang variabel terbaik untuk model prediktif.

Model Regresi

1) Regresi Linear Berganda.

Dalam praktiknya, kita sering memiliki lebih dari satu prediktor. Sebagai contoh, dengan kumpulan data yang digunakan dalam penelitian ini, kita mungkin ingin memahami apakah variabel bebas (6 variabel bebas), (secara linier) berhubungan dengan variabel terikat (biaya). Ini disebut sebagai model regresi linier berganda (MLR) [10]. Model MLR dengan variabel independen X_1, X_2, ...X_t dan Y hasilnya dapat dihitung seperti pada persamaan berikut

$$Y = \alpha_0 X_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_t X_t + u$$

Dalam persamaan di atas, u adalah regresi residual sedangkan α adalah bobot dari setiap variabel independen atau parameter yang diberikan.

2) *Random Forest*

Random Forest mencerminkan pergeseran ke pohon keputusan yang dikantongi yang menciptakan sejumlah besar pohon terkait dekorasi sehingga efisiensi prediksi dapat ditingkatkan lebih lanjut. Mereka adalah algoritme pembelajaran 'off-the-box' atau off-the-shelf yang sangat populer, dengan kinerja prediktif yang baik dan hyperparameter yang relatif sedikit.

Ada beberapa implementasi *random forest* yang ada, tetapi algoritma Leo Breiman (Breiman 2001) [12] sekarang sebagian besar bersifat otoritatif. *Random forest* menciptakan nilai prediksi rata-rata sebagai hasil dari seluruh regresi masing-masing pohon. *Random forest* memutuskan untuk overfit [10]. Seperti pada persamaan berikut, model acak untuk regressor hutan dapat dinyatakan.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots + f_n(x)$$

di mana g adalah model akhir yang merupakan jumlah dari semua model. Setiap model f(x) adalah pohon keputusan

3) *XGBoost*.

Baru-baru ini perangkat lunak pembelajaran ansambel baru bernama XGBoost telah diusulkan [13]. Merupakan model peningkatan pohon baru yang memberikan pembelajaran out-of-core yang efektif dan memori yang jarang. Oleh karena itu, XGBoost adalah algoritme pembelajaran terawasi, yang akan sangat berguna untuk masalah prediksi argumen dengan data pelatihan yang luas dan nilai yang hilang. Nilai yang hilang masih belum dapat dikelola dengan pendekatan yang paling populer, seperti *random forest* dan *neural network*. Metode memerlukan kerangka kerja tambahan untuk mengelola nilai yang hilang.

Kekuatan XGBoost meningkatkan penggunaan alat ini di banyak aplikasi lainnya. Misalnya, dalam pemisahan matahari langsung, Aler et al. [14] mengembangkan dua versi XGBoost. Yang pertama adalah model tidak langsung, yang menggunakan XGBoost untuk mempelajari model pemisahan radiasi matahari dari berbagai sumber literatur dalam kumpulan data dari model instruksi level 1 tradisional. Model lain adalah model langsung yang langsung cocok dengan XGboost dalam kumpulan data. Kasus tambahan adalah [15], yang menggunakan XGBoost untuk merekomendasikan sesuatu kepada pengguna menggunakan fungsi yang diturunkan dari pasangan pengguna menggunakan rekayasa fitur yang rumit dalam kerangka rekomendasi. Dalam penelitian ini menganalisis XGBoost sebagai model prediktor untuk biaya asuransi kesehatan.

4) *Support Vector Machine (SVM)*

SVM dapat digeneralisasi untuk masalah dengan regresi (yaitu, ketika hasilnya kontinu seperti variabel target kami dalam penelitian kami). Pada dasarnya, SVM mencari hyperplane dalam ruang fungsi yang diperluas yang biasanya menghasilkan batas keputusan nonlinier dengan efisiensi generalisasi yang kuat di ruang fitur aslinya. Fungsi khusus yang disebut fungsi kernel digunakan untuk membangun fungsionalitas yang diperluas dan terpisah ini.

5) *K-Nearest Neighbors (K-NN)*

K-NN adalah model prediktif yang sangat sederhana yang memprediksi nilai berdasarkan "kemungkinan" mereka dari nilai lain. Berlawanan dengan kebanyakan pendekatan pembelajaran mesin lainnya, KNN bergantung pada memori dan tidak dapat disimpulkan sebagai algoritma tertutup. Ini menyiratkan bahwa data pelatihan diperlukan selama operasi dan prakiraan dihasilkan segera dari hubungan data pelatihan. KNNs juga diidentifikasi sebagai *lazy learning* [16] dan juga tidak efisien secara komputasi. Namun demikian, KNN telah berhasil dalam beberapa masalah pasar [17][18].

6) *Decision Tree* (DT)
DT bersifat langsung, sangat populer [22], pelatihan cepat, dan model yang mudah dibaca dengan metode pembelajaran komparatif atau lainnya dari data. Mereka cukup kompeten tetapi rentan terhadap overfitting dalam prediksi mereka. Mereka dapat diperkuat dengan meningkatkan kinerjanya [23].
7) *Neural Network*

Algoritma ML umumnya mencari representasi data yang optimal dalam bentuk fungsi tujuan dengan menggunakan sinyal umpan balik. Namun, sebagian besar algoritme ML hanya dapat menggunakan maksimal dua lapisan transformasi data untuk mempelajari representasi output. Kami menyebut model dangkal ini karena hanya menggunakan 1–2 representasi ruang fungsional. Karena kumpulan data terus tumbuh dalam ukuran ruang, tidak selalu mungkin untuk menemukan representasi output yang optimal dengan model yang dangkal. Pembelajaran mendalam menawarkan pendekatan multi-lapisan yang umumnya terjadi melalui jaringan saraf berlapis-lapis. Deep neural network (DNN), seperti algoritme pembelajaran mesin lainnya, membuat pembelajaran dengan memetakan fungsi ke target melalui transformasi data sederhana dan indikator umpan balik, DNN menekankan pembelajaran berbagai lapisan representasi yang bermakna.

Penelitian Sebelumnya

Pada bagian ini dibahas upaya penelitian dari eksplorasi informasi dan teknik pembelajaran mesin. Beberapa makalah telah membahas masalah prediksi klaim. Jessica Pesantez-Narvaez menyarankan, "Memprediksi klaim asuransi kendaraan menggunakan data telematika" pada tahun 2019. Penelitian ini membandingkan kinerja regresi logistik dan teknik XGBoost untuk meramalkan adanya klaim kecelakaan dengan jumlah kecil dan hasilnya menunjukkan bahwa karena interpretasinya dan kuat prediktabilitas [3], regresi logistik adalah model yang efektif daripada XGBoost.

Sistem lainnya diusulkan oleh Ranjodh Singh pada tahun 2019. Sistem ini mengambil gambar mobil yang rusak sebagai masukan dan menghasilkan detail yang relevan, seperti biaya perbaikan untuk memutuskan jumlah klaim asuransi dan lokasi kerusakan. Dengan demikian prediksi klaim asuransi mobil tidak diperhitungkan dalam analisis ini tetapi difokuskan pada perhitungan biaya perbaikan [4]. Tujuan dari penelitian ini [5] adalah untuk mempelajari prediksi churn. *Random Forest* dianggap sebagai model terbaik (akurasi 74 persen). Di beberapa bidang, kumpulan data memiliki nilai yang hilang. Mengikuti analisis distribusi, keputusan diambil untuk mengganti variabel yang hilang dengan atribut tambahan yang menunjukkan bahwa data ini tidak ada. Hal ini diperbolehkan hanya jika data benar-benar hilang secara acak, sehingga mekanisme data yang hilang yang memutuskan pendekatan yang tepat untuk pemrosesan data harus ditetapkan terlebih dahulu[6][7].

Pada tahun 2018, Muhammad Arief Fauzan et al. Dalam makalah ini, ketepatan XGBoost diterapkan untuk memprediksi pernyataan. Dengan membandingkan hasil kinerja XGBoost, AdaBoost, Random Forest, Neural Network. XGBoost menawarkan akurasi terstruktur yang lebih baik. Menggunakan dataset Porto Seguro ke Kaggle yang dapat diakses publik. Kumpulan data mencakup nilai NaN dalam jumlah besar tetapi makalah ini mengelola nilai yang hilang dengan penggantian median dan median. Namun, metode sederhana dan tidak berprinsip ini juga terbukti bias[7]. Oleh karena itu, mereka berkonsentrasi untuk mengeksplorasi metode ML yang sangat sesuai untuk masalah beberapa nilai yang hilang, seperti XGboost[8].

G. Kowshalya, M. Nandhini. pada tahun 2018 telah dikembangkan studi untuk memprediksi dan memperkirakan klaim penipuan dan persentase premi untuk berbagai pelanggan berdasarkan data pribadi dan keuangan mereka. Untuk klasifikasi, algoritma Random Forest, J48, dan Naïve Bayes dipilih. Temuan menunjukkan bahwa Random Forest melebihi teknik yang tersisa tergantung pada dataset sintetik. Oleh karena itu, makalah ini tidak mencakup perkiraan klaim asuransi, melainkan berfokus pada klaim palsu [9]. Pekerjaan sebelumnya di atas tidak mempertimbangkan prediksi biaya atau keparahan klaim, mereka hanya membuat klasifikasi untuk masalah klaim (apakah klaim diajukan untuk pemegang polis tersebut atau tidak).

Penelitian ini fokus pada metode statistik lanjutan dan algoritma *machine learning* untuk memprediksi biaya asuransi kesehatan.

III. METODOLOGI

Penelitian ini bertujuan untuk melakukan analisa perbadingan metode *Neural Network, Gradient Boosting, Random Forest, k-Nearest Neighbors, Decision tree, Multiple Linear Regression*, dan *Support Vector Machine* yang digunakan untuk memprediksi biaya asuransi. Untuk aplikasi yang digunakan untuk simulasi adalah Orange Data Mining yaitu aplikasi data mining open source yang terbukti mampu membantu peneliti menganalisa datanya. Tahapan proses pada riset ini bisa dilihat pada Gambar 1.

Gambar 1. Tahapan penelitian

Sesuai Gambar 1 tersebut, langkah pertama adalah identifikasi masalah, perumusan dan kajian pustaka hal ini dilakukan pertama kali untuk menyusun tujuan riset dan kontribusi riset. Kedua adalah proses *collecting data* yaitu menyusun data latih dan data uji sebagai sumber prediksi data. Ketiga adalah proses perancangan *widget* orange data mining untuk proses prediksi biaya dan perbandingan metode. Keempat adalah proses prediksi biaya asuransi mengggunakan model *Neural Network, Gradient Boosting, Random Forest, k-Nearest Neighbors, Decision tree, Multiple Linear Regression*, dan *Support Vector Machine*. Kelima adalah proses evaluasi kinerja metode prediksi dan menganalisa hasil perbandingan metode tersebut.

Atribut Penelitian

Dataset diperoleh dari dataset publik *Kaggle.com* . Atribut awal dari data ini adalah 1 atribut tujuan dan 6 *variable independen*. Deskripsi data set dijelaskan pada Tabel 1.

Tabel 1. *Dataset Overview*

Nama kolom	Deskripsi
age	Usia klien
BMI	Body Mass Index
The Number of Kids	Jumlah anak klien
gender	Laki/Perempuan
smoker	Klient merokok atau tidak
region	Tempat tinggal klien di southwest, southeast, northwest or northeast
Charges(target variable)	Biaya kesehatan yang klien bayar

Data Selection Process / Prepocessing

Kumpulan data mencakup tujuh variabel, seperti yang ditunjukkan pada tabel 1. setiap atribut ini memiliki beberapa kontribusi untuk memperkirakan biaya asuransi, yang merupakan variabel dependen. Pada tahap ini, data diperiksa dan diperbarui dengan benar untuk menerapkan data secara efisien ke algoritme ML.

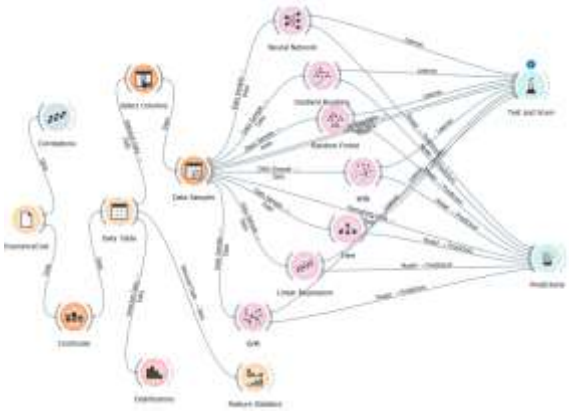
Variabel kategori diterjemahkan ke dalam nilai numerik atau biner untuk mewakili 0 atau 1. Misalnya, jenis kelamin "SEX" dengan laki-laki atau perempuan, variabel "Pria" akan benar (1) jika orangnya laki-laki. Dan "perempuan" adalah (0) dengan melihat tabel 2.

Tabel 2. Konversi Numerik v*ariabel kategori*

Kids	have
gender	Male / Female 1=Male 0=Female
smoker	Klien perokok atau bukan 1=yes 0=no
region	Tempat tinggal klien 1= southwest 2= southeast 3= northwest 4= northeast

Data Mining Process

Dalam menganalisa performa beberapa model pada orange tool, dilakukan perbandingan beberapa metode data mining untuk memilih metode yang terbaik dengan akurasi yang tinggi, dalam prediksi biaya seperti terlihat pada Gambar 2.



Gambar 2. Design *widget* model prediksi biaya asuransi

Pada Gambar 2 merupakan perancangan widget memakai model klasifikasi pada software Orange data mining berupa *Neural Network, Gradient Boosting, Random Forest, k-Nearest Neighbors, Decision tree, Multiple Linear Regression,* dan *Support Vector Machine* yang diinputkan dataset yang telah diolah sebelumnya. Kemudian dataset tersebut diproses kedalam mode prediksi.

Proses Pengujian Model Prediksi

Dalam proses pengujian model prediksi yang telah dibuat sebelumnya, dibutuhkan kumpulan data uji untuk mengetahui hasil prediksi.

Proses Evaluasi Hasil Perbandingan Model Klasifikasi

Proses selanjutnya adalah melakukan proses perbandingan model prediksi dengan menggunakan *Test and Score* diperlukan untuk menghitung tingkat keberhasilan antara masing- masing model prediksi di *data mining Orange* seperti terlihat di Gambar 2.
Pada Gambar 2 merupakan desain *widget* yang telah ditambahkan proses perhitungan tingkat keberhasilan model prediksi dengan menggunakan *widget Test and Score* yang selanjutnya akan dilakukan evaluasi akurasi menggunakan *MSE, RMSE, MAE dan R2*.

IV. HASIL DAN PEMBAHASAN

Hasil simuasi model regresi

Hasil simulasi model prediksi dilakukan dengan menggunakan kumpulan data uji dengan 1 atribut sebagai target, 6 atribute yaitu *age,BMI,The Number of Kids, gender, smoker,* dan *region,* sehingga diperoleh hasil test score seperti terlihat pada Gambar 3.

Model	MSE	RMSE	MAE	R2
kNN	110755750.010	10524.056	7109.119	0.231
Tree	28657962.055	5353.313	2725.346	0.801
SVM	155142241.028	12455.611	10439.623	-0.077
Random Forest	22951470.973	4790.769	2684.412	0.841
Neural Network	304488655.581	17449.603	12934.880	-1.114
Linear Regression	36354923.988	6029.504	4196.534	0.748
Gradient Boosting	20500507.211	4527.749	2460.358	0.858

Gambar 3. Hasil *widget Test and Score*

Berdasarkan data yang telah diuji, diperoleh hasil perhitungan *MSE, RMSE, MAE dan R2* dari masing-masing model seperti terlihat pada Gambar 3. Hasil prediksi model *Neural Network, Gradient Boosting, Random Forest, k-Nearest Neighbors, Decision tree, Multiple Linear Regression,* dan *Support Vector Machine* menunjukkan bahwa nilai akurasi *Gradient Boosting* paling tinggi karena nilai *R2* yang paling mendekati 1 dan nilai *MSE, RMSE, MAE* yang paling kecil.

V. PENUTUP

Penelitian ini menggunakan berbagai model regresi *machine learning* untuk memperkirakan biaya asuransi kesehatan berdasarkan atribut tertentu, pada kumpulan data pribadi biaya medis dari Kaggle.com. Temuan dirangkum dalam gambar 3 yang menunjukkan bahwa Gradient Boosting menawarkan efisiensi terbaik, dengan nilai RMSE sebesar 4527.749, nilai MAE sebesar 2460.358, nilai MSE sebesar 20500507.210865 dan nilai R2 sebesar 0.858. *Gradien Boosting* dapat digunakan dalam estimasi biaya asuransi dengan kinerja yang lebih baik daripada model regresi lainnya. Peramalan biaya asuransi berdasarkan faktor-faktor tertentu membantu penyedia polis asuransi untuk menarik konsumen dan menghemat waktu dalam

merumuskan rencana untuk setiap individu. *Machine Learning* dapat secara signifikan meminimalkan upaya individu dalam pembuatan kebijakan ini, karena model ML dapat melakukan perhitungan biaya dalam waktu singkat, sementara manusia membutuhkan waktu lama untuk melakukan tugas yang sama. Ini akan membantu bisnis meningkatkan profitabilitas mereka. Model ML juga dapat mengelola data dalam jumlah besar.

DAFTAR PUSTAKA

[1] Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE.

[2] Kaggle Medical Cost Personal Datasets. Kaggle Inc. https://www.kaggle.com/mirichoi0218/insurance.

[3] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70

[4] Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

[5] Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data.

[6] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.

[7] Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

[8] Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2).

[9] Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.

[10] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.

[11] Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions. 10.1007/978-3-030-25820-7.

[12] Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.

[13] Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system 22nd ACM SIGKDD Int. In Conf. on Knowledge Discovery and Data Mining.

[14] Aler, R., Galván, I.M., Ruiz-Arias, J.A., Gueymard, C.A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. In Solar Energy vol. 150, pp. 558-569.

[15] Volkovs, M., Yu, G. W., & Poutanen, T. (2017). Content-based neighbor models for cold start in recommender systems. In Proceedings of the Recommender Systems Challenge 2017 (pp. 1-6).

[16] Cunningham, Padraig, and Sarah Jane Delany. 2007. ―K-Nearest Neighbour Classifiers.‖ Multiple Classifier Systems 34 (8). Springer New York, NY, USA: 1–17

[17] Jiang, Shengyi, Guansong Pang, Meiling Wu, and Limin Kuang. 2012. ―An Improved K-Nearest-Neighbor Algorithm for Text Categorization.‖ Expert Systems with Applications 39 (1). Elsevier: 1503 9.

[18] Mccord, Michael, and M Chuah. 2011. ―Spam Detection on Twitter Using Traditional Classifiers.‖ In International

Conference on Autonomic and Trusted Computing, 175–86. Springer.
[19] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140
[20] Breiman, Leo, and others. 2001. ―Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).‖ Statistical Science 16 (3). Institute of Mathematical Statistics: 199–231.
[21] Friedman. 2002. ―Stochastic Gradient Boosting.‖ Computational Statistics & Data Analysis 38 (4). Elsevier: 367–78.
[22] Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. International Journal of Advanced Computer Science and Applications, 9(2).
[23] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press.
[24] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.
[25] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.
[26] Kansara, Dhvani & Singh, Rashika & Sanghvi, Deep & Kanani, Pratik. (2018). Improving Accuracy of Real Estate Valuation Using Stacked Regression. Int. J. Eng. Dev. Res. (IJEDR) 6(3), 571–577 (2018)
[27] Yerpude, P., Gudur, V.: Predictive modelling of crime dataset using data mining. Int. J. Data Min. Knowl. Manag. Process (IJDKP) 7(4) (2017)
[28] Grosan, C., Abraham, A.: Intelligent Systems: A Modern Approach, Intelligent Systems Reference Library Series. Springer, Cham (2011)