BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Dalam tinjauan pustaka merupakan bagian yang menjelaskan tentang macam konsep dan teori-teori yang dilakukan dalam implementasi metode Regresi Linier dan bahasa pemrograman Python serta R.

Pada penelitian pertama yang dilakukan oleh (Boy, 2020) melakukan penelitian tentang memprediksi harga minyak, jenis minyak sawit/Crude Palm Oil (CPO) menggunakan metode Regresi Linier Berganda dengan studi kasus Dinas Perkebunan provinsi Sumatra Utara yang mendapatkan hasil akhir prediksi pada tahun 2018 dengan kenaikan harga pada minyak berkisar RP. 9.152,-/Kg. Kemudian dalam penelitian tersebut juga memberikan produk aplikasi yang bisa diakses melalui komputer maupun laptop untuk melihat perhitungan hasil, serta dapat dicetak sebagai laporan kepada pihak Dinas Perkebunan tersebut.

Pada penelitian kedua yang dilakukan oleh (Sulistyono & Sulistiyowati, 2017) membahas perencanaan peramalan produksi pada mesin pendingin dengan menganalisa metode Regresi Linier Berganda, kemudian dari penelitian tersebut ditemukan hasil dengan total produksi mesin pendingin sebesar 500.300 yang dipengaruhi variabel (x1) sebagai kerusakan mesin, (x2) harga bahan baku, (x3) jumlah tenaga kerja, bertujuan untuk memanajemen kebutuhan pasar dan meminimalisir adanya kerugian pabrik terhadap produksi yang tidak dibutuhkan.

Pada penelitian yang hampir serupa yang dilakukan oleh (Puteri & Silvanie, 2020) juga melakukan analisa implementasi metode Regesi Linier Berganda menggunakan bahasa pemrograman Python dengan membuat aplikasi web dari Django. Tentang memprediksi harga sembako dengan dataset total 90.945 baris dengan didukung variabel tanggal (x1), komoditas (x2), pasar (x3) untuk mencari harga (y') dari 9 sembako. Dari penelitian tersebut berhasil mendapatkan hasil 84.2% yang didapatkan dari variabel yang dipengaruhi.

Dari penelitian terkait yang sudah dijelaskan, maka bisa disimpulkan bahwa bahasa pemrograman Python dan R akan melakukan langkah penelitian dengan menyiapkan data CSV sebanyak 6 karena ada 6 jenis minyak sayur (minyak sawit, minyak kedelai, minyak kacang tanah, minyak bunga matahari, minyak kelapa dan minyak ikan) kemudian melakukan training pada keenam tabel CSV di bahasa Python dan R, hasil training tersebut berupa persamaan garis yang nantinya akan digunakan untuk uji coba di data baru, kemudian melakukan analisa menghitung MSE, dan RMSE atau MAE. Tabel 2.1 merupakan state of the art dari beberapa penelitian terkait bahasa pemrograman dan prediksi peramalan.

Tabel 2.1 State Of The Art

No	Judul	Tahun	Objek Penelitian	Metode	Hasil
1.	Implementasi data mining dalam memprediksi harga Crude Palm Oil (CPO) Pasar Domestik menggunakan Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatra Utara)	2020	Minyak Sawit.	Regresi Linier Berganda.	Berdasarkan hasil dan pembahasan diperoleh hasil cukup memuaskan dengan prediksi harga sebesar Rp.9.152,-/Kg yang berarti mengalami kenaikan serta membuat suatu aplikasi berbasis desktop untuk membantu dinas dalam membuat laporan.
2.	Peramalan produksi dengan metode Regresi Linier Berganda.	2017	Mesin Pendingin.	Regresi Linier Berganda.	Berdasarkan hasil dan pembahasan diperoleh peramalan hasil dari produksi mesin pendingin sebanyak 500.300 unit.

Tabel 2.2 (Lanjutan)

No	Judul	Tahun	Objek Penelitian	Metode	Hasil
	Peramalan produksi dengan metode Regresi Linier Berganda.	2017	Mesin Pendingin.	Regresi Linier Berganda.	Dengan variabel Input meliputi kerusakan mesin (x1), harga barang baku (x2), jumlah tenaga kerja (x3).
3.	Implementasi data mining untuk memprediksi prestasi siswa berdasarkan status sosial dan kedisiplinan pada SMK Bayu Pertiwi menggunakan metode Regresi Linier Berganda.	2018	Sekolah SMK Bayu Pertiwi.	Regresi Linier Berganda.	Berdasarkan hasil dan pembahasan, peneliti berhasil membantu instansi tersebut dalam melakukan peramalan prestasi dalam rangka memajukan SMK Bayu Pertiwi lebih baik dan lebih berkompeten, bahkan hal ini telah diterapkan oleh para pegawai tata usaha instansi tersebut.
4.	Analisis permintaan konsumen terhadap konsumsi minyak tanah rumah tangga di desa Pelauw dengan menggunakan analisis Regresi Linier Berganda.	2017	Minyak tanah.	Regresi Linier Berganda.	Berdasarkan hasil riset penelitian diperoleh model terbaik untuk permintaan minyak tanah menggunakan Regresi Linier Berganda dipengaruhi oleh variabel pendapatan (x1), jumlah anggota keluarga (x2), harga (x3).

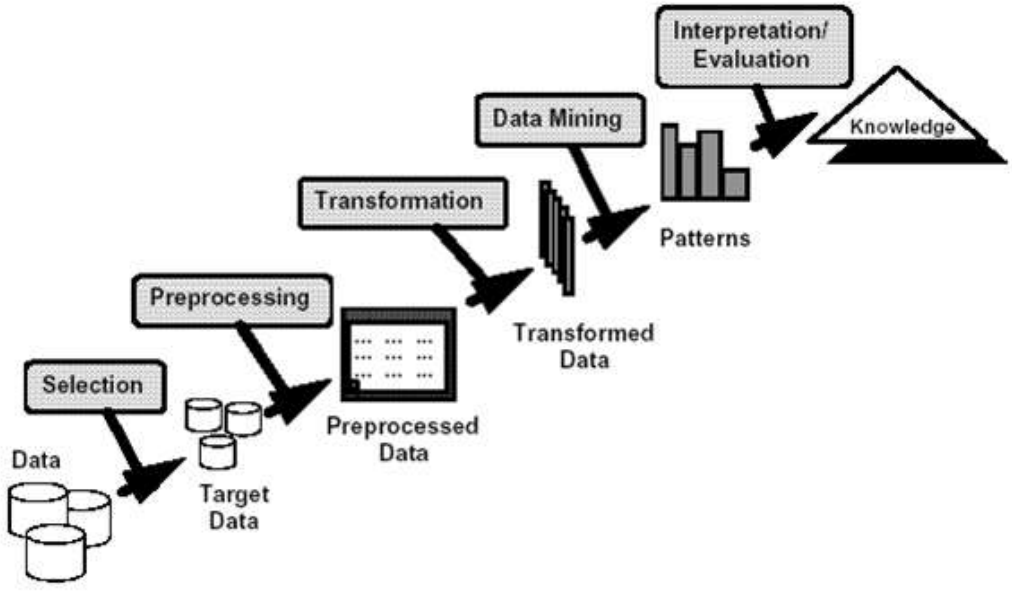
Tabel 2.3 (Lanjutan)

No	Judul	Tahun	Objek Penelitian	Metode	Hasil
		2017	Minyak tanah.	Regresi Linier Berganda.	Harga gas elpiji (x4), penggunaan energi rumah tangga (x5), dengan hasil persamaan garis adalah $\hat{Y}$=-0.794+0.001χ1+1.211χ2+0.164χ3-0.048χ4.
5.	Machine learning untuk model prediksi harga sembako dengan metode Regresi Linier Berganda.	2020	Sembako.	Regresi Linier Berganda.	Hasil riset penelitian diperoleh, sistem machine learning terbukti dapat membantu dalam melakukan peramalan menggunakan metode tersebut dengan hasil prediksi sebesar 84.2%. Dengan variabel inputan meliputi tanggal (x1), komoditas (x2), pasar (x3).
Berdasarkan hasil review table state of the art maka peneliti mengambil kesimpulan dalam metode algoritma penelitian selanjutnya adalah metode Regresi Linier Berganda dan menggunakan bahasa pemrograman Python dan R sebagai pendukungnya, karena metode dan bahasa pemrograman tersebut cocok dijadikan sebagai prediksi suatu studi kasus penelitian dataset minyak sayur.					

2.2 Data Mining

Data mining menurut (Adinugroho & Sari, 2018) membahas tentang serangkaian teknik pengumpulan data baru berupa informasi untuk keperluan peramalan data. Istilah data mining mulai populer di komunitas pengguna basis data pada tahun 1990-an. Namun, teori dan metode dasar dari data mining telah lahir jauh sebelum era 90.

Data mining berasal dari berbagai disiplin ilmu, 2 ilmu yang paling mendasari adalah statistika dan machine learning namun tetap membutuhkan database serta visualisasi data. Data mining juga mempunyai operasi dasar yang terbagi menjadi 2 kategori menurut Kantardzic J.B., M., yaitu pertama metode Deskriptif dengan tujuan untuk menghasilkan pola, relasi atau anomali ke dalam data yang mudah difahami oleh manusia Clustering, Association Rule dan Sequential Pattern Discovery merupakan contoh dari metode Deskriptif. Sedangkan metode kedua adalah metode Prediktif yang bertujuan untuk memprediksi atau memperkirakan nilai variabel berdasarkan nilai variabel yang lainnya, contoh adalah Klasifikasi dan Regresi. Dalam aktivitas manusia sehari-hari tanpa disadari selalu menggunakan data dan tujuan dari data mining adalah dapat menemukan atau menentukan pola atau model dalam analisa sederhana, sehingga dari data mining manusia mendapatkan pengetahuan umum dan mampu membuat sistem keputusan. Berikut adalah proses melakukan data mining:



Gambar 2.1 Tahapan Proses Data Mining

(Sumber: Isal, 2019)

Gambar 2.1 merupakan serangkaian tahapan data mining, yang dilakukan secara berurutan dimulai dari seleksi pemilihan data, kemudian melakukan perbaikan kualitas data pada tahap preprocessing, dilanjut melakukan transformation data ke bentuk standar seperti spreadsheet, selanjutnya melakukan data mining, hingga melakukan evaluasi untuk mendapatkan ilmu pengetahuan baru. Di bawah ini merupakan peran dari data mining antara lain:

1. Klastering/*Clustering*

Berasal dari metode Deskriptif atau metode *Unsupervised Learning*, yang berperan untuk mengelompokkan data dengan ciri kelas atau ciri kelompok ataupun karakteristik yang sama begitu pula pada Klastering yang lain dengan data berbeda, contoh algoritma metode: K-Means Clustering, K-NN, Hierarchical Clustering (Muflikhah, Ratnawati, & Putri, 2018).

Algoritma K-Means Clustering, ditemukan oleh Lloyd 1957-1982, Forgey 1965, Friedman dan Rubin 1967.

kemudian pada tahun 1967 oleh MacQueen dengan tujuan mengelompokkan data berdasarkan cluster/kelompok tertentu, kemudian mengelompokkan/centroid dari cluster-cluster yang berbeda dalam kelompok yang lain (Syafnidawaty, 2020).

Algoritma Hierarchical Clustering merupakan teknik data mining dalam pengelompokkan data yang memiliki kemiripan dalam hirarki/tingakatan dan tidak mempunyai kemiripan dalam hirarki/tingakatan yang jauh (Muflikhah, Ratnawati, & Putri, 2018).

Metode ini dapat dibedakan cara perhitungannya antara lain:

- Single linkage atau jarak minimum.
- Complete linkage atau jarak maksimum.
- Average linkage atau jarak rata-rata.

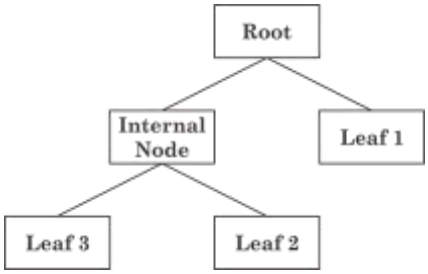
2. Klasifikasi/*Classification*

Berasal dari metode prediktif atau metode *Supervised Learning*, yang berperan mengelompokkan data atau kelas untuk memprediksi kelas objek di mana tidak diketahui kelasnya, contoh algoritma metode: Nearest-neighbor (K-NN), Naïve Bayes, Decision Tree, Support Vector Machines (SVM), Neural Network, Random Forest.

Algoritma K-NN merupakan algoritma sederhana teknik klasifikasi data berdasarkan karakteristik yang sudah diketahui. Cara kerja dari algoritma K-NN antara lain: menentukan parameter K terlebih dahulu, menghitung jarak pada data garis titik plot yang tersedia, mengurutkan dari nilai tertinggi ke rendah, mengumpulkan kategori berdasarkan nilai atau tetangga K, kemudian menggunakan K-NN yang cenderung dominan sama.

Algoritma Naïve Bayes merupakan teknik klasifikasi dengan menggunakan probabilitas keanggotaan kelas. Cara kerja dari algoritma tersebut dengan menggunakan probabilitas bersyarat kemudian disempurnakan dengan teorima bayes yang bertujuan untuk melakukan evaluasi guna untuk mendapatkan tambahan informasi atau pengetahuan (Muflikhah, Ratnawati, & Putri, 2018).

Algoritma Decision Tree merupakan teknik klasifikasi menggunakan struktur pohon atau dalam istilah lain pohon keputusan. Metode ini merupakan gabungan dari Classification Tree dan Regression Tree, sebagaimana alur kerja dalam gambar:



Gambar 2.2 Decision Tree

(Sumber: Mubarok, 2018)

Gambar 2.2 Root merupakan node teratas dan memiliki output yang lebih dari satu, internal Node merupakan bentuk percabangan, leaf atau daun merupakan node terakhir yang tidak memiliki output dan hanya memiliki 1 input.

3. Asosiasi/*Association*

Merupakan teknik data mining yang digunakan untuk mencari model atau pola antar hubungan dalam dataset.

4. Estimasi/*Estimation*

Merupakan teknik data mining untuk memprediksi nilai yang belum diketahui misalkan menerka-nerka informasi dari seorang siswa SMA.

5. Prediksi/*Predictions*

Merupakan teknik data mining bertujuan untuk memprediksi suatu variabel tertentu berdasarkan variabel lain yang sudah diketahui.

Contoh dari algoritma metode prediksi adalah metode Regresi Linier yang akan dijelaskan pasa sub bab 2.4 pada penelitian ini.

2.3 Manfaat Data Mining

Banyak sekali saat ini para peneliti yang telah menerapkan data mining dalam riset mereka karena manfaat data mining yang telah mereka rasakan. Metode klasifikasi dapat diimplementasikan untuk penginderaan jarak jauh melalui citra satelit, dengan menggunakan citra satelit tersebut, manusia dapat mengetahui perubahan lahan dari perubahan waktu ke waktu dan menentukan lokasi.

Data mining juga bermanfaat dalam dunia bisnis dikarenakan dapat memperkirakan pola-pola perbelanjaan konsumen hingga memprediksi harga suatu produk contoh seperti supermarket, E-commerce, transaksi bank dll (Gunawan, 2018).

Dalam dunia kedokteran data mining bermanfaat sebagai rekomendasi seperti mengetahui klasfikasi penyakit pada pasien, rekomendasi obat sesuai keluhan pasien, membantu perusahaan asuransi dalam mengawasi kecurangan dll. Dalam dunia pendidikan bertujuan untuk mengembangkan SDM lebih maju contoh penerapan data mining adalah rekomendasi dalam memberikan beasiswa kepada siswa, rekomendasi memilih jurusan yang diminati siswa dll.

Dalam perusahaan besar contoh penerapannya seperti klasifikasi data pada nasabah, rekomendasi tabungan untuk nasabah dan lain-lain, rata-rata data mining sangat berpotensi baik dalam penelitian tersebut yang sudah dijelaskan (Adinugroho & Sari, 2018).

2.4 Regresi Linier

Analisis Regresi merupakan teknik perhitungan statistika yang bertujuan menemukan hubungan atau persamaan dari variabel dependen (y) dengan variabel independen (x) (Widarjono, 2018). Statistika berbeda dengan statistik, statistika merupakan ilmu yang mempelajari pengumpulan data, menganalisa, memprediksi, memperdiksi data. Sedangkan statistik merupakan hasil kinerja ilmu statistika pada suatu data. Memasuki abad ke-19 dan awal abad ke-20 statistika merupakan bagian dari matematika yang mulai banyak diterapkan mulai dari metode ilmiah atau penelitian, peluang dan dalam bidang komputasi bisa diterapkan dalam pengenalan pola pada kecerdasan buatan. Berdasarkan metodenya statistika terbagi menjadi 2 macam pertama statistik deskriptif digunakan untuk menganalisa kumpulan data yang diterapkan dalam bentuk tabel atau grafik, sedangkan yang kedua statistik inferensi digunakan untuk melakukan pengujian hipotesis berdasarkan sampel atau berdasarkan analisa data (Jimy, 2019).

Di mana dalam kebanyakan penelitian yang sering digunakan adalah Regresi Sederhana dan kali ini, peneliti akan menggunakan Regresi Linier Berganda dalam algoritma metodenya.

Algoritma Regresi Linier adalah teknik statistika yang bertujuan mencari nilai hasil persamaan dari variabel dependen yang dipengaruhi oleh variabel bebas bisa ada 1variabel atau bahkan lebih (Boy, 2020). Regresi Linier sangat sesuai diimplementasikan untuk prediksi atau peramalan suatu data yang akan diolah. Dalam Regresi Linier Sederhana hanya ada 1 variabel bebas atau variabel independen dengan simbol *(x)* yang mempengaruhi 1 variabel tak bebas atau dependen dengan simbol *(y)*.

Dalam Persamaan 2.1 Regresi Linier Sederhana:

$y'=a+bx$……..……..(2.1)

Atau bisa juga menggunakan penulisan rumus seperti pada Persamaan 2.2 ini namun pada dasarnya adalah sama.

$\gamma = \beta 0 + \beta 1\chi$ ……...……..(2.2)

- *y*=variabel dependen/terikat
- x=variabel independen/bebas
- b=koefisien variabel x, untuk mendapatkan nilai koefisien dengan rumus di Persamaan 2.3
- a=konstanta, untuk mendapatkan nilai konstanta dengan rumus Persamaan 2.4

$$b=\frac{n\left(\sum \chi\gamma\right)-\left(\sum \chi\right)\left(\sum \gamma\right)}{n\left(\sum \chi^2\right)-\left(\sum \chi\right)^2} \text{……..…….(2.3)}$$

$$\alpha=\frac{\sum \gamma - b\left(\sum \chi\right)}{n} \text{……..…….....(2.4)}$$

Sedangkan dalam Regresi Linier Berganda variabel bebas atau variabel independen dengan simbol *(x)* bisa lebih dari dua variabel lainnya, yang mempengaruhi satu variabel tidak bebas atau dependen dengan simbol *(y)*.

Dalam persamaan 2.5 Regresi Linier Berganda:

$\gamma = \beta 0 + \beta 1\chi 1 + \beta 2\chi 2 + \varepsilon$ ……..……..(2.5)

β0 merupakan: nilai konstanta, β1 dan β2: nilai koefisien, sedangkan nilai variabelnya χ1: tahun dan χ2: bulan.

Regresi Linier dari machine learning akan melakukan training pada data input dengan variabel $\chi 1$: tahun dan $\chi 2$: bulan, kemudian variabel outputnya terdiri dari $\gamma 1$: harga minyak sawit, $\gamma 2$: harga minyak kedelai, $\gamma 3$: harga minyak kacang tanah, $\gamma 4$: harga minyak bunga matahari, $\gamma 5$: harga minyak kelapa, $\gamma 6$: harga minyak ikan.

1) Cara Kinerja Regresi Linier Berganda Pada Jupyter Bahasa Python

Tentukan dataset yang sudah tersimpan dalam file CSV, selanjutnya melakukan import library pada Jupyter Python seperti pandas, numpy, matplotib dll, kemudian melakukan training, selanjutnya mencari model atau persamaan dengan Regresi Linier. (keterangann lebih mendalam diterangkan pada bab 4).

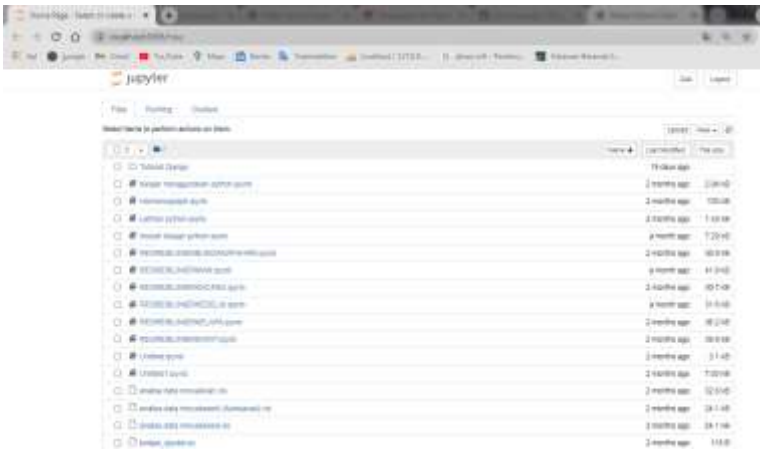
2) Cara Kinerja Regresi Linier Berganda pada RStudio Bahasa R

Tentukan dataset yang akan dianalisa dalam tipe file CSV, selanjutnya membuat model hubungan dengan fungsi lm() pada R, untuk melihat hasil dapatkan rangkuman dengan pemanggilan fungsi summary() sehingga bisa melihat rata-rata error proses prediksi, min, median, max. (keterangan lebih mendalam diterangkan pada bab 4).

2.5 Bahasa Pemrograman Python

Bahasa pemrograman Python merupakan bidang ilmu yang berfokus pada kode-kode atau sintaks algoritma yang mudah dipelajari bahkan oleh pemula sekalipun, Python juga bersifat open source dan mudah dalam melakukan penginstallan baik di windows, mac oc x, dan linux. Dalam catatan sejarah (Enterprise, 2017) Python dikembangkan oleh Guido van Rossum 1990 di CWI Amsterdam. Pada penelitian ini menggunakan Python versi 3.8.5 dan Anaconda tipe 64-Bit dalam mendukung penyelesaian studi kasus ini. Untuk membuat program perhitungannya dibantu dengan tool Jupyter Notebook yang sudah otomatis terinstall bersamaan dengan Python dan Anaconda.

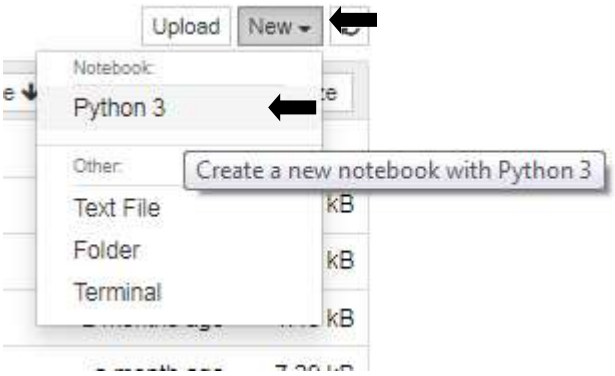
Sedangkan untuk menjalankannya bisa mengetikan sintaks *jupyter notebook* di Command Prompt (windows) kemudian akan menampilkan pada browser lokasi folder untuk menjalankan program atau dashboard jupyter notebook.



Gambar 2.3 Tampilan Dashboard Jupyter Notebook

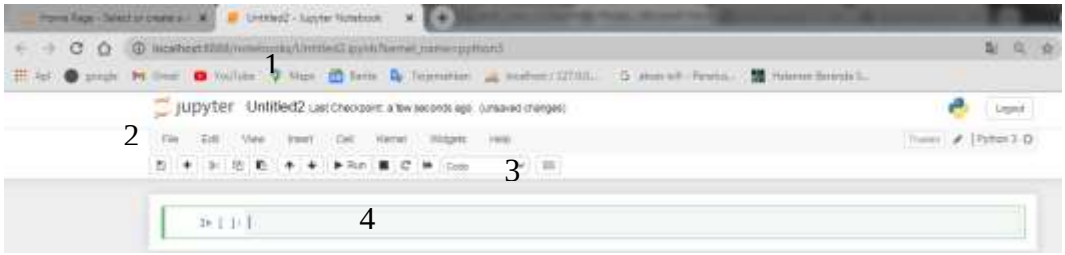
Gambar 2.3 merupakan tampilan utama dashboard Jupyter Notebook atau interface yang siap untuk eksekusi program. Untuk lebih mendalami bahasa Python harus mempelajari dasa-dasar pada Python seperti pengunaan indentation, pengenalan variable, string, memanfaatkan operator, mengenal fungsi/ fungction pada Python, perulangan, pengenalan kondisinal if, pengenalan stuktur data, belajar module, hingga Api database Python yang bisa mendukung bermacam-macam server seperti: MysQl, PostgreSQL, SQLite, mSQL, Oracle, Sybase, Interbase, Informix, Microsoft SQL Server 2000, GadFly.

Cara membuat notebook baru pada Jupyter Notebook lakukan klik pada tombol **New** pilih **Python 3**



Gambar 2.4 Langkah Membuat Notebook Baru

Kemudian untuk melakukan penyuntingan, kenali terlebih dahulu untuk tampilan utama atau antara muka sebelum eksekusi program.



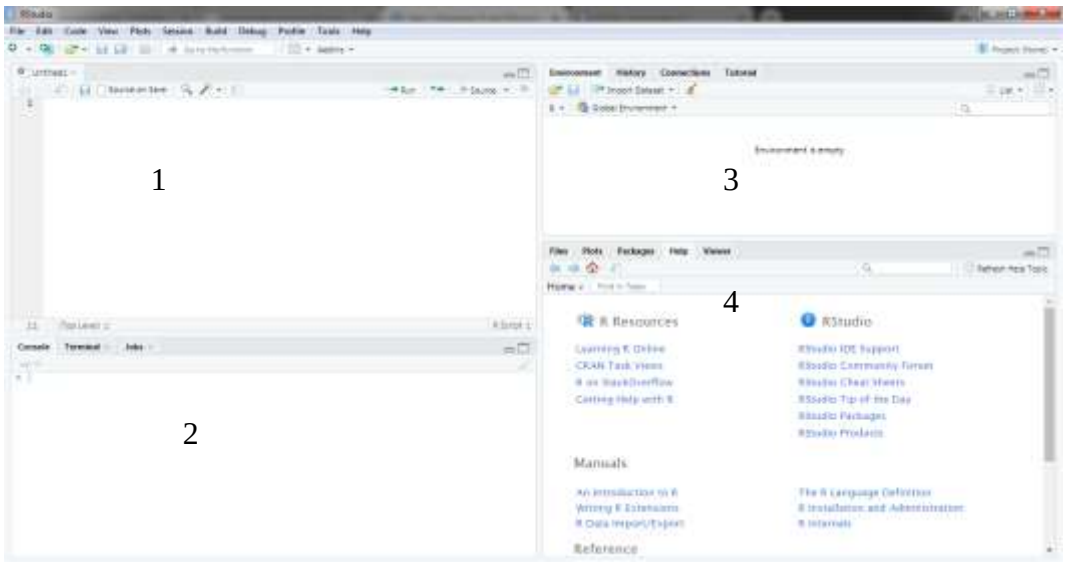
Gambar 2.5 Pengenalan Interface Notebook Baru

Gambar 2.5 merupakan tampilan dari notebook baru dan tiap nomer mempunyai keterangan yakni nomer satu merupakan tempat untuk memasukan tittle atau judul dari program yang akan dikerjakan, nomer dua merupakan barisan menu dengan berbagai fungsi atau perintah yang berbeda, nomer tiga merupakan dropdown jenis sel yang bisa dirubah sebelum melakukan eksekusi program, nomer empat merupakan sell atau tempat untuk pengisian kodingan dengan kunci pintas pada keyboard untuk membuat sell baru atau menambahkan gunakan keyword **shift+enter** maka, akan menampilkan sell baru, kemudian Jupyter Notebook siap untuk melakukan program eksekusi sesuai kebutuhan.

2.6 Bahasa Pemrograman R

Bahasa pemrograman R merupakan bidang ilmu yang mengolah analisis data statistik dan grafik, pada awalnya bahasa ini diciptakan oleh Ross Ihaka dan Robert Gentlemean. Sifatnya yang open source menjadikan mudah diakses oleh siapapun secara gratis untuk dikembangkan kembali oleh para developers.

Bahasa pemrograman R saat ini sudah sampai pada versi 4.0.3 dapat pula diunduh melalui windows, mac os x, dan linux, setelah menginstall pada situs resmi lakukan pula pengunduhan environmentnya yakni install editor RStudio pada website resminya, pada peneliti ini mengunakan versi 1.4.1103 (Gio & Effendie, 2017). Bahasa pemrograman Python dan R memiliki tujuan yang sama yakni menghasilkan persamaan garis. Sebelum melakukan eksekusi program pada RStudio dengan bahasa R kenali terlebih dahulu interface pada RStudio.

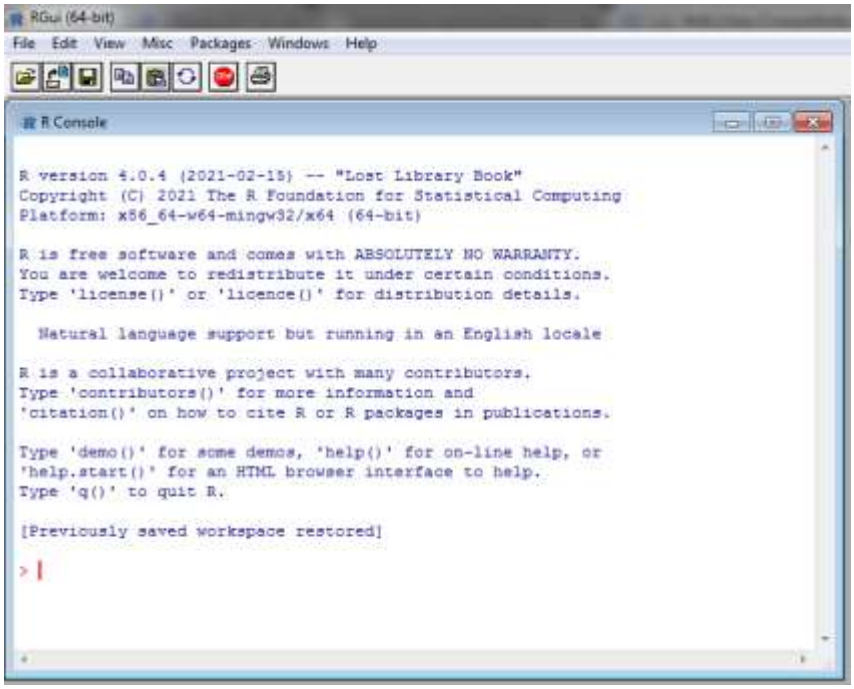


Gambar 2.6 Tampilan Utama RStudio

Gambar 2.6 merupakan tampilan dari RStudio di mana pada tool ini akan dilakukan program perhitungan studi kasus, pada nomer satu merupakan bagian scripts editor untuk pengisian perintah sama seperti dengan Python, nomer dua merupakan console juga bisa diterapkan untuk pengisian perintah kode.

Nomer tiga merupakan environment jendela dari program yang memilki tab environment, history, connections, Tutorial, nomer empat merupakan jendela yang mampu memanajemn file, menampilkan output command plot, informasi dan bantuan script.

Sebelum melakukan program pelajari dasar-dasar pemrograman R seperti: assignment, matriks, array, dataframe, list, pengenalan function, pengenalan package, perulangan, stuktur data, import data dan export data, dan pengenalan visualisasi data. Berikut adalah tampilan GUI pada R, namun yang digunakan pada penelitian ini menggunakan RStudio agar mudah difahami dalam pembelajaran.



Gambar 2.7 Tampilan GUI R

Gambar 2.7 merupakan tampilan GUI pada R, namun untuk kenyamanan dalam belajar bisa mengunakan GUI R atau RStudio untuk eksekusi program.

2.7 Microsoft Excel

Ms.Excel merupakan pengolahan angka yang sangat sering diterapkan baik pelajar, pengajar atau karyawan dalam pekerjaan mereka dan penggunaannya yang relative mudah. Dengan menggunakan Ms.Excel, dapat membantu dalam pengerjaan input data, menganalisa data, manipulasi data, membuat grafik dan mengolah data-data besar sehingga mendapatkan laporan seperti yang diinginkan. Sampai saat ini sudah banyak sekali versi dari Microsoft Excel yang beredar. Ms.Excel diperlukan untuk menyimpan data-data dalam bentuk format file CSV.