Perbandingan Metode Klasifikasi *Random Forest* dan SVM Pada Analisis Sentimen PSBB

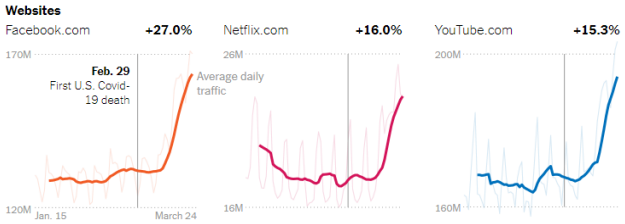
M. R. Adrian[1], M. P. Putra[2] M. H. Rafialdy[3], N. A. Rakhmawati[4]
[1,2,3]*Departemen Sistem Informasi, Fakultas Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember*
Gedung Departemen Sistem Informasi ITS, Kampus ITS, Keputih, Kec. Sukolilo, Kota SBY, Jawa Timur 60117
E-mail : mr.adrian40@gmail.com[1], vitopapuan.vonworks@gmail.com[2], hilmanrfd@gmail.com[3], nur.aini@is.its.ac.id[4]

Abstract—With the continuing increase in the spread of COVID-19 in Indonesia, has made the local government not remain silent. Several local governments in Indonesia have enacted regulations to reduce the growth of COVID-19 victims by limiting public meetings with Large-Scale Social Restrictions or LSSR. However, the implementation of this LSSR has received many comments from social media users, especially from Twitter. This research was conducted with the aim of analyzing the sentiment of implementing the LSSR with media tweets on the Twitter social media platform. The data that were successfully extracted were 466 tweet data with training data and test data having a ratio of 7 to 3. Then the data was calculated into 2 different algorithms to be compared, the first algorithm used was the Support Vector Machine (SVM) algorithm and Random Forest with the aim get the most accurate sentiment analysis results.

Abstrak—Terusnya bertambah angka persebaran *COVID-19* di Indonesia membuat pemerintah tidak tinggal diam. Beberapa pemerintah daerah di Indonesia menetapkan peraturan untuk mengurangi laju pertumbuhan korban *COVID-19* dengan membatasi pertemuan di publik dengan Pembatasan Sosial Berskala Besar atau PSBB. Namun, penerapan PSBB ini ternyata menerima banyak komentar dari pengguna media sosial khususnya melalui media sosial Twitter. Penelitian ini dilakukan dengan tujuan untuk menganalisis sentimen publik mengenai penerapan PSBB dengan medium *tweet* pada *platform* media sosial Twitter. Data yang berhasil di gali sebanyak 466 data *tweet* dengan data latih dan data tes memiliki perbandingan 7 banding 3. Kemudian data tersebut dikalkulasikan kedalam 2 algoritma yang berbeda untuk dikomparasikan, algoritma pertama yang digunakan adalah algoritma *Support Vector Machine* (SVM) dan *Random Forest* dengan objektif mendapatkan hasil analisis sentimen terakurat.
Kata Kunci—Sentiment Analysis, *COVID-19*, PSBB, *Support Vector Machine, Random Forest*

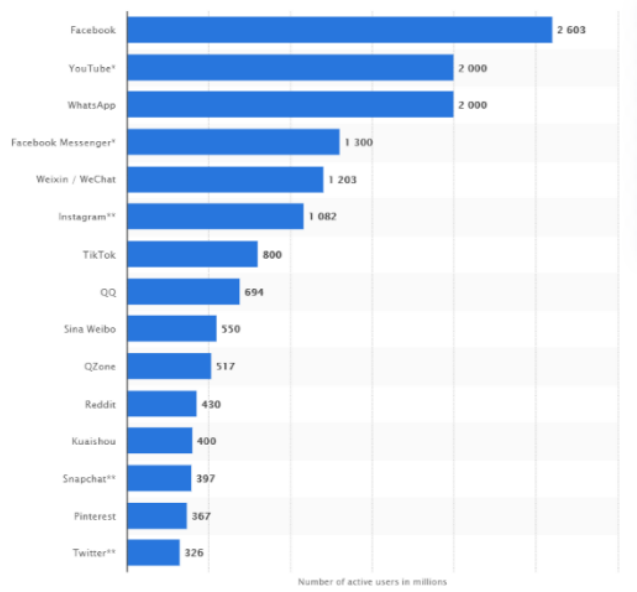
I. PENDAHULUAN

perkembangan teknologi informasi dan komunikasi di era revolusi industri masa ini sangat cepat. Hal ini didukung dengan adanya pandemi *COVID-19* yang membuat segala kegiatan yang bisa dilakukan di tempat, dipaksa untuk berkembang dan dilakukan secara daring. Alhasil, internet menjadi hal yang wajib untuk dimiliki oleh setiap orang. The New York Times, sebuah media koran harian yang diterbitkan di New York dan dipublikasikan secara internasional menganalisis penggunaan internet penduduk di Amerika Serikat bersama SimilarWeb dan Apptopia, berhasil menemukan bahwa setelah adanya pandemi *COVID-19*, kebiasaan kita dalam berinternet pun berubah. Berikut adalah grafik perbandingan frekuensi penggunaan media sosial sebelum dan sesudah pandemi menyerang.



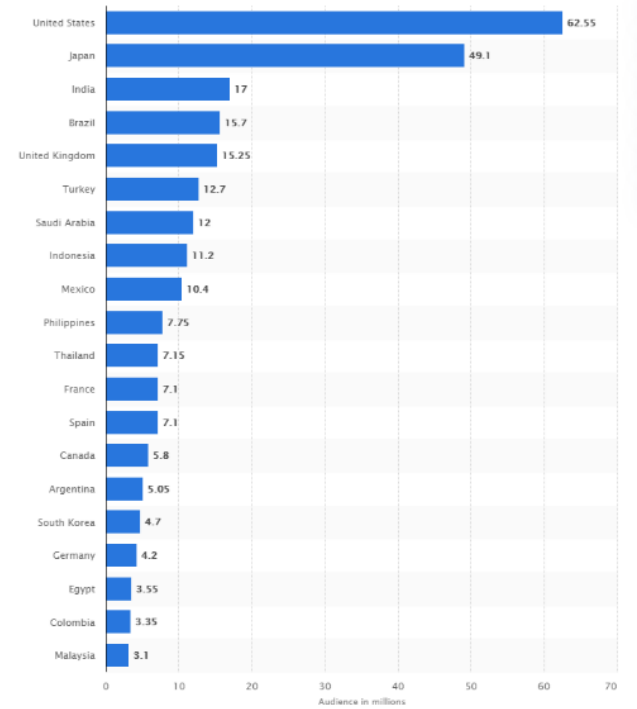
Gambar 1 menjelaskan Perbandingan Frekuensi Penggunaan Media Sosial di Internet setelah Pandemi (Sumber : www.nytimes.com)

Dapat dilihat dari grafik di atas bahwa pengguna media sosial di Amerika Serikat meningkat pesat seiring pandemi datang. Data yang diteliti dari 15 Januari hingga 24 Maret tersebut menunjukkan bahwa orang-orang selama pandemi ini lebih aktif dalam menggunakan media sosial. Dalam penelitian ini, kami menggunakan salah satu media sosial yaitu Twitter. Berikut adalah statistik jumlah pengguna Twitter di dunia per Juli 2020.



Gambar 2 adalah grafik soal Sosial Media Terpopuler secara Internasional per Juli 2020 (Sumber : www.statista.com)

Dari statistik di atas, dapat dilihat bahwa jumlah pengguna aktif Twitter mencapai 326 juta pengguna. Kemudian kita bandingkan data tersebut dengan jumlah pengguna media sosial Twitter berdasarkan negaranya.



Gambar 3 Negara dengan Jumlah User Twitter terbanyak per Juli 2020
(Sumber : www.statista.com)

Indonesia menempati posisi ke-8 di urutan pengguna Twitter terbanyak. Hal ini tidak aneh mengingat dengan adanya pandemi banyak warga Indonesia yang membutuhkan hiburan dan interaksi sosial yang sangat sulit didapatkan secara offline.

Twitter sering kali digunakan sebagai media penyampaian aspirasi pribadi, baik senang, sedih, maupun yang kontroversial. Dengan adanya pandemi yang mewajibkan kita untuk melakukan kegiatan di rumah dan tidak boleh berkegiatan banyak di publik, para pengguna Twitter pun sering kali menyampaikan keluh kesah mereka mengenai hal tersebut. Kebijakan dari pemerintah mengenai tidak bolehnya orang orang berkegiatan dalam jumlah banyak di publik disebut dengan PSBB.

PSBB merupakan singkatan dari Pembatasan Sosial Berskala Besar. Penerapan PSBB dilakukan dengan tujuan memberikan jaminan bahwa rantai penularan *COVID-19* ini bisa diputuskan jika dijalani secara bersamaan. Di antaranya mencegah terjadinya berkumpul orang. Baik dalam konteks untuk berkumpul alasan kesenian, alasan budaya ataupun alasan olahraga lainnya. Lalu juga dengan cara tidak melakukan mobilisasi sosial untuk kepentingan apapun apabila tidak diperlukan [1].

Tidak sedikit orang orang yang merasa disulitkan dengan kondisi seperti ini yang kemudian mengutarakan pendapatnya di banyak media sosial, termasuk Twitter. Karena banyak warga Indonesia yang merasa susah dalam melaksanaan pekerjaannya apabila tidak bisa ke tempat pekerjaannya langsung. Tetapi, tidak sedikit juga warga Indonesia yang merasa diuntungkan dengan kondisi seperti ini.

Berdasarkan latar belakang tersebut, tujuan dari penelitian ini adalah untuk menganalisa klasifikasi sentimen pengguna media sosial twitter terkait PSBB. Output dari penelitian ini adalah ingin membuktikan apakah persepsi warga Indonesia terhadap kebijakan PSBB itu positif atau negatif. Oleh karena itu, diharapkan setelah analisis pemerintah bisa mengetahui apakah kebijakan PSBB merupakan kebijakan yang benar-benar tepat untuk diterapkan selama pandemi berdasarkan opini masyarakat.

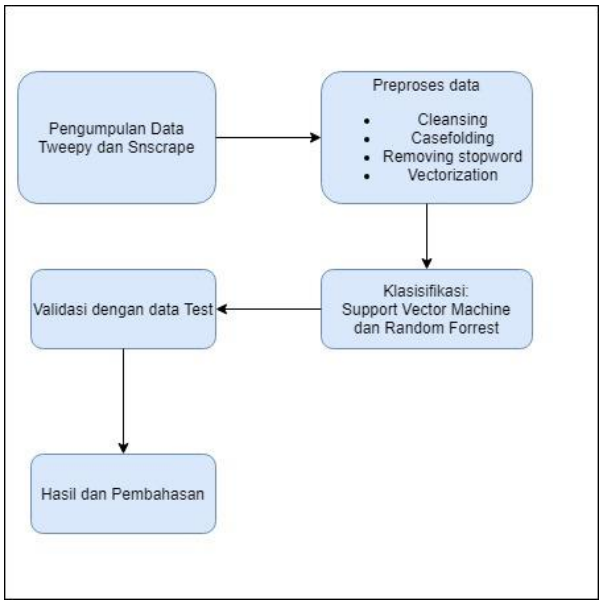
II. TINJAUAN PUSTAKA

Sentimen menurut

Istilah PSBB atau Pembatasan Sosial Berskala Besar pertama kali diusulkan melalui Peraturan Pemerintah Nomor 21 Tahun 2020 Tentang Pembatasan Sosial Berskala Besar dalam Rangka Percepatan Penanganan *Corona Virus Disease 2019 (COVID-19)*. Peraturan ini bertujuan untuk menekan angka peningkatan kasus penyebaran *COVID-19*

dengan membatasi kegiatan tertentu penduduk dalam suatu wilayah yang diduga terinfeksi *COVID-19* menurut Pasal 1 Peraturan Pemerintah Nomor 21 Tahun 2020. Dalam peraturan ini, hal-hal yang diatur, antara lain pembatasan pergerakan orang dan barang pada satu provinsi atau kabupaten/kota tertentu, kriteria sebagai penentu suatu provinsi atau kabupaten/kota harus melakukan PSBB, dan kegiatan yang dibatasi.

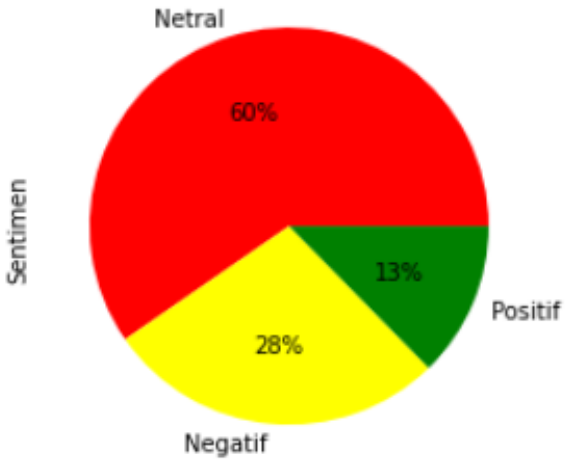
III. METODE PENELITIAN



Gambar 4 menjelaskan mengenai metodologi peneltian analisis sentiment dari mulai penggalian data sampai pembahasan kesimpulan.

A. *Pengumpulan data*

Pada awalnya dilakukan pengambilan data *tweet* dari platform media sosial twitter menggunakan API Tweepy dan Snscrape dengan bahasa pemrograman *python* [2]. Pengumpulan data ini diambil dari media sosial twitter dengan *query* kata 'PSBB' diambil pada rentang waktu tertentu yang sudah ditentukan sebelumnya, Snscrape berperan untuk mencari tweet dengan blok waktu tertentu rentang waktu ini dipertimbangakn dari kebijakan PSBB yang dilakukan di Jakarta sebagai Ibu kota, kemudian diteruskan oleh tweepy untuk mencari dengan acuan nomor id dari tweet tersebut, data yang terkumpul sebanyak 390 *tweet* yang nantinya akan dibagi menjadi data training dan test sebagai validasi klasifikasi yang dilakukan.



Gambar 5 Pie Chart hasil scrapping dan labeling sentimen dengan total data 466 *tweet*.

B. *Pra proses data*

Pra proses data harus dilakukan untuk memudahkan pembelajaran yang dilakukan mesin, dengan membersihkan *tweet* dari kata–kata yang tidak diperlukan seperti kata hubung, tanda titik, koma, dan sebagainya akan mempermudah dan meningkatkan akurasi dari pembelajaran mesin yang akan dirancang.

1) *Cleansing*: Pembersihan dilakukan untuk menghilangkan tanda titik, koma, garis miring dan sebagainya pada *tweet*.

2) *Casefolding*: Casefolding bertujuan untuk menyeragamkan huruf menjadi kecil untuk semua tweet tanpa terkecuali.

3) *Removing stop words*: Menghilangkan kata sambung pada *tweet* supaya kata – kata yang diproses adalah kata inti dari *tweet* tersebut, beberapa kata yang dihapus seperti "yang", "di", "dari", dan masih banyak lagi.

4) Vectorization: Mengubah teks yang sudah bersih menjadi angka represntatif dari teks tersebut menggunakan *package* Sklearn python. Bertujuan untuk inputan pada pembelajaran mesin yang hanya mengerti angka untuk dapat diproses [3].

C. *Klasifikasi dan Validasi*

Metode klasifikasi yang dipakai adalah *Support Vector Machine* karena diketahaui algoritma ini sangat baik untuk dilakukan pada klasifikasi teks dan tidak memerlukan kemampuan komputasi yang berat, metode ini sangat mudah di implementasi pada perangkat yang tidak terlalu mumpuni untuk melakukan pembelajaran mesin dan sangat sering

menjadi benchmark untuk metode – metode pembelaran mesin lainnya [4].

Kemudian kami melakukan teknis serupa dengan menggunakan *Random Forest Classifier*, luaran dari kedua algoritma ini nantinya akan kami bandingkan hasilnya dan di evaluasi.

Kemudian validasi dilakukan dengan menggunakan dataset yang sudah disiapkan sebagai data test, validasi berguna untuk mengevaluasi model pembelajaran mesin yang sudah dibuat apakah sudah memenuhi ekspektasi dari segi akurasinya atau belum.

IV. HASIL DAN PEMBAHASAN

A. Support Vector Machine

Support Vector Machine atau SVM merupakan sekumpulan metode *supervised learning* yang membuat *hyperlane* atau sekumpulan hyperlane pada proses klasifikasi, regresi, dan *outlier detection* [5]. Salah satu penggunaanya adalah dalam mengelompokkan *text* dan *hypertext* [6]. Kelebihan pada SVM ini adalah (1) efektif pada *high dimensional space*, (2) efektif dalam kasus dengan jumlah dimensi yang lebih banyak daripada jumlah sampelnya, (3) menggunakan subset titik pelatihan sehingga lebih memori efisien [5].

B. Random Forest

Selain menggunakan SVM atau *Support Vector Machine*, penelitian ini dilakukan juga dengan menggunakan metode *Random Forest*. *Random Forest*, merupakan sebuah metode yang dikembangkan dari metode CART (*Classification and Regression Trees*), yang juga merupakan metode atau algoritma dari teknik pohon keputusan [7]. Yang membedakan metode *random forest* dari metode CART adalah *Random Forest* menerapkan metode *bootstrap aggregating* (*bagging*) dan juga seleksi fitur *random* atau bisa disebut *random feature selection* [8].

Random Forest adalah kombinasi dari masing masing teknik pohon keputusan yang ada, lalu kemudian digabung dan dikombinasikan kedalam suatu model. Ada tiga poin utama dalam metode *Random Forest*, tiga poin utama tersebut yaitu (1) melakukan *bootstrap sampling* untuk membangun pohon prediksi; (2) masing-masing pohon keputusan memprediksi dengan prediktor acak; (3) kemudian *Random Forest* melakukan prediksi dengan mengombinasikan hasil dari tiap tiap pohon keputusan dengan cara *majority vote* untuk klasifikasi atau rata-rata untuk regresi [9].

Analogi dari penerapan *Random Forest* adalah sebagai berikut: seorang karyawan mendapatkan jatah cuti 2 minggu dan ingin berlibur ke suatu tempat wisata tetapi tidak tahu harus kemana. Karyawan tersebut memutuskan untuk bertanya kepada koleganya. Koleganya kemudian memberikan beberapa pertanyaan untuk memutuskan rekomendasi tempat wisata yang cocok untuk karyawan tersebut. Ketika sang karyawan menjawab pertanyaan dari koleganya, ia akan memberikan rekomendasi tempat. Hal ini adalah salah satu contoh pendekatan dari *decision tree algorithm*, dimana sang kolega dari karyawan tersebut menggambarkan pohon keputusan yang dibuat untuk membantu karyawan menentukan tempat wisata untuk berlibur. Tetapi model yang muncul dari pohon keputusan tersebut termasuk bias karena hanya berasal dari bertanya pada 1 kolega. Karyawan tersebut pun bertanya kembali menanyakan kolega-kolega lainnya mengenai rekomendasi tempat wisata. Setelah bertanya kepada beberapa kolega, karyawan tersebut mendapatkan tempat yang selalu direkomendasikan oleh banyak koleganya. Tempat yang direkomendasikan oleh banyak kolega sang karyawan ini disebut juga sebagai *Target Prediction.* Tempat wisata tersebut di anggap *high votes* karena banyak yang memilih dan karyawan tersebut pun memilih tempat wisata yang terkait. Hal ini adalah salah satu contoh pendekatan *Random Forest* karena *Random Forest* dimana pohon pohon keputusan yang telah terbentuk akan memutuskan sebuah keputusan, yang kemudian keputusan akhir akan ditentukan dengan hasil keputusan paling banyak. Konsep *voting* yang terjadi tanpa disadari ini dikenal dengan nama *majority voting* [10].

C. Evaluasi dan Perbandingan

Berikut hasil evaluasi dari masing–masing algoritma yang dipakai. *Precision, recall, f1-score,* dan *accuracy* menjadi indikator evaluasi

```
[[ 0 38  0]
 [ 1 81  2]
 [ 0 18  0]]
              precision    recall  f1-score   support

     Negatif       0.00      0.00      0.00        38
      Netral       0.59      0.96      0.73        84
     Positif       0.00      0.00      0.00        18

    accuracy                           0.58       140
   macro avg       0.20      0.32      0.24       140
weighted avg       0.35      0.58      0.44       140

0.5785714285714286
```

Gambar 6 Confusion matrix dari algoritma Random Forest.

Akurasi dari algoritma *Random Forest* pada data yang di tes sebesar 0.578.

```
[[ 0 37  1]
 [ 3 77  4]
 [ 1 16  1]]
              precision    recall  f1-score   support

     Negatif       0.00      0.00      0.00        38
      Netral       0.59      0.92      0.72        84
     Positif       0.17      0.06      0.08        18

    accuracy                           0.56       140
   macro avg       0.25      0.32      0.27       140
weighted avg       0.38      0.56      0.44       140

0.5571428571428572
```

Gambar 7 Confusion matrix dari algoritma Support Vector Machine

Akurasi dari algoritma *Support Vector Machine* pada data yang di tes sebesar 0.557.

Dari 466 data yang kami ambil dari twitter, dengan mempertimbangkan 3 tanggal mulai dari PSBB yang ada di Jakarta, kami membagi data menjadi data latih dan data tes dengan perbandingan 7 banding 3.
Dari tes yang dilakukan untuk masing–masing model didapatkan model *Random Forest* memiliki akurasi yang lebih tinggi namun tidak mampu mendeteksi label "Positif", dan akurasi pada model *Support Vector Machine* memang lebih rendah namun dapat mendeteksi label "Positif".

V. KESIMPULAN

Dari objek tweet dan model algoritma yang diuji, dapat disimpulkan bahwa beragamnya bahasa pada twitter mampu menurunkan kemampuan model untuk memprediksi suatu sentimen pada tweet terkait PSBB. *Support Vector Machine* dianggap lebih baik karena mampu mengenali tweet dengan label "Positif". Kurangnya data penelitian menjadi salah satu evaluasi pada penelitian ini, jumlah data yang cukup banyak akan membantu model untuk mengenali sentimen sebuah tweet lebih baik lagi, kemudian *text vectorization* khusus bahasa Indonesia akan sangat membantu pra proses data yang nantinya akan memperkaya informasi input dari sebuah model yang akan dilatih, terakhir algoritma *deep learning* juga mampu membuat pelatihan model yang dibangun lebih baik lagi.

DAFTAR PUSTAKA

[1] P. Batubara, "OkeZone," 7 April 2020. [Online]. Available: https://nasional.okezone.com/read/2020/04/07/337/2195637/pemerintah-ungkap-tujuan-dan-manfaat-status-psbb-di-jakarta. [Accessed 20 October 2020].

[2] A. Hernandez-Suarez, G. Sanchez-Perez, T.-M. K., V. Martinez-Hernandez, V. Sanchez and H. Perez-Meana, "A Web Scraping Methodology for Bypassing Twitter API," arXiv, Warwick, 2018.

[3] R. Bartusiak, Ł. Augustyniak, T. Kajdanowicz, P. Kazienko and M. Piasecki, "WordNet2Vec: Corpora agnostic word vectorization method," *Neurocomputing*, Vols. 326 - 327, pp. 141-150, 2019.

[4] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali and Z. Nawaz, "SVM Optimization for Sentiment Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 9, pp. 393 - 398, 2018.

[5] J. d. Boisberranger, " 1.4. Support Vector Machines — scikit-learn 0.20.2 documentation," Scikit-Learn, 3 August 2020. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html. [Accessed 31 October 2020].

[6] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, Berlin, 1998.

[7] A. Hartati, I. Zain and B. S. Suprih Ulama, "Analisis CART (Classification And Regression," *JURNAL SAINS DAN SENI ITS*, vol. 1, no. 1, pp. 101-105, 2012.

[8] L. Binarwati, I. Mukhlash and S. , "Implementasi Algoritma Genetika untuk Optimalisasi Random Forest Dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru: Studi Kasus PT.XYZ," *JURNAL SAINS DAN SENI ITS*, vol. 6, no. 2, pp. 78-82, 2017.

[9] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, vol. 1, no. 1, pp. 27-31, 2018.

[10] S. Polamuri, "Dataaspirant," Dataaspirant, 22 May 2017. [Online]. Available: https://dataaspirant.com/random-forest-algorithm-machine-learing/. [Accessed 29 October 2020].